Soc504: Causal Inference Topics

Brandon Stewart[1]

Princeton

April 10 - April 19, 2017

[1]This lecture draws from slides by Matt Blackwell, Jens Hainmueller, Erin Hartman and Gary King

Readings

- Monday

Readings

- Monday
  - King, Gary and Langche Zeng. "The Dangers of Extreme Counterfactuals," Political Analysis, 14, 2, (2007): 131-159.

Readings

- Monday
  - King, Gary and Langche Zeng. "The Dangers of Extreme Counterfactuals," Political Analysis, 14, 2, (2007): 131-159.
  - King, Gary and Langche Zeng. "When Can History be Our Guide? The Pitfalls of Counterfactual Inference," International Studies Quarterly, 2006, 51 (March, 2007): 183–210.

Readings

- Monday
  - King, Gary and Langche Zeng. "The Dangers of Extreme Counterfactuals," Political Analysis, 14, 2, (2007): 131-159.
  - King, Gary and Langche Zeng. "When Can History be Our Guide? The Pitfalls of Counterfactual Inference," International Studies Quarterly, 2006, 51 (March, 2007): 183–210.
- Wednesday

Readings

- Monday
  - King, Gary and Langche Zeng. "The Dangers of Extreme Counterfactuals," Political Analysis, 14, 2, (2007): 131-159.
  - King, Gary and Langche Zeng. "When Can History be Our Guide? The Pitfalls of Counterfactual Inference," International Studies Quarterly, 2006, 51 (March, 2007): 183–210.
- Wednesday
  - Daniel Ho, Kosuke Imai, Gary King, and Elizabeth Stuart." Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference," Political Analysis, 15 (2007): 199-236.

Readings

- Monday
  - King, Gary and Langche Zeng. "The Dangers of Extreme Counterfactuals," Political Analysis, 14, 2, (2007): 131-159.
  - King, Gary and Langche Zeng. "When Can History be Our Guide? The Pitfalls of Counterfactual Inference," International Studies Quarterly, 2006, 51 (March, 2007): 183–210.
- Wednesday
  - Daniel Ho, Kosuke Imai, Gary King, and Elizabeth Stuart." Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference," Political Analysis, 15 (2007): 199-236.
- Monday

Readings

- Monday
  - King, Gary and Langche Zeng. "The Dangers of Extreme Counterfactuals," Political Analysis, 14, 2, (2007): 131-159.
  - King, Gary and Langche Zeng. "When Can History be Our Guide? The Pitfalls of Counterfactual Inference," International Studies Quarterly, 2006, 51 (March, 2007): 183–210.
- Wednesday
  - Daniel Ho, Kosuke Imai, Gary King, and Elizabeth Stuart." Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference," Political Analysis, 15 (2007): 199-236.
- Monday
  - Review Morgan and Winship Potential Outcomes Chapter

Readings

- Monday
  - King, Gary and Langche Zeng. "The Dangers of Extreme Counterfactuals," Political Analysis, 14, 2, (2007): 131-159.
  - King, Gary and Langche Zeng. "When Can History be Our Guide? The Pitfalls of Counterfactual Inference," International Studies Quarterly, 2006, 51 (March, 2007): 183–210.
- Wednesday
  - Daniel Ho, Kosuke Imai, Gary King, and Elizabeth Stuart." Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference," Political Analysis, 15 (2007): 199-236.
- Monday
  - Review Morgan and Winship Potential Outcomes Chapter
  - Kosuke Imai, Gary King, and Elizabeth Stuart. Misunderstandings Among Experimentalists and Observationalists About Causal Inference. Journal of the Royal Statistical Society, Series A, (2008) 171, part 2: 481502

Readings

- Monday
  - King, Gary and Langche Zeng. "The Dangers of Extreme Counterfactuals," Political Analysis, 14, 2, (2007): 131-159.
  - King, Gary and Langche Zeng. "When Can History be Our Guide? The Pitfalls of Counterfactual Inference," International Studies Quarterly, 2006, 51 (March, 2007): 183–210.
- Wednesday
  - Daniel Ho, Kosuke Imai, Gary King, and Elizabeth Stuart."Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference," Political Analysis, 15 (2007): 199-236.
- Monday
  - Review Morgan and Winship Potential Outcomes Chapter
  - Kosuke Imai, Gary King, and Elizabeth Stuart. Misunderstandings Among Experimentalists and Observationalists About Causal Inference. Journal of the Royal Statistical Society, Series A, (2008) 171, part 2: 481502
- Wednesday

Readings

- Monday
  - King, Gary and Langche Zeng. "The Dangers of Extreme Counterfactuals," Political Analysis, 14, 2, (2007): 131-159.
  - King, Gary and Langche Zeng. "When Can History be Our Guide? The Pitfalls of Counterfactual Inference," International Studies Quarterly, 2006, 51 (March, 2007): 183–210.
- Wednesday
  - Daniel Ho, Kosuke Imai, Gary King, and Elizabeth Stuart." Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference," Political Analysis, 15 (2007): 199-236.
- Monday
  - Review Morgan and Winship Potential Outcomes Chapter
  - Kosuke Imai, Gary King, and Elizabeth Stuart. Misunderstandings Among Experimentalists and Observationalists About Causal Inference. Journal of the Royal Statistical Society, Series A, (2008) 171, part 2: 481502
- Wednesday
  - Optional: Imai, Keele, Tingley and Yamamoto. "Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies" American Political Science Review (2011)

Readings

- Monday
  - King, Gary and Langche Zeng. "The Dangers of Extreme Counterfactuals," Political Analysis, 14, 2, (2007): 131-159.
  - King, Gary and Langche Zeng. "When Can History be Our Guide? The Pitfalls of Counterfactual Inference," International Studies Quarterly, 2006, 51 (March, 2007): 183–210.
- Wednesday
  - Daniel Ho, Kosuke Imai, Gary King, and Elizabeth Stuart." Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference," Political Analysis, 15 (2007): 199-236.
- Monday
  - Review Morgan and Winship Potential Outcomes Chapter
  - Kosuke Imai, Gary King, and Elizabeth Stuart. Misunderstandings Among Experimentalists and Observationalists About Causal Inference. Journal of the Royal Statistical Society, Series A, (2008) 171, part 2: 481502
- Wednesday
  - Optional: Imai, Keele, Tingley and Yamamoto. "Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies" American Political Science Review (2011)
  - Optional: Acharya, Blackwell and Sen. "Explaining Causal Findings Without Bias: Detecting and Assessing Direct Effects." American Political Science Review. (2016).

Counterfactuals

Counterfactuals

- Three types:

Counterfactuals

- Three types:
  1. Forecasts Will Donald Trump win reelection?

Counterfactuals

- Three types:
    1. Forecasts Will Donald Trump win reelection?
    2. What-if Questions What would have happened if the U.S. had not invaded Iraq?

Counterfactuals

- Three types:
  1. Forecasts Will Donald Trump win reelection?
  2. What-if Questions What would have happened if the U.S. had not invaded Iraq?
  3. Causal Effects What is the causal effect of the Iraq war on U.S. Supreme Court decision making? (a factual minus a counterfactual)

Counterfactuals

- Three types:
  1. Forecasts Will Donald Trump win reelection?
  2. What-if Questions What would have happened if the U.S. had not invaded Iraq?
  3. Causal Effects What is the causal effect of the Iraq war on U.S. Supreme Court decision making? (a factual minus a counterfactual)
- Counterfactuals are some part of most research, absolutely essential in the context quantities of interest

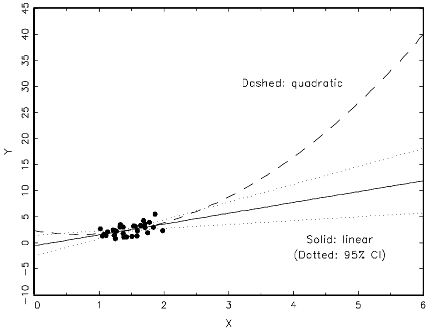
Counterfactuals

- Three types:
  1. Forecasts Will Donald Trump win reelection?
  2. What-if Questions What would have happened if the U.S. had not invaded Iraq?
  3. Causal Effects What is the causal effect of the Iraq war on U.S. Supreme Court decision making? (a factual minus a counterfactual)
- Counterfactuals are some part of most research, absolutely essential in the context quantities of interest
- The model will always give an answer- so how do identify reasonable counterfactuals?

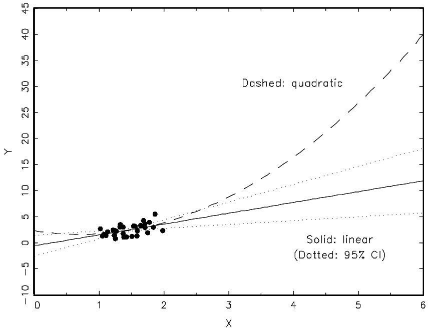
Counterfactuals

- Three types:
  1. Forecasts Will Donald Trump win reelection?
  2. What-if Questions What would have happened if the U.S. had not invaded Iraq?
  3. Causal Effects What is the causal effect of the Iraq war on U.S. Supreme Court decision making? (a factual minus a counterfactual)
- Counterfactuals are some part of most research, absolutely essential in the context quantities of interest
- The model will always give an answer- so how do identify reasonable counterfactuals?
- Summary of Today: don't ask your model unreasonable questions. (remember the Momentous Sprint?)

Which model would you choose? (Both fit the data well.)

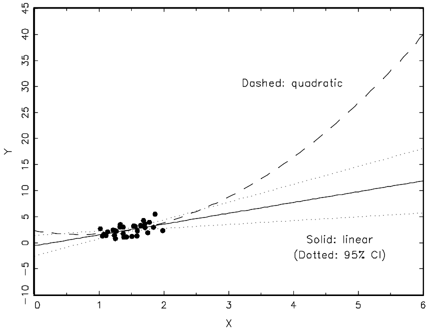


Which model would you choose? (Both fit the data well.)



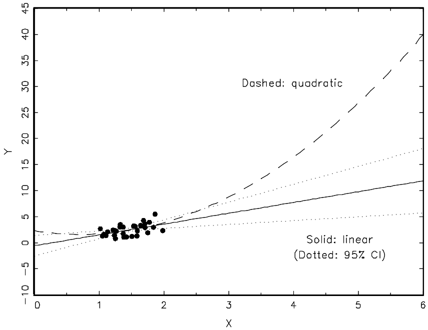
- Compare prediction at $x = 1.5$ to prediction at $x = 5$

Which model would you choose? (Both fit the data well.)



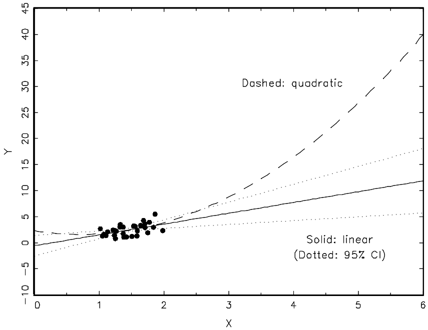
- Compare prediction at $x = 1.5$ to prediction at $x = 5$
- How do you choose a model?

Which model would you choose? (Both fit the data well.)



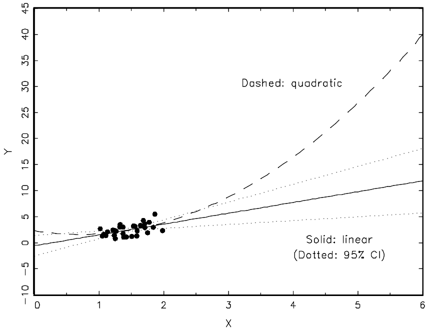
- Compare prediction at $x = 1.5$ to prediction at $x = 5$
- How do you choose a model? R^2?

Which model would you choose? (Both fit the data well.)



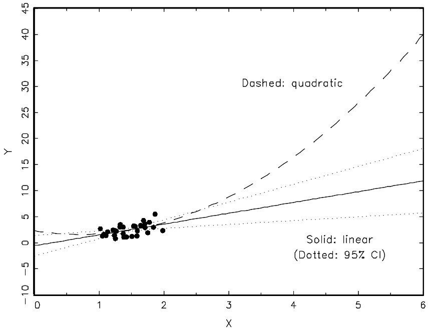
- Compare prediction at $x = 1.5$ to prediction at $x = 5$
- How do you choose a model? R^2? Some "test"?

Which model would you choose? (Both fit the data well.)



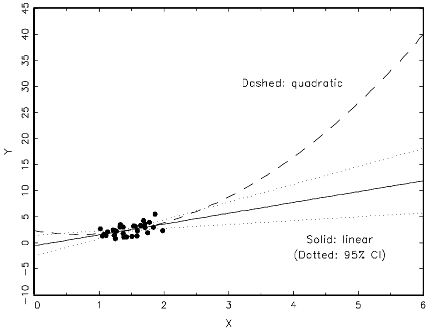
- Compare prediction at $x = 1.5$ to prediction at $x = 5$
- How do you choose a model? R^2? Some "test"? "Theory"?

Which model would you choose? (Both fit the data well.)



- Compare prediction at $x = 1.5$ to prediction at $x = 5$
- How do you choose a model? R^2? Some "test"? "Theory"?
- The bottom line: answers to some questions don't exist in the data.

Which model would you choose? (Both fit the data well.)



- Compare prediction at $x = 1.5$ to prediction at $x = 5$
- How do you choose a model? R^2? Some "test"? "Theory"?
- The bottom line: answers to some questions don't exist in the data.
- Our estimate of certain quantities of interest is highly model dependent

Model Dependence Proof

Model Dependence Proof

Model Free Inference

Model Dependence Proof

Model Free Inference

To estimate $E(Y|X = x)$ at x, average many observed Y with value x

Model Dependence Proof

Model Free Inference

To estimate $E(Y|X = x)$ at x, average many observed Y with value x

Assumptions (Model-Based Inference)

Model Dependence Proof

Model Free Inference

To estimate $E(Y|X = x)$ at x, average many observed Y with value x

Assumptions (Model-Based Inference)

1. Definition: model dependence at x is the difference between predicted outcomes for any two models that fit about equally well.

Model Dependence Proof

Model Free Inference

To estimate $E(Y|X = x)$ at x, average many observed Y with value x

Assumptions (Model-Based Inference)

1. Definition: model dependence at x is the difference between predicted outcomes for any two models that fit about equally well.
2. The functional form follows strong continuity (think smoothness, although it is less restrictive)

Model Dependence Proof

Model Free Inference

To estimate $E(Y|X = x)$ at x, average many observed Y with value x

Assumptions (Model-Based Inference)

1. Definition: model dependence at x is the difference between predicted outcomes for any two models that fit about equally well.
2. The functional form follows strong continuity (think smoothness, although it is less restrictive)

Result

Model Dependence Proof

Model Free Inference

To estimate $E(Y|X = x)$ at x, average many observed Y with value x

Assumptions (Model-Based Inference)

1. Definition: model dependence at x is the difference between predicted outcomes for any two models that fit about equally well.
2. The functional form follows strong continuity (think smoothness, although it is less restrictive)

Result

The maximum degree of model dependence: solely a function of the distance from the counterfactual to the data

Detecting Model Dependence

Detecting Model Dependence

A (Hypothethical) Research Design

Detecting Model Dependence

A (Hypothethical) Research Design

- Randomly select a large number of infants

Detecting Model Dependence

A (Hypothethical) Research Design

- Randomly select a large number of infants
- Randomly assign them to $\mathbf{0,6,8,10,12,16}$ years of education

Detecting Model Dependence

A (Hypothethical) Research Design

- Randomly select a large number of infants
- Randomly assign them to **0,6,8,10,12,16** years of education
- Assume 100% compliance, and no measurement error, omitted variables, or missing data

Detecting Model Dependence

A (Hypothethical) Research Design

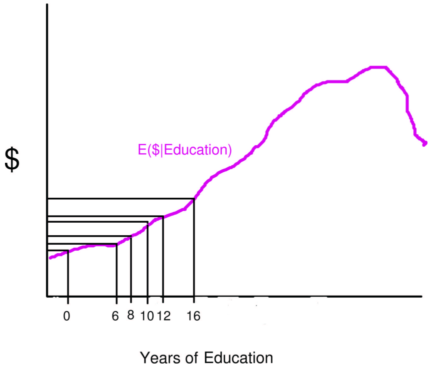
- Randomly select a large number of infants
- Randomly assign them to $\mathbf{0,6,8,10,12,16}$ years of education
- Assume 100% compliance, and no measurement error, omitted variables, or missing data
- Regress cumulative salary in year 17 on education

Detecting Model Dependence

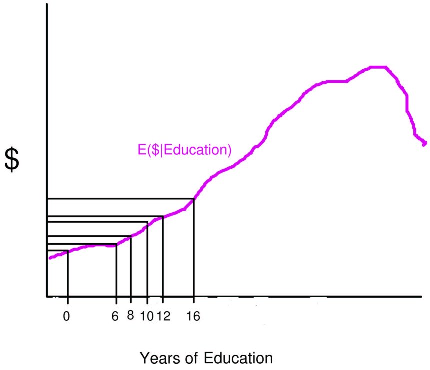
A (Hypothethical) Research Design

- Randomly select a large number of infants
- Randomly assign them to **0,6,8,10,12,16** years of education
- Assume 100% compliance, and no measurement error, omitted variables, or missing data
- Regress cumulative salary in year 17 on education
- We find a coefficient of $\hat{\beta} = \$1,000$, big t-statistics, narrow confidence intervals, and pass every test for auto-correlation, fit, normality, linearity, homoskedasticity, etc.

What Inferences Would You Be Willing to Make?

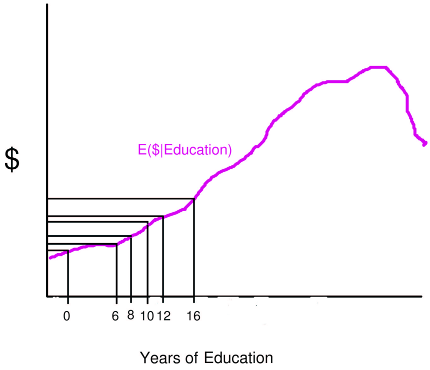


What Inferences Would You Be Willing to Make?



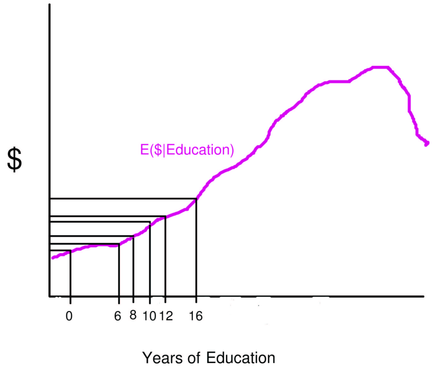
- A Factual Question: How much salary would someone receive with 12 years of education (a high school degree)?

What Inferences Would You Be Willing to Make?



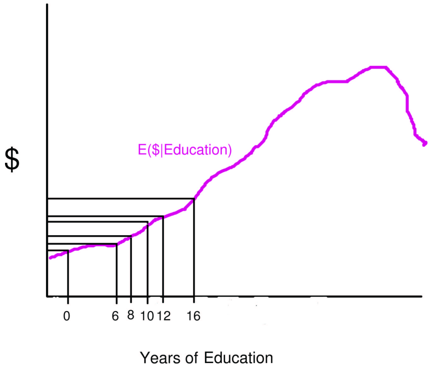
- A Factual Question: How much salary would someone receive with 12 years of education (a high school degree)?
- The model-free estimate: mean(Y) among those with $X = 12$.

What Inferences Would You Be Willing to Make?

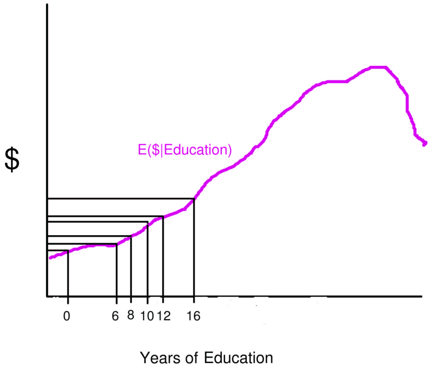


- A Factual Question: How much salary would someone receive with 12 years of education (a high school degree)?
- The model-free estimate: mean(Y) among those with $X = 12$.
- The model-based estimate: $\hat{Y} = X\hat{\beta} = 12 \times \$1,000 = \$12,000$

Counterfactual Inferences with Interpolation

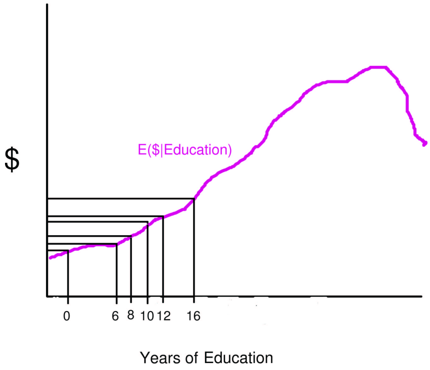


Counterfactual Inferences with Interpolation



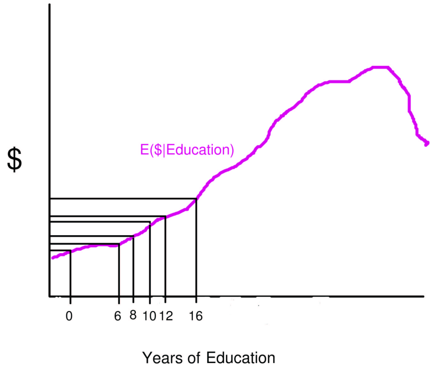
- How much salary would someone receive with 14 years of education (an Associates Degree)?

Counterfactual Inferences with Interpolation



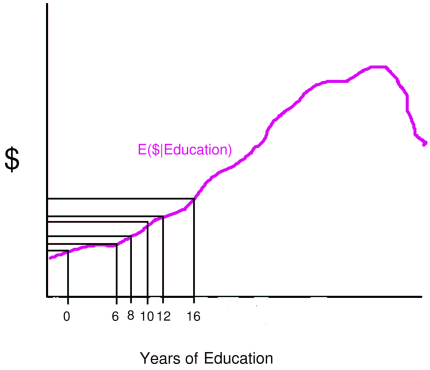
- How much salary would someone receive with 14 years of education (an Associates Degree)?
- Model free estimate: impossible

Counterfactual Inferences with Interpolation

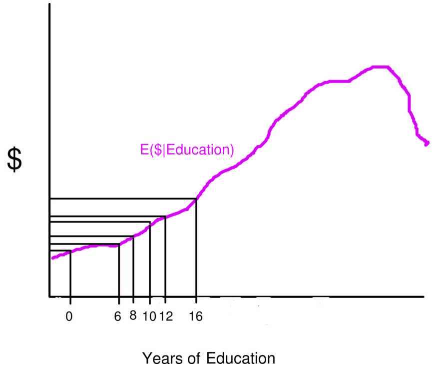


- How much salary would someone receive with 14 years of education (an Associates Degree)?
- Model free estimate: impossible
- Model-based estimate: $\hat{Y} = X\hat{\beta} = 14 \times \$1,000 = \$14,000$

Counterfactual Inference with Extrapolation

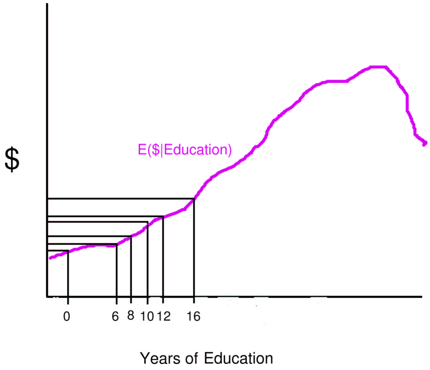


Counterfactual Inference with Extrapolation



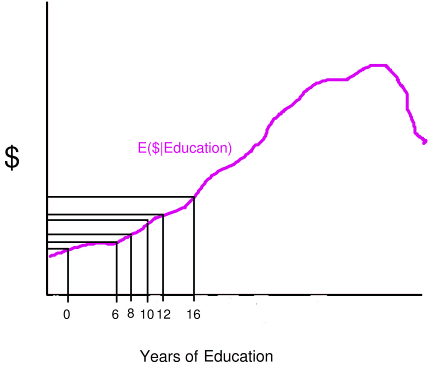
- How much salary would someone receive with 24 years of education (a Ph.D.)?

Counterfactual Inference with Extrapolation

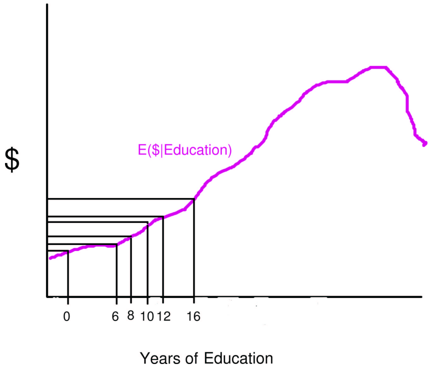


- How much salary would someone receive with 24 years of education (a Ph.D.)?
- $\hat{Y} = X\hat{\beta} = 24 \times \$1,000 = \$24,000$

Another Counterfactual Inference with Extrapolation

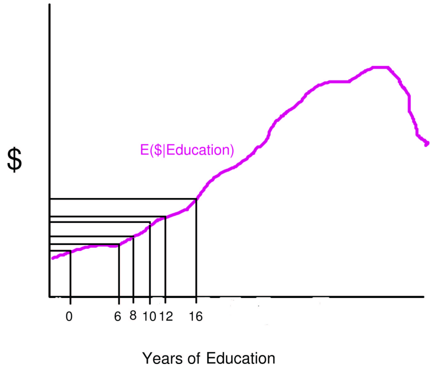


Another Counterfactual Inference with Extrapolation



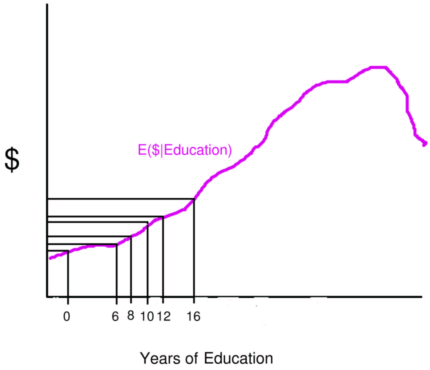
- How much salary would someone receive with 53 years of education?

Another Counterfactual Inference with Extrapolation



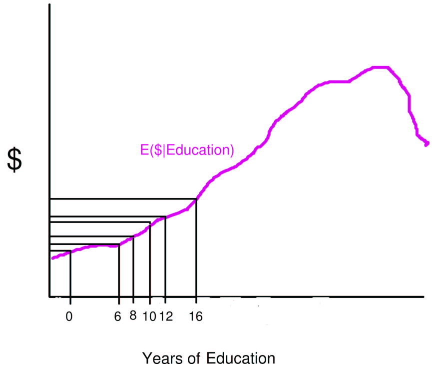
- How much salary would someone receive with 53 years of education?
- $\hat{Y} = X\hat{\beta} = 53 \times \$1,000 = \$53,000$

Another Counterfactual Inference with Extrapolation



- How much salary would someone receive with 53 years of education?
- $\hat{Y} = X\hat{\beta} = 53 \times \$1,000 = \$53,000$
- Recall: the regression passed every test and met every assumption; identical calculations worked for the other questions.

Another Counterfactual Inference with Extrapolation



- How much salary would someone receive with 53 years of education?
- $\hat{Y} = X\hat{\beta} = 53 \times \$1,000 = \$53,000$
- Recall: the regression passed every test and met every assumption; identical calculations worked for the other questions.
- What's changed? How would we recognize it when the example is less extreme or multidimensional?

Model Dependence with One Explanatory Variable

Model Dependence with One Explanatory Variable

- Suppose Y is starting salary; X is education in 10 categories.

Model Dependence with One Explanatory Variable

- Suppose Y is starting salary; X is education in 10 categories.
- To estimate $E(Y|X)$: we need 10 parameters, $E(Y|X = x_j)$, $j = 1, \ldots, 10$.

Model Dependence with One Explanatory Variable

- Suppose Y is starting salary; X is education in 10 categories.
- To estimate $E(Y|X)$: we need 10 parameters, $E(Y|X = x_j)$, $j = 1, \ldots, 10$.
- Model-free method: average 50 observations on Y for each value of X

Model Dependence with One Explanatory Variable

- Suppose Y is starting salary; X is education in 10 categories.
- To estimate $E(Y|X)$: we need 10 parameters, $E(Y|X = x_j)$, $j = 1, \ldots, 10$.
- Model-free method: average 50 observations on Y for each value of X
- Model-based method: regress Y on X, summarizing 10 parameters with 2 (intercept and slope).

Model Dependence with One Explanatory Variable

- Suppose Y is starting salary; X is education in 10 categories.
- To estimate $E(Y|X)$: we need 10 parameters, $E(Y|X = x_j)$, $j = 1, \ldots, 10$.
- Model-free method: average 50 observations on Y for each value of X
- Model-based method: regress Y on X, summarizing 10 parameters with 2 (intercept and slope).
- The difference between the 10 we need and the 2 we estimate with regression is pure assumption.

Model Dependence with One Explanatory Variable

- Suppose Y is starting salary; X is education in 10 categories.
- To estimate $E(Y|X)$: we need 10 parameters, $E(Y|X = x_j)$, $j = 1, \ldots, 10$.
- Model-free method: average 50 observations on Y for each value of X
- Model-based method: regress Y on X, summarizing 10 parameters with 2 (intercept and slope).
- The difference between the 10 we need and the 2 we estimate with regression is pure assumption.
- (If X were continuous, we would be reducing ∞ to 2, also by assumption)

Model Dependence with Two Explanatory Variables

Model Dependence with Two Explanatory Variables

Variables: X (education) and Z, parent's income, both with 10 categories

Model Dependence with Two Explanatory Variables

Variables: X (education) and Z, parent's income, both with 10 categories

- How many parameters do we now need to estimate?

Model Dependence with Two Explanatory Variables

Variables: X (education) and Z, parent's income, both with 10 categories

- How many parameters do we now need to estimate? 20?

Model Dependence with Two Explanatory Variables

Variables: X (education) and Z, parent's income, both with 10 categories

- How many parameters do we now need to estimate? 20? Nope.

Model Dependence with Two Explanatory Variables

Variables: X (education) and Z, parent's income, both with 10 categories

- How many parameters do we now need to estimate? 20? Nope. Its $10 \times 10 = 100$.

Model Dependence with Two Explanatory Variables

Variables: X (education) and Z, parent's income, both with 10 categories

- How many parameters do we now need to estimate? 20? Nope. Its $10 \times 10 = 100$. This is the curse of dimensionality: the number of parameters goes up geometrically, not additively.

Model Dependence with Two Explanatory Variables

Variables: X (education) and Z, parent's income, both with 10 categories

- How many parameters do we now need to estimate? 20? Nope. Its $10 \times 10 = 100$. This is the curse of dimensionality: the number of parameters goes up geometrically, not additively.
- If we run a regression, we are summarizing 100 parameters with 3 (an intercept and two slopes).

Model Dependence with Two Explanatory Variables

Variables: X (education) and Z, parent's income, both with 10 categories

- How many parameters do we now need to estimate? 20? Nope. Its $10 \times 10 = 100$. This is the curse of dimensionality: the number of parameters goes up geometrically, not additively.
- If we run a regression, we are summarizing 100 parameters with 3 (an intercept and two slopes).
- But what about including an interaction? Right, so now we're summarizing 100 parameters with 4.

Model Dependence with Two Explanatory Variables

Variables: X (education) and Z, parent's income, both with 10 categories

- How many parameters do we now need to estimate? 20? Nope. Its $10 \times 10 = 100$. This is the curse of dimensionality: the number of parameters goes up geometrically, not additively.
- If we run a regression, we are summarizing 100 parameters with 3 (an intercept and two slopes).
- But what about including an interaction? Right, so now we're summarizing 100 parameters with 4.
- The difference: an enormous assumption based on convenience, not evidence or theory.

Model Dependence with Many Explanatory Variables

Model Dependence with Many Explanatory Variables

- Suppose: 15 explanatory variables, with 10 categories each.

Model Dependence with Many Explanatory Variables

- Suppose: 15 explanatory variables, with 10 categories each.
  - need to estimate 10^{15} (a quadrillion) parameters with how many observations?

Model Dependence with Many Explanatory Variables

- Suppose: 15 explanatory variables, with 10 categories each.
  - need to estimate 10^{15} (a quadrillion) parameters with how many observations?
  - Regression reduces this to 16 parameters; quite an assumption!

Model Dependence with Many Explanatory Variables

- Suppose: 15 explanatory variables, with 10 categories each.
  - need to estimate 10^{15} (a quadrillion) parameters with how many observations?
  - Regression reduces this to 16 parameters; quite an assumption!
- Suppose: 80 explanatory variables.

Model Dependence with Many Explanatory Variables

- Suppose: 15 explanatory variables, with 10 categories each.
  - need to estimate 10^{15} (a quadrillion) parameters with how many observations?
  - Regression reduces this to 16 parameters; quite an assumption!
- Suppose: 80 explanatory variables.
  - 10^{80} is more than the number of atoms in the universe.

Model Dependence with Many Explanatory Variables

- Suppose: 15 explanatory variables, with 10 categories each.
  - need to estimate 10^{15} (a quadrillion) parameters with how many observations?
  - Regression reduces this to 16 parameters; quite an assumption!
- Suppose: 80 explanatory variables.
  - 10^{80} is more than the number of atoms in the universe.
  - Yet, with a few simple assumptions, we can still run a regression and estimate only 81 parameters.

Model Dependence with Many Explanatory Variables

- Suppose: 15 explanatory variables, with 10 categories each.
  - need to estimate 10^{15} (a quadrillion) parameters with how many observations?
  - Regression reduces this to 16 parameters; quite an assumption!
- Suppose: 80 explanatory variables.
  - 10^{80} is more than the number of atoms in the universe.
  - Yet, with a few simple assumptions, we can still run a regression and estimate only 81 parameters.
- The curse of dimensionality introduces huge assumptions, often unrecognized.

How Factual is your Counterfactual?

How Factual is your Counterfactual?

- Is your counterfactual close enough to data so that statistical methods provide *empirical* answers?

How Factual is your Counterfactual?

- Is your counterfactual close enough to data so that statistical methods provide *empirical* answers?
- If not, the same calculations will be based on indefensible model assumptions. With the curse of dimensionality, its too easy to fall into this trap.

How Factual is your Counterfactual?

- Is your counterfactual close enough to data so that statistical methods provide *empirical* answers?
- If not, the same calculations will be based on indefensible model assumptions. With the curse of dimensionality, its too easy to fall into this trap.
- A good existing approach: *Sensitivity testing*, but this requires the user to specify a class of models and then to estimate them all and check how much inferences change

How Factual is your Counterfactual?

- Is your counterfactual close enough to data so that statistical methods provide *empirical* answers?
- If not, the same calculations will be based on indefensible model assumptions. With the curse of dimensionality, its too easy to fall into this trap.
- A good existing approach: *Sensitivity testing*, but this requires the user to specify a class of models and then to estimate them all and check how much inferences change
- King/Zeng "Convex Hull" approach:

How Factual is your Counterfactual?

- Is your counterfactual close enough to data so that statistical methods provide *empirical* answers?
- If not, the same calculations will be based on indefensible model assumptions. With the curse of dimensionality, its too easy to fall into this trap.
- A good existing approach: *Sensitivity testing*, but this requires the user to specify a class of models and then to estimate them all and check how much inferences change
- King/Zeng "Convex Hull" approach:
  - Specify your explanatory variables, X.

How Factual is your Counterfactual?

- Is your counterfactual close enough to data so that statistical methods provide *empirical* answers?
- If not, the same calculations will be based on indefensible model assumptions. With the curse of dimensionality, its too easy to fall into this trap.
- A good existing approach: *Sensitivity testing*, but this requires the user to specify a class of models and then to estimate them all and check how much inferences change
- King/Zeng "Convex Hull" approach:
  - Specify your explanatory variables, X.
  - Assume $\text{E}(Y|X)$ is (minimally) smooth in X

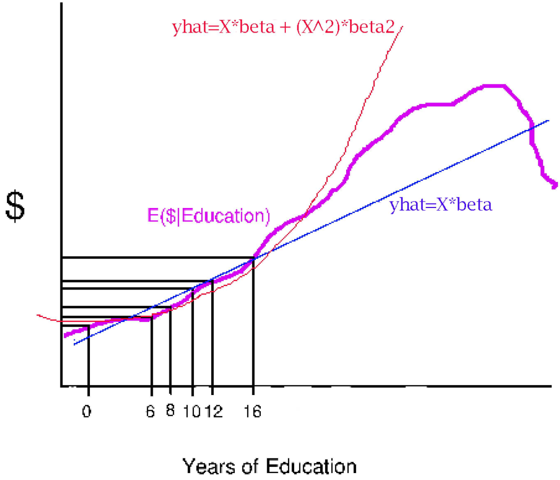
How Factual is your Counterfactual?

- Is your counterfactual close enough to data so that statistical methods provide *empirical* answers?
- If not, the same calculations will be based on indefensible model assumptions. With the curse of dimensionality, its too easy to fall into this trap.
- A good existing approach: *Sensitivity testing*, but this requires the user to specify a class of models and then to estimate them all and check how much inferences change
- King/Zeng "Convex Hull" approach:
  - Specify your explanatory variables, X.
  - Assume $\mathrm{E}(Y|X)$ is (minimally) smooth in X
  - No need to specify models (or a class of models), estimators, or dependent variables.

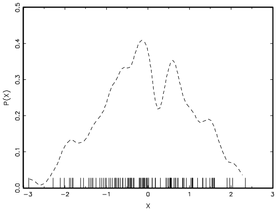
How Factual is your Counterfactual?

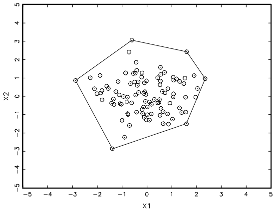
- Is your counterfactual close enough to data so that statistical methods provide *empirical* answers?
- If not, the same calculations will be based on indefensible model assumptions. With the curse of dimensionality, its too easy to fall into this trap.
- A good existing approach: *Sensitivity testing*, but this requires the user to specify a class of models and then to estimate them all and check how much inferences change
- King/Zeng "Convex Hull" approach:
  - Specify your explanatory variables, X.
  - Assume $\mathrm{E}(Y|X)$ is (minimally) smooth in X
  - No need to specify models (or a class of models), estimators, or dependent variables.
  - Results of one run apply to the class of all models, all estimators, and all dependent variables.

Interpolation vs Extrapolation in one Dimension

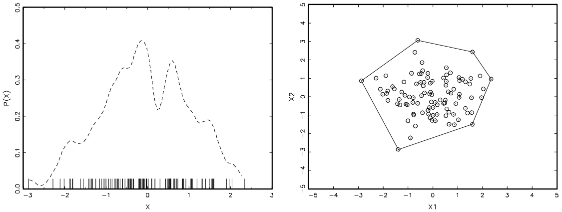


Interpolation or Extrapolation in One and Two Dimensions



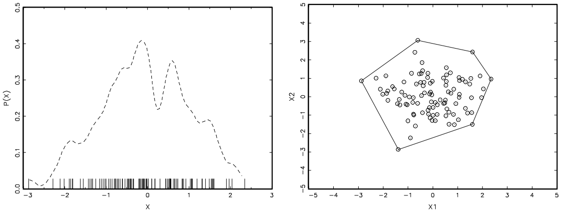


Interpolation or Extrapolation in One and Two Dimensions



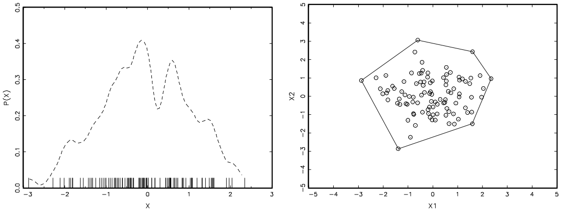
- Interpolation: Inside the convex hull

Interpolation or Extrapolation in One and Two Dimensions



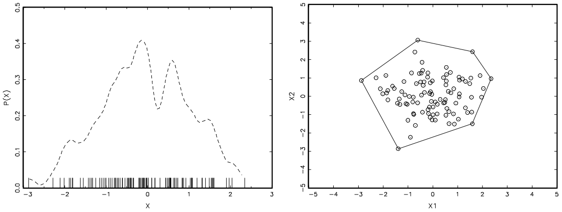
- Interpolation: Inside the convex hull
- Extrapolation: Outside the convex hull

Interpolation or Extrapolation in One and Two Dimensions



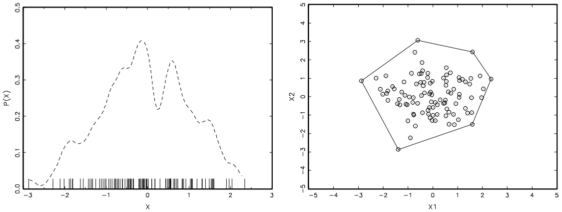
- Interpolation: Inside the convex hull
- Extrapolation: Outside the convex hull
- Calculating the convex hull would take forever in high-dimensions

Interpolation or Extrapolation in One and Two Dimensions



- Interpolation: Inside the convex hull
- Extrapolation: Outside the convex hull
- Calculating the convex hull would take forever in high-dimensions
- `WhatIf` package uses linear programming to check if a candidate point is inside the hull

Interpolation or Extrapolation in One and Two Dimensions



- Interpolation: Inside the convex hull
- Extrapolation: Outside the convex hull
- Calculating the convex hull would take forever in high-dimensions
- `WhatIf` package uses linear programming to check if a candidate point is inside the hull
- The key idea is making sure your counterfactual is near the data!

Replication: Doyle and Sambanis, APSR 2000

Replication: Doyle and Sambanis, APSR 2000

- Data: 124 Post-World War II civil wars

Replication: Doyle and Sambanis, APSR 2000

- Data: 124 Post-World War II civil wars
- Dependent variable: peacebuilding success

Replication: Doyle and Sambanis, APSR 2000

- Data: 124 Post-World War II civil wars
- Dependent variable: peacebuilding success
- Treatment variable: multilateral UN peacekeeping intervention (0/1)

Replication: Doyle and Sambanis, APSR 2000

- Data: 124 Post-World War II civil wars
- Dependent variable: peacebuilding success
- Treatment variable: multilateral UN peacekeeping intervention (0/1)
- Control variables: war type, severity, and duration; development status; etc...

Replication: Doyle and Sambanis, APSR 2000

- Data: 124 Post-World War II civil wars
- Dependent variable: peacebuilding success
- Treatment variable: multilateral UN peacekeeping intervention (0/1)
- Control variables: war type, severity, and duration; development status; etc...
- Counterfactuals: UN intervention switched (0/1 to 1/0) for each observation

Replication: Doyle and Sambanis, APSR 2000

- Data: 124 Post-World War II civil wars
- Dependent variable: peacebuilding success
- Treatment variable: multilateral UN peacekeeping intervention (0/1)
- Control variables: war type, severity, and duration; development status; etc...
- Counterfactuals: UN intervention switched (0/1 to 1/0) for each observation
- Percent of counterfactuals in the convex hull:

Replication: Doyle and Sambanis, APSR 2000

- Data: 124 Post-World War II civil wars
- Dependent variable: peacebuilding success
- Treatment variable: multilateral UN peacekeeping intervention (0/1)
- Control variables: war type, severity, and duration; development status; etc...
- Counterfactuals: UN intervention switched (0/1 to 1/0) for each observation
- Percent of counterfactuals in the convex hull: 0%

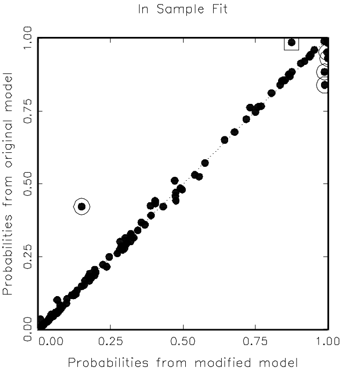
Replication: Doyle and Sambanis, APSR 2000

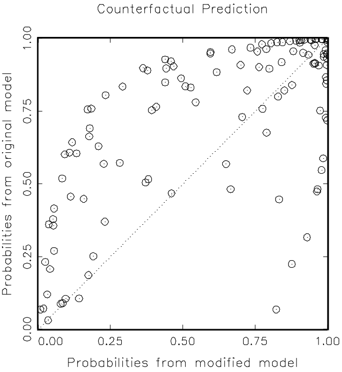
- Data: 124 Post-World War II civil wars
- Dependent variable: peacebuilding success
- Treatment variable: multilateral UN peacekeeping intervention (0/1)
- Control variables: war type, severity, and duration; development status; etc...
- Counterfactuals: UN intervention switched (0/1 to 1/0) for each observation
- Percent of counterfactuals in the convex hull: 0%
- Thus, without estimating any models, we know inferences will be model dependent; for illustration, here is an example. . . .

Doyle and Sambanis, Logit Model

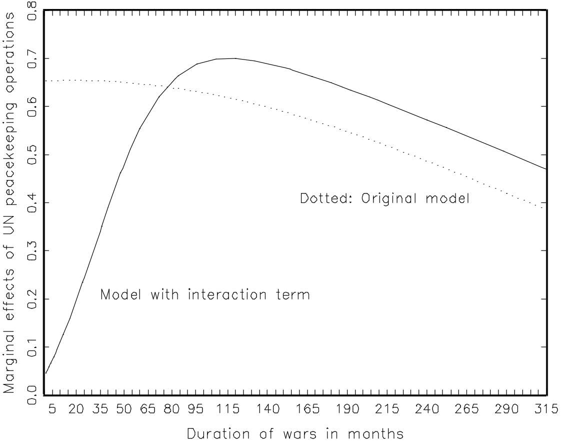
	Original Model			Modified Model		
Variables	Coeff	SE	P-val	Coeff	SE	P-val
Wartype	−1.742	.609	.004	−1.666	.606	.006
Logdead	−.445	.126	.000	−.437	.125	.000
Wardur	.006	.006	.258	.006	.006	.342
Factnum	−1.259	.703	.073	−1.045	.899	.245
Factnum2	.062	.065	.346	.032	.104	.756
Trnsfcap	.004	.002	.010	.004	.002	.017
Develop	.001	.000	.065	.001	.000	.068
Exp	−6.016	3.071	.050	−6.215	3.065	.043
Decade	−.299	.169	.077	−0.284	.169	.093
Treaty	2.124	.821	.010	2.126	.802	.008
UNOP4	3.135	1.091	.004	.262	1.392	.851
Wardur*UNOP4	—	—	—	.037	.011	.001
Constant	8.609	2.157	0.000	7.978	2.350	.000
N		122			122	
Log-likelihood		-45.649			-44.902	
Pseudo R^2		.423			.433	

Doyle and Sambanis: Model Dependence





UN Peacekeeping Operations



Another Example

Remember our negative binomial model?

```
mod <- zelig(repdeaths ~ cathunemp + protunemp,
        data = troubles, model = "negbin")
summary(mod)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.5943     0.1718   3.459 0.000541 ***
cathunemp     7.9323     0.9150   8.669  < 2e-16 ***
protunemp   -19.1683     2.3713  -8.084 6.29e-16 ***
```

Proposed Counterfactuals

Proposed Counterfactuals

Let's consider two first differences we might plausibly estimate.

Proposed Counterfactuals

Let's consider two first differences we might plausibly estimate. At the baseline, both variables are assumed fixed at their sample means.

Proposed Counterfactuals

Let's consider two first differences we might plausibly estimate. At the baseline, both variables are assumed fixed at their sample means.

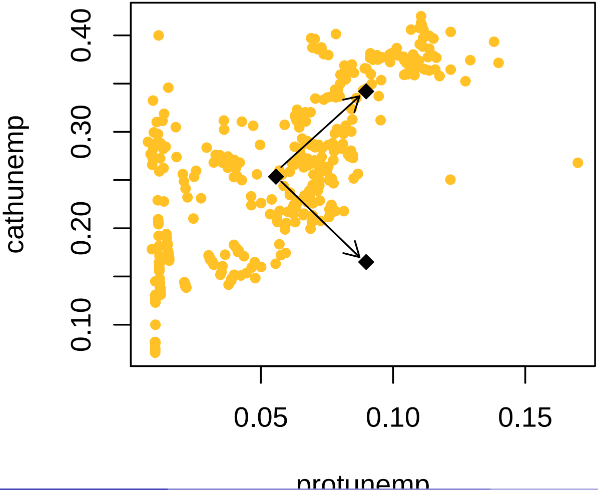
1. Counterfactual 1: Catholic unemployment increases by one standard deviation and Protestant unemployment increases by one standard deviation.

Proposed Counterfactuals

Let's consider two first differences we might plausibly estimate. At the baseline, both variables are assumed fixed at their sample means.

1. Counterfactual 1: Catholic unemployment increases by one standard deviation and Protestant unemployment increases by one standard deviation.
2. Counterfactual 2: Catholic unemployment decreases by one standard deviation and Protestant unemployment increases by one standard deviation.

Proposed Counterfactuals Plotted



Checking the Convex Hull

```
library(WhatIf)
cf1 <- cbind(mean(cathunemp) + sd(cathunemp),
        mean(protunemp) + sd(protunemp))
cf2 <- cbind(mean(cathunemp) - sd(cathunemp),
        mean(protunemp) + sd(protunemp))

cf.res1 <- whatif(data = mod, cfact = cf1)
> cf.res1$in.hull
[1] TRUE

cf.res2 <- whatif(data = mod, cfact = cf2)
cf.res2$in.hull
[1] FALSE
```

A Measure of Distance

A Measure of Distance

The `whatif` function also tells us the percentage of data points within 1 geometric variance of the counterfactual.

```
> cf.res1$sum.stat
        1
0.2608696

> cf.res2$sum.stat
         1
0.04603581
```

A Measure of Distance

The `whatif` function also tells us the percentage of data points within 1 geometric variance of the counterfactual.

```
> cf.res1$sum.stat
        1
0.2608696

> cf.res2$sum.stat
         1
0.04603581
```

The geometric variance is a generalization of the usual variance which is more suitable to discrete and continuous variables- essentially it is the average pairwise Gower distance in the data. The number of GV's away can be altered with the `nearby` argument.

Biases in Regression: A Decomposition

$$d = \text{mean}(Y|D = 1) - \text{mean}(Y|D = 0)$$

Biases in Regression: A Decomposition

$$d = \text{mean}(Y|D = 1) - \text{mean}(Y|D = 0)$$

Biases in Regression: A Decomposition

$$d = \text{mean}(Y|D = 1) - \text{mean}(Y|D = 0)$$

$$\text{bias} \equiv E(d) - \theta$$

Biases in Regression: A Decomposition

$$d = \text{mean}(Y|D=1) - \text{mean}(Y|D=0)$$

$$\text{bias} \equiv E(d) - \theta = \Delta_o + \Delta_p + \Delta_i + \Delta_e$$

Biases in Regression: A Decomposition

$$d = \text{mean}(Y|D = 1) - \text{mean}(Y|D = 0)$$

$$\text{bias} \equiv E(d) - \theta = \Delta_o + \Delta_p + \Delta_i + \Delta_e$$

- Δ_o Omitted variable bias (ignorability)

Biases in Regression: A Decomposition

$$d = \text{mean}(Y|D=1) - \text{mean}(Y|D=0)$$

$$\text{bias} \equiv E(d) - \theta = \Delta_o + \Delta_p + \Delta_i + \Delta_e$$

- Δ_o Omitted variable bias (ignorability)
- Δ_p Post-treatment bias (check this with theory!)

Biases in Regression: A Decomposition

$$d = \text{mean}(Y|D = 1) - \text{mean}(Y|D = 0)$$

$$\text{bias} \equiv E(d) - \theta = \Delta_o + \Delta_p + \Delta_i + \Delta_e$$

- Δ_o Omitted variable bias (ignorability)
- Δ_p Post-treatment bias (check this with theory!)
- Δ_i Interpolation bias (use models or matching)

Biases in Regression: A Decomposition

$$d = \text{mean}(Y|D=1) - \text{mean}(Y|D=0)$$

$$\text{bias} \equiv E(d) - \theta = \Delta_o + \Delta_p + \Delta_i + \Delta_e$$

- Δ_o Omitted variable bias (ignorability)
- Δ_p Post-treatment bias (check this with theory!)
- Δ_i Interpolation bias (use models or matching)
- Δ_e Extrapolation bias (check this with data!)

Counterfactuals Summary

- When the model is true, we can extrapolate or interpolate as desired.

Counterfactuals Summary

- When the model is true, we can extrapolate or interpolate as desired.
- In practical settings we do not believe the model is true, even if it is locally accurate

Counterfactuals Summary

- When the model is true, we can extrapolate or interpolate as desired.
- In practical settings we do not believe the model is true, even if it is locally accurate
- Thus we may get wildly different counterfactuals from different models when we are far from the data, we call this model dependence

Counterfactuals Summary

- When the model is true, we can extrapolate or interpolate as desired.
- In practical settings we do not believe the model is true, even if it is locally accurate
- Thus we may get wildly different counterfactuals from different models when we are far from the data, we call this model dependence
- The convex hull provides a way to check for extrapolation

Counterfactuals Summary

- When the model is true, we can extrapolate or interpolate as desired.
- In practical settings we do not believe the model is true, even if it is locally accurate
- Thus we may get wildly different counterfactuals from different models when we are far from the data, we call this model dependence
- The convex hull provides a way to check for extrapolation
- This is a great way of assessing the reasonableness of our simulated quantities of interest

Conditioning

Conditioning

- In observational data we need to understand assignment into treatment

Conditioning

- In observational data we need to understand assignment into treatment
- We control for confounders to achieve identification under selection on observables

Conditioning

- In observational data we need to understand assignment into treatment
- We control for confounders to achieve identification under selection on observables
- Loosely this means that we have sufficient background covariates that we can explain away the common causes of treatment and outcome

Conditioning

- In observational data we need to understand assignment into treatment
- We control for confounders to achieve identification under selection on observables
- Loosely this means that we have sufficient background covariates that we can explain away the common causes of treatment and outcome
- When we have the right set of observed confounders, matching is a strategy that helps reduce model dependence in this conditioning

Conditioning

- In observational data we need to understand assignment into treatment
- We control for confounders to achieve identification under selection on observables
- Loosely this means that we have sufficient background covariates that we can explain away the common causes of treatment and outcome
- When we have the right set of observed confounders, matching is a strategy that helps reduce model dependence in this conditioning
- Matching itself is not an identification strategy, nor is it fundamentally different than alternatives like weighting or regression adjustment

Repeat After Me

Repeat After Me

Matching is not an identification strategy.

Repeat After Me

Matching is not an identification strategy.

So what is?

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)*

Identification Result

Given selection on observables we have

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

Therefore, under the common support condition:

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

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)*

Identification Result

Similarly,

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

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	$\mathbf{E}[Y_1 \mid X = 0, D = 1]$	$\mathbf{E}[Y_0 \mid X = 0, D = 1]$	1	0
2			1	0
3	$\mathbf{E}[Y_1 \mid X = 0, D = 0]$	$\mathbf{E}[Y_0 \mid X = 0, D = 0]$	0	0
4			0	0
5	$\mathbf{E}[Y_1 \mid X = 1, D = 1]$	$\mathbf{E}[Y_0 \mid X = 1, D = 1]$	1	1
6			1	1
7	$\mathbf{E}[Y_1 \mid X = 1, D = 0]$	$\mathbf{E}[Y_0 \mid X = 1, D = 0]$	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	$\mathbf{E}[Y_1 \mid X=0, D=1]$	$\mathbf{E}[Y_0 \mid X=0, D=1]=$	1	0
2		$\mathbf{E}[Y_0 \mid X=0, D=0]$	1	0
3	$\mathbf{E}[Y_1 \mid X=0, D=0]$	$\mathbf{E}[Y_0 \mid X=0, D=0]$	0	0
4			0	0
5	$\mathbf{E}[Y_1 \mid X=1, D=1]$	$\mathbf{E}[Y_0 \mid X=1, D=1]=$	1	1
6		$\mathbf{E}[Y_0 \mid X=1, D=0]$	1	1
7	$\mathbf{E}[Y_1 \mid X=1, D=0]$	$\mathbf{E}[Y_0 \mid X=1, D=0]$	0	1
8			0	1

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

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

Identification Under Selection on Observables

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

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

$$\begin{aligned} \mathbf{E}[Y_1|X=0, D=1] &= \mathbf{E}[Y_1|X=0, D=0] \text{ and} \\ \mathbf{E}[Y_1|X=1, D=1] &= \mathbf{E}[Y_1|X=1, D=0] \end{aligned}$$

Why Match?

Why Match?

- When selection on observables holds, we still need to adjust for X_i

Why Match?

- When selection on observables holds, we still need to adjust for X_i
- Common solution: write a parametric model for $\mathbf{E}[Y_i(d)|X_i]$

Why Match?

- When selection on observables holds, we still need to adjust for X_i
- Common solution: write a parametric model for $\mathbf{E}[Y_i(d)|X_i]$
  - For example, could assume it is linear: $\mathbf{E}[Y_i(d)|X_i] = X_i'\beta$

Why Match?

- When selection on observables holds, we still need to adjust for X_i
- Common solution: write a parametric model for $\mathbf{E}[Y_i(d)|X_i]$
  - For example, could assume it is linear: $\mathbf{E}[Y_i(d)|X_i] = X_i'\beta$
  - Regression, MLE, Bayes, etc.

Why Match?

- When selection on observables holds, we still need to adjust for X_i
- Common solution: write a parametric model for $\mathbf{E}[Y_i(d)|X_i]$
  - For example, could assume it is linear: $\mathbf{E}[Y_i(d)|X_i] = X_i'\beta$
  - Regression, MLE, Bayes, etc.
  - But this model might be wrong $\rightsquigarrow$ wrong causal estimates.

Why Match?

- When selection on observables holds, we still need to adjust for X_i
- Common solution: write a parametric model for $\mathbf{E}[Y_i(d)|X_i]$
  - For example, could assume it is linear: $\mathbf{E}[Y_i(d)|X_i] = X_i'\beta$
  - Regression, MLE, Bayes, etc.
  - But this model might be wrong $\rightsquigarrow$ wrong causal estimates.
- Matching has two benefits:

Why Match?

- When selection on observables holds, we still need to adjust for X_i
- Common solution: write a parametric model for $\mathbf{E}[Y_i(d)|X_i]$
  - For example, could assume it is linear: $\mathbf{E}[Y_i(d)|X_i] = X_i'\beta$
  - Regression, MLE, Bayes, etc.
  - But this model might be wrong $\rightsquigarrow$ wrong causal estimates.
- Matching has two benefits:
  1. Can simplify the analysis of causal effects

Why Match?

- When selection on observables holds, we still need to adjust for X_i
- Common solution: write a parametric model for $\mathbf{E}[Y_i(d)|X_i]$
  - For example, could assume it is linear: $\mathbf{E}[Y_i(d)|X_i] = X_i'\beta$
  - Regression, MLE, Bayes, etc.
  - But this model might be wrong $\rightsquigarrow$ wrong causal estimates.
- Matching has two benefits:
  1. Can simplify the analysis of causal effects
  2. Reduces dependence of estimates on parametric models.

Overview of Matching

Overview of Matching

- Goal: reduce model dependence in our matching approach

Overview of Matching

- Goal: reduce model dependence in our matching approach
- Makes parametric models work better rather than substitute for them (i.e,. matching is not an estimator; its a preprocessing method)

Overview of Matching

- Goal: reduce model dependence in our matching approach
- Makes parametric models work better rather than substitute for them (i.e,. matching is not an estimator; its a preprocessing method)
- Might be better called pruning (no bias is introduced if pruning is a function of T and X, but not Y, although quantity of interest changes)

Overview of Matching

- Goal: reduce model dependence in our matching approach
- Makes parametric models work better rather than substitute for them (i.e,. matching is not an estimator; its a preprocessing method)
- Might be better called pruning (no bias is introduced if pruning is a function of T and X, but not Y, although quantity of interest changes)
- Apply model to preprocessed (pruned) rather than raw data

Overview of Matching

- Goal: reduce model dependence in our matching approach
- Makes parametric models work better rather than substitute for them (i.e,. matching is not an estimator; its a preprocessing method)
- Might be better called pruning (no bias is introduced if pruning is a function of T and X, but not Y, although quantity of interest changes)
- Apply model to preprocessed (pruned) rather than raw data
- Violates the "more data is better" principle, but that only applies when you know the DGP

Overview of Matching

- Goal: reduce model dependence in our matching approach
- Makes parametric models work better rather than substitute for them (i.e,. matching is not an estimator; its a preprocessing method)
- Might be better called pruning (no bias is introduced if pruning is a function of T and X, but not Y, although quantity of interest changes)
- Apply model to preprocessed (pruned) rather than raw data
- Violates the "more data is better" principle, but that only applies when you know the DGP
- Overall idea:

Overview of Matching

- Goal: reduce model dependence in our matching approach
- Makes parametric models work better rather than substitute for them (i.e,. matching is not an estimator; its a preprocessing method)
- Might be better called pruning (no bias is introduced if pruning is a function of T and X, but not Y, although quantity of interest changes)
- Apply model to preprocessed (pruned) rather than raw data
- Violates the "more data is better" principle, but that only applies when you know the DGP
- Overall idea:
  - If each treated unit exactly matches a control unit w.r.t. X, then:

Overview of Matching

- Goal: reduce model dependence in our matching approach
- Makes parametric models work better rather than substitute for them (i.e,. matching is not an estimator; its a preprocessing method)
- Might be better called pruning (no bias is introduced if pruning is a function of T and X, but not Y, although quantity of interest changes)
- Apply model to preprocessed (pruned) rather than raw data
- Violates the "more data is better" principle, but that only applies when you know the DGP
- Overall idea:
  - If each treated unit exactly matches a control unit w.r.t. X, then:

Overview of Matching

- Goal: reduce model dependence in our matching approach
- Makes parametric models work better rather than substitute for them (i.e,. matching is not an estimator; its a preprocessing method)
- Might be better called pruning (no bias is introduced if pruning is a function of T and X, but not Y, although quantity of interest changes)
- Apply model to preprocessed (pruned) rather than raw data
- Violates the "more data is better" principle, but that only applies when you know the DGP
- Overall idea:
    - If each treated unit exactly matches a control unit w.r.t. X, then: (1) treated and control groups are identical,

Overview of Matching

- Goal: reduce model dependence in our matching approach
- Makes parametric models work better rather than substitute for them (i.e,. matching is not an estimator; its a preprocessing method)
- Might be better called pruning (no bias is introduced if pruning is a function of T and X, but not Y, although quantity of interest changes)
- Apply model to preprocessed (pruned) rather than raw data
- Violates the "more data is better" principle, but that only applies when you know the DGP
- Overall idea:
  - If each treated unit exactly matches a control unit w.r.t. X, then: (1) treated and control groups are identical, (2) X is no longer a confounder,

Overview of Matching

- Goal: reduce model dependence in our matching approach
- Makes parametric models work better rather than substitute for them (i.e,. matching is not an estimator; its a preprocessing method)
- Might be better called pruning (no bias is introduced if pruning is a function of T and X, but not Y, although quantity of interest changes)
- Apply model to preprocessed (pruned) rather than raw data
- Violates the "more data is better" principle, but that only applies when you know the DGP
- Overall idea:
  - If each treated unit exactly matches a control unit w.r.t. X, then: (1) treated and control groups are identical, (2) X is no longer a confounder, (3) no need to worry about the functional form ($\bar{Y}_T - \bar{Y}_C$ is good enough).

Overview of Matching

- Goal: reduce model dependence in our matching approach
- Makes parametric models work better rather than substitute for them (i.e,. matching is not an estimator; its a preprocessing method)
- Might be better called pruning (no bias is introduced if pruning is a function of T and X, but not Y, although quantity of interest changes)
- Apply model to preprocessed (pruned) rather than raw data
- Violates the "more data is better" principle, but that only applies when you know the DGP
- Overall idea:
  - If each treated unit exactly matches a control unit w.r.t. X, then: (1) treated and control groups are identical, (2) X is no longer a confounder, (3) no need to worry about the functional form ($\bar{Y}_T - \bar{Y}_C$ is good enough).
  - If treated and control groups are better balanced than when you started, due to pruning, model dependence is reduced

Matching as Preprocessing

Matching as Preprocessing

- Y_i dep var, T_i (1=treated, 0=control), X_i confounders

Matching as Preprocessing

- Y_i dep var, T_i (1=treated, 0=control), X_i confounders
- Treatment Effect for <u>treated</u> observation i:

Matching as Preprocessing

- Y_i dep var, T_i (1=treated, 0=control), X_i confounders
- Treatment Effect for treated observation i:

$$\text{TE}_i = Y_i(1) - Y_i(0)$$

Matching as Preprocessing

- Y_i dep var, T_i (1=treated, 0=control), X_i confounders
- Treatment Effect for treated observation i:

$$\begin{aligned}\text{TE}_i &= Y_i(1) - Y_i(0) \\ &= \text{observed} - \textit{unobserved}\end{aligned}$$

Matching as Preprocessing

- Y_i dep var, T_i (1=treated, 0=control), X_i confounders
- Treatment Effect for treated observation i:

$$\begin{aligned}\text{TE}_i &= Y_i - Y_i(0) \\ &= \text{observed} - \textit{unobserved}\end{aligned}$$

Matching as Preprocessing

- Y_i dep var, T_i (1=treated, 0=control), X_i confounders
- Treatment Effect for <u>treated</u> observation i:

$$\begin{aligned}\text{TE}_i &= Y_i - Y_i(0) \\ &= \text{observed} - \textit{unobserved}\end{aligned}$$

- Estimate $Y_i(0)$ with Y_j from matched ($X_i \approx X_j$) control

Matching as Preprocessing

- Y_i dep var, T_i (1=treated, 0=control), X_i confounders
- Treatment Effect for <u>treated</u> observation i:

$$\begin{aligned} \text{TE}_i &= Y_i - {\color{red}Y_i(0)} \\ &= \text{observed} - {\color{red}\textit{unobserved}} \end{aligned}$$

- Estimate $Y_i(0)$ with Y_j from matched ($X_i \approx X_j$) control
- Prune nonmatches: reduces imbalance & model dependence

Matching as Preprocessing

- Y_i dep var, T_i (1=treated, 0=control), X_i confounders
- Treatment Effect for treated observation i:

$$\begin{aligned} \text{TE}_i &= Y_i - Y_i(0) \\ &= \text{observed} - \textit{unobserved} \end{aligned}$$

- Estimate $Y_i(0)$ with Y_j from matched ($X_i \approx X_j$) control
- Prune nonmatches: reduces imbalance & model dependence
- Follow preprocessing with whatever statistical method you'd have used without matching

Matching as Preprocessing

- Y_i dep var, T_i (1=treated, 0=control), X_i confounders
- Treatment Effect for treated observation i:

$$\begin{aligned} \text{TE}_i &= Y_i - Y_i(0) \\ &= \text{observed} - \textit{unobserved} \end{aligned}$$

- Estimate $Y_i(0)$ with Y_j from matched ($X_i \approx X_j$) control
- Prune nonmatches: reduces imbalance & model dependence
- Follow preprocessing with whatever statistical method you'd have used without matching

Matching as Preprocessing

- Y_i dep var, T_i (1=treated, 0=control), X_i confounders
- Treatment Effect for treated observation i:

$$\begin{aligned} \text{TE}_i &= Y_i - Y_i(0) \\ &= \text{observed} - \textit{unobserved} \end{aligned}$$

- Estimate $Y_i(0)$ with Y_j from matched ($X_i \approx X_j$) control
- Prune nonmatches: reduces imbalance & model dependence
- Follow preprocessing with whatever statistical method you'd have used without matching

Warning: Pruning nonmatches can change your estimand.

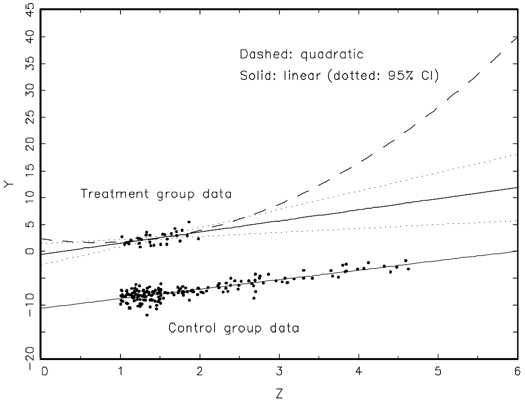
How Matching Helps with Model Dependence

How Matching Helps with Model Dependence

(King and Zeng, 2006: fig.4 *Political Analysis*)

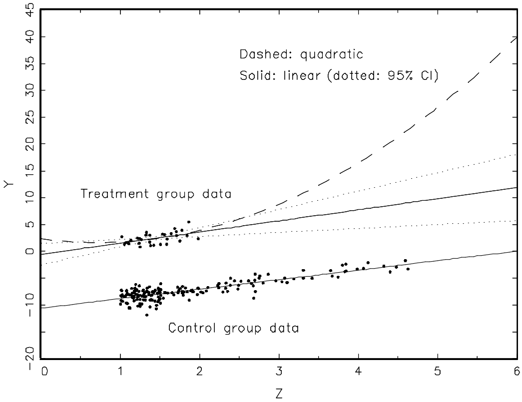
How Matching Helps with Model Dependence

(King and Zeng, 2006: fig.4 *Political Analysis*)



How Matching Helps with Model Dependence

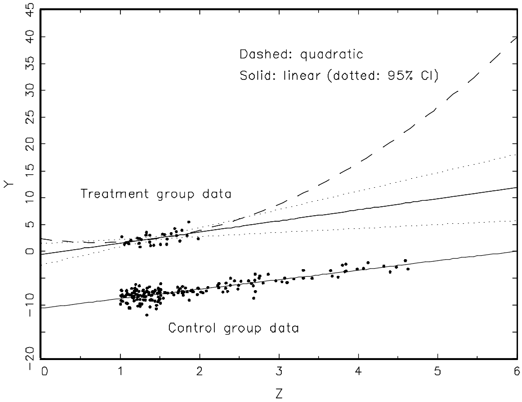
(King and Zeng, 2006: fig.4 *Political Analysis*)



What to do?

How Matching Helps with Model Dependence

(King and Zeng, 2006: fig.4 *Political Analysis*)

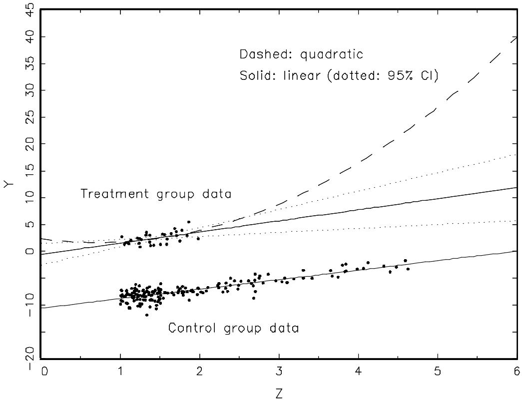


What to do?

- Preprocess I: Eliminate extrapolation region

How Matching Helps with Model Dependence

(King and Zeng, 2006: fig.4 *Political Analysis*)

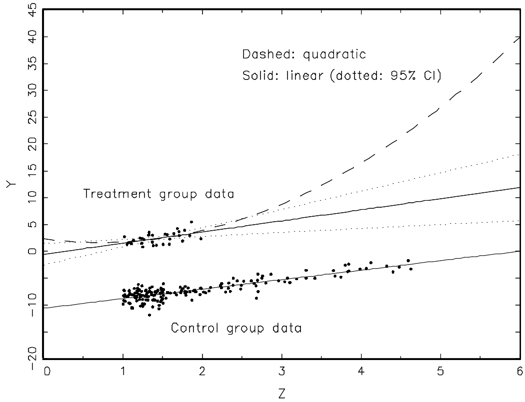


What to do?

- Preprocess I: Eliminate extrapolation region
- Preprocess II: Match (prune) within interpolation region

How Matching Helps with Model Dependence

(King and Zeng, 2006: fig.4 *Political Analysis*)

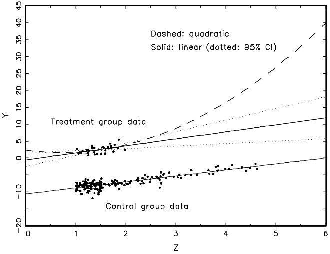


What to do?

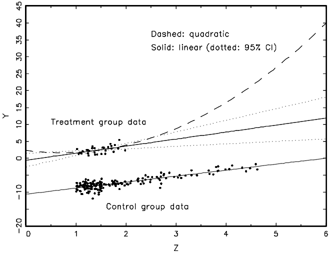
- Preprocess I: Eliminate extrapolation region
- Preprocess II: Match (prune) within interpolation region
- Model remaining imbalance (as you would w/o matching)

Remove Extrapolation Region, then Match

Remove Extrapolation Region, then Match

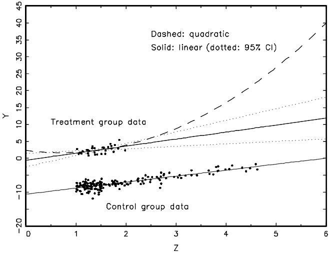


Remove Extrapolation Region, then Match



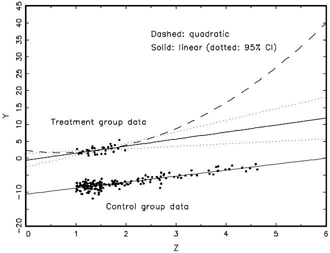
- Must remove data (selecting on X) to avoid extrapolation.

Remove Extrapolation Region, then Match



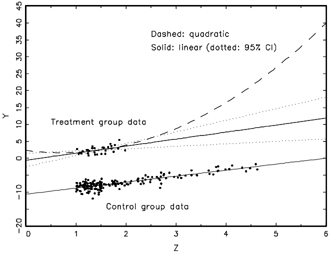
- Must remove data (selecting on X) to avoid extrapolation.
- Options to find "common support" of $P(X|T=1)$ and $P(X|T=0)$

Remove Extrapolation Region, then Match



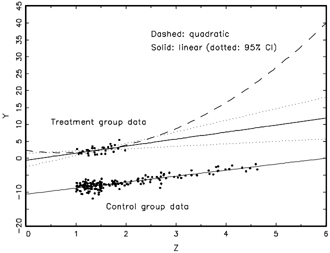
- Must remove data (selecting on X) to avoid extrapolation.
- Options to find "common support" of $P(X|T=1)$ and $P(X|T=0)$
  1. Exact match, so support is defined only at data points

Remove Extrapolation Region, then Match



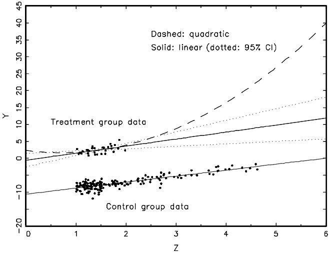
- Must remove data (selecting on X) to avoid extrapolation.
- Options to find "common support" of $P(X|T=1)$ and $P(X|T=0)$
  1. Exact match, so support is defined only at data points
  2. Less but still conservative: convex hull approach

Remove Extrapolation Region, then Match



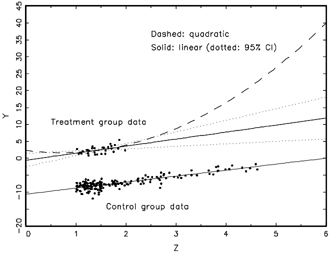
- Must remove data (selecting on X) to avoid extrapolation.
- Options to find "common support" of $P(X|T = 1)$ and $P(X|T = 0)$
  1. Exact match, so support is defined only at data points
  2. Less but still conservative: convex hull approach
     - ★ let T^* and X^* denote subsets of T and X s.t. $\{1 - T^*, X^*\}$ falls within the convex hull of $\{T, X\}$

Remove Extrapolation Region, then Match



- Must remove data (selecting on X) to avoid extrapolation.
- Options to find "common support" of $P(X|T=1)$ and $P(X|T=0)$
  1. Exact match, so support is defined only at data points
  2. Less but still conservative: convex hull approach
     - ★ let T^* and X^* denote subsets of T and X s.t. $\{1-T^*, X^*\}$ falls within the convex hull of $\{T, X\}$
     - ★ use X^* as estimate of common support (deleting remaining observations)

Remove Extrapolation Region, then Match



- Must remove data (selecting on X) to avoid extrapolation.
- Options to find "common support" of $P(X|T=1)$ and $P(X|T=0)$
  1. Exact match, so support is defined only at data points
  2. Less but still conservative: convex hull approach
     - ★ let T^* and X^* denote subsets of T and X s.t. $\{1-T^*, X^*\}$ falls within the convex hull of $\{T, X\}$
     - ★ use X^* as estimate of common support (deleting remaining observations)
  3. Many other possible matching approaches

Matching within the Interpolation Region

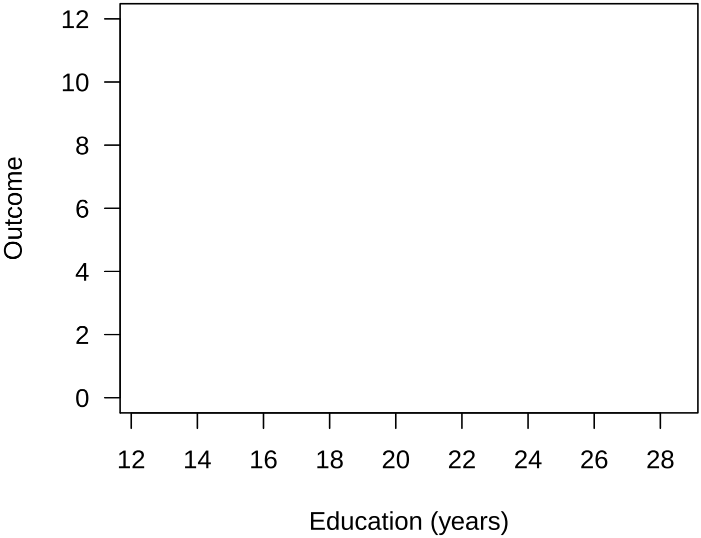
(Ho, Imai, King, Stuart, 2007: fig.1, *Political Analysis*)

Matching within the Interpolation Region

(Ho, Imai, King, Stuart, 2007: fig.1, *Political Analysis*)

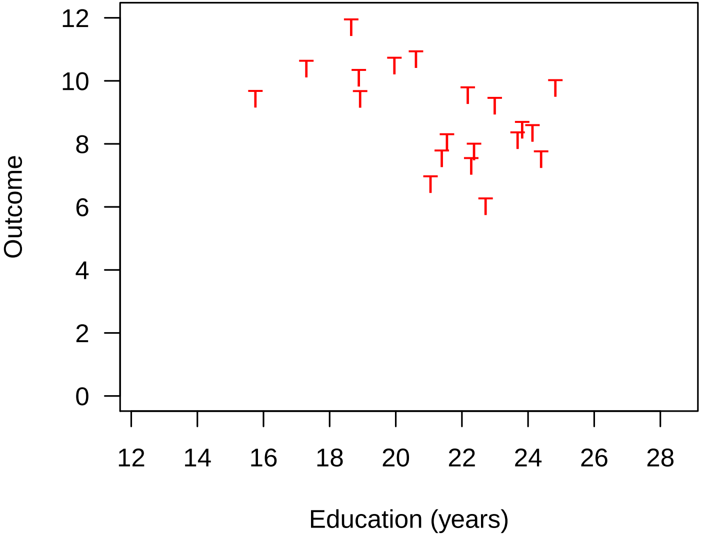
Matching within the Interpolation Region

(Ho, Imai, King, Stuart, 2007: fig.1, *Political Analysis*)



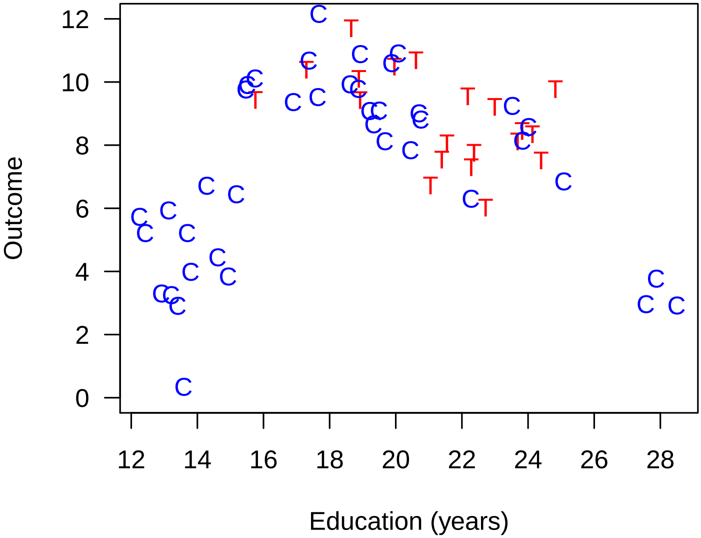
Matching within the Interpolation Region

(Ho, Imai, King, Stuart, 2007: fig.1, *Political Analysis*)



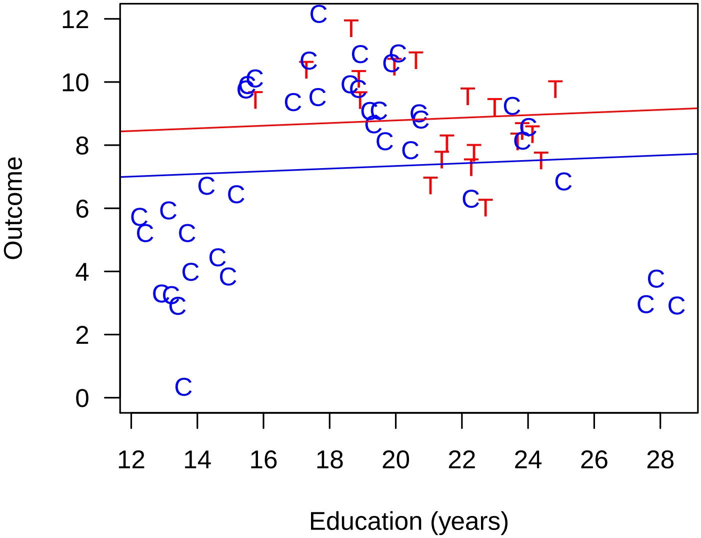
Matching within the Interpolation Region

(Ho, Imai, King, Stuart, 2007: fig.1, *Political Analysis*)



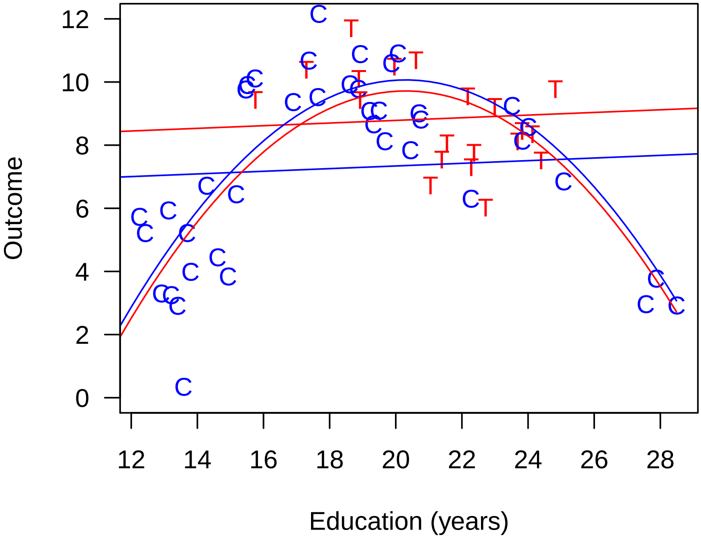
Matching within the Interpolation Region

(Ho, Imai, King, Stuart, 2007: fig.1, *Political Analysis*)



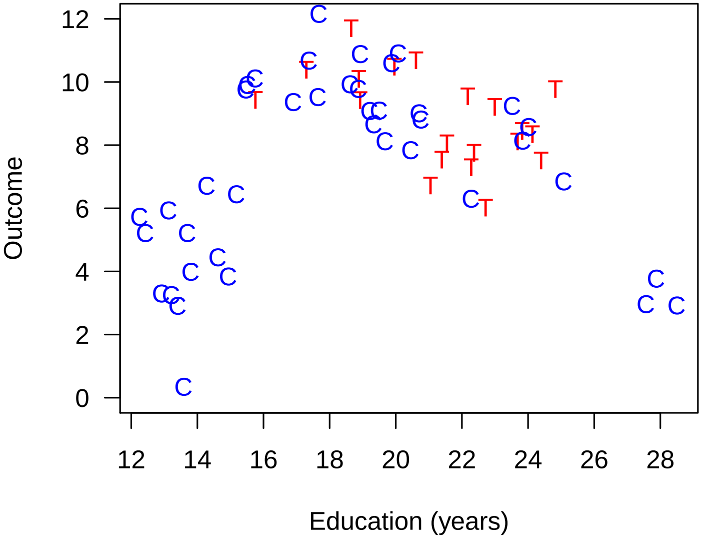
Matching within the Interpolation Region

(Ho, Imai, King, Stuart, 2007: fig.1, *Political Analysis*)



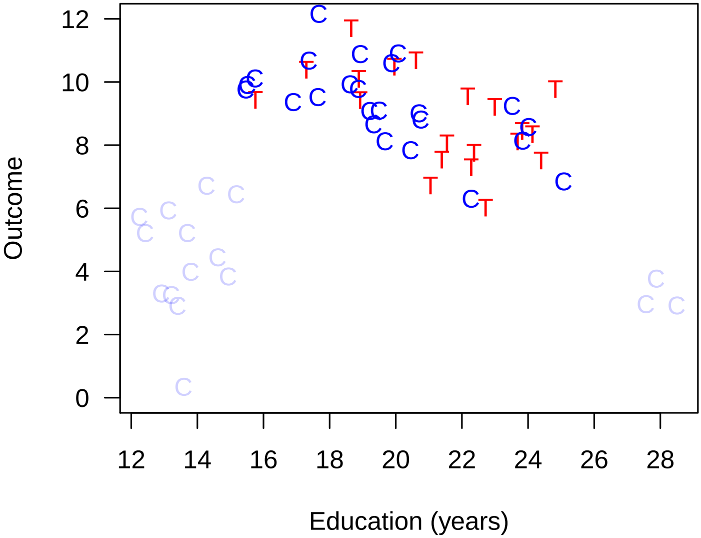
Matching within the Interpolation Region

(Ho, Imai, King, Stuart, 2007: fig.1, *Political Analysis*)



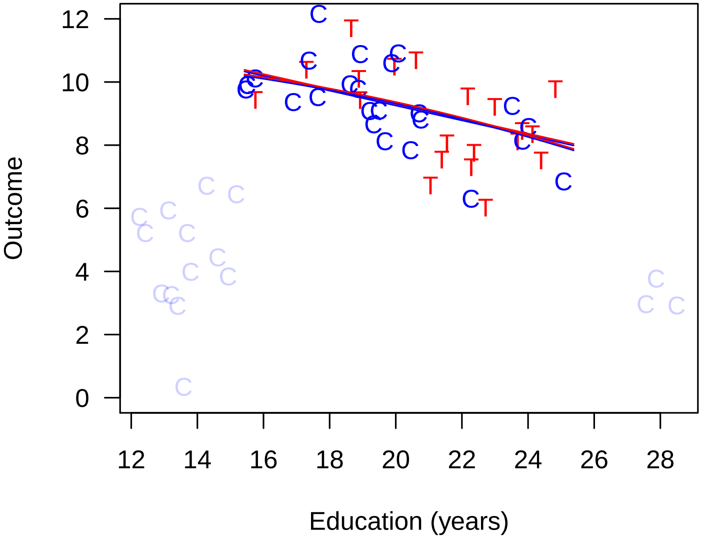
Matching within the Interpolation Region

(Ho, Imai, King, Stuart, 2007: fig.1, *Political Analysis*)



Matching within the Interpolation Region

(Ho, Imai, King, Stuart, 2007: fig.1, *Political Analysis*)



Matching within the Interpolation Region

(Ho, Imai, King, Stuart, 2007: fig.1, *Political Analysis*)

Matching reduces model dependence, bias, and variance

Empirical Illustration: Carpenter, AJPS, 2002

Empirical Illustration: Carpenter, AJPS, 2002

- Hypothesis: Democratic senate majorities slow FDA drug approval time

Empirical Illustration: Carpenter, AJPS, 2002

- Hypothesis: Democratic senate majorities slow FDA drug approval time
- $n = 408$ new drugs (262 approved, 146 pending).

Empirical Illustration: Carpenter, AJPS, 2002

- Hypothesis: Democratic senate majorities slow FDA drug approval time
- $n = 408$ new drugs (262 approved, 146 pending).
- lognormal survival model.

Empirical Illustration: Carpenter, AJPS, 2002

- Hypothesis: Democratic senate majorities slow FDA drug approval time
- $n = 408$ new drugs (262 approved, 146 pending).
- lognormal survival model.
- seven oversight variables (median adjusted ADA scores for House and Senate Committees as well as for House and Senate floors, Democratic Majority in House and Senate, and Democratic Presidency).

Empirical Illustration: Carpenter, AJPS, 2002

- Hypothesis: Democratic senate majorities slow FDA drug approval time
- $n = 408$ new drugs (262 approved, 146 pending).
- lognormal survival model.
- seven oversight variables (median adjusted ADA scores for House and Senate Committees as well as for House and Senate floors, Democratic Majority in House and Senate, and Democratic Presidency).
- 18 control variables (clinical factors, firm characteristics, media variables, etc.)

Evaluating Reduction in Model Dependence

Evaluating Reduction in Model Dependence

- Focus on the causal effect of a Democratic majority in the Senate (identified by Carpenter as not robust).

Evaluating Reduction in Model Dependence

- Focus on the causal effect of a Democratic majority in the Senate (identified by Carpenter as not robust).
- Match: prune 49 units (2 treated, 17 control units).

Evaluating Reduction in Model Dependence

- Focus on the causal effect of a Democratic majority in the Senate (identified by Carpenter as not robust).
- Match: prune 49 units (2 treated, 17 control units).
- run 262,143 possible specifications and calculates ATE for each.

Evaluating Reduction in Model Dependence

- Focus on the causal effect of a Democratic majority in the Senate (identified by Carpenter as not robust).
- Match: prune 49 units (2 treated, 17 control units).
- run 262,143 possible specifications and calculates ATE for each.
- Look at *variability* in ATE estimate across specifications.

Evaluating Reduction in Model Dependence

- Focus on the causal effect of a Democratic majority in the Senate (identified by Carpenter as not robust).
- Match: prune 49 units (2 treated, 17 control units).
- run 262,143 possible specifications and calculates ATE for each.
- Look at *variability* in ATE estimate across specifications.
- (Normal applications would only use one or a few specifications.)

Reducing Model Dependence

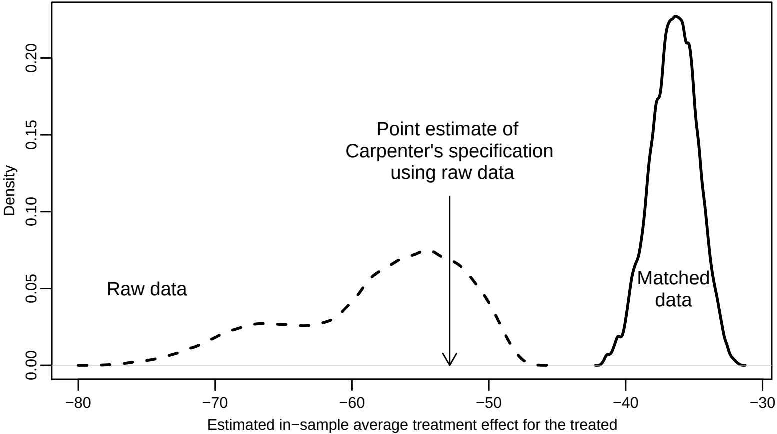


Figure: SATT Histogram: Effect of Democratic Senate majority on FDA drug approval time, across 262, 143 specifications.

The Advantage of Matching

Without Matching:

The Advantage of Matching

Without Matching:

Imbalance

The Advantage of Matching

Without Matching:

Imbalance $\rightsquigarrow$ Model Dependence

The Advantage of Matching

Without Matching:

Imbalance $\rightsquigarrow$ Model Dependence $\rightsquigarrow$ Researcher discretion

The Advantage of Matching

Without Matching:

Imbalance $\rightsquigarrow$ Model Dependence $\rightsquigarrow$ Researcher discretion $\rightsquigarrow$ Bias

The Advantage of Matching

Without Matching:

~~Im~~balance $\rightsquigarrow$ Model Dependence $\rightsquigarrow$ Researcher discretion $\rightsquigarrow$ Bias

The Advantage of Matching

Without Matching:

~~Im~~balance $\rightsquigarrow$ ~~Model Dependence~~ $\rightsquigarrow$ Researcher discretion $\rightsquigarrow$ Bias

The Advantage of Matching

Without Matching:

Imbalance ⇝ Model Dependence ⇝ Researcher discretion ⇝ Bias

The Advantage of Matching

Without Matching:

~~Im~~balance $\rightsquigarrow$ ~~Model Dependence~~ $\rightsquigarrow$ ~~Researcher discretion~~ $\rightsquigarrow$ ~~Bias~~

Preliminary Final Matching Thoughts

- Matching is helpful in a conditioning strategy for identification using selection on observables

Preliminary Final Matching Thoughts

- Matching is helpful in a conditioning strategy for identification using selection on observables
- It does not help you if there are unobservable common causes of treatment and outcomes

Preliminary Final Matching Thoughts

- Matching is helpful in a conditioning strategy for identification using selection on observables
- It does not help you if there are unobservable common causes of treatment and outcomes
- We can (and should!) verify that we have improved balance on the observed covariates, but this does not imply we have improved balanced on unobserved covariates.

Assumptions

1. No unmeasured confounders:

$$D_i \perp\!\!\!\perp \big(Y_i(0), Y_i(1)\big) | X_i$$

Assumptions

1. No unmeasured confounders:

$$D_i \perp\!\!\!\perp \big(Y_i(0), Y_i(1)\big) | X_i$$

2. Positivity/overlap:

$$0 < \mathbb{P}(D_i = 1 | X_i = x) < 1$$

Exact matching

- Let X_i take on a finite number of values, $x \in \mathcal{X}$.

Exact matching

- Let X_i take on a finite number of values, $x \in \mathcal{X}$.
- Let $\mathbb{I}_t = \{1, 2, \ldots, N_t\}$ be the set of treated units.

Exact matching

- Let X_i take on a finite number of values, $x \in \mathcal{X}$.
- Let $\mathbb{I}_t = \{1, 2, \ldots, N_t\}$ be the set of treated units.
- Exact matching. For each treated unit, $i \in \mathbb{I}_t$:

Exact matching

- Let X_i take on a finite number of values, $x \in \mathcal{X}$.
- Let $\mathbb{I}_t = \{1, 2, \ldots, N_t\}$ be the set of treated units.
- Exact matching. For each treated unit, $i \in \mathbb{I}_t$:
  - Find the set of unmatched control units j such that $X_i = X_j$

Exact matching

- Let X_i take on a finite number of values, $x \in \mathcal{X}$.
- Let $\mathbb{I}_t = \{1, 2, \ldots, N_t\}$ be the set of treated units.
- Exact matching. For each treated unit, $i \in \mathbb{I}_t$:
  - Find the set of unmatched control units j such that $X_i = X_j$
  - Randomly select one of these control units to be the match, indicated $j(i)$.

Exact matching

- Let X_i take on a finite number of values, $x \in \mathcal{X}$.
- Let $\mathbb{I}_t = \{1, 2, \ldots, N_t\}$ be the set of treated units.
- Exact matching. For each treated unit, $i \in \mathbb{I}_t$:
  - Find the set of unmatched control units j such that $X_i = X_j$
  - Randomly select one of these control units to be the match, indicated $j(i)$.
- Let $\mathbb{I}_c = \{j(1), \ldots, j(N_t)\}$ be the set of matched controls.

Exact matching

- Let X_i take on a finite number of values, $x \in \mathcal{X}$.
- Let $\mathbb{I}_t = \{1, 2, \dots, N_t\}$ be the set of treated units.
- Exact matching. For each treated unit, $i \in \mathbb{I}_t$:
  - Find the set of unmatched control units j such that $X_i = X_j$
  - Randomly select one of these control units to be the match, indicated $j(i)$.
- Let $\mathbb{I}_c = \{j(1), \dots, j(N_t)\}$ be the set of matched controls.
- Last, discard all unmatched control units.

Exact matching

- Let X_i take on a finite number of values, $x \in \mathcal{X}$.
- Let $\mathbb{I}_t = \{1, 2, \ldots, N_t\}$ be the set of treated units.
- Exact matching. For each treated unit, $i \in \mathbb{I}_t$:
  - Find the set of unmatched control units j such that $X_i = X_j$
  - Randomly select one of these control units to be the match, indicated $j(i)$.
- Let $\mathbb{I}_c = \{j(1), \ldots, j(N_t)\}$ be the set of matched controls.
- Last, discard all unmatched control units.
- The distribution of X_i will be exactly the same for treated and matched control:

$$\mathbb{P}(X_i = x | D_i = 1) = \mathbb{P}(X_i = x | D_i = 0, \mathbb{I}_c)$$

Weakening the identification assumptions

- No unmeasured confounders, consistency, and exact matches $\leadsto$ identifying the ATT.

Weakening the identification assumptions

- No unmeasured confounders, consistency, and exact matches $\rightsquigarrow$ identifying the ATT.
- Can weaken no unmeasured confounders to conditional mean independence (CMI):

$$E[Y_i(0)|X_i, D_i = 1] = E[Y_i(0)|X_i, D_i = 0]$$

Weakening the identification assumptions

- No unmeasured confounders, consistency, and exact matches $\rightsquigarrow$ identifying the ATT.
- Can weaken no unmeasured confounders to conditional mean independence (CMI):

$$E[Y_i(0)|X_i, D_i = 1] = E[Y_i(0)|X_i, D_i = 0]$$

- Two nice features of CMI:

Weakening the identification assumptions

- No unmeasured confounders, consistency, and exact matches $\rightsquigarrow$ identifying the ATT.
- Can weaken no unmeasured confounders to conditional mean independence (CMI):

$$E[Y_i(0)|X_i, D_i = 1] = E[Y_i(0)|X_i, D_i = 0]$$

- Two nice features of CMI:
  1. Only have to make assumptions about $Y_i(0)$ not $Y_i(1)$

Weakening the identification assumptions

- No unmeasured confounders, consistency, and exact matches $\rightsquigarrow$ identifying the ATT.
- Can weaken no unmeasured confounders to conditional mean independence (CMI):

$$E[Y_i(0)|X_i, D_i = 1] = E[Y_i(0)|X_i, D_i = 0]$$

- Two nice features of CMI:
  1. Only have to make assumptions about $Y_i(0)$ not $Y_i(1)$
  2. Only places restrictions on the means, not other parts of the distribution (variance, skew, kurtosis, etc)

Analyzing exactly matched data

- How do we analyze the exactly matched data?

Analyzing exactly matched data

- How do we analyze the exactly matched data?
- Dead simple difference in means:

$$\widehat{\tau}_m = \frac{1}{N_t} \sum_{i=1}^{N_t} Y_i - \frac{1}{N_c} \sum_{j \in \mathbb{I}_c} Y_j$$

Analyzing exactly matched data

- How do we analyze the exactly matched data?
- Dead simple difference in means:

$$\widehat{\tau}_m = \frac{1}{N_t}\sum_{i=1}^{N_t} Y_i - \frac{1}{N_c}\sum_{j\in\mathbb{I}_c} Y_j$$

- Notice that we matched 1 treated to 1 control exactly, so we have:

$$\widehat{\tau}_m = \frac{1}{N_t}\sum_{i=1}^{N_t}(Y_i - Y_{j(i)})$$

Analyzing exactly matched data

- How do we analyze the exactly matched data?
- Dead simple difference in means:

$$\widehat{\tau}_m = \frac{1}{N_t}\sum_{i=1}^{N_t} Y_i - \frac{1}{N_c}\sum_{j \in \mathbb{I}_c} Y_j$$

- Notice that we matched 1 treated to 1 control exactly, so we have:

$$\widehat{\tau}_m = \frac{1}{N_t}\sum_{i=1}^{N_t}(Y_i - Y_{j(i)})$$

- $\rightsquigarrow$ average of the within matched-pair differences.

Variance with exact matches

- Notice that with 1:1 treated/control matching, similar to a matched-pair experiment.

Variance with exact matches

- Notice that with 1:1 treated/control matching, similar to a matched-pair experiment.
- Variance estimators are a little different for these.

Variance with exact matches

- Notice that with 1:1 treated/control matching, similar to a matched-pair experiment.
- Variance estimators are a little different for these.
- Variance estimator:

$$\widehat{\text{Var}}(\widehat{\tau}_m) = \frac{1}{N_t} \sum_{i=1}^{N_t} \left(Y_i - Y_{j(i)} - \widehat{\tau}_m \right)^2$$

Variance with exact matches

- Notice that with 1:1 treated/control matching, similar to a matched-pair experiment.
- Variance estimators are a little different for these.
- Variance estimator:

$$\widehat{\text{Var}}(\widehat{\tau}_m) = \frac{1}{N_t} \sum_{i=1}^{N_t} \left(Y_i - Y_{j(i)} - \widehat{\tau}_m \right)^2$$

- In-sample variance of the within-pair differences.

Beyond exact matching

- With high-dimensional X_i, not feasible to exact match.

Beyond exact matching

- With high-dimensional X_i, not feasible to exact match.
- Let S be a matching solution: a subset of the data produced by the matching procedure: $(\mathbb{I}_t, \mathbb{I}_c)$.

Beyond exact matching

- With high-dimensional X_i, not feasible to exact match.
- Let S be a matching solution: a subset of the data produced by the matching procedure: $(\mathbb{I}_t, \mathbb{I}_c)$.
- Suppose that this procedure produces balance:

$$D_i \perp\!\!\!\perp X_i | S$$

Beyond exact matching

- With high-dimensional X_i, not feasible to exact match.
- Let S be a matching solution: a subset of the data produced by the matching procedure: $(\mathbb{I}_t, \mathbb{I}_c)$.
- Suppose that this procedure produces balance:

$$D_i \perp\!\!\!\perp X_i | S$$

- With no unmeasured confounders we have:

$$\big(Y_i(0), Y_i(1)\big) \perp\!\!\!\perp D_i | S$$

Beyond exact matching

- With high-dimensional X_i, not feasible to exact match.
- Let S be a matching solution: a subset of the data produced by the matching procedure: $(\mathbb{I}_t, \mathbb{I}_c)$.
- Suppose that this procedure produces balance:

$$D_i \perp\!\!\!\perp X_i | S$$

- With no unmeasured confounders we have:

$$\big(Y_i(0), Y_i(1)\big) \perp\!\!\!\perp D_i | S$$

- Balance is checkable $\rightsquigarrow$ are D_i and X_i related in the matched data?

The matching procedure

1. Choose a number of matches

The matching procedure

1. Choose a number of matches
2. Choose a distance metric

The matching procedure

1. Choose a number of matches
2. Choose a distance metric
3. Find matches (drop non-matches)

The matching procedure

1. Choose a number of matches
2. Choose a distance metric
3. Find matches (drop non-matches)
4. Check balance

The matching procedure

1. Choose a number of matches
2. Choose a distance metric
3. Find matches (drop non-matches)
4. Check balance
5. Repeat (1)-(4) until balance is acceptable

The matching procedure

1. Choose a number of matches
2. Choose a distance metric
3. Find matches (drop non-matches)
4. Check balance
5. Repeat (1)-(4) until balance is acceptable
6. Calculate the effect of the treatment on the outcome in the matched dataset.

More than 1 control match

- What if we have enough controls to have M matched controls per treated?

More than 1 control match

- What if we have enough controls to have M matched controls per treated?
  - $\mathbb{P}(X_i = x | D_i = 1) = \mathbb{P}(X_i = x | D_i = 0, \mathbb{I}_c)$ because M is constant across treated units.

More than 1 control match

- What if we have enough controls to have M matched controls per treated?
  - $\mathbb{P}(X_i = x | D_i = 1) = \mathbb{P}(X_i = x | D_i = 0, \mathbb{I}_c)$ because M is constant across treated units.
- Now, $J_M(i)$ is a set of M control matches. Use these to "impute" missing potential outcome.

More than 1 control match

- What if we have enough controls to have M matched controls per treated?
  - $\mathbb{P}(X_i = x | D_i = 1) = \mathbb{P}(X_i = x | D_i = 0, \mathbb{I}_c)$ because M is constant across treated units.
- Now, $J_M(i)$ is a set of M control matches. Use these to "impute" missing potential outcome.
- For $i \in \mathbb{I}_t$ define:

$$\widehat{Y}_i(0) = \frac{1}{M} \sum_{j \in J_M(i)} Y_j$$

More than 1 control match

- What if we have enough controls to have M matched controls per treated?
  - $\mathbb{P}(X_i = x | D_i = 1) = \mathbb{P}(X_i = x | D_i = 0, \mathbb{I}_c)$ because M is constant across treated units.
- Now, $J_M(i)$ is a set of M control matches. Use these to "impute" missing potential outcome.
- For $i \in \mathbb{I}_t$ define:

$$\widehat{Y}_i(0) = \frac{1}{M} \sum_{j \in J_M(i)} Y_j$$

- New estimator for the effect:

$$\widehat{\tau}_m = \frac{1}{N_t} \sum_{i=1}^{N_t} (Y_i - \widehat{Y}_i(0))$$

More than 1 control match

- What if we have enough controls to have M matched controls per treated?
  - $\mathbb{P}(X_i = x | D_i = 1) = \mathbb{P}(X_i = x | D_i = 0, \mathbb{I}_c)$ because M is constant across treated units.
- Now, $J_M(i)$ is a set of M control matches. Use these to "impute" missing potential outcome.
- For $i \in \mathbb{I}_t$ define:

$$\widehat{Y}_i(0) = \frac{1}{M} \sum_{j \in J_M(i)} Y_j$$

- New estimator for the effect:

$$\widehat{\tau}_m = \frac{1}{N_t} \sum_{i=1}^{N_t} (Y_i - \widehat{Y}_i(0))$$

- Under no unmeasured confounding, $\widehat{Y}_i(0)$ is a good predictor of the true potential outcome under control, Y_i.

Number of matches

- How many control matches should we include?

Number of matches

- How many control matches should we include?
  - Small $M \rightsquigarrow$ small sample sizes

Number of matches

- How many control matches should we include?
  - Small $M \rightsquigarrow$ small sample sizes
  - Large $M \rightsquigarrow$ worse matches (each additional match is further away).

Number of matches

- How many control matches should we include?
  - Small $M \rightsquigarrow$ small sample sizes
  - Large $M \rightsquigarrow$ worse matches (each additional match is further away).
- If M varies by treated unit, need to weight observations to ensure balance.

With or without replacement

- Matching with replacement: a single control unit can be matched to multiple treated units

With or without replacement

- Matching with replacement: a single control unit can be matched to multiple treated units
- Benefits:

With or without replacement

- Matching with replacement: a single control unit can be matched to multiple treated units
- Benefits:
  - Better matches!

With or without replacement

- Matching with replacement: a single control unit can be matched to multiple treated units
- Benefits:
  - Better matches!
  - Order of matching does not matter.

With or without replacement

- Matching with replacement: a single control unit can be matched to multiple treated units
- Benefits:
  - Better matches!
  - Order of matching does not matter.
- Drawbacks:

With or without replacement

- Matching with replacement: a single control unit can be matched to multiple treated units
- Benefits:
  - Better matches!
  - Order of matching does not matter.
- Drawbacks:
  - Inference is more complicated.

With or without replacement

- Matching with replacement: a single control unit can be matched to multiple treated units
- Benefits:
  - Better matches!
  - Order of matching does not matter.
- Drawbacks:
  - Inference is more complicated.
  - $\rightsquigarrow$ need to account for multiple appearances with weights.

With or without replacement

- Matching with replacement: a single control unit can be matched to multiple treated units
- Benefits:
  - Better matches!
  - Order of matching does not matter.
- Drawbacks:
  - Inference is more complicated.
  - $\rightsquigarrow$ need to account for multiple appearances with weights.
  - Potentially higher uncertainty (using the same data multiple times = relying on less data).

Defining closeness

- We want to find control observations that are similar to the treated unit on X_i.

Defining closeness

- We want to find control observations that are similar to the treated unit on X_i.
- How do we define distance/similarity on X_i if it is high dimensional?

Defining closeness

- We want to find control observations that are similar to the treated unit on X_i.
- How do we define distance/similarity on X_i if it is high dimensional?
- We need a distance metric which maps two covariates vectors into a single number.

Defining closeness

- We want to find control observations that are similar to the treated unit on X_i.
- How do we define distance/similarity on X_i if it is high dimensional?
- We need a distance metric which maps two covariates vectors into a single number.
  - Lower values $\rightsquigarrow$ more similar values of X_i.

Defining closeness

- We want to find control observations that are similar to the treated unit on X_i.
- How do we define distance/similarity on X_i if it is high dimensional?
- We need a distance metric which maps two covariates vectors into a single number.
  - Lower values $\rightsquigarrow$ more similar values of X_i.
  - Choice of distance metric will lead to different matches.

Exact distance metric

- Exact: only match units to other units that have the same exact values of X_i.

$$D_{ij} = \begin{cases} 0 & \text{if } X_i = X_j \\ \infty & \text{if } X_i \neq X_j \end{cases}$$

Euclidean distance

- The normalized Euclidean distance metric just uses the sum of the normalized distances for each covariate.

Euclidean distance

- The normalized Euclidean distance metric just uses the sum of the normalized distances for each covariate.
  - “Closeness” is standardized across covariates.

Euclidean distance

- The normalized Euclidean distance metric just uses the sum of the normalized distances for each covariate.
  - "Closeness" is standardized across covariates.
- Suppose that $X_i = (X_{i1}, \ldots, X_{iK})'$, so that there are K covariates.

Euclidean distance

- The normalized Euclidean distance metric just uses the sum of the normalized distances for each covariate.
  - "Closeness" is standardized across covariates.
- Suppose that $X_i = (X_{i1}, \dots, X_{iK})'$, so that there are K covariates.
- Then the Euclidean distance metric is:

$$D_{ij} = \sqrt{\sum_{k=1}^{K} \frac{(X_{ik} - X_{jk})^2}{\widehat{\sigma}_k^2}}$$

Euclidean distance

- The normalized Euclidean distance metric just uses the sum of the normalized distances for each covariate.
  - "Closeness" is standardized across covariates.
- Suppose that $X_i = (X_{i1}, \ldots, X_{iK})'$, so that there are K covariates.
- Then the Euclidean distance metric is:

$$D_{ij} = \sqrt{\sum_{k=1}^{K} \frac{(X_{ik} - X_{jk})^2}{\widehat{\sigma}_k^2}}$$

- Here, $\widehat{\sigma}_k^2$ is the variance of the kth variable:

$$\widehat{\sigma}_k^2 = \frac{1}{N-1} \sum_{i=1}^{N} (X_{ik} - \bar{X}_k)$$

Mahalanobis distance

- Mahalanobis distance: Euclidean distance adjusted for covariance in the data.

Mahalanobis distance

- Mahalanobis distance: Euclidean distance adjusted for covariance in the data.
- Intuition: if X_{ik} and $X_{ik'}$ are two covariates that are highly correlated, then their contribution to the distances should be lower.

Mahalanobis distance

- Mahalanobis distance: Euclidean distance adjusted for covariance in the data.
- Intuition: if X_{ik} and $X_{ik'}$ are two covariates that are highly correlated, then their contribution to the distances should be lower.
  - Easy to get close on correlated covariates $\rightsquigarrow$ downweight.

Mahalanobis distance

- Mahalanobis distance: Euclidean distance adjusted for covariance in the data.
- Intuition: if X_{ik} and $X_{ik'}$ are two covariates that are highly correlated, then their contribution to the distances should be lower.
  - Easy to get close on correlated covariates $\rightsquigarrow$ downweight.
  - Harder to get close on uncorrelated covariates $\rightsquigarrow$ upweight.

Mahalanobis distance

- Mahalanobis distance: Euclidean distance adjusted for covariance in the data.
- Intuition: if X_{ik} and $X_{ik'}$ are two covariates that are highly correlated, then their contribution to the distances should be lower.
  - Easy to get close on correlated covariates $\rightsquigarrow$ downweight.
  - Harder to get close on uncorrelated covariates $\rightsquigarrow$ upweight.
- Metric:

$$D_{ij} = \sqrt{(X_i - X_j)'\widehat{\Sigma}^{-1}(X_i - X_j)}$$

Mahalanobis distance

- Mahalanobis distance: Euclidean distance adjusted for covariance in the data.
- Intuition: if X_{ik} and $X_{ik'}$ are two covariates that are highly correlated, then their contribution to the distances should be lower.
  - Easy to get close on correlated covariates $\leadsto$ downweight.
  - Harder to get close on uncorrelated covariates $\leadsto$ upweight.
- Metric:

$$D_{ij} = \sqrt{(X_i - X_j)'\widehat{\Sigma}^{-1}(X_i - X_j)}$$

- $\widehat{\Sigma}$ is the estimated variance-covariance matrix of the observations:

$$\widehat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})(X_i - \bar{X})^T$$

Complications

- Combining distance metrics:

Complications

- Combining distance metrics:
  - Exact on race/gender, Mahalanobis on the rest.

Complications

- Combining distance metrics:
  - Exact on race/gender, Mahalanobis on the rest.
- Some matches are too far on the distance metric.

Complications

- Combining distance metrics:
  - Exact on race/gender, Mahalanobis on the rest.
- Some matches are too far on the distance metric.
  - Dropping those matches (treated and control) improves balance.

Complications

- Combining distance metrics:
  - Exact on race/gender, Mahalanobis on the rest.
- Some matches are too far on the distance metric.
  - Dropping those matches (treated and control) improves balance.
  - Dropping treated units changes the quantity of interest.

Complications

- Combining distance metrics:
  - Exact on race/gender, Mahalanobis on the rest.
- Some matches are too far on the distance metric.
  - Dropping those matches (treated and control) improves balance.
  - Dropping treated units changes the quantity of interest.
- Implementation: a caliper, which is the maximum distance we would accept

Estimands

- Matching easiest to justify for the ATT.

Estimands

- Matching easiest to justify for the ATT.
  - Dropping control units doesn't affect this identification.

Estimands

- Matching easiest to justify for the ATT.
  - Dropping control units doesn't affect this identification.
- Can also identify the ATC by finding matched treated units for the controls.

Estimands

- Matching easiest to justify for the ATT.
  - Dropping control units doesn't affect this identification.
- Can also identify the ATC by finding matched treated units for the controls.
- Combine the two to get the ATE:

$$\tau = \tau_{ATT}\mathbb{P}(D_i = 1) + \tau_{ATC}\mathbb{P}(D_i = 0)$$

Estimands

- Matching easiest to justify for the ATT.
  - Dropping control units doesn't affect this identification.
- Can also identify the ATC by finding matched treated units for the controls.
- Combine the two to get the ATE:

$$\tau = \tau_{ATT}\mathbb{P}(D_i = 1) + \tau_{ATC}\mathbb{P}(D_i = 0)$$

- Estimated:

$$\widehat{\tau} = \widehat{\tau}_{ATT}\left(\frac{N_t}{N}\right) + \widehat{\tau}_{ATC}\left(\frac{N_c}{N}\right)$$

Moving the goalposts

- Common support: finding the subspace of X_i where there is overlap between the treated and control groups.

Moving the goalposts

- Common support: finding the subspace of X_i where there is overlap between the treated and control groups.
    - Have to extrapolate outside is region.

Moving the goalposts

- Common support: finding the subspace of X_i where there is overlap between the treated and control groups.
  - Have to extrapolate outside is region.
  - Theoretical: effect of voting for those under 18 $(\mathbb{P}(D_i = 1 | X_i < 18) = 0)$.

Moving the goalposts

- Common support: finding the subspace of X_i where there is overlap between the treated and control groups.
  - Have to extrapolate outside is region.
  - Theoretical: effect of voting for those under 18 ($\mathbb{P}(D_i = 1 | X_i < 18) = 0$).
  - Empirical: no/extremely few treated units in a sea of controls.

Moving the goalposts

- Common support: finding the subspace of X_i where there is overlap between the treated and control groups.
    - Have to extrapolate outside is region.
    - Theoretical: effect of voting for those under 18 ($\mathbb{P}(D_i = 1 | X_i < 18) = 0$).
    - Empirical: no/extremely few treated units in a sea of controls.
    - Solution: restrict analysis to common support (dropping treated and controls).

Moving the goalposts

- Common support: finding the subspace of X_i where there is overlap between the treated and control groups.
  - Have to extrapolate outside is region.
  - Theoretical: effect of voting for those under 18 ($\mathbb{P}(D_i = 1|X_i < 18) = 0$).
  - Empirical: no/extremely few treated units in a sea of controls.
  - Solution: restrict analysis to common support (dropping treated and controls).
- Moving the goalposts: dropping treated units.

Moving the goalposts

- Common support: finding the subspace of X_i where there is overlap between the treated and control groups.
  - Have to extrapolate outside is region.
  - Theoretical: effect of voting for those under 18 ($\mathbb{P}(D_i = 1|X_i < 18) = 0$).
  - Empirical: no/extremely few treated units in a sea of controls.
  - Solution: restrict analysis to common support (dropping treated and controls).
- Moving the goalposts: dropping treated units.
  - We move away from being able to identify the ATT.

Moving the goalposts

- Common support: finding the subspace of X_i where there is overlap between the treated and control groups.
  - Have to extrapolate outside is region.
  - Theoretical: effect of voting for those under 18 ($\mathbb{P}(D_i = 1|X_i < 18) = 0$).
  - Empirical: no/extremely few treated units in a sea of controls.
  - Solution: restrict analysis to common support (dropping treated and controls).
- Moving the goalposts: dropping treated units.
  - We move away from being able to identify the ATT.
  - Now it's the ATT in the matched subsample (sometimes called the feasible ATT).

Moving the goalposts

- Common support: finding the subspace of X_i where there is overlap between the treated and control groups.
  - Have to extrapolate outside is region.
  - Theoretical: effect of voting for those under 18 ($\mathbb{P}(D_i = 1|X_i < 18) = 0$).
  - Empirical: no/extremely few treated units in a sea of controls.
  - Solution: restrict analysis to common support (dropping treated and controls).
- Moving the goalposts: dropping treated units.
  - We move away from being able to identify the ATT.
  - Now it's the ATT in the matched subsample (sometimes called the feasible ATT).
  - Good to be clear about this.

Matching methods

- Now that we have distances between all units, we just need to match!

Matching methods

- Now that we have distances between all units, we just need to match!
- For a particular unit, easy:

$$j(i) = \arg\min_{j \in \mathbb{J}_c} D_{ij}$$

Matching methods

- Now that we have distances between all units, we just need to match!
- For a particular unit, easy:

$$j(i) = \arg\min_{j \in \mathbb{J}_c} D_{ij}$$

  - $\mathbb{J}_c$ are the available controls for matching.

Matching methods

- Now that we have distances between all units, we just need to match!
- For a particular unit, easy:

$$j(i) = \arg\min_{j \in \mathbb{J}_c} D_{ij}$$

  - $\mathbb{J}_c$ are the available controls for matching.
- This is nearest neighbor: "Find the control unit with the smallest distance metric."

Matching methods

- Now that we have distances between all units, we just need to match!
- For a particular unit, easy:

$$j(i) = \arg\min_{j \in \mathbb{J}_c} D_{ij}$$

  - $\mathbb{J}_c$ are the available controls for matching.
- This is nearest neighbor: "Find the control unit with the smallest distance metric."
- Do the same for all treated units.

Matching methods

- Now that we have distances between all units, we just need to match!
- For a particular unit, easy:

$$j(i) = \arg\min_{j \in \mathbb{J}_c} D_{ij}$$

  - $\mathbb{J}_c$ are the available controls for matching.
- This is nearest neighbor: "Find the control unit with the smallest distance metric."
- Do the same for all treated units.
- What about ties?

Matching methods

- Now that we have distances between all units, we just need to match!
- For a particular unit, easy:

$$j(i) = \arg\min_{j \in \mathbb{J}_c} D_{ij}$$

  - $\mathbb{J}_c$ are the available controls for matching.
- This is nearest neighbor: "Find the control unit with the smallest distance metric."
- Do the same for all treated units.
- What about ties?
  - Randomly choose between them.

Matching methods

- Now that we have distances between all units, we just need to match!
- For a particular unit, easy:

$$j(i) = \arg\min_{j \in \mathbb{J}_c} D_{ij}$$

  - $\mathbb{J}_c$ are the available controls for matching.
- This is nearest neighbor: "Find the control unit with the smallest distance metric."
- Do the same for all treated units.
- What about ties?
  - Randomly choose between them.
- Note: in nearest neighbor without replacement the order matters!

Assessing balance

- All matching methods seek to maximize balance:

$$\mathbb{P}(X_i = x | D_i = 1, S) = \mathbb{P}(X_i = x | D_i = 0, S)$$

Assessing balance

- All matching methods seek to maximize balance:

$$\mathbb{P}(X_i = x | D_i = 1, S) = \mathbb{P}(X_i = x | D_i = 0, S)$$

- Choice of balance metric will determine which matching method does better.

Assessing balance

- All matching methods seek to maximize balance:

$$\mathbb{P}(X_i = x|D_i = 1, S) = \mathbb{P}(X_i = x|D_i = 0, S)$$

- Choice of balance metric will determine which matching method does better.
  - If you use Mahalanobis distance as the balance metric, then matching on the Mahalanobis score will do well because that's what it's designed to do.

Assessing balance

- All matching methods seek to maximize balance:

$$\mathbb{P}(X_i = x | D_i = 1, S) = \mathbb{P}(X_i = x | D_i = 0, S)$$

- Choice of balance metric will determine which matching method does better.
  - If you use Mahalanobis distance as the balance metric, then matching on the Mahalanobis score will do well because that's what it's designed to do.
- Options:

Assessing balance

- All matching methods seek to maximize balance:

$$\mathbb{P}(X_i = x | D_i = 1, S) = \mathbb{P}(X_i = x | D_i = 0, S)$$

- Choice of balance metric will determine which matching method does better.
  - If you use Mahalanobis distance as the balance metric, then matching on the Mahalanobis score will do well because that's what it's designed to do.
- Options:
  - Differences-in-means/medians, standardized.

Assessing balance

- All matching methods seek to maximize balance:

$$\mathbb{P}(X_i = x | D_i = 1, S) = \mathbb{P}(X_i = x | D_i = 0, S)$$

- Choice of balance metric will determine which matching method does better.
  - If you use Mahalanobis distance as the balance metric, then matching on the Mahalanobis score will do well because that's what it's designed to do.
- Options:
  - Differences-in-means/medians, standardized.
  - Quantile-quantile plots/KS statistics for comparing the entire distribution of X_i.

Assessing balance

- All matching methods seek to maximize balance:

$$\mathbb{P}(X_i = x | D_i = 1, S) = \mathbb{P}(X_i = x | D_i = 0, S)$$

- Choice of balance metric will determine which matching method does better.
  - If you use Mahalanobis distance as the balance metric, then matching on the Mahalanobis score will do well because that's what it's designed to do.
- Options:
  - Differences-in-means/medians, standardized.
  - Quantile-quantile plots/KS statistics for comparing the entire distribution of X_i.
  - L_1: multivariate histogram.

Three Approaches to Matching

Three Approaches to Matching

- There are many approaches to matching. We will cover just three for the sake of time.

Three Approaches to Matching

- There are many approaches to matching. We will cover just three for the sake of time.
- This isn't a statement that these are the best three, just a set which are straightforward to learn.

Three Approaches to Matching

- There are many approaches to matching. We will cover just three for the sake of time.
- This isn't a statement that these are the best three, just a set which are straightforward to learn.
- Which is the best method? The one that produces the best balance!

Method 1: Mahalanobis Distance Matching

Method 1: Mahalanobis Distance Matching

(Approximates Fully Blocked Experiment)

Method 1: Mahalanobis Distance Matching

(Approximates Fully Blocked Experiment)

1. Preprocess (Matching)
2. Checking Measure imbalance, tweak, repeat, . . .
3. Estimation Difference in means or a model

Method 1: Mahalanobis Distance Matching

(Approximates Fully Blocked Experiment)

1. Preprocess (Matching)
   - Distance$(X_i, X_j) = \sqrt{(X_i - X_j)' S^{-1} (X_i - X_j)}$
2. Checking Measure imbalance, tweak, repeat, ...
3. Estimation Difference in means or a model

Method 1: Mahalanobis Distance Matching

(Approximates Fully Blocked Experiment)

1. Preprocess (Matching)
   - Distance$(X_i, X_j) = \sqrt{(X_i - X_j)' S^{-1} (X_i - X_j)}$
   - Match each treated unit to the nearest control unit
2. Checking Measure imbalance, tweak, repeat, . . .
3. Estimation Difference in means or a model

Method 1: Mahalanobis Distance Matching

(Approximates Fully Blocked Experiment)

1. Preprocess (Matching)
   - Distance$(X_i, X_j) = \sqrt{(X_i - X_j)' S^{-1} (X_i - X_j)}$
   - Match each treated unit to the nearest control unit
   - Control units: not reused; pruned if unused
2. Checking Measure imbalance, tweak, repeat, ...
3. Estimation Difference in means or a model

Method 1: Mahalanobis Distance Matching

(Approximates Fully Blocked Experiment)

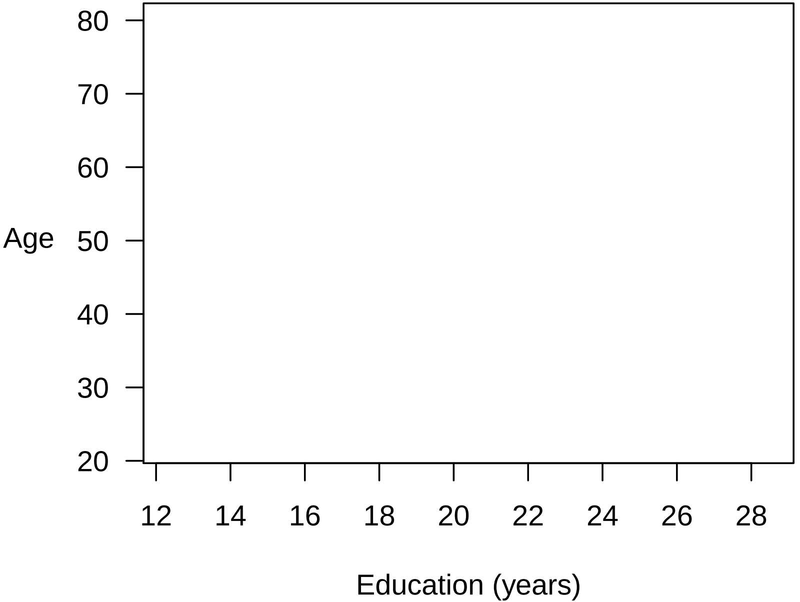
1. Preprocess (Matching)
   - Distance$(X_i, X_j) = \sqrt{(X_i - X_j)' S^{-1} (X_i - X_j)}$
   - Match each treated unit to the nearest control unit
   - Control units: not reused; pruned if unused
   - Prune matches if Distance$>$*caliper*
2. Checking Measure imbalance, tweak, repeat, . . .
3. Estimation Difference in means or a model

Method 1: Mahalanobis Distance Matching

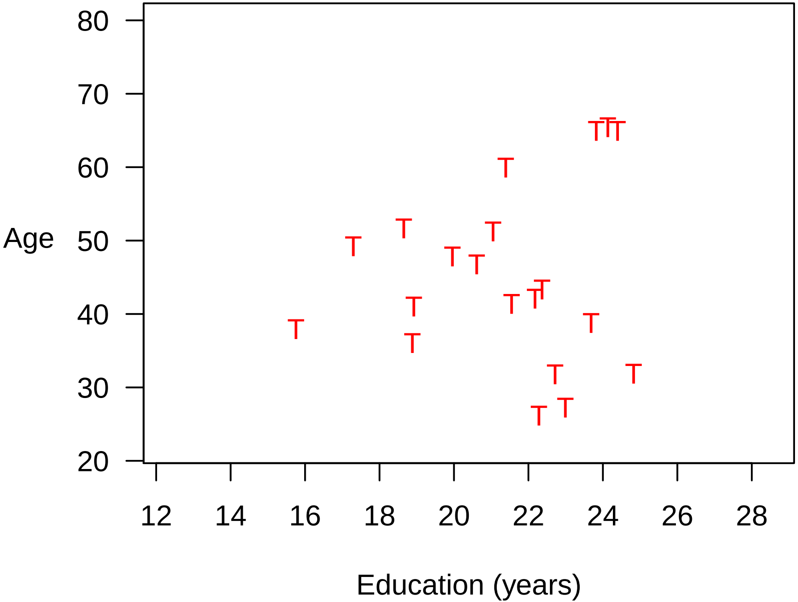
(Approximates Fully Blocked Experiment)

1. Preprocess (Matching)
   - Distance$(X_i, X_j) = \sqrt{(X_i - X_j)' S^{-1} (X_i - X_j)}$
   - Match each treated unit to the nearest control unit
   - Control units: not reused; pruned if unused
   - Prune matches if Distance$>$*caliper*
2. Checking Measure imbalance, tweak, repeat, ...
3. Estimation Difference in means or a model

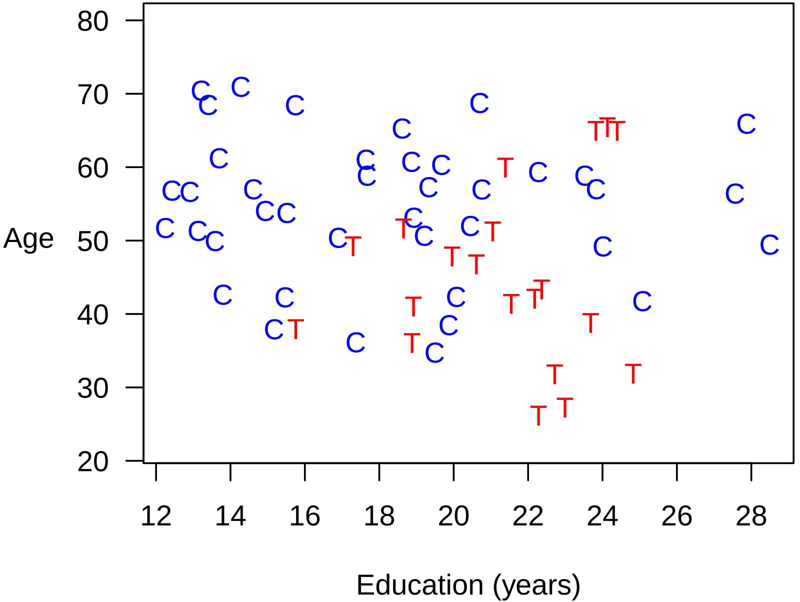
Mahalanobis Distance Matching



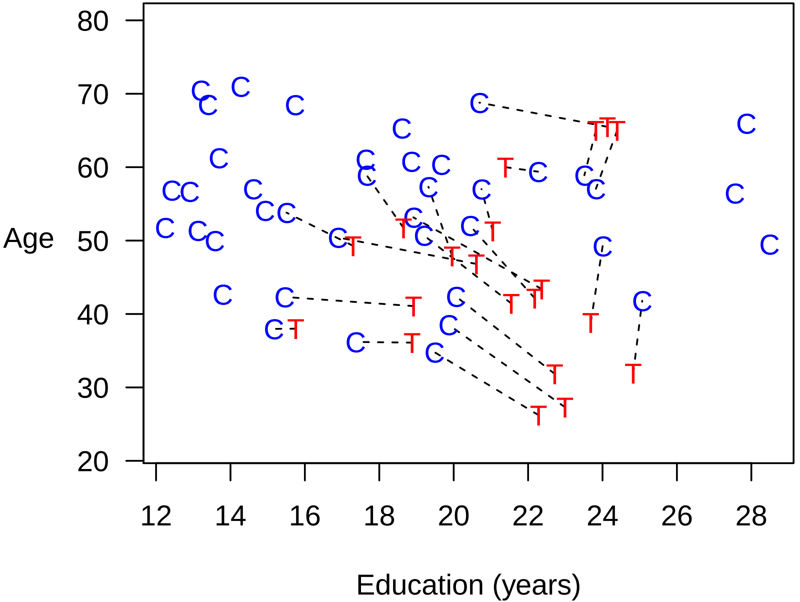
Mahalanobis Distance Matching



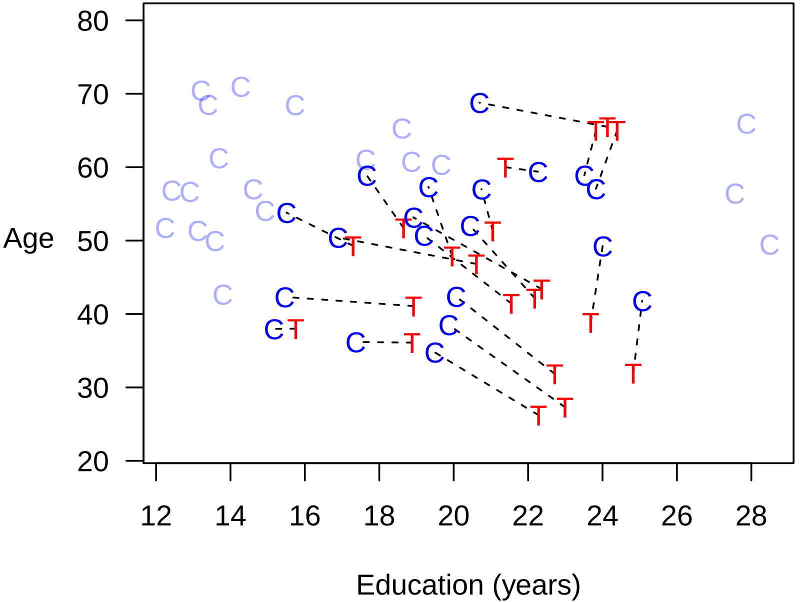
Mahalanobis Distance Matching



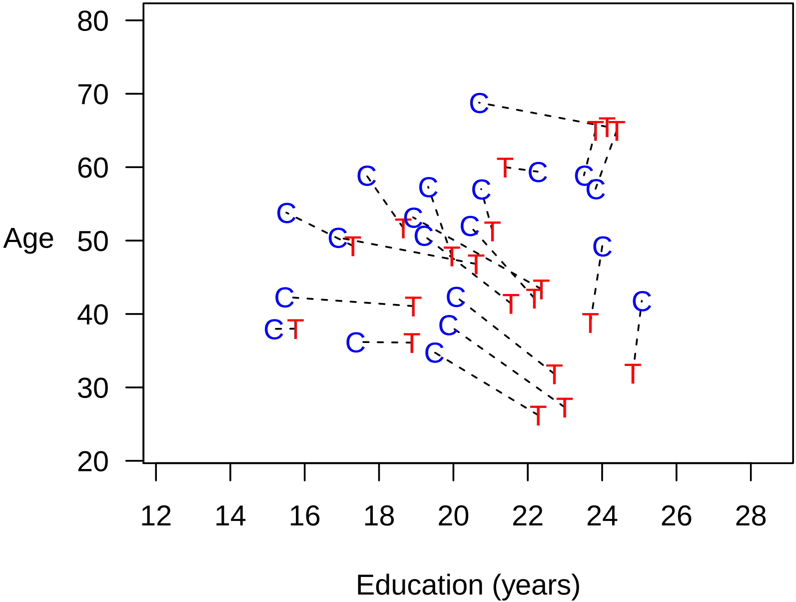
Mahalanobis Distance Matching



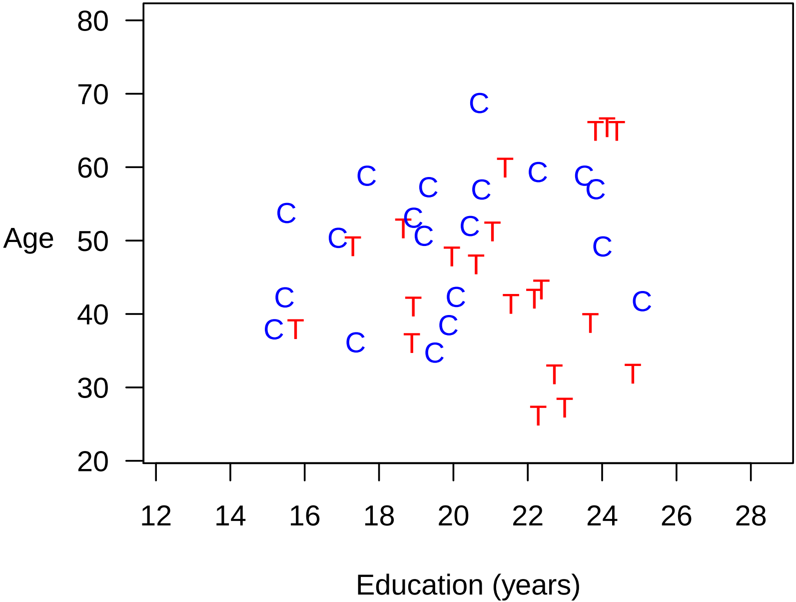
Mahalanobis Distance Matching



Mahalanobis Distance Matching



Mahalanobis Distance Matching



Method 2: Coarsened Exact Matching

Method 2: Coarsened Exact Matching

(Approximates Fully Blocked Experiment)

Method 2: Coarsened Exact Matching

(Approximates Fully Blocked Experiment)

1. Preprocess (Matching)

2. Checking Determine matched sample size, tweak, repeat, . . .

3. Estimation Difference in means or a model

Method 2: Coarsened Exact Matching

(Approximates Fully Blocked Experiment)

1. Preprocess (Matching)
   - Temporarily coarsen X as much as you're willing
2. Checking Determine matched sample size, tweak, repeat, ...
3. Estimation Difference in means or a model

Method 2: Coarsened Exact Matching

(Approximates Fully Blocked Experiment)

1. Preprocess (Matching)
   - Temporarily coarsen X as much as you're willing
     - e.g., Education (grade school, high school, college, graduate)

2. Checking Determine matched sample size, tweak, repeat, . . .

3. Estimation Difference in means or a model

Method 2: Coarsened Exact Matching

(Approximates Fully Blocked Experiment)

1. Preprocess (Matching)
   - Temporarily coarsen X as much as you're willing
     - e.g., Education (grade school, high school, college, graduate)
   - Apply exact matching to the coarsened X, $C(X)$
2. Checking Determine matched sample size, tweak, repeat, . . .
3. Estimation Difference in means or a model

Method 2: Coarsened Exact Matching

(Approximates Fully Blocked Experiment)

1. Preprocess (Matching)
   - Temporarily coarsen X as much as you're willing
     - e.g., Education (grade school, high school, college, graduate)
   - Apply exact matching to the coarsened X, $C(X)$
     - Sort observations into strata, each with unique values of $C(X)$
2. Checking Determine matched sample size, tweak, repeat, ...
3. Estimation Difference in means or a model

Method 2: Coarsened Exact Matching

(Approximates Fully Blocked Experiment)

1. Preprocess (Matching)
   - Temporarily coarsen X as much as you're willing
     - e.g., Education (grade school, high school, college, graduate)
   - Apply exact matching to the coarsened X, $C(X)$
     - Sort observations into strata, each with unique values of $C(X)$
     - Prune any stratum with 0 treated or 0 control units
2. Checking Determine matched sample size, tweak, repeat, ...
3. Estimation Difference in means or a model

Method 2: Coarsened Exact Matching

(Approximates Fully Blocked Experiment)

1. Preprocess (Matching)
   - Temporarily coarsen X as much as you're willing
     - e.g., Education (grade school, high school, college, graduate)
   - Apply exact matching to the coarsened X, $C(X)$
     - Sort observations into strata, each with unique values of $C(X)$
     - Prune any stratum with 0 treated or 0 control units
   - Pass on original (uncoarsened) units except those pruned
2. Checking Determine matched sample size, tweak, repeat, ...
3. Estimation Difference in means or a model

Method 2: Coarsened Exact Matching

(Approximates Fully Blocked Experiment)

1. Preprocess (Matching)
   - Temporarily coarsen X as much as you're willing
     - e.g., Education (grade school, high school, college, graduate)
   - Apply exact matching to the coarsened X, $C(X)$
     - Sort observations into strata, each with unique values of $C(X)$
     - Prune any stratum with 0 treated or 0 control units
   - Pass on original (uncoarsened) units except those pruned
2. Checking Determine matched sample size, tweak, repeat, ...
   - Easier, but still iterative
3. Estimation Difference in means or a model

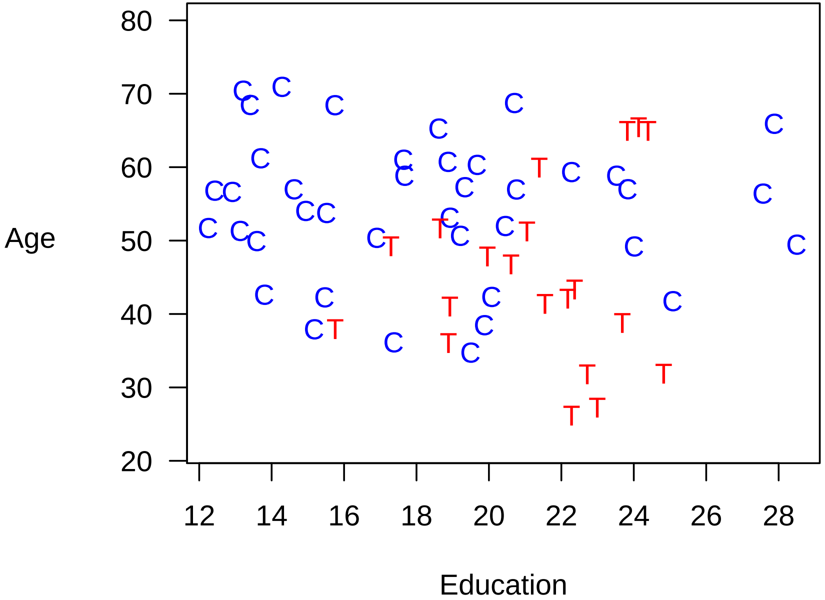
Method 2: Coarsened Exact Matching

(Approximates Fully Blocked Experiment)

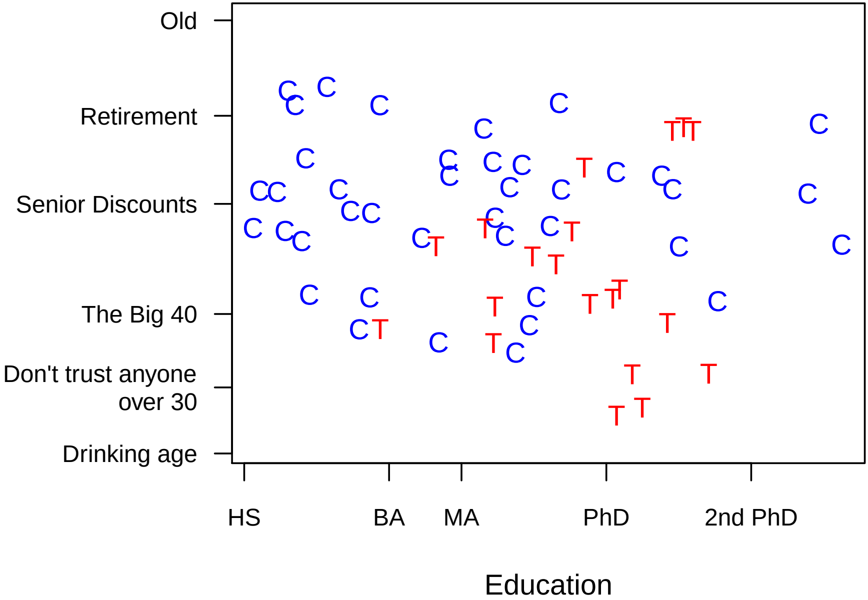
1. Preprocess (Matching)
   - Temporarily coarsen X as much as you're willing
     - e.g., Education (grade school, high school, college, graduate)
   - Apply exact matching to the coarsened X, $C(X)$
     - Sort observations into strata, each with unique values of $C(X)$
     - Prune any stratum with 0 treated or 0 control units
   - Pass on original (uncoarsened) units except those pruned
2. Checking Determine matched sample size, tweak, repeat, ...
   - Easier, but still iterative
3. Estimation Difference in means or a model
   - Need to weight controls in each stratum to equal treateds

Coarsened Exact Matching

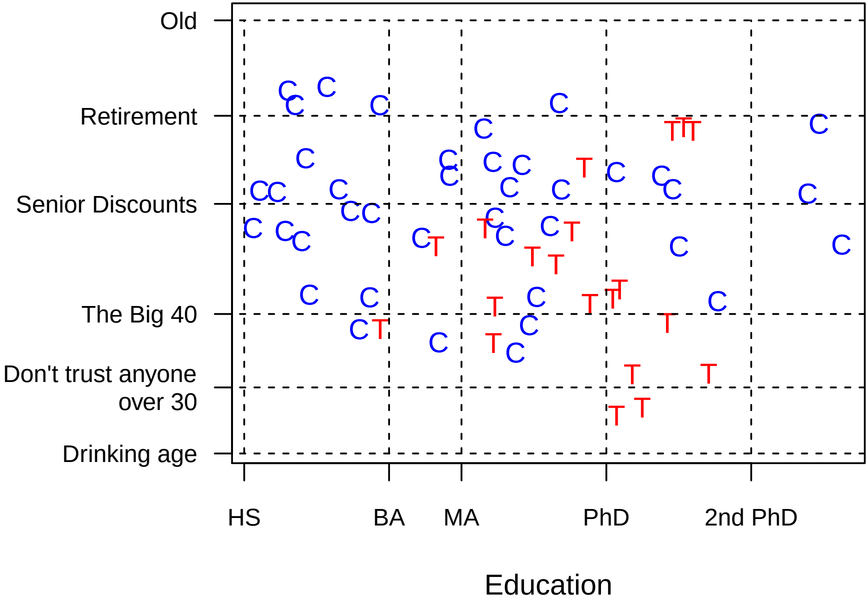
Coarsened Exact Matching



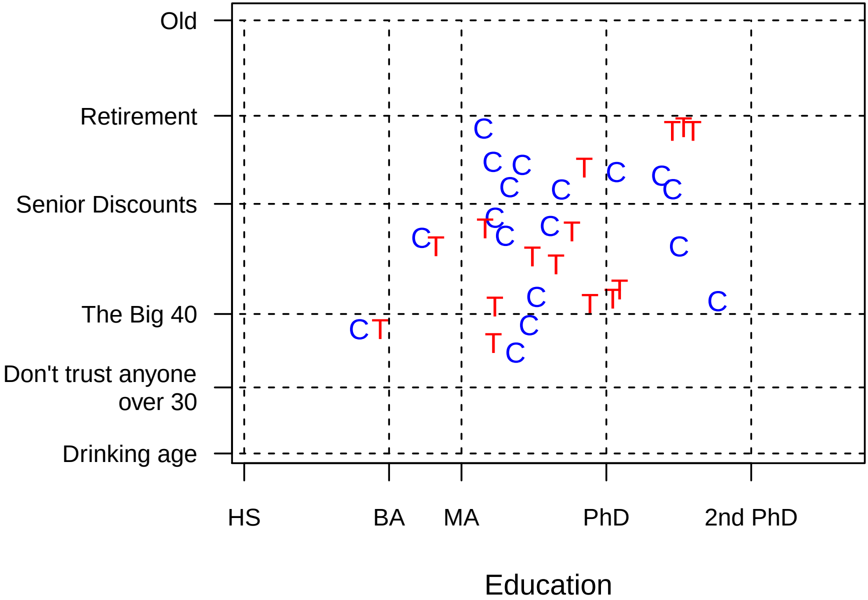
Coarsened Exact Matching



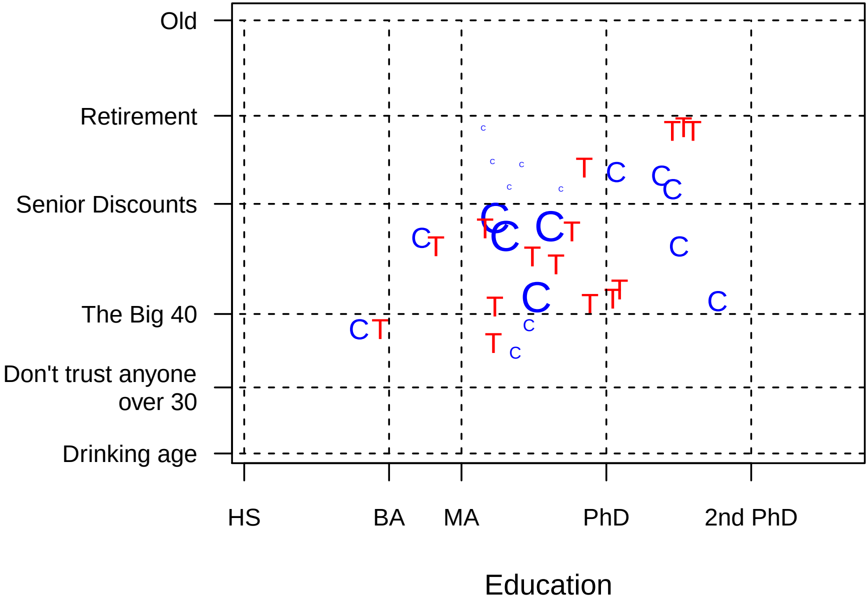
Coarsened Exact Matching



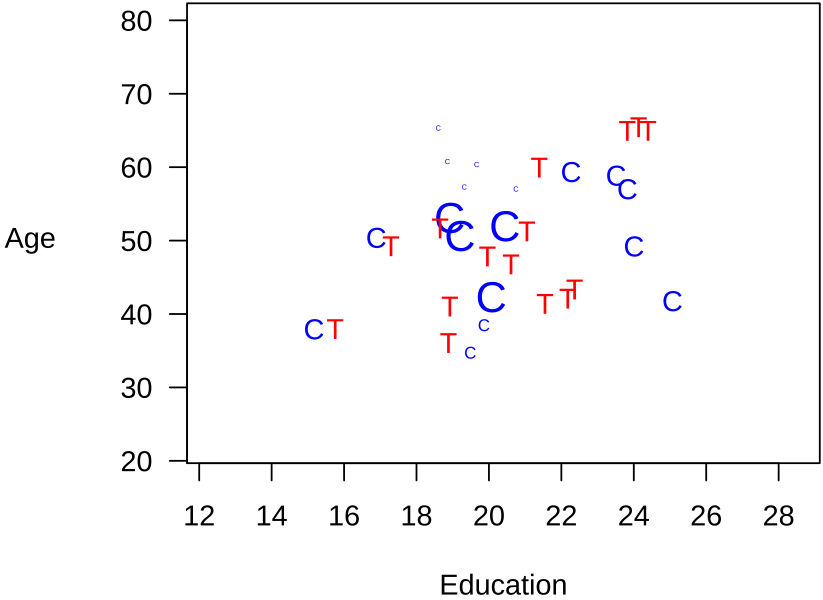
Coarsened Exact Matching



Coarsened Exact Matching



Coarsened Exact Matching



Method 3: Propensity Score Matching

Method 3: Propensity Score Matching

(Approximates Completely Randomized Experiment)

Method 3: Propensity Score Matching

(Approximates Completely Randomized Experiment)

1. Preprocess (Matching)

2. Checking Measure imbalance, tweak, repeat, ...
3. Estimation Difference in means or a model

Method 3: Propensity Score Matching

(Approximates Completely Randomized Experiment)

1. Preprocess (Matching)
   - Reduce k elements of X to scalar $\pi_i \equiv \Pr(T_i = 1|X) = \frac{1}{1+e^{-X_i\beta}}$
2. Checking Measure imbalance, tweak, repeat, ...
3. Estimation Difference in means or a model

Method 3: Propensity Score Matching

(Approximates Completely Randomized Experiment)

1. Preprocess (Matching)
   - Reduce k elements of X to scalar $\pi_i \equiv \Pr(T_i = 1|X) = \frac{1}{1+e^{-X_i\beta}}$
   - $\text{Distance}(X_i, X_j) = |\pi_i - \pi_j|$
2. Checking Measure imbalance, tweak, repeat, ...
3. Estimation Difference in means or a model

Method 3: Propensity Score Matching

(Approximates Completely Randomized Experiment)

1. Preprocess (Matching)
   - Reduce k elements of X to scalar $\pi_i \equiv \Pr(T_i = 1|X) = \frac{1}{1+e^{-X_i\beta}}$
   - $\text{Distance}(X_i, X_j) = |\pi_i - \pi_j|$
   - Match each treated unit to the nearest control unit
2. Checking Measure imbalance, tweak, repeat, ...
3. Estimation Difference in means or a model

Method 3: Propensity Score Matching

(Approximates Completely Randomized Experiment)

1. Preprocess (Matching)
    - Reduce k elements of X to scalar $\pi_i \equiv \Pr(T_i = 1|X) = \frac{1}{1+e^{-X_i \beta}}$
    - $\text{Distance}(X_i, X_j) = |\pi_i - \pi_j|$
    - Match each treated unit to the nearest control unit
    - Control units: not reused; pruned if unused
2. Checking Measure imbalance, tweak, repeat, . . .
3. Estimation Difference in means or a model

Method 3: Propensity Score Matching

(Approximates Completely Randomized Experiment)

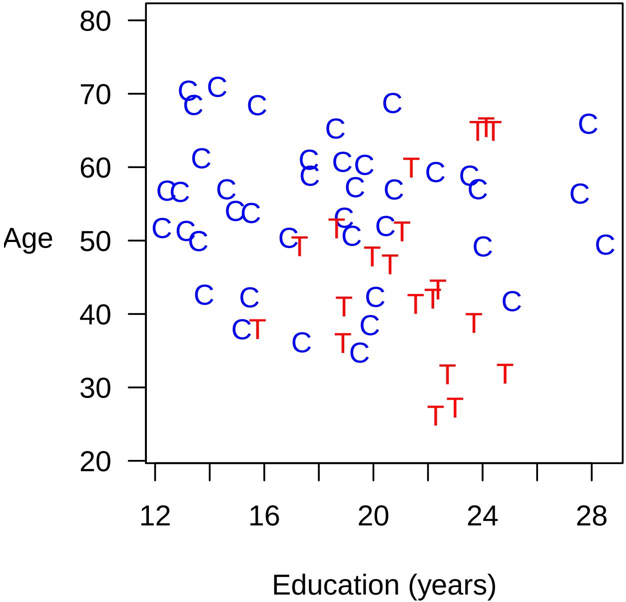
1. Preprocess (Matching)
   - Reduce k elements of X to scalar $\pi_i \equiv \Pr(T_i = 1|X) = \frac{1}{1+e^{-X_i\beta}}$
   - $\text{Distance}(X_i, X_j) = |\pi_i - \pi_j|$
   - Match each treated unit to the nearest control unit
   - Control units: not reused; pruned if unused
   - Prune matches if Distance$>$*caliper*
2. Checking Measure imbalance, tweak, repeat, ...
3. Estimation Difference in means or a model

Method 3: Propensity Score Matching

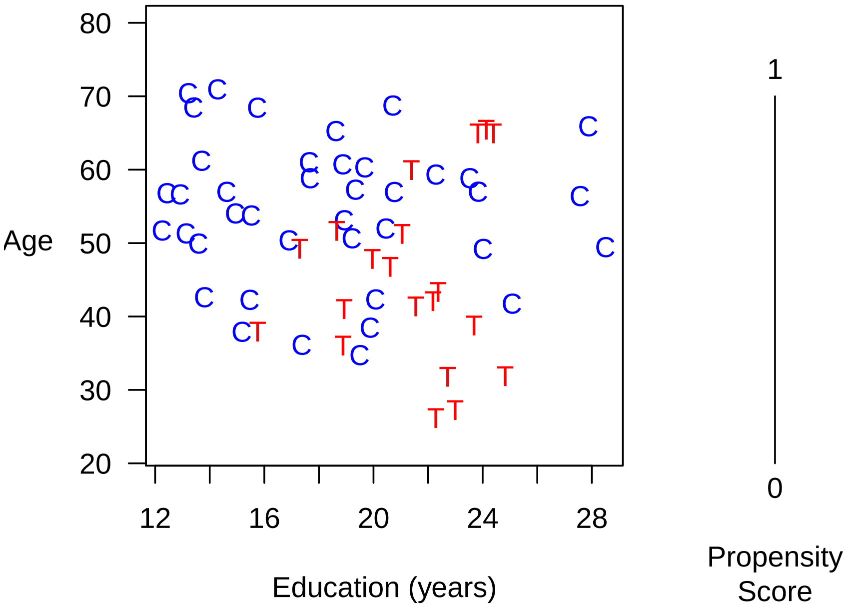
(Approximates Completely Randomized Experiment)

1. Preprocess (Matching)
   - Reduce k elements of X to scalar $\pi_i \equiv \Pr(T_i = 1|X) = \frac{1}{1+e^{-X_i\beta}}$
   - $\text{Distance}(X_i, X_j) = |\pi_i - \pi_j|$
   - Match each treated unit to the nearest control unit
   - Control units: not reused; pruned if unused
   - Prune matches if Distance>*caliper*
2. Checking Measure imbalance, tweak, repeat, . . .
3. Estimation Difference in means or a model

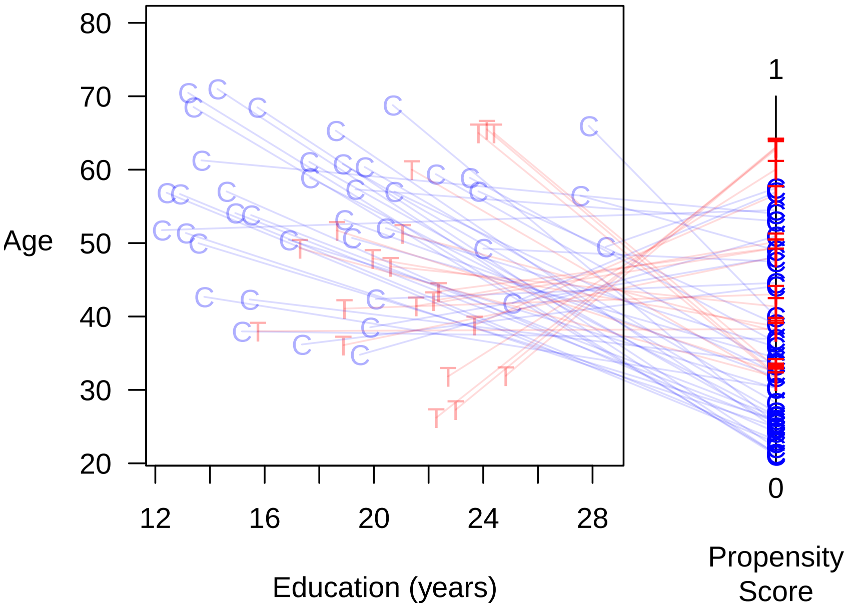
Propensity Score Matching



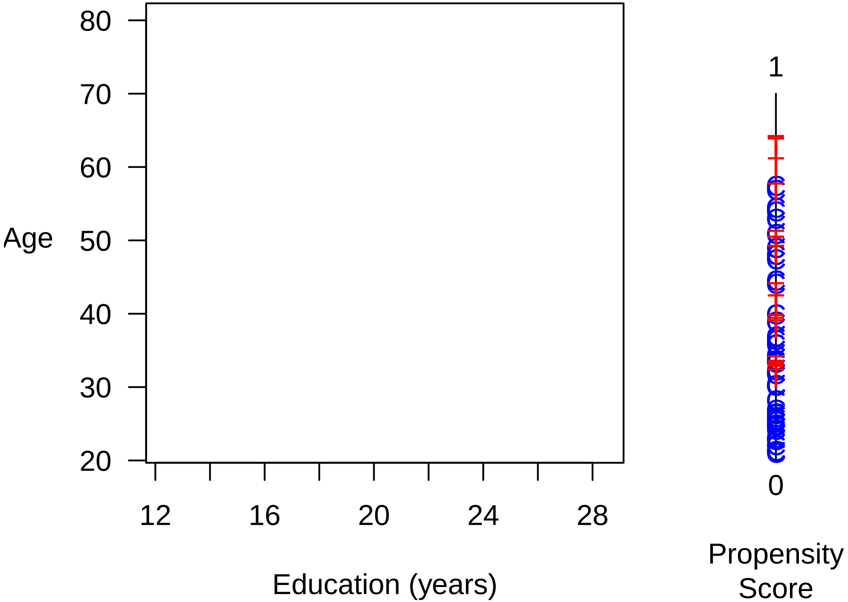
Propensity Score Matching



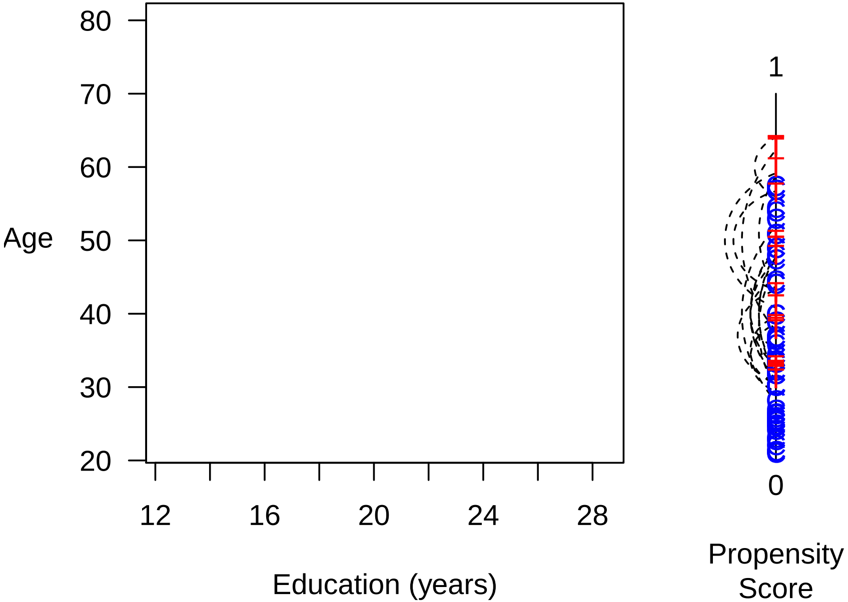
Propensity Score Matching



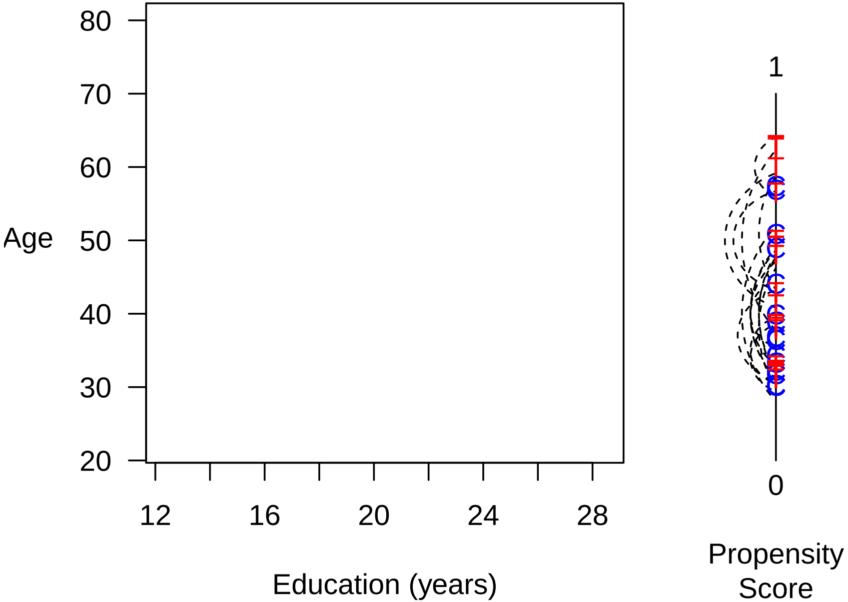
Propensity Score Matching



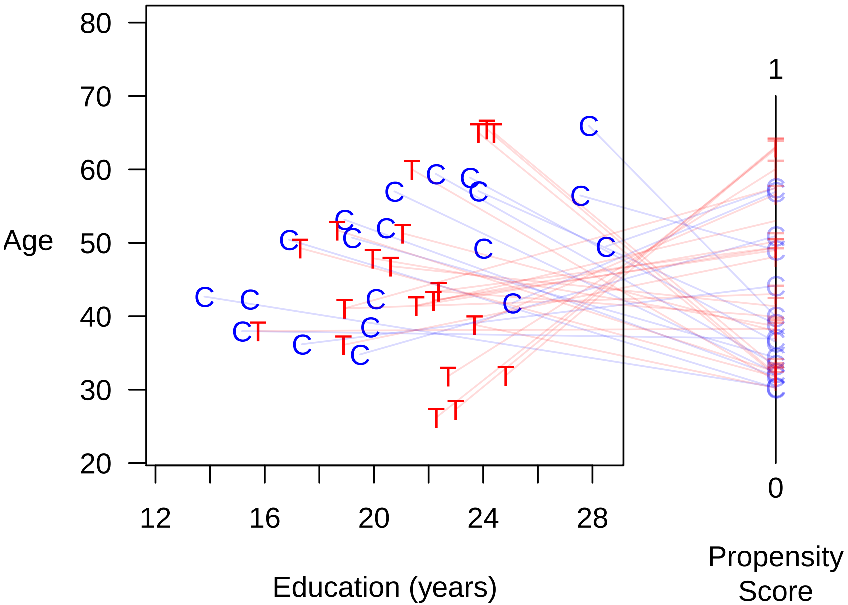
Propensity Score Matching



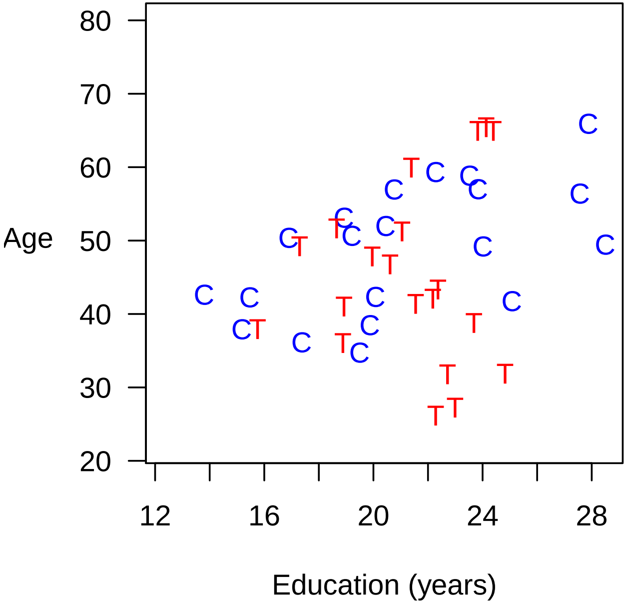
Propensity Score Matching



Propensity Score Matching



Propensity Score Matching



What to do with matched data?

- You matched and pruned the data of non-matches, now what?

What to do with matched data?

- You matched and pruned the data of non-matches, now what?
- Exact matching: simple difference in means.

What to do with matched data?

- You matched and pruned the data of non-matches, now what?
- Exact matching: simple difference in means.
- Inexact matching: there will be matching discrepancy:

$$W_i = X_i - X_{j(i)}$$

What to do with matched data?

- You matched and pruned the data of non-matches, now what?
- Exact matching: simple difference in means.
- Inexact matching: there will be matching discrepancy:

$$W_i = X_i - X_{j(i)}$$

- If balance is good then W_i should be quite small, but could still be large and produce bias.

What to do with matched data?

- You matched and pruned the data of non-matches, now what?
- Exact matching: simple difference in means.
- Inexact matching: there will be matching discrepancy:

$$W_i = X_i - X_{j(i)}$$

- If balance is good then W_i should be quite small, but could still be large and produce bias.
- Matching discrepancy will grow with the dimension of X_i

Bias of inexact matching

- Let $\mu_c(x) = \mathbf{E}[Y_i(0)|X_i = x]$ be how the mean of $Y_i(0)$ changes as a function of X_i.

Bias of inexact matching

- Let $\mu_c(x) = \mathbf{E}[Y_i(0)|X_i = x]$ be how the mean of $Y_i(0)$ changes as a function of X_i.
- Take a single matched pair produced by matching:

$$\widehat{\tau}_{mi} = Y_i - Y_{j(i)}$$

Bias of inexact matching

- Let $\mu_c(x) = \mathbf{E}[Y_i(0)|X_i = x]$ be how the mean of $Y_i(0)$ changes as a function of X_i.
- Take a single matched pair produced by matching:

$$\widehat{\tau}_{mi} = Y_i - Y_{j(i)}$$

- We hope this estimates $\tau(X_i)$, but there is actually bias:

$$\mathbf{E}[\widehat{\tau}_{mi}|D_i = 1, X_i, X_{j(i)}] = \tau(X_i) + \underbrace{(\mu_c(X_i) - \mu_c(X_{j(i)}))}_{\text{unit-level bias}}$$

Bias of inexact matching

- Let $\mu_c(x) = \mathbf{E}[Y_i(0)|X_i = x]$ be how the mean of $Y_i(0)$ changes as a function of X_i.
- Take a single matched pair produced by matching:

$$\widehat{\tau}_{mi} = Y_i - Y_{j(i)}$$

- We hope this estimates $\tau(X_i)$, but there is actually bias:

$$\mathbf{E}[\widehat{\tau}_{mi}|D_i = 1, X_i, X_{j(i)}] = \tau(X_i) + \underbrace{(\mu_c(X_i) - \mu_c(X_{j(i)}))}_{\text{unit-level bias}}$$

- If X_i has a big effect on the mean of $Y_i(0)$ then this bias could be big!

Bias-corrected estimators

$$B_i = \mu_c(X_i) - \mu_c(X_{j(i)})$$

- How do we get rid of this bias?

Bias-corrected estimators

$$B_i = \mu_c(X_i) - \mu_c(X_{j(i)})$$

- How do we get rid of this bias?
  - Estimate it, $\widehat{B}_i$, and subtract it off, $(Y_i - Y_{j(i)}) - \widehat{B}_i$

Bias-corrected estimators

$$B_i = \mu_c(X_i) - \mu_c(X_{j(i)})$$

- How do we get rid of this bias?
  - Estimate it, $\widehat{B}_i$, and subtract it off, $(Y_i - Y_{j(i)}) - \widehat{B}_i$
- Specify a parametric model for $\mu_c(x) = \alpha_c + x'\beta_c$ and estimate $\widehat{\beta}_c$ from the control data:

$$\widehat{B}_i = \widehat{\mu}_c(X_i) - \widehat{\mu}_c(X_{j(i)}) = (X_i - X_{j(i)})'\widehat{\beta}_c$$

Bias-corrected estimators

$$B_i = \mu_c(X_i) - \mu_c(X_{j(i)})$$

- How do we get rid of this bias?
  - Estimate it, $\widehat{B}_i$, and subtract it off, $(Y_i - Y_{j(i)}) - \widehat{B}_i$
- Specify a parametric model for $\mu_c(x) = \alpha_c + x'\beta_c$ and estimate $\widehat{\beta}_c$ from the control data:

$$\widehat{B}_i = \widehat{\mu}_c(X_i) - \widehat{\mu}_c(X_{j(i)}) = (X_i - X_{j(i)})'\widehat{\beta}_c$$

- Specification of $\mu_c(x)$ will matter less after matching.

Bias-corrected estimators

$$B_i = \mu_c(X_i) - \mu_c(X_{j(i)})$$

- How do we get rid of this bias?
  - Estimate it, $\widehat{B}_i$, and subtract it off, $(Y_i - Y_{j(i)}) - \widehat{B}_i$
- Specify a parametric model for $\mu_c(x) = \alpha_c + x'\beta_c$ and estimate $\widehat{\beta}_c$ from the control data:

$$\widehat{B}_i = \widehat{\mu}_c(X_i) - \widehat{\mu}_c(X_{j(i)}) = (X_i - X_{j(i)})'\widehat{\beta}_c$$

- Specification of $\mu_c(x)$ will matter less after matching.
- Create bias-corrected/adjusted imputations for $Y_i(0)$:

$$\widehat{Y}_i(0) = Y_{j(i)} + (X_i - X_{j(i)})'\widehat{\beta}_c$$

Bias-corrected inference

$$\widehat{Y}_i(0) = Y_{j(i)} + (X_i - X_{j(i)})'\widehat{\beta}_c$$

Bias-corrected inference

$$\widehat{Y}_i(0) = Y_{j(i)} + (X_i - X_{j(i)})'\widehat{\beta}_c$$

- Plug this into the same estimator:

Bias-corrected inference

$$\widehat{Y}_i(0) = Y_{j(i)} + (X_i - X_{j(i)})'\widehat{\beta}_c$$

- Plug this into the same estimator:

Bias-corrected inference

$$\widehat{Y}_i(0) = Y_{j(i)} + (X_i - X_{j(i)})'\widehat{\beta}_c$$

- Plug this into the same estimator:

$$\widehat{\tau}_{m,bc} = \frac{1}{N_t}\sum_{i=1}^{N_t}\left(Y_i - \widehat{Y}_i(0)\right)$$

Bias-corrected inference

$$\widehat{Y}_i(0) = Y_{j(i)} + (X_i - X_{j(i)})'\widehat{\beta}_c$$

- Plug this into the same estimator:

$$\widehat{\tau}_{m,bc} = \frac{1}{N_t}\sum_{i=1}^{N_t}\left(Y_i - \widehat{Y}_i(0)\right)$$

- Variance estimation for this quantity is easiest without replacement.

Bias-corrected inference

$$\widehat{Y}_i(0) = Y_{j(i)} + (X_i - X_{j(i)})'\widehat{\beta}_c$$

- Plug this into the same estimator:

$$\widehat{\tau}_{m,bc} = \frac{1}{N_t} \sum_{i=1}^{N_t} \left(Y_i - \widehat{Y}_i(0) \right)$$

- Variance estimation for this quantity is easiest without replacement.
- Simply take the variance of the within-match differences:

$$\widehat{\mathrm{Var}}[\widehat{\tau}_m] = \frac{1}{N_t} \sum_{i=1}^{N_t} \left(Y_i - \widehat{Y}_i(0) - \widehat{\tau}_{m,bc} \right)^2$$

Fully pooled model

- What if we simply run our original analysis model on the pooled, matching data:

$$\tilde{Y}_i = \alpha_p + \tau_p \cdot \tilde{D}_i + \tilde{X}_i' \beta_p + \nu_i$$

Fully pooled model

- What if we simply run our original analysis model on the pooled, matching data:
$$\tilde{Y}_i = \alpha_p + \tau_p \cdot \tilde{D}_i + \tilde{X}_i' \beta_p + \nu_i$$
- $\tilde{Y}_i$ is the $2 \times N_t$ matched treated and control units stacked.

Fully pooled model

- What if we simply run our original analysis model on the pooled, matching data:

$$\tilde{Y}_i = \alpha_p + \tau_p \cdot \tilde{D}_i + \tilde{X}_i' \beta_p + \nu_i$$

- $\tilde{Y}_i$ is the $2 \times N_t$ matched treated and control units stacked.
- $\widehat{\tau}_p$ from OLS on this model is a bias-corrected estimate where we assume that:

$$\mu_c(x) = \mu_t(x)$$

Fully pooled model

- What if we simply run our original analysis model on the pooled, matching data:

$$\tilde{Y}_i = \alpha_p + \tau_p \cdot \tilde{D}_i + \tilde{X}_i' \beta_p + \nu_i$$

- $\tilde{Y}_i$ is the $2 \times N_t$ matched treated and control units stacked.
- $\widehat{\tau}_p$ from OLS on this model is a bias-corrected estimate where we assume that:

$$\mu_c(x) = \mu_t(x)$$

- Still corrects for some of the residual bias left over from the matching.

Fully pooled model

- What if we simply run our original analysis model on the pooled, matching data:
$$\tilde{Y}_i = \alpha_p + \tau_p \cdot \tilde{D}_i + \tilde{X}_i' \beta_p + \nu_i$$
- $\tilde{Y}_i$ is the $2 \times N_t$ matched treated and control units stacked.
- $\widehat{\tau}_p$ from OLS on this model is a bias-corrected estimate where we assume that:
$$\mu_c(x) = \mu_t(x)$$
- Still corrects for some of the residual bias left over from the matching.
- SEs from these models might make additional assumptions (homoskedasticity, etc).

Final Thoughts on Matching

- Causal inference is hard due to the fundamental problem of causal inference

Final Thoughts on Matching

- Causal inference is hard due to the fundamental problem of causal inference
- You always have to make some assumption, there is no magical panacea.

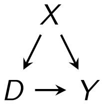
Final Thoughts on Matching

- Causal inference is hard due to the fundamental problem of causal inference
- You always have to make some assumption, there is no magical panacea.
- Even if you aren't running an experiment, thinking through the ideal experiment will help you.

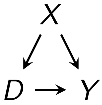
Final Thoughts on Matching

- Causal inference is hard due to the fundamental problem of causal inference
- You always have to make some assumption, there is no magical panacea.
- Even if you aren't running an experiment, thinking through the ideal experiment will help you.
- and remember: Matching is not an identification strategy

Selection

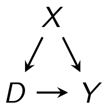


Selection



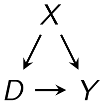
- In selection on observables approaches we are trying to block open backdoor paths.

Selection



- In selection on observables approaches we are trying to block open backdoor paths.
- A useful way to think about this problem is what causes selection into treatment.

Selection



- In selection on observables approaches we are trying to block open backdoor paths.
- A useful way to think about this problem is what causes selection into treatment.
- Propensity score methods emphasize this interpretation by focusing on estimating the probability that a unit will take the treatment.

Propensity Score as a Low-Dimensional Summary

- Stratification is hard when X has has many dimensions

Propensity Score as a Low-Dimensional Summary

- Stratification is hard when X has has many dimensions
- Curse of dimensionality: there will be very few, if any, units in a given stratum of X_i.

Propensity Score as a Low-Dimensional Summary

- Stratification is hard when X has has many dimensions
- Curse of dimensionality: there will be very few, if any, units in a given stratum of X_i.
- We can instead stratify on a low-dimensional summary, the propensity score:

$$e(x) = \mathbb{P}[D_i = 1 | X_i = x]$$

Propensity Score as a Low-Dimensional Summary

- Stratification is hard when X has has many dimensions
- Curse of dimensionality: there will be very few, if any, units in a given stratum of X_i.
- We can instead stratify on a low-dimensional summary, the propensity score:

$$e(x) = \mathbb{P}[D_i = 1 | X_i = x]$$

  - PS = unit's probability of being treated, conditional on X_i

Propensity Score as a Low-Dimensional Summary

- Stratification is hard when X has has many dimensions
- Curse of dimensionality: there will be very few, if any, units in a given stratum of X_i.
- We can instead stratify on a low-dimensional summary, the propensity score:

$$e(x) = \mathbb{P}[D_i = 1 | X_i = x]$$

  - PS = unit's probability of being treated, conditional on X_i
- For a particular unit, this is $e(X_i) = \mathbb{P}[D_i = 1 | X_i]$

Propensity Score as a Low-Dimensional Summary

- Stratification is hard when X has has many dimensions
- Curse of dimensionality: there will be very few, if any, units in a given stratum of X_i.
- We can instead stratify on a low-dimensional summary, the propensity score:

$$e(x) = \mathbb{P}[D_i = 1 | X_i = x]$$

  - PS = unit's probability of being treated, conditional on X_i
- For a particular unit, this is $e(X_i) = \mathbb{P}[D_i = 1 | X_i]$
- Rosenbaum and Rubin (1983) showed that:

$$D_i \perp\!\!\!\perp \big(Y_i(0), Y_i(1)\big) \mid X_i \implies D_i \perp\!\!\!\perp \big(Y_i(0), Y_i(1)\big) \mid e(X_i)$$

Propensity Score as a Low-Dimensional Summary

- Stratification is hard when X has has many dimensions
- Curse of dimensionality: there will be very few, if any, units in a given stratum of X_i.
- We can instead stratify on a low-dimensional summary, the propensity score:

$$e(x) = \mathbb{P}[D_i = 1 | X_i = x]$$

  - PS = unit's probability of being treated, conditional on X_i
- For a particular unit, this is $e(X_i) = \mathbb{P}[D_i = 1 | X_i]$
- Rosenbaum and Rubin (1983) showed that:

$$D_i \perp\!\!\!\perp \big(Y_i(0), Y_i(1)\big) \mid X_i \implies D_i \perp\!\!\!\perp \big(Y_i(0), Y_i(1)\big) \mid e(X_i)$$

  - $\rightsquigarrow$ stratifying on e_i is the same in expectation as stratifying on the full X_i.

Propensity score as balancing score

- The propensity score is actually a balancing score, which means that

$$D_i \perp\!\!\!\perp X_i \mid e(X_i)$$

Propensity score as balancing score

- The propensity score is actually a balancing score, which means that

$$D_i \perp\!\!\!\perp X_i \mid e(X_i)$$

- Conditional on the propensity score, treatment is independent of the covariates.

Propensity score as balancing score

- The propensity score is actually a balancing score, which means that

$$D_i \perp\!\!\!\perp X_i \mid e(X_i)$$

- Conditional on the propensity score, treatment is independent of the covariates.
  - Treatment status is said to be balanced

Propensity score as balancing score

- The propensity score is actually a balancing score, which means that

$$D_i \perp\!\!\!\perp X_i \mid e(X_i)$$

- Conditional on the propensity score, treatment is independent of the covariates.
    - Treatment status is said to be balanced
    - $f(X_i|D_i = 1, e(X_i)) = f(X_i|D_i = 0, e(X_i))$

Propensity score as balancing score

- The propensity score is actually a balancing score, which means that

$$D_i \perp\!\!\!\perp X_i \mid e(X_i)$$

- Conditional on the propensity score, treatment is independent of the covariates.
    - Treatment status is said to be balanced
    - $f(X_i|D_i = 1, e(X_i)) = f(X_i|D_i = 0, e(X_i))$
- However, we have to know the true PS to have all these results work!

Estimating the propensity score

- Of course, in observational studies, we don't know the propensity score.

Estimating the propensity score

- Of course, in observational studies, we don't know the propensity score.
- We would run a parametric model with parameters γ to estimate the propensity scores:

Estimating the propensity score

- Of course, in observational studies, we don't know the propensity score.
- We would run a parametric model with parameters γ to estimate the propensity scores:

1. Estimate $\hat{\gamma}$

Estimating the propensity score

- Of course, in observational studies, we don't know the propensity score.
- We would run a parametric model with parameters γ to estimate the propensity scores:

1. Estimate $\hat{\gamma}$
2. Create $\hat{e}_i = \Pr[D_i = 1 | X_i; \hat{\gamma}]$

Estimating the propensity score

- Of course, in observational studies, we don't know the propensity score.
- We would run a parametric model with parameters γ to estimate the propensity scores:

1. Estimate $\hat{\gamma}$
2. Create $\hat{e}_i = \Pr[D_i = 1 | X_i; \hat{\gamma}]$

- For instance, in R, we could easily calculate the propensity scores using the `glm` function:

Estimating the propensity score

- Of course, in observational studies, we don't know the propensity score.
- We would run a parametric model with parameters γ to estimate the propensity scores:

1. Estimate $\hat{\gamma}$
2. Create $\hat{e}_i = \Pr[D_i = 1 | X_i; \hat{\gamma}]$

- For instance, in R, we could easily calculate the propensity scores using the `glm` function:

Propensity score specifics

- What variables do we include in the propensity score model?

Propensity score specifics

- What variables do we include in the propensity score model?
  - Any set of variables that blocks all the backdoor paths from D_i to Y_i.

Propensity score specifics

- What variables do we include in the propensity score model?
  - Any set of variables that blocks all the backdoor paths from D_i to Y_i.
- Check balance within strata of $\hat{e}_i$. Covariates should be balanced:

$$f(X_i|D_i = 1, \hat{e}_i) = f(X_i|D_i = 0, \hat{e}_i)$$

Propensity score specifics

- What variables do we include in the propensity score model?
  - Any set of variables that blocks all the backdoor paths from D_i to Y_i.
- Check balance within strata of $\hat{e}_i$. Covariates should be balanced:

$$f(X_i|D_i = 1, \hat{e}_i) = f(X_i|D_i = 0, \hat{e}_i)$$

- Can also use automated/nonparametric tools for estimating $\hat{e}_i$.

Propensity score specifics

- What variables do we include in the propensity score model?
  - Any set of variables that blocks all the backdoor paths from D_i to Y_i.
- Check balance within strata of $\hat{e}_i$. Covariates should be balanced:

$$f(X_i|D_i = 1, \hat{e}_i) = f(X_i|D_i = 0, \hat{e}_i)$$

- Can also use automated/nonparametric tools for estimating $\hat{e}_i$.
  - Covariate Balancing Propensity Scores (Imai and Ratkovic)

Using the propensity score

How do we use the propensity score?

- Propensity score can be used in many contexts: weighting, matching, regression or even just stratification

Using the propensity score

How do we use the propensity score?

- Propensity score can be used in many contexts: weighting, matching, regression or even just stratification
- It also shows up in a number of more advanced methods for heterogeneous treatment effects, causal inference in longitudinal data etc.

Using the propensity score

How do we use the propensity score?

- Propensity score can be used in many contexts: weighting, matching, regression or even just stratification
- It also shows up in a number of more advanced methods for heterogeneous treatment effects, causal inference in longitudinal data etc.
- We will focus on settings where the propensity score is a tool to achieve balance.

Using the propensity score

How do we use the propensity score?

- Propensity score can be used in many contexts: weighting, matching, regression or even just stratification
- It also shows up in a number of more advanced methods for heterogeneous treatment effects, causal inference in longitudinal data etc.
- We will focus on settings where the propensity score is a tool to achieve balance.
- However, note that the propensity score only achieves balance in expectation

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)*

Identification 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 $\pi(X)$.

$$\Pr(D = 1|Y_1, Y_0, \pi(X)) = \mathbf{E}[D|Y_1, Y_0, \pi(X)]$$

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 $\pi(X)$.

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

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 $\pi(X)$.

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

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 $\pi(X)$.

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

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 $\pi(X)$.

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

Using a similar argument

$$\Pr(D = 1|\pi(X)) = \mathbf{E}[D|\pi(X)] = \mathbf{E}[\mathbf{E}[D|X]|\pi(x)]$$

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 $\pi(X)$.

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

Using a similar argument

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

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

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:

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

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: $P(X|D = 1, \widehat{\pi}(X)) = P(X|D = 0, \widehat{\pi}(X))$

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

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.)

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

Weighting with the Propensity Score

Intuition

- Treated and control samples are unrepresentative of the overall population.

Weighting with the Propensity Score

Intuition

- Treated and control samples are unrepresentative of the overall population.
- Leads to imbalance in the covariates.

Weighting with the Propensity Score

Intuition

- Treated and control samples are unrepresentative of the overall population.
- Leads to imbalance in the covariates.
- Reweight them to be more representative.

Survey samples

- Useful to review survey samples to understand the logic

Survey samples

- Useful to review survey samples to understand the logic
- Finite population: $\{1, \ldots, N\}$

Survey samples

- Useful to review survey samples to understand the logic
- Finite population: $\{1, \ldots, N\}$
- Suppose that we wanted estimate the population mean of Y_i:

$$\bar{Y}_N = \frac{1}{N} \sum_{i=1}^{N} Y_i$$

Survey samples

- Useful to review survey samples to understand the logic
- Finite population: $\{1, \dots, N\}$
- Suppose that we wanted estimate the population mean of Y_i:

$$\bar{Y}_N = \frac{1}{N} \sum_{i=1}^{N} Y_i$$

- We have a sample of size n, where $Z_i = 1$ indicates that i is included in the sample.

Survey samples

- Useful to review survey samples to understand the logic
- Finite population: $\{1, \ldots, N\}$
- Suppose that we wanted estimate the population mean of Y_i:

$$\bar{Y}_N = \frac{1}{N} \sum_{i=1}^{N} Y_i$$

- We have a sample of size n, where $Z_i = 1$ indicates that i is included in the sample.
- Unequal sampling probability: $\mathbb{P}(Z_i = 1) = \pi_i$

Survey samples

- Useful to review survey samples to understand the logic
- Finite population: $\{1, \dots, N\}$
- Suppose that we wanted estimate the population mean of Y_i:

$$\bar{Y}_N = \frac{1}{N} \sum_{i=1}^{N} Y_i$$

- We have a sample of size n, where $Z_i = 1$ indicates that i is included in the sample.
- Unequal sampling probability: $\mathbb{P}(Z_i = 1) = \pi_i$
  - $\rightsquigarrow$ sample is not representative.

Survey samples

- Useful to review survey samples to understand the logic
- Finite population: $\{1, \dots, N\}$
- Suppose that we wanted estimate the population mean of Y_i:

$$\bar{Y}_N = \frac{1}{N} \sum_{i=1}^{N} Y_i$$

- We have a sample of size n, where $Z_i = 1$ indicates that i is included in the sample.
- Unequal sampling probability: $\mathbb{P}(Z_i = 1) = \pi_i$
  - $\rightsquigarrow$ sample is not representative.
  - $\sum_{i=1}^{N} \pi_i = n$

Survey weights

- Sample mean is biased:

$$\mathbf{E}\left[\frac{1}{n}\sum_{i=1}^{N} Z_i Y_i\right] = \frac{1}{n}\sum_{i=1} \pi_i Y_i$$

Survey weights

- Sample mean is biased:

$$\mathbf{E}\left[\frac{1}{n}\sum_{i=1}^{N} Z_i Y_i\right] = \frac{1}{n}\sum_{i=1} \pi_i Y_i$$

- Inverse probability weighting: To correct, weight each unit by the reciprocal of the probability of being included in the sample: Y_i/π_i.

Survey weights

- Sample mean is biased:

$$\mathbf{E}\left[\frac{1}{n}\sum_{i=1}^{N} Z_i Y_i\right] = \frac{1}{n}\sum_{i=1} \pi_i Y_i$$

- Inverse probability weighting: To correct, weight each unit by the reciprocal of the probability of being included in the sample: Y_i/π_i.
- Horvitz-Thompson estimator is unbiased:

Survey weights

- Sample mean is biased:

$$\mathbf{E}\left[\frac{1}{n}\sum_{i=1}^{N} Z_i Y_i\right] = \frac{1}{n}\sum_{i=1} \pi_i Y_i$$

- Inverse probability weighting: To correct, weight each unit by the reciprocal of the probability of being included in the sample: Y_i/π_i.
- Horvitz-Thompson estimator is unbiased:

$$\mathbf{E}\left[\frac{1}{N}\sum_{i=1}^{N} \frac{Z_i Y_i}{\pi_i}\right]$$

Survey weights

- Sample mean is biased:

$$\mathbf{E}\left[\frac{1}{n}\sum_{i=1}^{N} Z_i Y_i\right] = \frac{1}{n}\sum_{i=1} \pi_i Y_i$$

- Inverse probability weighting: To correct, weight each unit by the reciprocal of the probability of being included in the sample: Y_i/π_i.
- Horvitz-Thompson estimator is unbiased:

$$\mathbf{E}\left[\frac{1}{N}\sum_{i=1}^{N} \frac{Z_i Y_i}{\pi_i}\right] = \frac{1}{N}\sum_{i=1}^{N} \frac{\mathbf{E}[Z_i] Y_i}{\pi_i}$$

Survey weights

- Sample mean is biased:

$$\mathbf{E}\left[\frac{1}{n}\sum_{i=1}^{N} Z_i Y_i\right] = \frac{1}{n}\sum_{i=1} \pi_i Y_i$$

- Inverse probability weighting: To correct, weight each unit by the reciprocal of the probability of being included in the sample: Y_i/π_i.
- Horvitz-Thompson estimator is unbiased:

$$\mathbf{E}\left[\frac{1}{N}\sum_{i=1}^{N} \frac{Z_i Y_i}{\pi_i}\right] = \frac{1}{N}\sum_{i=1}^{N} \frac{\mathbf{E}[Z_i] Y_i}{\pi_i} = \frac{1}{N}\sum_{i=1}^{N} \frac{\pi_i Y_i}{\pi_i}$$

Survey weights

- Sample mean is biased:

$$\mathbf{E}\left[\frac{1}{n}\sum_{i=1}^{N} Z_i Y_i\right] = \frac{1}{n}\sum_{i=1} \pi_i Y_i$$

- Inverse probability weighting: To correct, weight each unit by the reciprocal of the probability of being included in the sample: Y_i/π_i.
- Horvitz-Thompson estimator is unbiased:

$$\mathbf{E}\left[\frac{1}{N}\sum_{i=1}^{N} \frac{Z_i Y_i}{\pi_i}\right] = \frac{1}{N}\sum_{i=1}^{N} \frac{\mathbf{E}[Z_i] Y_i}{\pi_i} = \frac{1}{N}\sum_{i=1}^{N} \frac{\pi_i Y_i}{\pi_i} = \bar{Y}_N$$

- Reweights the sample to be representative of the population.

Back to causal effects

- With a completely randomized experiment, we can just use the simple differences in means:

$$\mathbf{E}[Y_i|D_i = 1] - \mathbf{E}[Y_i|D_i = 0] = \mathbf{E}[Y_i(1)] - \mathbf{E}[Y_i(0)]$$

Back to causal effects

- With a completely randomized experiment, we can just use the simple differences in means:

$$\mathbf{E}[Y_i|D_i = 1] - \mathbf{E}[Y_i|D_i = 0] = \mathbf{E}[Y_i(1)] - \mathbf{E}[Y_i(0)]$$

- With no unmeasured confounders, we need to adjust for X_i.

Back to causal effects

- With a completely randomized experiment, we can just use the simple differences in means:

$$\mathbf{E}[Y_i|D_i = 1] - \mathbf{E}[Y_i|D_i = 0] = \mathbf{E}[Y_i(1)] - \mathbf{E}[Y_i(0)]$$

- With no unmeasured confounders, we need to adjust for X_i.

$$\mathbf{E}[Y_i(d)] = \mathbf{E}\left[\mathbf{E}[Y_i(d)|X_i]\right]$$

Back to causal effects

- With a completely randomized experiment, we can just use the simple differences in means:

$$\mathbf{E}[Y_i|D_i = 1] - \mathbf{E}[Y_i|D_i = 0] = \mathbf{E}[Y_i(1)] - \mathbf{E}[Y_i(0)]$$

- With no unmeasured confounders, we need to adjust for X_i.

$$\begin{aligned}\mathbf{E}[Y_i(d)] &= \mathbf{E}\left[\mathbf{E}[Y_i(d)|X_i]\right] \\ &= \sum_{x \in \mathcal{X}} \mathbf{E}[Y_i(d)|X_i = x]\mathbb{P}(X_i = x)\end{aligned}$$

Back to causal effects

- With a completely randomized experiment, we can just use the simple differences in means:

$$\mathbf{E}[Y_i|D_i = 1] - \mathbf{E}[Y_i|D_i = 0] = \mathbf{E}[Y_i(1)] - \mathbf{E}[Y_i(0)]$$

- With no unmeasured confounders, we need to adjust for X_i.

$$\begin{aligned}
\mathbf{E}[Y_i(d)] &= \mathbf{E}\left[\mathbf{E}[Y_i(d)|X_i]\right] \\
&= \sum_{x\in\mathcal{X}} \mathbf{E}[Y_i(d)|X_i = x]\mathbb{P}(X_i = x) \\
&= \sum_{x\in\mathcal{X}} \mathbf{E}[Y_i(d)|D_i = d, X_i = x]\mathbb{P}(X_i = x)
\end{aligned}$$

Back to causal effects

- With a completely randomized experiment, we can just use the simple differences in means:

$$\mathbf{E}[Y_i|D_i = 1] - \mathbf{E}[Y_i|D_i = 0] = \mathbf{E}[Y_i(1)] - \mathbf{E}[Y_i(0)]$$

- With no unmeasured confounders, we need to adjust for X_i.

$$\begin{aligned}
\mathbf{E}[Y_i(d)] &= \mathbf{E}\left[\mathbf{E}[Y_i(d)|X_i]\right] \\
&= \sum_{x \in \mathcal{X}} \mathbf{E}[Y_i(d)|X_i = x]\mathbb{P}(X_i = x) \\
&= \sum_{x \in \mathcal{X}} \mathbf{E}[Y_i(d)|D_i = d, X_i = x]\mathbb{P}(X_i = x) \\
&= \sum_{x \in \mathcal{X}} \mathbf{E}[Y_i|D_i = d, X_i = x]\mathbb{P}(X_i = x)
\end{aligned}$$

Back to causal effects

- With a completely randomized experiment, we can just use the simple differences in means:

$$\mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0] = \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)]$$

- With no unmeasured confounders, we need to adjust for X_i.

$$\begin{aligned}
\mathbb{E}[Y_i(d)] &= \mathbb{E}\left[\mathbb{E}[Y_i(d)|X_i]\right] \\
&= \sum_{x \in \mathcal{X}} \mathbb{E}[Y_i(d)|X_i = x]\mathbb{P}(X_i = x) \\
&= \sum_{x \in \mathcal{X}} \mathbb{E}[Y_i(d)|D_i = d, X_i = x]\mathbb{P}(X_i = x) \\
&= \sum_{x \in \mathcal{X}} \mathbb{E}[Y_i|D_i = d, X_i = x]\mathbb{P}(X_i = x)
\end{aligned}$$

- With subclassification, we binned X_i, calclulated within-bin differences and then averaged across the bins, just like this.

Searching for the weights

$$\mathbf{E}[Y_i(d)] = \sum_{x \in \mathcal{X}} \mathbf{E}[Y_i | D_i = d, X_i = x] \mathbb{P}(X_i = x)$$

- Compare this to the the within treatment group average:

$$\mathbf{E}[Y_i | D_i = d] = \sum_{x \in \mathcal{X}} \mathbf{E}[Y_i | D_i = d, X_i = x] \mathbb{P}(X_i = x | D_i = d)$$

Searching for the weights

$$\mathbb{E}[Y_i(d)] = \sum_{x \in \mathcal{X}} \mathbb{E}[Y_i | D_i = d, X_i = x] \mathbb{P}(X_i = x)$$

- Compare this to the the within treatment group average:

$$\begin{aligned} \mathbb{E}[Y_i | D_i = d] &= \sum_{x \in \mathcal{X}} \mathbb{E}[Y_i | D_i = d, X_i = x] \mathbb{P}(X_i = x | D_i = d) \\ &= \sum_{x \in \mathcal{X}} \mathbb{E}[Y_i | D_i = d, X_i = x] \frac{\mathbb{P}(D_i = d | X_i = x)\mathbb{P}(X_i = x)}{\mathbb{P}(D_i = d)} \end{aligned}$$

Searching for the weights

$$\mathbf{E}[Y_i(d)] = \sum_{x \in \mathcal{X}} \mathbf{E}[Y_i | D_i = d, X_i = x] \mathbb{P}(X_i = x)$$

- Compare this to the the within treatment group average:

$$\begin{aligned}\mathbf{E}[Y_i | D_i = d] &= \sum_{x \in \mathcal{X}} \mathbf{E}[Y_i | D_i = d, X_i = x] \mathbb{P}(X_i = x | D_i = d) \\ &= \sum_{x \in \mathcal{X}} \mathbf{E}[Y_i | D_i = d, X_i = x] \frac{\mathbb{P}(D_i = d | X_i = x) \mathbb{P}(X_i = x)}{\mathbb{P}(D_i = d)}\end{aligned}$$

- How should we reweight the data from an observational study?

Searching for the weights

$$\mathbf{E}[Y_i(d)] = \sum_{x \in \mathcal{X}} \mathbf{E}[Y_i | D_i = d, X_i = x] \mathbb{P}(X_i = x)$$

- Compare this to the the within treatment group average:

$$\begin{aligned}
\mathbf{E}[Y_i | D_i = d] &= \sum_{x \in \mathcal{X}} \mathbf{E}[Y_i | D_i = d, X_i = x] \mathbb{P}(X_i = x | D_i = d) \\
&= \sum_{x \in \mathcal{X}} \mathbf{E}[Y_i | D_i = d, X_i = x] \frac{\mathbb{P}(D_i = d | X_i = x) \mathbb{P}(X_i = x)}{\mathbb{P}(D_i = d)}
\end{aligned}$$

- How should we reweight the data from an observational study?
- If we were to reweight the data by $W_i = 1/\mathbb{P}(D_i = d | X_i)$, then we would break the relationship between D_i and X_i.

Weights

- Single binary covariate. Define the weight function:

$$w(d, x) = \frac{1}{e(x)^d (1 - e(x))^{1-d}}$$

Weights

- Single binary covariate. Define the weight function:

$$w(d, x) = \frac{1}{e(x)^d (1 - e(x))^{1-d}}$$

- To get the weight for i, plug in observed treatment, covariate: $W_i = w(D_i, X_i)$

Weights

- Single binary covariate. Define the weight function:

$$w(d,x) = \frac{1}{e(x)^d(1-e(x))^{1-d}}$$

- To get the weight for i, plug in observed treatment, covariate: $W_i = w(D_i, X_i)$
- If $(D_i, X_i) = (1, 1)$,

$$W_i = \frac{1}{e(1)} = \frac{1}{\mathbb{P}(D_i = 1 | X_i = 1)}$$

Weights

- Single binary covariate. Define the weight function:

$$w(d, x) = \frac{1}{e(x)^d (1 - e(x))^{1-d}}$$

- To get the weight for i, plug in observed treatment, covariate: $W_i = w(D_i, X_i)$
- If $(D_i, X_i) = (1, 1)$,

$$W_i = \frac{1}{e(1)} = \frac{1}{\mathbb{P}(D_i = 1 | X_i = 1)}$$

- If $(D_i, X_i) = (0, 0)$:

$$W_i = \frac{1}{1 - e(0)} = \frac{1}{\mathbb{P}(D_i = 0 | X_i = 0)}$$

Example

	$X_i = 0$	$X_i = 1$
$D_i = 0$	4	3
$D_i = 1$	4	9

- $\mathbb{P}(D_i = 1 | X_i = 0) = 0.5$

Example

	$X_i = 0$	$X_i = 1$
$D_i = 0$	4	3
$D_i = 1$	4	9

- $\mathbb{P}(D_i = 1 | X_i = 0) = 0.5$
- $\mathbb{P}(D_i = 1 | X_i = 1) = 0.75$

Example

	$X_i = 0$	$X_i = 1$
$D_i = 0$	4	3
$D_i = 1$	4	9

- $\mathbb{P}(D_i = 1 | X_i = 0) = 0.5$
- $\mathbb{P}(D_i = 1 | X_i = 1) = 0.75$
- Weights:

Example

	$X_i = 0$	$X_i = 1$
$D_i = 0$	4	3
$D_i = 1$	4	9

- $\mathbb{P}(D_i = 1 | X_i = 0) = 0.5$
- $\mathbb{P}(D_i = 1 | X_i = 1) = 0.75$
- Weights:

	$X_i = 0$	$X_i = 1$
$D_i = 0$	$1/0.5$	$1/0.25$
$D_i = 1$	$1/0.5$	$1/0.75$

Example

	$X_i = 0$	$X_i = 1$
$D_i = 0$	4	3
$D_i = 1$	4	9

- $\mathbb{P}(D_i = 1 | X_i = 0) = 0.5$
- $\mathbb{P}(D_i = 1 | X_i = 1) = 0.75$
- Weights:

	$X_i = 0$	$X_i = 1$
$D_i = 0$	2	4
$D_i = 1$	2	$4/3$

Example

	$X_i = 0$	$X_i = 1$
$D_i = 0$	4	3
$D_i = 1$	4	9

- $\mathbb{P}(D_i = 1 | X_i = 0) = 0.5$
- $\mathbb{P}(D_i = 1 | X_i = 1) = 0.75$
- Weights:

	$X_i = 0$	$X_i = 1$
$D_i = 0$	2	4
$D_i = 1$	2	$4/3$

- Weighted data (the pseudo-population):

	$X_i = 0$	$X_i = 1$
$D_i = 0$	8	12
$D_i = 1$	8	12

Example

	$X_i = 0$	$X_i = 1$
$D_i = 0$	4	3
$D_i = 1$	4	9

- $\mathbb{P}(D_i = 1 | X_i = 0) = 0.5$
- $\mathbb{P}(D_i = 1 | X_i = 1) = 0.75$
- Weights:

	$X_i = 0$	$X_i = 1$
$D_i = 0$	2	4
$D_i = 1$	2	$4/3$

- Weighted data (the pseudo-population):

	$X_i = 0$	$X_i = 1$
$D_i = 0$	8	12
$D_i = 1$	8	12

- $\mathbb{P}_W(D_i = 1 | X_i = x) = 0.5$ for all x

Properties of reweighted data

- Let's calculate the weighted probability that $D_i = 1$.

Properties of reweighted data

- Let's calculate the weighted probability that $D_i = 1$.

$$\mathbb{P}_W[D_i = 1|X_i = x]$$

Properties of reweighted data

- Let's calculate the weighted probability that $D_i = 1$.

$$\mathbb{P}_W[D_i = 1 | X_i = x] = \frac{w(1, x) \cdot \mathbb{P}[D_i = 1 | X_i = x]}{\omega^*}$$

Properties of reweighted data

- Let's calculate the weighted probability that $D_i = 1$.

$$\begin{aligned} & \mathbb{P}_W[D_i = 1|X_i = x] \\ &= \frac{w(1,x) \cdot \mathbb{P}[D_i = 1|X_i = x]}{\omega^*} \\ &= \frac{\frac{1}{\mathbb{P}[D_i=1|X_i=x]} \cdot \mathbb{P}[D_i = 1|X_i = x]}{\omega^*} \end{aligned}$$

Properties of reweighted data

- Let's calculate the weighted probability that $D_i = 1$.

$$\begin{aligned}
&\mathbb{P}_W[D_i = 1|X_i = x] \\
&= \frac{w(1,x) \cdot \mathbb{P}[D_i = 1|X_i = x]}{\omega^*} \\
&= \frac{\frac{1}{\mathbb{P}[D_i=1|X_i=x]} \cdot \mathbb{P}[D_i = 1|X_i = x]}{\omega^*} \\
&= \frac{1}{\omega^*}.
\end{aligned}$$

Properties of reweighted data

- Let's calculate the weighted probability that $D_i = 1$.

$$
\begin{aligned}
&\mathbb{P}_W[D_i = 1 | X_i = x] \\
&= \frac{w(1, x) \cdot \mathbb{P}[D_i = 1 | X_i = x]}{\omega^*} \\
&= \frac{\frac{1}{\mathbb{P}[D_i = 1 | X_i = x]} \cdot \mathbb{P}[D_i = 1 | X_i = x]}{\omega^*} \\
&= \frac{1}{\omega^*}.
\end{aligned}
$$

- ω^* is a normalization factor to make sure probabilities sum to 1.

Properties of reweighted data

- Let's calculate the weighted probability that $D_i = 1$.

$$
\begin{aligned}
&\mathbb{P}_W[D_i = 1 | X_i = x] \\
&= \frac{w(1,x) \cdot \mathbb{P}[D_i = 1 | X_i = x]}{\omega^*} \\
&= \frac{\frac{1}{\mathbb{P}[D_i = 1 | X_i = x]} \cdot \mathbb{P}[D_i = 1 | X_i = x]}{\omega^*} \\
&= \frac{1}{\omega^*}.
\end{aligned}
$$

- ω^* is a normalization factor to make sure probabilities sum to 1.
- Important point: $\mathbb{P}_W(D_i = 1 | X_i = 1) = \mathbb{P}_W(D_i = 1 | X_i = 0) = \frac{1}{\omega^*}$

Properties of reweighted data

- Let's calculate the weighted probability that $D_i = 1$.

$$\begin{aligned}
&\mathbb{P}_W[D_i = 1|X_i = x] \\
&= \frac{w(1,x) \cdot \mathbb{P}[D_i = 1|X_i = x]}{\omega^*} \\
&= \frac{\frac{1}{\mathbb{P}[D_i=1|X_i=x]} \cdot \mathbb{P}[D_i = 1|X_i = x]}{\omega^*} \\
&= \frac{1}{\omega^*}.
\end{aligned}$$

- ω^* is a normalization factor to make sure probabilities sum to 1.
- Important point: $\mathbb{P}_W(D_i = 1|X_i = 1) = \mathbb{P}_W(D_i = 1|X_i = 0) = \frac{1}{\omega^*}$
- $\rightsquigarrow D_i$ independent of X_i in the reweighted data.

Overall mean

- What is the weighted mean for the treated group?

Overall mean

- What is the weighted mean for the treated group?
- Use a similar approach to survey weights, where D_i is the "sampling indicator":

$$\bar{Y}_i^w = \frac{1}{N} \sum_{i=1}^{N} D_i W_i Y_i$$

Overall mean

- What is the weighted mean for the treated group?
- Use a similar approach to survey weights, where D_i is the "sampling indicator":

$$\bar{Y}_i^w = \frac{1}{N}\sum_{i=1}^{N} D_i W_i Y_i$$

- $W_i Y_i$ is the weighted outcome, D_i is there to select out the treated observations.

Overall mean

- What is the weighted mean for the treated group?
- Use a similar approach to survey weights, where D_i is the "sampling indicator":

$$\bar{Y}_i^w = \frac{1}{N}\sum_{i=1}^{N} D_i W_i Y_i$$

- $W_i Y_i$ is the weighted outcome, D_i is there to select out the treated observations.
- We want to see what the conditional weighted mean identifies:

$$\mathbf{E}\left[\frac{1}{N}\sum_{i=1}^{N} W_i D_i Y_i\right] = \frac{1}{N}\sum_{i=1}^{N}\mathbf{E}[W_i D_i Y_i] = \mathbf{E}[W_i D_i Y_i]$$

Proving unbiasedness

- Weighted mean of treated units is mean of potential outcome:

Proving unbiasedness

- Weighted mean of treated units is mean of potential outcome:

$$\mathbf{E}[W_i D_i Y_i] = \mathbf{E}\left[\frac{D_i Y_i}{e(X_i)}\right]$$

Proving unbiasedness

- Weighted mean of treated units is mean of potential outcome:

$$\mathbf{E}[W_i D_i Y_i] = \mathbf{E}\left[\frac{D_i Y_i}{e(X_i)}\right] \qquad \text{(Weight Def.)}$$

$$= E\left[\frac{D_i Y_i(1)}{e(X_i)}\right] \qquad \text{(Consistency)}$$

Proving unbiasedness

- Weighted mean of treated units is mean of potential outcome:

$$\begin{aligned}
\mathbf{E}[W_i D_i Y_i] &= \mathbf{E}\left[\frac{D_i Y_i}{e(X_i)}\right] && \text{(Weight Def.)} \\
&= E\left[\frac{D_i Y_i(1)}{e(X_i)}\right] && \text{(Consistency)} \\
&= E\left[E\left[\frac{D_i Y_i(1)}{e(X_i)}\middle| X_i\right]\right] && \text{(Iterated Expectations)}
\end{aligned}$$

Proving unbiasedness

- Weighted mean of treated units is mean of potential outcome:

$$
\begin{aligned}
\mathbf{E}[W_i D_i Y_i] &= \mathbf{E}\left[\frac{D_i Y_i}{e(X_i)}\right] && \text{(Weight Def.)} \\
&= E\left[\frac{D_i Y_i(1)}{e(X_i)}\right] && \text{(Consistency)} \\
&= E\left[E\left[\frac{D_i Y_i(1)}{e(X_i)}\Big| X_i\right]\right] && \text{(Iterated Expectations)} \\
&= E\left[\frac{E[D_i|X_i]E[Y_i(1)|X_i]}{e(X_i)}\right] && \text{(n.u.c.)}
\end{aligned}
$$

Proving unbiasedness

- Weighted mean of treated units is mean of potential outcome:

$$
\begin{aligned}
\mathbf{E}[W_i D_i Y_i] &= \mathbf{E}\left[\frac{D_i Y_i}{e(X_i)}\right] && \text{(Weight Def.)} \\
&= E\left[\frac{D_i Y_i(1)}{e(X_i)}\right] && \text{(Consistency)} \\
&= E\left[E\left[\frac{D_i Y_i(1)}{e(X_i)}\middle| X_i\right]\right] && \text{(Iterated Expectations)} \\
&= E\left[\frac{E[D_i|X_i]E[Y_i(1)|X_i]}{e(X_i)}\right] && \text{(n.u.c.)} \\
&= E\left[\frac{e(X_i)E[Y_i(1)|X_i]}{e(X_i)}\right] && \text{(Propensity Score Definition)}
\end{aligned}
$$

Proving unbiasedness

- Weighted mean of treated units is mean of potential outcome:

$$
\begin{aligned}
\mathbf{E}[W_i D_i Y_i] &= \mathbf{E}\left[\frac{D_i Y_i}{e(X_i)}\right] && \text{(Weight Def.)} \\
&= E\left[\frac{D_i Y_i(1)}{e(X_i)}\right] && \text{(Consistency)} \\
&= E\left[E\left[\frac{D_i Y_i(1)}{e(X_i)}\middle| X_i\right]\right] && \text{(Iterated Expectations)} \\
&= E\left[\frac{E[D_i|X_i]E[Y_i(1)|X_i]}{e(X_i)}\right] && \text{(n.u.c.)} \\
&= E\left[\frac{e(X_i)E[Y_i(1)|X_i]}{e(X_i)}\right] && \text{(Propensity Score Definition)} \\
&= E[Y_i(1)] && \text{(Iterated Expectations)}
\end{aligned}
$$

Proving unbiasedness

- Weighted mean of treated units is mean of potential outcome:

$$
\begin{aligned}
\mathbf{E}[W_i D_i Y_i] &= \mathbf{E}\left[\frac{D_i Y_i}{e(X_i)}\right] && \text{(Weight Def.)} \\
&= E\left[\frac{D_i Y_i(1)}{e(X_i)}\right] && \text{(Consistency)} \\
&= E\left[E\left[\frac{D_i Y_i(1)}{e(X_i)}\middle| X_i\right]\right] && \text{(Iterated Expectations)} \\
&= E\left[\frac{E[D_i|X_i]E[Y_i(1)|X_i]}{e(X_i)}\right] && \text{(n.u.c.)} \\
&= E\left[\frac{e(X_i)E[Y_i(1)|X_i]}{e(X_i)}\right] && \text{(Propensity Score Definition)} \\
&= E[Y_i(1)] && \text{(Iterated Expectations)}
\end{aligned}
$$

Putting it all together

- The same logic would give us the mean potential outcomes under control:

$$E\left[\frac{(1-D_i)Y_i}{1-e(X_i)}\right] = E[Y_i(0)]$$

Putting it all together

- The same logic would give us the mean potential outcomes under control:
$$E\left[\frac{(1-D_i)Y_i}{1-e(X_i)}\right] = E[Y_i(0)]$$
- These two facts provide an estimator for the average treatment effect:

$$\hat{\tau} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{D_i Y_i}{e(X_i)} - \frac{(1-D_i)Y_i}{1-e(X_i)}\right)$$

Putting it all together

- The same logic would give us the mean potential outcomes under control:

$$E\left[\frac{(1-D_i)Y_i}{1-e(X_i)}\right] = E[Y_i(0)]$$

- These two facts provide an estimator for the average treatment effect:

$$\hat{\tau} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{D_i Y_i}{e(X_i)} - \frac{(1-D_i)Y_i}{1-e(X_i)}\right)$$

- The above two results give us that this esimator is unbiased.

Putting it all together

- The same logic would give us the mean potential outcomes under control:
$$E\left[\frac{(1-D_i)Y_i}{1-e(X_i)}\right] = E[Y_i(0)]$$
- These two facts provide an estimator for the average treatment effect:

$$\hat{\tau} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{D_i Y_i}{e(X_i)} - \frac{(1-D_i)Y_i}{1-e(X_i)}\right)$$

- The above two results give us that this esimator is unbiased.
- This is sometimes called the Horvitz-Thompson estimator due to the close connection to the survey sampling estimator.

Estimation of the propensity score

$$\hat{\tau} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{D_i Y_i}{e(X_i)} - \frac{(1-D_i)Y_i}{1-e(X_i)}\right)$$

- Need to know or estimate the propensity score, $e(X_i)$. How do we do that?

Estimation of the propensity score

$$\hat{\tau} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{D_i Y_i}{e(X_i)} - \frac{(1-D_i)Y_i}{1-e(X_i)}\right)$$

- Need to know or estimate the propensity score, $e(X_i)$. How do we do that?
- Discrete covariates estimate the within-strata propensity scores

$$\hat{e}(x) = \frac{N_{xd}}{N_x}$$

Estimation of the propensity score

$$\hat{\tau} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{D_i Y_i}{e(X_i)} - \frac{(1-D_i)Y_i}{1-e(X_i)}\right)$$

- Need to know or estimate the propensity score, $e(X_i)$. How do we do that?
- Discrete covariates estimate the within-strata propensity scores

$$\hat{e}(x) = \frac{N_{xd}}{N_x}$$

  - Non-parametric estimate of the propensity score in each stratum of the data.

Estimation of the propensity score

$$\hat{\tau} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{D_i Y_i}{e(X_i)} - \frac{(1-D_i)Y_i}{1-e(X_i)}\right)$$

- Need to know or estimate the propensity score, $e(X_i)$. How do we do that?
- Discrete covariates estimate the within-strata propensity scores

$$\hat{e}(x) = \frac{N_{xd}}{N_x}$$

  - Non-parametric estimate of the propensity score in each stratum of the data.
- Continuous covariates $\rightsquigarrow$ Logistic regression of D_i on X_i.

Why use estimated pscores?

- Obviously we often don't have the true propensity score, but in some circumstances, the estimate propensity score can be better than the true propensity score. Why?

Why use estimated pscores?

- Obviously we often don't have the true propensity score, but in some circumstances, the estimate propensity score can be better than the true propensity score. Why?
- Removing chance variations using $\hat{e}(X_i)$ adjusts for any small imbalances that arise because of a finite sample.

Why use estimated pscores?

- Obviously we often don't have the true propensity score, but in some circumstances, the estimate propensity score can be better than the true propensity score. Why?
- Removing chance variations using $\hat{e}(X_i)$ adjusts for any small imbalances that arise because of a finite sample.
- The true p-score only adjusts for the expected differences between samples.

Positivity violations

- Remember the positivity assumption:

$$0 < p(D_i = 1 | X_i) < 1$$

Positivity violations

- Remember the positivity assumption:

$$0 < p(D_i = 1|X_i) < 1$$

- What happens to the weights if this is violated? Then, $\hat{e}(x) = 0$ or $\hat{e}(x) = 1$ and

$$\frac{1}{\hat{e}(x)} = \frac{1}{0} = \infty$$

Positivity violations

- Remember the positivity assumption:

$$0 < p(D_i = 1|X_i) < 1$$

- What happens to the weights if this is violated? Then, $\hat{e}(x) = 0$ or $\hat{e}(x) = 1$ and

$$\frac{1}{\hat{e}(x)} = \frac{1}{0} = \infty$$

- Structural $\rightsquigarrow$ population probability is 0.

Positivity violations

- Remember the positivity assumption:

$$0 < p(D_i = 1|X_i) < 1$$

- What happens to the weights if this is violated? Then, $\hat{e}(x) = 0$ or $\hat{e}(x) = 1$ and

$$\frac{1}{\hat{e}(x)} = \frac{1}{0} = \infty$$

- Structural $\rightsquigarrow$ population probability is 0.
- Random $\rightsquigarrow$ sample probability is 0.

Positivity violations

- Remember the positivity assumption:

$$0 < p(D_i = 1|X_i) < 1$$

- What happens to the weights if this is violated? Then, $\hat{e}(x) = 0$ or $\hat{e}(x) = 1$ and

$$\frac{1}{\hat{e}(x)} = \frac{1}{0} = \infty$$

- Structural $\rightsquigarrow$ population probability is 0.
- Random $\rightsquigarrow$ sample probability is 0.
  - Need to "borrow" information from other values of X_i to estimate $e(X_i)$

Positivity violations

- Remember the positivity assumption:

$$0 < p(D_i = 1|X_i) < 1$$

- What happens to the weights if this is violated? Then, $\hat{e}(x) = 0$ or $\hat{e}(x) = 1$ and

$$\frac{1}{\hat{e}(x)} = \frac{1}{0} = \infty$$

- Structural $\rightsquigarrow$ population probability is 0.
- Random $\rightsquigarrow$ sample probability is 0.
  - Need to "borrow" information from other values of X_i to estimate $e(X_i)$
  - $\rightsquigarrow$ modeling via logit, etc.

Automated approaches

- Challenge: specifying the propensity score model.

$$\hat{e}(X_i) = \text{logit}^{-1}(X_i'\beta)$$

Automated approaches

- Challenge: specifying the propensity score model.

$$\hat{e}(X_i) = \text{logit}^{-1}(X_i'\beta)$$

- What terms should we include?

Automated approaches

- Challenge: specifying the propensity score model.

$$\hat{e}(X_i) = \text{logit}^{-1}(X_i'\beta)$$

- What terms should we include?
- Big problem for weights: small changes to PS model lead to big changes in the weights.

Automated approaches

- Challenge: specifying the propensity score model.

$$\hat{e}(X_i) = \text{logit}^{-1}(X_i'\beta)$$

- What terms should we include?
- Big problem for weights: small changes to PS model lead to big changes in the weights.
- Entropy balancing (Hainmueller 2012):

Automated approaches

- Challenge: specifying the propensity score model.

$$\hat{e}(X_i) = \text{logit}^{-1}(X_i'\beta)$$

- What terms should we include?
- Big problem for weights: small changes to PS model lead to big changes in the weights.
- Entropy balancing (Hainmueller 2012):
  - Choose weights for each observation that maximize the balance between treatment and control groups.

Automated approaches

- Challenge: specifying the propensity score model.

$$\hat{e}(X_i) = \text{logit}^{-1}(X_i'\beta)$$

- What terms should we include?
- Big problem for weights: small changes to PS model lead to big changes in the weights.
- Entropy balancing (Hainmueller 2012):
  - Choose weights for each observation that maximize the balance between treatment and control groups.
- Covariate Balancing Propensity Scores (Imai and Ratkovic):

Automated approaches

- Challenge: specifying the propensity score model.

$$\hat{e}(X_i) = \text{logit}^{-1}(X_i'\beta)$$

- What terms should we include?
- Big problem for weights: small changes to PS model lead to big changes in the weights.
- Entropy balancing (Hainmueller 2012):
  - Choose weights for each observation that maximize the balance between treatment and control groups.
- Covariate Balancing Propensity Scores (Imai and Ratkovic):
  - Estimate the propensity score subject to the additional constraint of maximizing balance.

Boostrapping to get the SEs

- How to get the standard error for $\hat{\tau}$?

Boostrapping to get the SEs

- How to get the standard error for $\hat{\tau}$?
- Variance estimators are messy $\rightsquigarrow$ use the bootstrap!

Boostrapping to get the SEs

- How to get the standard error for $\hat{\tau}$?
- Variance estimators are messy $\leadsto$ use the bootstrap!
  1. Draw a sample of the data with replacement, call this, S_b.

Boostrapping to get the SEs

- How to get the standard error for $\hat{\tau}$?
- Variance estimators are messy $\leadsto$ use the bootstrap!
  1. Draw a sample of the data with replacement, call this, S_b.
  2. Estimate the propensity scores in this sample, $\hat{e}_b$ and create weights, W_b.

Boostrapping to get the SEs

- How to get the standard error for $\hat{\tau}$?
- Variance estimators are messy $\rightsquigarrow$ use the bootstrap!
  1. Draw a sample of the data with replacement, call this, S_b.
  2. Estimate the propensity scores in this sample, $\hat{e}_b$ and create weights, W_b.
  3. Use the weights to get an estimate of the average treatment effect, τ_b in the sample S_b.

Boostrapping to get the SEs

- How to get the standard error for $\hat{\tau}$?
- Variance estimators are messy $\rightsquigarrow$ use the bootstrap!
  1. Draw a sample of the data with replacement, call this, S_b.
  2. Estimate the propensity scores in this sample, $\hat{e}_b$ and create weights, W_b.
  3. Use the weights to get an estimate of the average treatment effect, τ_b in the sample S_b.
  4. Repeat.

Boostrapping to get the SEs

- How to get the standard error for $\hat{\tau}$?
- Variance estimators are messy $\rightsquigarrow$ use the bootstrap!
  1. Draw a sample of the data with replacement, call this, S_b.
  2. Estimate the propensity scores in this sample, $\hat{e}_b$ and create weights, W_b.
  3. Use the weights to get an estimate of the average treatment effect, τ_b in the sample S_b.
  4. Repeat.
- The distribution of the estimates, $\hat{\tau}_b$, will give us the bootstrapped standard errors and confidence intervals.

Reducing weight variation

- $e(X_i)$ close to 0 or 1 lead to very large weights, high standard errors.

Reducing weight variation

- $e(X_i)$ close to 0 or 1 lead to very large weights, high standard errors.
- Potential solutions:

Reducing weight variation

- $e(X_i)$ close to 0 or 1 lead to very large weights, high standard errors.
- Potential solutions:

1. Trimming/Windsorizing the weights

Reducing weight variation

- $e(X_i)$ close to 0 or 1 lead to very large weights, high standard errors.
- Potential solutions:

1. Trimming/Windsorizing the weights
    - Pick some value w' and create trimmed weights which are:

$$W_i' = \begin{cases} W_i & \text{if } W_i < w' \\ w' & \text{if } W_i \geq w' \end{cases}$$

Reducing weight variation

- $e(X_i)$ close to 0 or 1 lead to very large weights, high standard errors.
- Potential solutions:

1. Trimming/Windsorizing the weights
    - Pick some value w' and create trimmed weights which are:

$$W_i' = \begin{cases} W_i & \text{if } W_i < w' \\ w' & \text{if } W_i \geq w' \end{cases}$$

2. Stabilized weights

Reducing weight variation

- $e(X_i)$ close to 0 or 1 lead to very large weights, high standard errors.
- Potential solutions:

1. Trimming/Windsorizing the weights
   - Pick some value w' and create trimmed weights which are:

$$W_i' = \begin{cases} W_i & \text{if } W_i < w' \\ w' & \text{if } W_i \geq w' \end{cases}$$

2. Stabilized weights
   - We can actually put any other function of the treatment vector in the numerator, which can reduce the variation in the weights.

Reducing weight variation

- $e(X_i)$ close to 0 or 1 lead to very large weights, high standard errors.
- Potential solutions:

1. Trimming/Windsorizing the weights
   - Pick some value w' and create trimmed weights which are:

$$W_i' = \begin{cases} W_i & \text{if } W_i < w' \\ w' & \text{if } W_i \geq w' \end{cases}$$

2. Stabilized weights
   - We can actually put any other function of the treatment vector in the numerator, which can reduce the variation in the weights.
   - We call these stabilized weights:

$$sw(d,x) = \frac{\mathbb{P}[D_i = 1]^d (1 - \mathbb{P}[D_i = 1])^{1-d}}{e(x)^d (1 - e(x))^{1-d}}$$

Stablized weights

- With a binary treatment, we can implement the stabilized weight by normalizing the weights:

$$SW_i = \frac{W_i}{\sum_{i=1}^{N} W_i}$$

Stablized weights

- With a binary treatment, we can implement the stabilized weight by normalizing the weights:

$$SW_i = \frac{W_i}{\sum_{i=1}^{N} W_i}$$

- This leads to the following estimator:

Stablized weights

- With a binary treatment, we can implement the stabilized weight by normalizing the weights:

$$SW_i = \frac{W_i}{\sum_{i=1}^N W_i}$$

- This leads to the following estimator:

$$\hat{\tau}_{IPTW} = \frac{1}{\sum_{i=1}^N W_i D_i} \sum_{i=1}^N W_i D_i Y_i - \frac{1}{\sum_{i=1}^N W_i(1 - D_i)} \sum_{i=1}^N W_i(1 - D_i) Y_i$$

Stablized weights

- With a binary treatment, we can implement the stabilized weight by normalizing the weights:

$$SW_i = \frac{W_i}{\sum_{i=1}^{N} W_i}$$

- This leads to the following estimator:

$$\begin{aligned}
\hat{\tau}_{IPTW} =& \frac{1}{\sum_{i=1}^{N} W_i D_i} \sum_{i=1}^{N} W_i D_i Y_i - \frac{1}{\sum_{i=1}^{N} W_i (1 - D_i)} \sum_{i=1}^{N} W_i (1 - D_i) Y_i \\
=& \frac{1}{\sum_{i=1}^{N} D_i / \hat{e}(X_i)} \sum_{i=1}^{N} \frac{D_i Y_i}{\hat{e}(X_i)} \\
& - \frac{1}{\sum_{i=1}^{N} (1 - D_i)/(1 - \hat{e}(X_i))} \sum_{i=1}^{N} \frac{(1 - D_i) Y_i}{1 - \hat{e}(X_i)}
\end{aligned}$$

Stablized weights

- With a binary treatment, we can implement the stabilized weight by normalizing the weights:

$$SW_i = \frac{W_i}{\sum_{i=1}^N W_i}$$

- This leads to the following estimator:

$$\begin{aligned}\hat{\tau}_{IPTW} =& \frac{1}{\sum_{i=1}^N W_i D_i} \sum_{i=1}^N W_i D_i Y_i - \frac{1}{\sum_{i=1}^N W_i(1-D_i)} \sum_{i=1}^N W_i(1-D_i)Y_i \\ =& \frac{1}{\sum_{i=1}^N D_i/\hat{e}(X_i)} \sum_{i=1}^N \frac{D_i Y_i}{\hat{e}(X_i)} \\ & - \frac{1}{\sum_{i=1}^N (1-D_i)/(1-\hat{e}(X_i))} \sum_{i=1}^N \frac{(1-D_i)Y_i}{1-\hat{e}(X_i)}\end{aligned}$$

- These are the means that the `weighted.mean()` function in R calculates. It normalizes the weights before calculating the mean.

Wrap-up on Propensity Scores

Wrap-up on Propensity Scores

- Propensity scores play an important role in causal inference as a form of dimension reduction and a conceptual tool for thinking about selection into treatment.

Wrap-up on Propensity Scores

- Propensity scores play an important role in causal inference as a form of dimension reduction and a conceptual tool for thinking about selection into treatment.
- In most basic selection on observables settings the goal is to use them to achieve balance. Thus we should ask, would we better off just pursuing directly with entropy balancing or CBPS.

Wrap-up on Propensity Scores

- Propensity scores play an important role in causal inference as a form of dimension reduction and a conceptual tool for thinking about selection into treatment.
- In most basic selection on observables settings the goal is to use them to achieve balance. Thus we should ask, would we better off just pursuing directly with entropy balancing or CBPS.
- Also remember, you don't need to include everything that predicts treatment, just variables on open backdoor paths (those that also influence the outcome)

Wrap-up on Propensity Scores

- Propensity scores play an important role in causal inference as a form of dimension reduction and a conceptual tool for thinking about selection into treatment.
- In most basic selection on observables settings the goal is to use them to achieve balance. Thus we should ask, would we better off just pursuing directly with entropy balancing or CBPS.
- Also remember, you don't need to include everything that predicts treatment, just variables on open backdoor paths (those that also influence the outcome) .
- Propensity scores have also been used to think about treatment effect heterogeneity (see work by Jennie Brand and Yu Xie).

Theory and causality

- Theory $\implies$ (or $\equiv$) causal effects

Theory and causality

- Theory $\implies$ (or $\equiv$) causal effects
- But they also tell us how those causes should impact the outcomes.

Theory and causality

- Theory $\implies$ (or $\equiv$) causal effects
- But they also tell us how those causes should impact the outcomes.
  - Theory A: causal effect is "due to" path A

Theory and causality

- Theory $\implies$ (or $\equiv$) causal effects
- But they also tell us how those causes should impact the outcomes.
  - Theory A: causal effect is "due to" path A
  - Theory B: causal effect is "due to" path B

Theory and causality

- Theory $\implies$ (or $\equiv$) causal effects
- But they also tell us how those causes should impact the outcomes.
  - Theory A: causal effect is "due to" path A
  - Theory B: causal effect is "due to" path B
- How do we adjudicate between these theories when they predict the same overall effect?

Theory and causality

- Theory $\implies$ (or $\equiv$) causal effects
- But they also tell us how those causes should impact the outcomes.
  - Theory A: causal effect is “due to” path A
  - Theory B: causal effect is “due to” path B
- How do we adjudicate between these theories when they predict the same overall effect?
- Put differently: what is the mechanism that drives a particular causal effect?

Theory and causality

- Theory $\implies$ (or $\equiv$) causal effects
- But they also tell us how those causes should impact the outcomes.
  - Theory A: causal effect is "due to" path A
  - Theory B: causal effect is "due to" path B
- How do we adjudicate between these theories when they predict the same overall effect?
- Put differently: what is the mechanism that drives a particular causal effect?
  - How do we get from cause to effect?

What is a causal mechanism?

- People mean different things by causal mechanism. Two biggest types are:

What is a causal mechanism?

- People mean different things by causal mechanism. Two biggest types are:
  1. Mediation (effect decomposition, indirect effects)

What is a causal mechanism?

- People mean different things by causal mechanism. Two biggest types are:
  1. Mediation (effect decomposition, indirect effects)
  2. Moderation (effect modification, subgroup effects)

What is a causal mechanism?

- People mean different things by causal mechanism. Two biggest types are:
  1. Mediation (effect decomposition, indirect effects)
  2. Moderation (effect modification, subgroup effects)
- We are going to focus on mediation today.

Notation

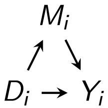
- Treatment variable D_i

Notation

- Treatment variable D_i
- Outcome variable Y_i

Notation

- Treatment variable D_i
- Outcome variable Y_i
- An intermediate, post-treatment variable, M_i, which we call a mediator.



Moderators vs. mediators

- Moderator: pretreatment variable that is correlated with the treatment effect.

$$\text{Cov}(\tau_i, X_i) \neq 0$$

Moderators vs. mediators

- Moderator: pretreatment variable that is correlated with the treatment effect.

$$\text{Cov}(\tau_i, X_i) \neq 0$$

- Mediator: a posttreatment variable that changes the effect of treatment.

Potential outcomes

- Mediators have potential outcomes $M_i(d)$: the value that the mediator takes when the treatment is d.

Potential outcomes

- Mediators have potential outcomes $M_i(d)$: the value that the mediator takes when the treatment is d.
- Potential outcomes $Y_i(d, m)$: the value that the outcome takes when the treatment has value d and the mediator takes the value m.

Potential outcomes

- Mediators have potential outcomes $M_i(d)$: the value that the mediator takes when the treatment is d.
- Potential outcomes $Y_i(d, m)$: the value that the outcome takes when the treatment has value d and the mediator takes the value m.
- Consistency assumption to connect the potential outcomes to the observed outcomes:

$$M_i = M_i(D_i)$$
$$Y_i = Y_i(D_i, M_i(D_i))$$

Potential outcomes example

- D_i is exercise, M_i is diet, and Y_i is weight.

Potential outcomes example

- D_i is exercise, M_i is diet, and Y_i is weight.
- d is "run 10 km/day" and m is "eat 1500 kcals"

Potential outcomes example

- D_i is exercise, M_i is diet, and Y_i is weight.
- d is "run 10 km/day" and m is "eat 1500 kcals"
- $Y_i(d, m)$ is the weight you would have if we forced you to run 10 km/day and eat 1500 kcals a day.

Estimands: Total causal effects

- We can recover "original" potential outcomes:

$$Y_i(d) = Y_i(d, M_i(d))$$

Estimands: Total causal effects

- We can recover "original" potential outcomes:

$$Y_i(d) = Y_i(d, M_i(d))$$

- Your weight if we force you to run 10 km/day, but don't intervene on your diet.

Estimands: Total causal effects

- We can recover "original" potential outcomes:

$$Y_i(d) = Y_i(d, M_i(d))$$

- Your weight if we force you to run 10 km/day, but don't intervene on your diet.
- We can define the typical individual causal effect, here called the total causal effect:

$$\tau_i = Y_i(1, M_i(1)) - Y_i(0, M_i(0))$$

Estimands: Total causal effects

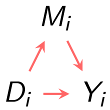
- We can recover "original" potential outcomes:

$$Y_i(d) = Y_i(d, M_i(d))$$

- Your weight if we force you to run 10 km/day, but don't intervene on your diet.
- We can define the typical individual causal effect, here called the total causal effect:

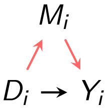
$$\tau_i = Y_i(1, M_i(1)) - Y_i(0, M_i(0))$$

- The total causal effect allows the effect of the treatment "propagate" through all causal pathways.



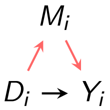
Direct and indirect effects

- The indirect effect is the part of the effect of treatment that "flows through" the mediator

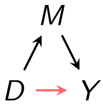


Direct and indirect effects

- The indirect effect is the part of the effect of treatment that “flows through” the mediator

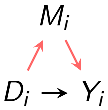


- The direct effect is the part of the effect that does not flow through the mediator.

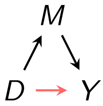


Direct and indirect effects

- The indirect effect is the part of the effect of treatment that "flows through" the mediator



- The direct effect is the part of the effect that does not flow through the mediator.



- These are loose definitions, let's be precise.

Indirect effects

- One estimand is the so-called "natural" indirect effect (NIE):

$$\delta_i(d) = Y_i(d, M_i(1)) - Y_i(d, M_i(0))$$

Indirect effects

- One estimand is the so-called "natural" indirect effect (NIE):

$$\delta_i(d) = Y_i(d, M_i(1)) - Y_i(d, M_i(0))$$

  - Fix treatment to d.

Indirect effects

- One estimand is the so-called "natural" indirect effect (NIE):

$$\delta_i(d) = Y_i(d, M_i(1)) - Y_i(d, M_i(0))$$

  - Fix treatment to d.
  - Vary M_i by the value that it (naturally) would take under treatment and control for unit i.

Indirect effects

- One estimand is the so-called "natural" indirect effect (NIE):

$$\delta_i(d) = Y_i(d, M_i(1)) - Y_i(d, M_i(0))$$

  - Fix treatment to d.
  - Vary M_i by the value that it (naturally) would take under treatment and control for unit i.
- If D_i doesn't affect M_i, so that $M_i(1) = M_i(0)$, then $\delta_i = 0$.

Indirect effects

- One estimand is the so-called "natural" indirect effect (NIE):

$$\delta_i(d) = Y_i(d, M_i(1)) - Y_i(d, M_i(0))$$

  - Fix treatment to d.
  - Vary M_i by the value that it (naturally) would take under treatment and control for unit i.
- If D_i doesn't affect M_i, so that $M_i(1) = M_i(0)$, then $\delta_i = 0$.
- Fundamental Problem of Causal Inference $\rightsquigarrow$ focus on the average natural indirect effect (ANIE):

$$\bar{\delta}(d) = \mathbf{E}[\delta_i(d)] = \mathbf{E}[Y_i(d, M_i(1)) - Y_i(d, M_i(0))]$$

How Natural?: Impossible counterfactuals

$$Y_i(1, M_i(1)) - Y_i(1, M_i(0))$$

- Compare your weight when we force you to run 10 km/day vs. weight when you run 10 km/day, but keep your diet as if you didn't run at all.

How Natural?: Impossible counterfactuals

$$Y_i(1, M_i(1)) - Y_i(1, M_i(0))$$

- Compare your weight when we force you to run 10 km/day vs. weight when you run 10 km/day, but keep your diet as if you didn't run at all.
- The second part, $Y_i(1, M_i(0))$, is logically unobservable.

How Natural?: Impossible counterfactuals

$$Y_i(1, M_i(1)) - Y_i(1, M_i(0))$$

- Compare your weight when we force you to run 10 km/day vs. weight when you run 10 km/day, but keep your diet as if you didn't run at all.
- The second part, $Y_i(1, M_i(0))$, is logically unobservable.
    - Need to see you in two states of the world simultaneously, running and not running.

How Natural?: Impossible counterfactuals

$$Y_i(1, M_i(1)) - Y_i(1, M_i(0))$$

- Compare your weight when we force you to run 10 km/day vs. weight when you run 10 km/day, but keep your diet as if you didn't run at all.
- The second part, $Y_i(1, M_i(0))$, is logically unobservable.
  - Need to see you in two states of the world simultaneously, running and not running.
  - Not just the fundamental problem of causal inference

How Natural?: Impossible counterfactuals

$$Y_i(1, M_i(1)) - Y_i(1, M_i(0))$$

- Compare your weight when we force you to run 10 km/day vs. weight when you run 10 km/day, but keep your diet as if you didn't run at all.
- The second part, $Y_i(1, M_i(0))$, is logically unobservable.
  - Need to see you in two states of the world simultaneously, running and not running.
  - Not just the fundamental problem of causal inference
  - Crossover experimental designs require strong no carry-over assumptions.

How Natural?: Impossible counterfactuals

$$Y_i(1, M_i(1)) - Y_i(1, M_i(0))$$

- Compare your weight when we force you to run 10 km/day vs. weight when you run 10 km/day, but keep your diet as if you didn't run at all.
- The second part, $Y_i(1, M_i(0))$, is logically unobservable.
  - Need to see you in two states of the world simultaneously, running and not running.
  - Not just the fundamental problem of causal inference
  - Crossover experimental designs require strong no carry-over assumptions.
- Leads some to dismiss mediation altogether.

How Natural?: Impossible counterfactuals

$$Y_i(1, M_i(1)) - Y_i(1, M_i(0))$$

- Compare your weight when we force you to run 10 km/day vs. weight when you run 10 km/day, but keep your diet as if you didn't run at all.
- The second part, $Y_i(1, M_i(0))$, is logically unobservable.
  - Need to see you in two states of the world simultaneously, running and not running.
  - Not just the fundamental problem of causal inference
  - Crossover experimental designs require strong no carry-over assumptions.
- Leads some to dismiss mediation altogether.
- However still an important topic for policy makers, theory-driven scholars etc.

Natural Direct Effects

- We can also define the natural direct effect (NDE) of the treatment:

$$\eta_i(d) = Y_i(1, M_i(d)) - Y_i(0, M_i(d))$$

Natural Direct Effects

- We can also define the natural direct effect (NDE) of the treatment:

$$\eta_i(d) = Y_i(1, M_i(d)) - Y_i(0, M_i(d))$$

- Thus, the natural direct effect is the effect of moving from control to treatment while holding the mediator fixed at the value it would have under treatment status d.

When are NDEs useful?

- The canonical example: D_i is smoking, M_i is tar intake, and Y_i is lung cancer.

When are NDEs useful?

- The canonical example: D_i is smoking, M_i is tar intake, and Y_i is lung cancer.
- We know that smoking increases tar consumption, $M_i(1) - M_i(0) > 0$,

When are NDEs useful?

- The canonical example: D_i is smoking, M_i is tar intake, and Y_i is lung cancer.
- We know that smoking increases tar consumption, $M_i(1) - M_i(0) > 0$,
- Also, smoking overall increases the likelihood of lung cancer, $Y_i(1, M_i(1)) - Y_i(0, M_i(0))$.

When are NDEs useful?

- The canonical example: D_i is smoking, M_i is tar intake, and Y_i is lung cancer.
- We know that smoking increases tar consumption, $M_i(1) - M_i(0) > 0$,
- Also, smoking overall increases the likelihood of lung cancer, $Y_i(1, M_i(1)) - Y_i(0, M_i(0))$.
- But what would happen if we created a tar-less cigarette?

When are NDEs useful?

- The canonical example: D_i is smoking, M_i is tar intake, and Y_i is lung cancer.
- We know that smoking increases tar consumption, $M_i(1) - M_i(0) > 0$,
- Also, smoking overall increases the likelihood of lung cancer, $Y_i(1, M_i(1)) - Y_i(0, M_i(0))$.
- But what would happen if we created a tar-less cigarette?
  - So that $M_i(1) = M_i(0)$ for all i.

When are NDEs useful?

- The canonical example: D_i is smoking, M_i is tar intake, and Y_i is lung cancer.
- We know that smoking increases tar consumption, $M_i(1) - M_i(0) > 0$,
- Also, smoking overall increases the likelihood of lung cancer, $Y_i(1, M_i(1)) - Y_i(0, M_i(0))$.
- But what would happen if we created a tar-less cigarette?
  - So that $M_i(1) = M_i(0)$ for all i.
- NDE answers this question.

Effect decomposition

- The total causal effect and the natural indirect and direct causal effects are related:

$$\tau_i = \delta_i(d) + \eta_i(1-d) = NIE_i(d) + NDE_i(1-d)$$

Effect decomposition

- The total causal effect and the natural indirect and direct causal effects are related:

$$\tau_i = \delta_i(d) + \eta_i(1-d) = NIE_i(d) + NDE_i(1-d)$$

- Thus, we know that the ATE, $\tau = \mathbf{E}[\tau_i]$, must be the sum of the average indirect and direct effects:

$$\tau = \bar{\delta}(d) + \bar{\eta}(1-d) = ANIE(d) + ANDE(1-d)$$

Effect decomposition

- The total causal effect and the natural indirect and direct causal effects are related:

$$\tau_i = \delta_i(d) + \eta_i(1-d) = NIE_i(d) + NDE_i(1-d)$$

- Thus, we know that the ATE, $\tau = \mathbf{E}[\tau_i]$, must be the sum of the average indirect and direct effects:

$$\tau = \bar{\delta}(d) + \bar{\eta}(1-d) = ANIE(d) + ANDE(1-d)$$

- The fact that we can decompose the total effect of treatment into the sum of a direct and indirect effect if very important to social science researchers.

A Different Estimand

- Another definition of direct effects is the controlled direct effect (CDE):

$$Y_i(1, m) - Y_i(0, m)$$

A Different Estimand

- Another definition of direct effects is the controlled direct effect (CDE):
$$Y_i(1, m) - Y_i(0, m)$$
- The effect of running 10 km/day if we fixed your diet to 1500 kcals/day.

A Different Estimand

- Another definition of direct effects is the controlled direct effect (CDE):
$$Y_i(1, m) - Y_i(0, m)$$
- The effect of running 10 km/day if we fixed your diet to 1500 kcals/day.
- ACDE is the average of these over the i units.

A Different Estimand

- Another definition of direct effects is the controlled direct effect (CDE):

$$Y_i(1, m) - Y_i(0, m)$$

- The effect of running 10 km/day if we fixed your diet to 1500 kcals/day.
- ACDE is the average of these over the i units.
- In general, this effect will be different than the NDE.

A Different Estimand

- Another definition of direct effects is the controlled direct effect (CDE):
$$Y_i(1, m) - Y_i(0, m)$$
- The effect of running 10 km/day if we fixed your diet to 1500 kcals/day.
- ACDE is the average of these over the i units.
- In general, this effect will be different than the NDE.
  - ACDE: set M_i to m for all units

A Different Estimand

- Another definition of direct effects is the controlled direct effect (CDE):

$$Y_i(1, m) - Y_i(0, m)$$

- The effect of running 10 km/day if we fixed your diet to 1500 kcals/day.
- ACDE is the average of these over the i units.
- In general, this effect will be different than the NDE.
  - ACDE: set M_i to m for all units
  - ANDE: set M_i to $M_i(0)$ for all units

A Different Estimand

- Another definition of direct effects is the controlled direct effect (CDE):
$$Y_i(1, m) - Y_i(0, m)$$
- The effect of running 10 km/day if we fixed your diet to 1500 kcals/day.
- ACDE is the average of these over the i units.
- In general, this effect will be different than the NDE.
  - ACDE: set M_i to m for all units
  - ANDE: set M_i to $M_i(0)$ for all units
- ACDE is identified under weaker conditions than the ANDE but it does not create a nice decomposition of effects.

Identifying indirect and direct effects

- What assumptions can identify the ANDE and ANIE?

Identifying indirect and direct effects

- What assumptions can identify the ANDE and ANIE?
- Imai et al use a sequential ignorability (SI) assumption, which has two parts. (Note: this same name means different things in different parts of the literature)

Identifying indirect and direct effects

- What assumptions can identify the ANDE and ANIE?
- Imai et al use a sequential ignorability (SI) assumption, which has two parts. (Note: this same name means different things in different parts of the literature)
- SI part 1: the treatment is independent of the potential outcomes and potential mediators, conditional on a set of covariates:

$$\{Y_i(d', m), M_i(d)\} \perp\!\!\!\perp D_i | X_i = x$$

Identifying indirect and direct effects

- What assumptions can identify the ANDE and ANIE?
- Imai et al use a sequential ignorability (SI) assumption, which has two parts. (Note: this same name means different things in different parts of the literature)
- SI part 1: the treatment is independent of the potential outcomes and potential mediators, conditional on a set of covariates:

$$\{Y_i(d', m), M_i(d)\} \perp\!\!\!\perp D_i | X_i = x$$

- This holds in a (well-executed) experiment

Identifying indirect and direct effects

- SI part 2: the mediator is ignorable with respect to the outcome, conditional on the treatment:

$$Y_i(d', m) \perp\!\!\!\perp M_i(d) | D_i = d, X_i = x$$

Identifying indirect and direct effects

- SI part 2: the mediator is ignorable with respect to the outcome, conditional on the treatment:

$$Y_i(d', m) \perp\!\!\!\perp M_i(d) | D_i = d, X_i = x$$

- This must hold for all values of d, d'.

Identifying indirect and direct effects

- SI part 2: the mediator is ignorable with respect to the outcome, conditional on the treatment:

$$Y_i(d', m) \perp\!\!\!\perp M_i(d) | D_i = d, X_i = x$$

- This must hold for all values of d, d'.
- Note that we have to believe ignorability in certain cross-world comparisons:

$$Y_i(1, m) \perp\!\!\!\perp M_i(0) | D_i = 0, X_i = x$$

Identifying indirect and direct effects

- SI part 2: the mediator is ignorable with respect to the outcome, conditional on the treatment:

$$Y_i(d', m) \perp\!\!\!\perp M_i(d) | D_i = d, X_i = x$$

- This must hold for all values of d, d'.
- Note that we have to believe ignorability in certain cross-world comparisons:

$$Y_i(1, m) \perp\!\!\!\perp M_i(0) | D_i = 0, X_i = x$$

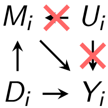
- Could be satisfied by randomizing M_i, but then the effect of D_i is not "natural." Also can hold if X includes all confounders

SI and posttreatment bias

- SI assumes that posttreatment bias is not a problem.

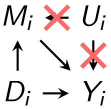
SI and posttreatment bias

- SI assumes that posttreatment bias is not a problem.
- The mediator is as-if random, so these situations can never happened:



SI and posttreatment bias

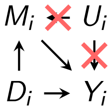
- SI assumes that posttreatment bias is not a problem.
- The mediator is as-if random, so these situations can never happened:



- Never any collider bias.

SI and posttreatment bias

- SI assumes that posttreatment bias is not a problem.
- The mediator is as-if random, so these situations can never happened:



- Never any collider bias.
- Is this plausible? It depends on the application.

Limitations of sequential ignorability

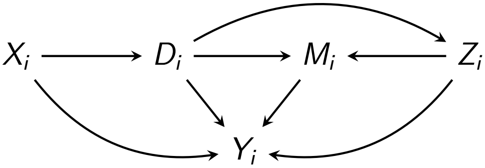
$$\{Y_i(d', m), M_i(d)\} \perp\!\!\!\perp D_i | X_i = x$$
$$Y_i(d', m) \perp\!\!\!\perp M_i(d) | D_i = d, X_i = x$$

- Conditioning set X_i is the same for both stages.

Limitations of sequential ignorability

$$\{Y_i(d', m), M_i(d)\} \perp\!\!\!\perp D_i | X_i = x$$
$$Y_i(d', m) \perp\!\!\!\perp M_i(d) | D_i = d, X_i = x$$

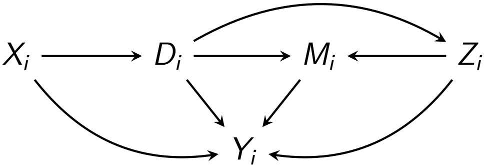
- Conditioning set X_i is the same for both stages.
- What if there are confounders for the relationship between M and Y that are affected by D? Too bad!



Limitations of sequential ignorability

$$\{Y_i(d', m), M_i(d)\} \perp\!\!\!\perp D_i | X_i = x$$
$$Y_i(d', m) \perp\!\!\!\perp M_i(d) | D_i = d, X_i = x$$

- Conditioning set X_i is the same for both stages.
- What if there are confounders for the relationship between M and Y that are affected by D? Too bad!



- More on this in a bit.

Identifying (in)direct effects

- Under SI and consistency, we can write the ANIE as a function of the observed data.

Identifying (in)direct effects

- Under SI and consistency, we can write the ANIE as a function of the observed data.
- With a binary mediator and a binary treatment:

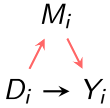
$$
\begin{aligned}
\bar{\delta}(d) =& \{\mathbb{P}[M_i = 1 | D_i = 1, X_i] - \mathbb{P}[M_i = 1 | D_i = 0, X_i]\} \\
& \cdot \{\mathbf{E}[Y_i | M_i = 1, D_i = d, X_i] - \mathbf{E}[Y_i | M_i = 0, D_i = d, X_i] \\
=& (\text{effect of } D_i \text{ on } M_i) \times (\text{effect of } M_i \text{ on } Y_i)
\end{aligned}
$$

Identifying (in)direct effects

- Under SI and consistency, we can write the ANIE as a function of the observed data.
- With a binary mediator and a binary treatment:

$$
\begin{aligned}
\bar{\delta}(d) = & \{\mathbb{P}[M_i = 1|D_i = 1, X_i] - \mathbb{P}[M_i = 1|D_i = 0, X_i]\} \\
& \cdot \{\mathbf{E}[Y_i|M_i = 1, D_i = d, X_i] - \mathbf{E}[Y_i|M_i = 0, D_i = d, X_i] \\
= & (\text{effect of } D_i \text{ on } M_i) \times (\text{effect of } M_i \text{ on } Y_i)
\end{aligned}
$$

- Intuitive given the DAG:



(In)direct effects with non-binary mediators

- Let's say that the mediator has J categories:

$$ANIE(d) = \sum_{m=0}^{J-1} \mathbb{E}[Y_i | M_i = m, D_i = d, X_i] \cdot \{\mathbb{P}[M_i = m | D_i = 1, X_i] - \mathbb{P}[M_i = m | D_i = 0, X_i]\}$$

(In)direct effects with non-binary mediators

- Let's say that the mediator has J categories:

$$ANIE(d) = \sum_{m=0}^{J-1} \mathbb{E}[Y_i | M_i = m, D_i = d, X_i] \cdot \{\mathbb{P}[M_i = m | D_i = 1, X_i] - \mathbb{P}[M_i = m | D_i = 0, X_i]\}$$

- The ANDE is the following:

$$ANDE(d) = \sum_{m=0}^{J-1} (\mathbb{E}[Y_i | M_i = m, D_i = 1, X_i] - \mathbb{E}[Y_i | M_i = m, D_i = 0, X_i]) \cdot \{\mathbb{P}[M_i = m | D_i = d, X_i]\}$$

(In)direct effects with non-binary mediators

- Let's say that the mediator has J categories:

$$ANIE(d) = \sum_{m=0}^{J-1} \mathbf{E}[Y_i | M_i = m, D_i = d, X_i] \cdot \{\mathbb{P}[M_i = m | D_i = 1, X_i] - \mathbb{P}[M_i = m | D_i = 0, X_i]\}$$

- The ANDE is the following:

$$ANDE(d) = \sum_{m=0}^{J-1} (\mathbf{E}[Y_i | M_i = m, D_i = 1, X_i] - \mathbf{E}[Y_i | M_i = m, D_i = 0, X_i]) \cdot \{\mathbb{P}[M_i = m | D_i = d, X_i]\}$$

- The ANDE is the effect of D_i for a fixed m, averaged over the distribution of M_i when $D_i = 0$.

Alternative identification

- Robins proposed a different identification strategy, based on a no-interactions assumption:

$$Y_i(1, m) - Y_i(0, m) = Y_i(1, m') - Y_i(0, m')$$

Alternative identification

- Robins proposed a different identification strategy, based on a no-interactions assumption:

$$Y_i(1, m) - Y_i(0, m) = Y_i(1, m') - Y_i(0, m')$$

- The CDE does not depend on m for any unit i.

Alternative identification

- Robins proposed a different identification strategy, based on a no-interactions assumption:

$$Y_i(1, m) - Y_i(0, m) = Y_i(1, m') - Y_i(0, m')$$

- The CDE does not depend on m for any unit i.
- $\leadsto$ ACDE = ANDE.

Alternative identification

- Robins proposed a different identification strategy, based on a no-interactions assumption:

$$Y_i(1, m) - Y_i(0, m) = Y_i(1, m') - Y_i(0, m')$$

- The CDE does not depend on m for any unit i.
- $\rightsquigarrow$ ACDE = ANDE.
- Strong assumption because it has to hold at the individual level (like monotonicity for IV).

Traditional estimation

Traditional estimation

- There a variety of other estimation strategies that rely on stronger assumptions and parametric models

Traditional estimation

- There a variety of other estimation strategies that rely on stronger assumptions and parametric models
- For the sake of saving time and not promoting confusion, I'm going to skip these alternatives.

Traditional estimation

- There a variety of other estimation strategies that rely on stronger assumptions and parametric models
- For the sake of saving time and not promoting confusion, I'm going to skip these alternatives.
- In general though: if someone says this is easy- they are fooling themselves.

Nonparametric estimation

- How far can we get with nonparametrics?

Nonparametric estimation

- How far can we get with nonparametrics?
- If the number of categories in M_i, D_i, and X_i are small, use plug-in estimator for the CEF of Y_i:

$$\widehat{\mathbb{E}}[Y_i|M_i = m, D_i = d, X_i = x] = \frac{\sum_i Y_i \mathbb{1}\{M_i = m, D_i = d, X_i = x\}}{\sum_i \mathbb{1}\{M_i = m, D_i = d, X_i = x\}}$$

Nonparametric estimation

- How far can we get with nonparametrics?
- If the number of categories in M_i, D_i, and X_i are small, use plug-in estimator for the CEF of Y_i:

$$\widehat{\mathbb{E}}[Y_i|M_i = m, D_i = d, X_i = x] = \frac{\sum_i Y_i \mathbb{1}\{M_i = m, D_i = d, X_i = x\}}{\sum_i \mathbb{1}\{M_i = m, D_i = d, X_i = x\}}$$

- Same for M_i:

$$\widehat{\mathbb{P}}[M_i = m|D_i = d, X_i = x] = \frac{\sum_i \mathbb{1}\{M_i = m, D_i = d, X_i = x\}}{\sum_i \mathbb{1}\{D_i = d, X_i = x\}}$$

What about more complicated scenarios?

- If the number of categories is large, then we can use nonparametric regressions for the outcome and the mediator.

$$\mu_{dm}(x) = \mathbf{E}[Y_i | M_i = m, D_i = d, X_i = x]$$

What about more complicated scenarios?

- If the number of categories is large, then we can use nonparametric regressions for the outcome and the mediator.

$$\mu_{dm}(x) = \mathbf{E}[Y_i|M_i = m, D_i = d, X_i = x]$$

- Flexibly estimate $\mu_{dm}(x)$ as a function of x using splines of x.

What about more complicated scenarios?

- If the number of categories is large, then we can use nonparametric regressions for the outcome and the mediator.

$$\mu_{dm}(x) = \mathbf{E}[Y_i | M_i = m, D_i = d, X_i = x]$$

- Flexibly estimate $\mu_{dm}(x)$ as a function of x using splines of x.
- To get the standard errors, we can use bootstrapping.

What about more complicated scenarios?

- If the number of categories is large, then we can use nonparametric regressions for the outcome and the mediator.

$$\mu_{dm}(x) = \mathbf{E}[Y_i | M_i = m, D_i = d, X_i = x]$$

- Flexibly estimate $\mu_{dm}(x)$ as a function of x using splines of x.
- To get the standard errors, we can use bootstrapping.
- Need to be careful with the curse of dimensionality in X_i. Use good nonparametric strategies (cross-validation, etc)

Continuous mediator, nonparametric

- What if the mediator is continuous? Things get tricky.

Continuous mediator, nonparametric

- What if the mediator is continuous? Things get tricky.
- Need to integrate over the distribution of the mediators to get the ANIE:

$$\bar{\delta}(d) = \int\int \mathbf{E}[Y_i | M_i = m, D_i = d, X_i = x] \\ \{dF_{M_i|D_i=1,X_i=x}(m) - dF_{M_i|D_i=0,X_i=x}(m)\} dF_{X_i}(x)$$

Continuous mediator, nonparametric

- What if the mediator is continuous? Things get tricky.
- Need to integrate over the distribution of the mediators to get the ANIE:

$$\bar{\delta}(d) = \int \int \mathbf{E}[Y_i | M_i = m, D_i = d, X_i = x] \\ \{dF_{M_i|D_i=1,X_i=x}(m) - dF_{M_i|D_i=0,X_i=x}(m)\} dF_{X_i}(x)$$

- Obviously, this is a much harder problem. In this case, we actually can use Monte Carlo simulation to take the integral.

Continuous mediator, nonparametric

- What if the mediator is continuous? Things get tricky.
- Need to integrate over the distribution of the mediators to get the ANIE:

$$\bar{\delta}(d) = \int \int \mathbf{E}[Y_i | M_i = m, D_i = d, X_i = x] \{dF_{M_i|D_i=1,X_i=x}(m) - dF_{M_i|D_i=0,X_i=x}(m)\} dF_{X_i}(x)$$

- Obviously, this is a much harder problem. In this case, we actually can use Monte Carlo simulation to take the integral.
- Modeling M_i probably appropriate here.

General Estimation Algorithm (Imai et al 2011)

General Estimation Algorithm (Imai et al 2011)

1. Model outcome and mediator
   - Outcome model: $p(Y_i|D_i, M_i, X_i)$
   - Mediator model: $p(M_i|D_i, X_i)$
   - These models can be of any form (linear, nonlinear, semiparametric etc.)

General Estimation Algorithm (Imai et al 2011)

1. Model outcome and mediator
   - Outcome model: $p(Y_i|D_i, M_i, X_i)$
   - Mediator model: $p(M_i|D_i, X_i)$
   - These models can be of any form (linear, nonlinear, semiparametric etc.)
2. Predict mediator for both treatment values $(M_i(1), M_i(0))$

General Estimation Algorithm (Imai et al 2011)

1. Model outcome and mediator
   - Outcome model: $p(Y_i|D_i, M_i, X_i)$
   - Mediator model: $p(M_i|D_i, X_i)$
   - These models can be of any form (linear, nonlinear, semiparametric etc.)
2. Predict mediator for both treatment values $(M_i(1), M_i(0))$
3. Predict outcome for $(D_i = 1, M_i = M_i(0))$ and $(D_i = 1, M_i = M_i(1))$

General Estimation Algorithm (Imai et al 2011)

1. Model outcome and mediator
   - Outcome model: $p(Y_i|D_i, M_i, X_i)$
   - Mediator model: $p(M_i|D_i, X_i)$
   - These models can be of any form (linear, nonlinear, semiparametric etc.)
2. Predict mediator for both treatment values $(M_i(1), M_i(0))$
3. Predict outcome for $(D_i = 1, M_i = M_i(0))$ and $(D_i = 1, M_i = M_i(1))$
4. Compute the average difference between two outcomes to obtain a consistent estimator of the average natural indirect effect.

General Estimation Algorithm (Imai et al 2011)

1. Model outcome and mediator
   - Outcome model: $p(Y_i|D_i, M_i, X_i)$
   - Mediator model: $p(M_i|D_i, X_i)$
   - These models can be of any form (linear, nonlinear, semiparametric etc.)
2. Predict mediator for both treatment values $(M_i(1), M_i(0))$
3. Predict outcome for $(D_i = 1, M_i = M_i(0))$ and $(D_i = 1, M_i = M_i(1))$
4. Compute the average difference between two outcomes to obtain a consistent estimator of the average natural indirect effect.
5. Bootstrap for the uncertainty

General Estimation Algorithm (Imai et al 2011)

1. Model outcome and mediator
    - Outcome model: $p(Y_i|D_i, M_i, X_i)$
    - Mediator model: $p(M_i|D_i, X_i)$
    - These models can be of any form (linear, nonlinear, semiparametric etc.)
2. Predict mediator for both treatment values ($M_i(1), M_i(0)$)
3. Predict outcome for ($D_i = 1, M_i = M_i(0)$) and ($D_i = 1, M_i = M_i(1)$)
4. Compute the average difference between two outcomes to obtain a consistent estimator of the average natural indirect effect.
5. Bootstrap for the uncertainty
6. Evaluate sensitivity to assumptions

See `mediation` package in R.

Extensions

Extensions

- Experimental Design
  - One puzzling part of all of this is that it is hard to think of the ideal experiment. The notion of the natural direct effect conflicts with our need to randomize for identification.

Extensions

- Experimental Design
  - One puzzling part of all of this is that it is hard to think of the ideal experiment. The notion of the natural direct effect conflicts with our need to randomize for identification.
  - Imai, Tingley and Yamamoto (2013, JRSS-A) provides an overview of some experimental designs for causal mechanisms (also includes great discussions)

Extensions

- Experimental Design
  - One puzzling part of all of this is that it is hard to think of the ideal experiment. The notion of the natural direct effect conflicts with our need to randomize for identification.
  - Imai, Tingley and Yamamoto (2013, JRSS-A) provides an overview of some experimental designs for causal mechanisms (also includes great discussions)
  - Won't discuss today for time, but see: parallel designs, encouragement designs, and crossover designs, crossover encouragement designs. Each requires different assumptions.

Extensions

- Experimental Design
  - One puzzling part of all of this is that it is hard to think of the ideal experiment. The notion of the natural direct effect conflicts with our need to randomize for identification.
  - Imai, Tingley and Yamamoto (2013, JRSS-A) provides an overview of some experimental designs for causal mechanisms (also includes great discussions)
  - Won't discuss today for time, but see: parallel designs, encouragement designs, and crossover designs, crossover encouragement designs. Each requires different assumptions.
- Multiple Mediators
  - This is all for one mediator, but multiple mediators common in applied settings

Extensions

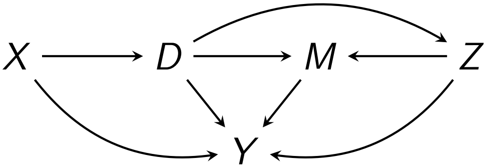
- Experimental Design
  - One puzzling part of all of this is that it is hard to think of the ideal experiment. The notion of the natural direct effect conflicts with our need to randomize for identification.
  - Imai, Tingley and Yamamoto (2013, JRSS-A) provides an overview of some experimental designs for causal mechanisms (also includes great discussions)
  - Won't discuss today for time, but see: parallel designs, encouragement designs, and crossover designs, crossover encouragement designs. Each requires different assumptions.
- Multiple Mediators
  - This is all for one mediator, but multiple mediators common in applied settings
  - These things are hard and relatively new. Adding another variable massively increases the assumptions. Adding a separate analysis messes with the decomposition.

Intermediate confounders

- Intermediate confounders are variables that confound the $M_i \to Y_i$ relationship, but are affected by D_i

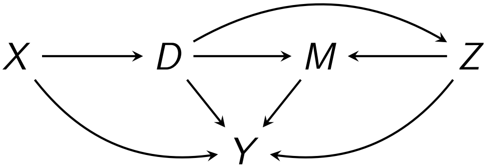
Intermediate confounders

- Intermediate confounders are variables that confound the $M_i \rightarrow Y_i$ relationship, but are affected by D_i
- Here we represent them as Z_i:



Intermediate confounders

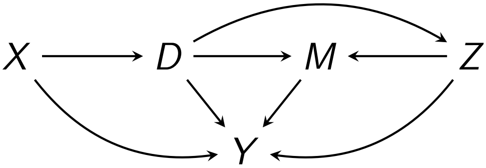
- Intermediate confounders are variables that confound the $M_i \rightarrow Y_i$ relationship, but are affected by D_i
- Here we represent them as Z_i:



- Can also be thought of as other mediators, about which we aren't directly interested.

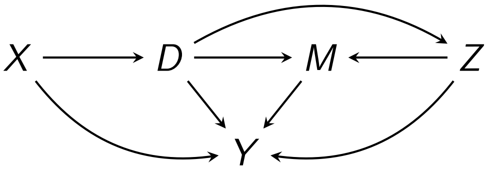
Intermediate confounders

- Intermediate confounders are variables that confound the $M_i \to Y_i$ relationship, but are affected by D_i
- Here we represent them as Z_i:



- Can also be thought of as other mediators, about which we aren't directly interested.
- Avin, Shpitser and Pearl (2003) showed that ANDE/ANIE identification not possible when SI incorporates intermediate confounders.

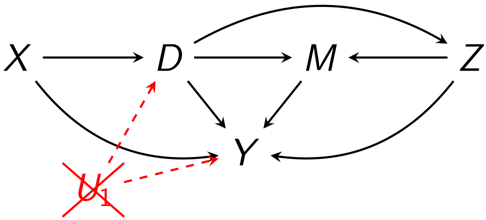
Sequential ignorability, II



- New version of sequential ignorability with intermediate confounders:

$$Y_i(d, m) \perp\!\!\!\perp D_i | X_i$$
$$Y_i(d, m) \perp\!\!\!\perp M_i | Z_i, D_i, X_i$$

Sequential ignorability, II

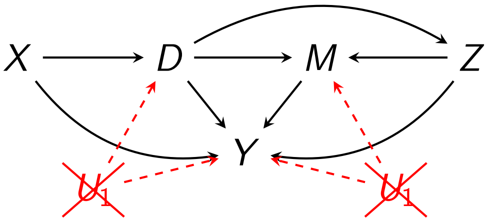


- New version of sequential ignorability with intermediate confounders:

$$Y_i(d, m) \perp\!\!\!\perp D_i | X_i$$
$$Y_i(d, m) \perp\!\!\!\perp M_i | Z_i, D_i, X_i$$

- No unmeasured confounders for D_i conditional on X_i

Sequential ignorability, II



- New version of sequential ignorability with intermediate confounders:

$$Y_i(d, m) \perp\!\!\!\perp D_i | X_i$$
$$Y_i(d, m) \perp\!\!\!\perp M_i | Z_i, D_i, X_i$$

- No unmeasured confounders for D_i conditional on X_i
- No unmeasured confounders for M_i conditional on Z_i, D_i, X_i

Sequential ignorability, II

$$Y_i(d, m) \perp\!\!\!\perp D_i | X_i$$
$$Y_i(d, m) \perp\!\!\!\perp M_i | Z_i, D_i, X_i$$

- Original Robins definition of sequential ignorability.

Sequential ignorability, II

$$Y_i(d,m) \perp\!\!\!\perp D_i | X_i$$
$$Y_i(d,m) \perp\!\!\!\perp M_i | Z_i, D_i, X_i$$

- Original Robins definition of sequential ignorability.
- No cross-world assumptions, allows for intermediate confounders.

Sequential ignorability, II

$$Y_i(d, m) \perp\!\!\!\perp D_i | X_i$$
$$Y_i(d, m) \perp\!\!\!\perp M_i | Z_i, D_i, X_i$$

- Original Robins definition of sequential ignorability.
- No cross-world assumptions, allows for intermediate confounders.
- Will only allow for the identification of the ACDE:

$$ACDE(m) = \mathbb{E}[Y_i(1, m) - Y_i(0, m)]$$

Sequential ignorability, II

$$Y_i(d, m) \perp\!\!\!\perp D_i | X_i$$
$$Y_i(d, m) \perp\!\!\!\perp M_i | Z_i, D_i, X_i$$

- Original Robins definition of sequential ignorability.
- No cross-world assumptions, allows for intermediate confounders.
- Will only allow for the identification of the ACDE:

$$ACDE(m) = \mathbb{E}[Y_i(1, m) - Y_i(0, m)]$$

- Require Robins's no-interaction assumption to connect ACDE to ANDE.

Identifying the ACDE

- Nonparametric idenfication of the ACDE:

Identifying the ACDE

- Nonparametric idenfication of the ACDE:

$$
\begin{aligned}
&\mathbb{E}[Y_i(d, m)] \\
&\quad = \int_x \mathbb{E}[Y_i(d, m) | x] dF_X(x) \quad \text{(LIE)}
\end{aligned}
$$

Identifying the ACDE

- Nonparametric idenfication of the ACDE:

$$
\begin{aligned}
&\mathbf{E}[Y_i(d,m)] \\
&\quad = \int_x \mathbf{E}[Y_i(d,m)|x]dF_X(x) \quad \text{(LIE)} \\
&\quad = \int_x \mathbf{E}[Y_i(d,m)|x,d]dF_X(x) \quad \text{(n.u.c for D)}
\end{aligned}
$$

Identifying the ACDE

- Nonparametric idenfication of the ACDE:

$$
\begin{aligned}
&\mathbf{E}[Y_i(d,m)] \\
&= \int_x \mathbf{E}[Y_i(d,m)|x]dF_X(x) \quad \text{(LIE)} \\
&= \int_x \mathbf{E}[Y_i(d,m)|x,d]dF_X(x) \quad \text{(n.u.c for D)} \\
&= \int_x \int_z \mathbf{E}[Y_i(d,m)|x,d,z]dF_{Z|D,X}(z|d,x)dF_X(x) \quad \text{(LIE)}
\end{aligned}
$$

Identifying the ACDE

- Nonparametric idenfication of the ACDE:

$$
\begin{aligned}
&\mathbf{E}[Y_i(d,m)] \\
&= \int_x \mathbf{E}[Y_i(d,m)|x] dF_X(x) \quad \text{(LIE)} \\
&= \int_x \mathbf{E}[Y_i(d,m)|x,d] dF_X(x) \quad \text{(n.u.c for D)} \\
&= \int_x \int_z \mathbf{E}[Y_i(d,m)|x,d,z] dF_{Z|D,X}(z|d,x) dF_X(x) \quad \text{(LIE)} \\
&= \int_x \int_z \mathbf{E}[Y_i(d,m)|x,d,z,m] dF_{Z|D,X}(z|d,x) dF_X(x) \quad \text{(n.u.c for M)}
\end{aligned}
$$

Identifying the ACDE

- Nonparametric idenfication of the ACDE:

$$
\begin{aligned}
&\mathbf{E}[Y_i(d,m)] \\
&= \int_x \mathbf{E}[Y_i(d,m)|x] dF_X(x) \quad \text{(LIE)} \\
&= \int_x \mathbf{E}[Y_i(d,m)|x,d] dF_X(x) \quad \text{(n.u.c for D)} \\
&= \int_x \int_z \mathbf{E}[Y_i(d,m)|x,d,z] dF_{Z|D,X}(z|d,x) dF_X(x) \quad \text{(LIE)} \\
&= \int_x \int_z \mathbf{E}[Y_i(d,m)|x,d,z,m] dF_{Z|D,X}(z|d,x) dF_X(x) \quad \text{(n.u.c for M)} \\
&= \int_x \int_z \mathbf{E}[Y_i|x,d,z,m] dF_{Z|D,X}(z|d,x) dF_X(x) \quad \text{(consistency)}
\end{aligned}
$$

Identifying the ACDE

- Nonparametric idenfication of the ACDE:

$$
\begin{aligned}
&\mathbb{E}[Y_i(d,m)] \\
&= \int_x \mathbb{E}[Y_i(d,m)|x]dF_X(x) \quad \text{(LIE)} \\
&= \int_x \mathbb{E}[Y_i(d,m)|x,d]dF_X(x) \quad \text{(n.u.c for D)} \\
&= \int_x \int_z \mathbb{E}[Y_i(d,m)|x,d,z]dF_{Z|D,X}(z|d,x)dF_X(x) \quad \text{(LIE)} \\
&= \int_x \int_z \mathbb{E}[Y_i(d,m)|x,d,z,m]dF_{Z|D,X}(z|d,x)dF_X(x) \quad \text{(n.u.c for M)} \\
&= \int_x \int_z \mathbb{E}[Y_i|x,d,z,m]dF_{Z|D,X}(z|d,x)dF_X(x) \quad \text{(consistency)}
\end{aligned}
$$

- Everything in the last line is identified from the data.

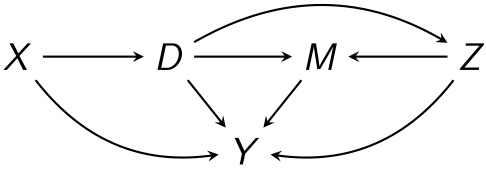
Identifying the ACDE

- Nonparametric idenfication of the ACDE:

$$\begin{aligned}
&\mathbb{E}[Y_i(d,m)] \\
&= \int_x \mathbb{E}[Y_i(d,m)|x]dF_X(x) \quad \text{(LIE)} \\
&= \int_x \mathbb{E}[Y_i(d,m)|x,d]dF_X(x) \quad \text{(n.u.c for D)} \\
&= \int_x \int_z \mathbb{E}[Y_i(d,m)|x,d,z]dF_{Z|D,X}(z|d,x)dF_X(x) \quad \text{(LIE)} \\
&= \int_x \int_z \mathbb{E}[Y_i(d,m)|x,d,z,m]dF_{Z|D,X}(z|d,x)dF_X(x) \quad \text{(n.u.c for M)} \\
&= \int_x \int_z \mathbb{E}[Y_i|x,d,z,m]dF_{Z|D,X}(z|d,x)dF_X(x) \quad \text{(consistency)}
\end{aligned}$$

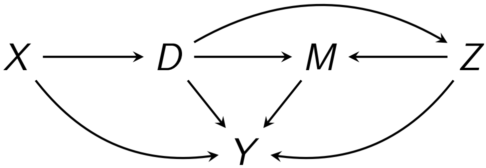
- Everything in the last line is identified from the data.
- Relationship can generalized to any number of treatments, and is called the g-formula by Robins.

Estimating direct effects



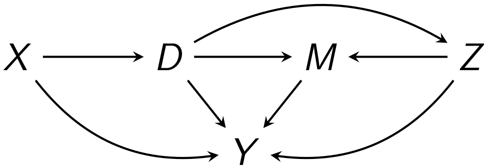
- Controlling for Z_i and $M_i \rightsquigarrow$ posttreatment bias

Estimating direct effects



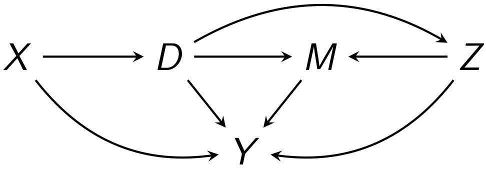
- Controlling for Z_i and $M_i \rightsquigarrow$ posttreatment bias
  - Conditioning on a collider $\rightsquigarrow$ selection bias

Estimating direct effects



- Controlling for Z_i and $M_i \rightsquigarrow$ posttreatment bias
  - Conditioning on a collider $\rightsquigarrow$ selection bias
  - Conditioning on $Z_i \rightsquigarrow$ masking part of the CDE

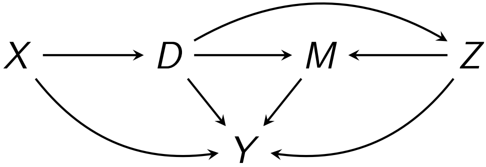
Estimating direct effects



- Controlling for Z_i and $M_i \rightsquigarrow$ posttreatment bias
  - Conditioning on a collider $\rightsquigarrow$ selection bias
  - Conditioning on $Z_i \rightsquigarrow$ masking part of the CDE
- Compare this conditioning approach:

$$\mathbf{E}[Y_i|x, d = 1, z, m] - E[Y_i|x, d = 0, z, m]$$

Estimating direct effects



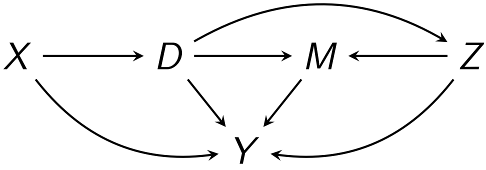
- Controlling for Z_i and $M_i \rightsquigarrow$ posttreatment bias
  - Conditioning on a collider $\rightsquigarrow$ selection bias
  - Conditioning on $Z_i \rightsquigarrow$ masking part of the CDE
- Compare this conditioning approach:

$$\mathbf{E}[Y_i|x, d = 1, z, m] - E[Y_i|x, d = 0, z, m]$$

- And the identification result from the g-formula:

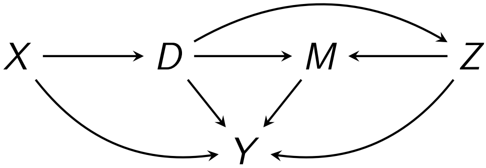
$$\int_x \int_z \mathbf{E}[Y_i|x, d = 1, z, m] dF_{Z|D,X}(z|d = 1, x) dF_X(x)$$
$$- \int_x \int_z \mathbf{E}[Y_i|x, d = 0, z, m] dF_{Z|D,X}(z|d = 0, x) dF_X(x)$$

Sequential g-estimation



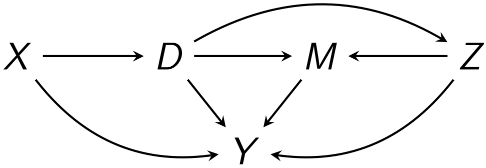
- Sequential g-estimation is one of many approaches in these settings.

Sequential g-estimation



- Sequential g-estimation is one of many approaches in these settings.
    - Other approaches include weighting.

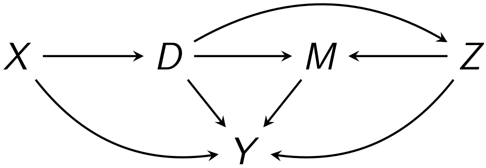
Sequential g-estimation



- Sequential g-estimation is one of many approaches in these settings.
  - Other approaches include weighting.
- Run the "long" regression:

$$Y_i = \gamma_0 + \gamma_1 D_i + \gamma_2 M_i + X_i' \gamma_3 + Z_i' \gamma_4 + \varepsilon_i$$

Sequential g-estimation

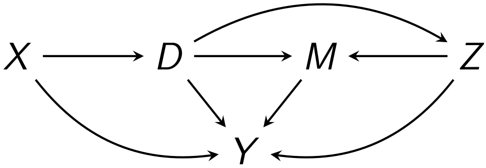


- Sequential g-estimation is one of many approaches in these settings.
  - Other approaches include weighting.
- Run the "long" regression:

$$Y_i = \gamma_0 + \gamma_1 D_i + \gamma_2 M_i + X_i' \gamma_3 + Z_i' \gamma_4 + \varepsilon_i$$

- γ_1 is not the CDE (posttreatment bias)

Sequential g-estimation

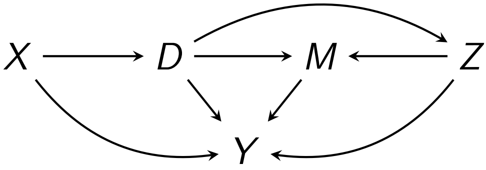


- Sequential g-estimation is one of many approaches in these settings.
  - Other approaches include weighting.
- Run the "long" regression:

$$Y_i = \gamma_0 + \gamma_1 D_i + \gamma_2 M_i + X_i' \gamma_3 + Z_i' \gamma_4 + \varepsilon_i$$

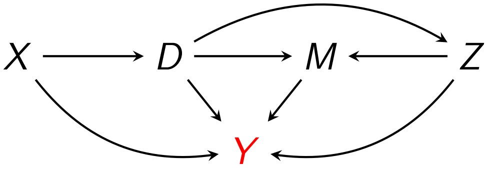
- γ_1 is not the CDE (posttreatment bias)
- γ_2 is the effect of M_i on Y_i

Blip down



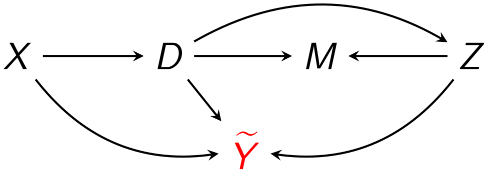
- Create a blipped down (or demediated) outcome: $\widetilde{Y}_i = Y_i - \widehat{\gamma}_2 M_i$

Blip down



- Create a blipped down (or demediated) outcome: $\widetilde{Y}_i = Y_i - \widehat{\gamma}_2 M_i$
- The blip-down removes the effect of M_i on Y_i from the outcome.

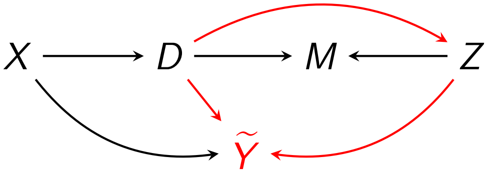
Blip down



- Create a blipped down (or demediated) outcome: $\widetilde{Y}_i = Y_i - \widehat{\gamma}_2 M_i$
- The blip-down removes the effect of M_i on Y_i from the outcome.
- Any remaining effect of D_i on Y_i is just the CDE:

$$\mathbf{E}[\widetilde{Y}_i | D_i = d, X_i] = \mathbf{E}[Y_i(d, 0) | X_i]$$

Blip down



- Create a blipped down (or demediated) outcome: $\widetilde{Y}_i = Y_i - \widehat{\gamma}_2 M_i$
- The blip-down removes the effect of M_i on Y_i from the outcome.
- Any remaining effect of D_i on Y_i is just the CDE:

$$\mathbf{E}[\widetilde{Y}_i | D_i = d, X_i] = \mathbf{E}[Y_i(d, 0) | X_i]$$

Sequential g-estimation

1. Run a regression of Y_i on M_i, Z_i, D_i, X_i.

$$Y_i = \gamma_0 + \gamma_1 D_i + \gamma_2 M_i + X_i' \gamma_3 + Z_i' \gamma_4 + \varepsilon_i$$

Sequential g-estimation

1. Run a regression of Y_i on M_i, Z_i, D_i, X_i.

$$Y_i = \gamma_0 + \gamma_1 D_i + \gamma_2 M_i + X_i' \gamma_3 + Z_i' \gamma_4 + \varepsilon_i$$

2. Subtract off the effect of M_i on Y_i:

$$\widetilde{Y}_i = Y_i - \widehat{\gamma}_2 M_i$$

Sequential g-estimation

1. Run a regression of Y_i on M_i, Z_i, D_i, X_i.

$$Y_i = \gamma_0 + \gamma_1 D_i + \gamma_2 M_i + X_i'\gamma_3 + Z_i'\gamma_4 + \varepsilon_i$$

2. Subtract off the effect of M_i on Y_i:

$$\widetilde{Y}_i = Y_i - \widehat{\gamma}_2 M_i$$

3. Regress blipped-down outcome on D_i and X_i:

$$\widetilde{Y}_i = \beta_0 + \beta_1 D_i + X_i'\beta_2 + \eta_i$$
$$CDE(0) = \mathbf{E}[Y_i(1,0) - Y_i(0,0)] = \beta_1$$

Sequential g-estimation

1. Run a regression of Y_i on M_i, Z_i, D_i, X_i.

$$Y_i = \gamma_0 + \gamma_1 D_i + \gamma_2 M_i + X_i'\gamma_3 + Z_i'\gamma_4 + \varepsilon_i$$

2. Subtract off the effect of M_i on Y_i:

$$\widetilde{Y}_i = Y_i - \widehat{\gamma}_2 M_i$$

3. Regress blipped-down outcome on D_i and X_i:

$$\widetilde{Y}_i = \beta_0 + \beta_1 D_i + X_i'\beta_2 + \eta_i$$
$$CDE(0) = \mathbf{E}[Y_i(1,0) - Y_i(0,0)] = \beta_1$$

4. Bootstrap or complicated variance estimator for SEs

Sequential g-estimation

1. Run a regression of Y_i on M_i, Z_i, D_i, X_i.

$$Y_i = \gamma_0 + \gamma_1 D_i + \gamma_2 M_i + X_i'\gamma_3 + Z_i'\gamma_4 + \varepsilon_i$$

2. Subtract off the effect of M_i on Y_i:

$$\widetilde{Y}_i = Y_i - \widehat{\gamma}_2 M_i$$

3. Regress blipped-down outcome on D_i and X_i:

$$\widetilde{Y}_i = \beta_0 + \beta_1 D_i + X_i'\beta_2 + \eta_i$$
$$CDE(0) = \mathbf{E}[Y_i(1,0) - Y_i(0,0)] = \beta_1$$

4. Bootstrap or complicated variance estimator for SEs
   - Second regression ignores the first regression.

Notes on sequential g-estimation

- Relies on a no (average) interaction assumption between CDE and intermediate confounders.

Notes on sequential g-estimation

- Relies on a no (average) interaction assumption between CDE and intermediate confounders.
- We can weaken this, but requires us to model the distribution of Z_i which might be very high dimensional:

$$\int_x \int_z \mathbb{E}[Y_i | x, d = 1, z, m] dF_{Z|D,X}(z | d = 1, x) dF_X(x)$$
$$- \int_x \int_z \mathbb{E}[Y_i | x, d = 0, z, m] dF_{Z|D,X}(z | d = 0, x) dF_X(x)$$

Notes on sequential g-estimation

- Relies on a no (average) interaction assumption between CDE and intermediate confounders.
- We can weaken this, but requires us to model the distribution of Z_i which might be very high dimensional:

$$\int_x \int_z \mathbb{E}[Y_i|x, d = 1, z, m] dF_{Z|D,X}(z|d = 1, x) dF_X(x)$$
$$- \int_x \int_z \mathbb{E}[Y_i|x, d = 0, z, m] dF_{Z|D,X}(z|d = 0, x) dF_X(x)$$

- Typical selection on observables: need correct model for covariates in both steps.

Notes on sequential g-estimation

- Relies on a no (average) interaction assumption between CDE and intermediate confounders.
- We can weaken this, but requires us to model the distribution of Z_i which might be very high dimensional:

$$\int_x \int_z \mathbf{E}[Y_i | x, d = 1, z, m] dF_{Z|D,X}(z | d = 1, x) dF_X(x) - \int_x \int_z \mathbf{E}[Y_i | x, d = 0, z, m] dF_{Z|D,X}(z | d = 0, x) dF_X(x)$$

- Typical selection on observables: need correct model for covariates in both steps.
- ATE - ACDE $\neq$ an indirect effect, but still can tell us something about mechanisms.

Notes on sequential g-estimation

- Relies on a no (average) interaction assumption between CDE and intermediate confounders.
- We can weaken this, but requires us to model the distribution of Z_i which might be very high dimensional:

$$\int_x \int_z \mathbb{E}[Y_i | x, d = 1, z, m] dF_{Z|D,X}(z | d = 1, x) dF_X(x)$$
$$- \int_x \int_z \mathbb{E}[Y_i | x, d = 0, z, m] dF_{Z|D,X}(z | d = 0, x) dF_X(x)$$

- Typical selection on observables: need correct model for covariates in both steps.
- ATE - ACDE $\neq$ an indirect effect, but still can tell us something about mechanisms.
- Acharya, Blackwell and Sen (2016) is a great paper on this.

Wrap-up

- Mechanisms are hard.

Wrap-up

- Mechanisms are hard.
- Mediation requires strong untestable assumptions.

Wrap-up

- Mechanisms are hard.
- Mediation requires strong untestable assumptions.
- Alternatives to mediation (like sequential g) lose the attractive property of decomposition.

Wrap-up

- Mechanisms are hard.
- Mediation requires strong untestable assumptions.
- Alternatives to mediation (like sequential g) lose the attractive property of decomposition.
- Use all techniques at your disposal to sort out competing mechanisms.

Wrap-up

- Mechanisms are hard.
- Mediation requires strong untestable assumptions.
- Alternatives to mediation (like sequential g) lose the attractive property of decomposition.
- Use all techniques at your disposal to sort out competing mechanisms.
  - Mediation

Wrap-up

- Mechanisms are hard.
- Mediation requires strong untestable assumptions.
- Alternatives to mediation (like sequential g) lose the attractive property of decomposition.
- Use all techniques at your disposal to sort out competing mechanisms.
  - Mediation
  - Controlled direct effects

Wrap-up

- Mechanisms are hard.
- Mediation requires strong untestable assumptions.
- Alternatives to mediation (like sequential g) lose the attractive property of decomposition.
- Use all techniques at your disposal to sort out competing mechanisms.
  - Mediation
  - Controlled direct effects
  - Effect modification

Wrap-up

- Mechanisms are hard.
- Mediation requires strong untestable assumptions.
- Alternatives to mediation (like sequential g) lose the attractive property of decomposition.
- Use all techniques at your disposal to sort out competing mechanisms.
  - Mediation
  - Controlled direct effects
  - Effect modification
  - Placebo tests

PSM's Statistical Properties (Nielsen and King 2016)

PSM's Statistical Properties (Nielsen and King 2016)

- PSM is Inefficient:

PSM's Statistical Properties (Nielsen and King 2016)

- PSM is Inefficient:
  - Efficient relative to complete randomization, but

PSM's Statistical Properties (Nielsen and King 2016)

- PSM is Inefficient:
  - Efficient relative to complete randomization, but
  - Inefficient relative to full blocking (Imai, King, and Nall: up to 600% difference in SEs in experiments)

PSM's Statistical Properties (Nielsen and King 2016)

- PSM is Inefficient:
  - Efficient relative to complete randomization, but
  - Inefficient relative to full blocking (Imai, King, and Nall: up to 600% difference in SEs in experiments)
- The PSM Paradox:

PSM's Statistical Properties (Nielsen and King 2016)

- PSM is Inefficient:
  - Efficient relative to complete randomization, but
  - Inefficient relative to full blocking (Imai, King, and Nall: up to 600% difference in SEs in experiments)
- The PSM Paradox:
  - If data are balanced to begin with, or after some pruning, $\hat{\pi} \approx 0.5$ (or constant within strata) $\rightsquigarrow$ matching is at random

PSM's Statistical Properties (Nielsen and King 2016)

- PSM is Inefficient:
  - Efficient relative to complete randomization, but
  - Inefficient relative to full blocking (Imai, King, and Nall: up to 600% difference in SEs in experiments)
- The PSM Paradox:
  - If data are balanced to begin with, or after some pruning, $\hat{\pi} \approx 0.5$ (or constant within strata) $\rightsquigarrow$ matching is at random
  - Random matching increases imbalance!

PSM's Statistical Properties (Nielsen and King 2016)

- PSM is Inefficient:
  - Efficient relative to complete randomization, but
  - Inefficient relative to full blocking (Imai, King, and Nall: up to 600% difference in SEs in experiments)
- The PSM Paradox:
  - If data are balanced to begin with, or after some pruning, $\hat{\pi} \approx 0.5$ (or constant within strata) $\rightsquigarrow$ matching is at random
  - Random matching increases imbalance!
  - Approximating complete randomization (by pruning) $\rightsquigarrow$ higher imbalance $\rightsquigarrow$ more inefficiency

PSM's Statistical Properties (Nielsen and King 2016)

- PSM is Inefficient:
  - Efficient relative to complete randomization, but
  - Inefficient relative to full blocking (Imai, King, and Nall: up to 600% difference in SEs in experiments)
- The PSM Paradox:
  - If data are balanced to begin with, or after some pruning, $\hat{\pi} \approx 0.5$ (or constant within strata) $\rightsquigarrow$ matching is at random
  - Random matching increases imbalance!
  - Approximating complete randomization (by pruning) $\rightsquigarrow$ higher imbalance $\rightsquigarrow$ more inefficiency
  - If the data have no good matches, the paradox won't be a problem but you're cooked anyway

PSM's Statistical Properties (Nielsen and King 2016)

- PSM is Inefficient:
  - Efficient relative to complete randomization, but
  - Inefficient relative to full blocking (Imai, King, and Nall: up to 600% difference in SEs in experiments)
- The PSM Paradox:
  - If data are balanced to begin with, or after some pruning, $\hat{\pi} \approx 0.5$ (or constant within strata) $\rightsquigarrow$ matching is at random
  - Random matching increases imbalance!
  - Approximating complete randomization (by pruning) $\rightsquigarrow$ higher imbalance $\rightsquigarrow$ more inefficiency
  - If the data have no good matches, the paradox won't be a problem but you're cooked anyway
- PSM + Researcher Degrees of Freedom is Biased:

PSM's Statistical Properties (Nielsen and King 2016)

- PSM is Inefficient:
  - Efficient relative to complete randomization, but
  - Inefficient relative to full blocking (Imai, King, and Nall: up to 600% difference in SEs in experiments)
- The PSM Paradox:
  - If data are balanced to begin with, or after some pruning, $\hat{\pi} \approx 0.5$ (or constant within strata) $\rightsquigarrow$ matching is at random
  - Random matching increases imbalance!
  - Approximating complete randomization (by pruning) $\rightsquigarrow$ higher imbalance $\rightsquigarrow$ more inefficiency
  - If the data have no good matches, the paradox won't be a problem but you're cooked anyway
- PSM + Researcher Degrees of Freedom is Biased:
  - Imbalance $\rightsquigarrow$ Inefficency $\rightsquigarrow$ Model dependence $\rightsquigarrow$ Bias

PSM's Statistical Properties (Nielsen and King 2016)

- PSM is Inefficient:
  - Efficient relative to complete randomization, but
  - Inefficient relative to full blocking (Imai, King, and Nall: up to 600% difference in SEs in experiments)
- The PSM Paradox:
  - If data are balanced to begin with, or after some pruning, $\hat{\pi} \approx 0.5$ (or constant within strata) $\rightsquigarrow$ matching is at random
  - Random matching increases imbalance!
  - Approximating complete randomization (by pruning) $\rightsquigarrow$ higher imbalance $\rightsquigarrow$ more inefficiency
  - If the data have no good matches, the paradox won't be a problem but you're cooked anyway
- PSM + Researcher Degrees of Freedom is Biased:
  - Imbalance $\rightsquigarrow$ Inefficency $\rightsquigarrow$ Model dependence $\rightsquigarrow$ Bias
- Curse of Dimensionality Problems:

PSM's Statistical Properties (Nielsen and King 2016)

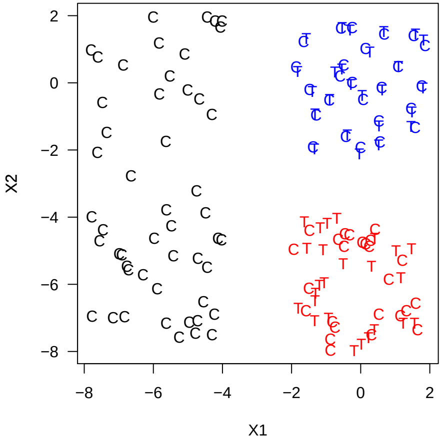
- PSM is Inefficient:
  - Efficient relative to complete randomization, but
  - Inefficient relative to full blocking (Imai, King, and Nall: up to 600% difference in SEs in experiments)
- The PSM Paradox:
  - If data are balanced to begin with, or after some pruning, $\hat{\pi} \approx 0.5$ (or constant within strata) $\rightsquigarrow$ matching is at random
  - Random matching increases imbalance!
  - Approximating complete randomization (by pruning) $\rightsquigarrow$ higher imbalance $\rightsquigarrow$ more inefficiency
  - If the data have no good matches, the paradox won't be a problem but you're cooked anyway
- PSM + Researcher Degrees of Freedom is Biased:
  - Imbalance $\rightsquigarrow$ Inefficency $\rightsquigarrow$ Model dependence $\rightsquigarrow$ Bias
- Curse of Dimensionality Problems:
  - The Promise: avoid it by balancing on π rather than X

PSM's Statistical Properties (Nielsen and King 2016)

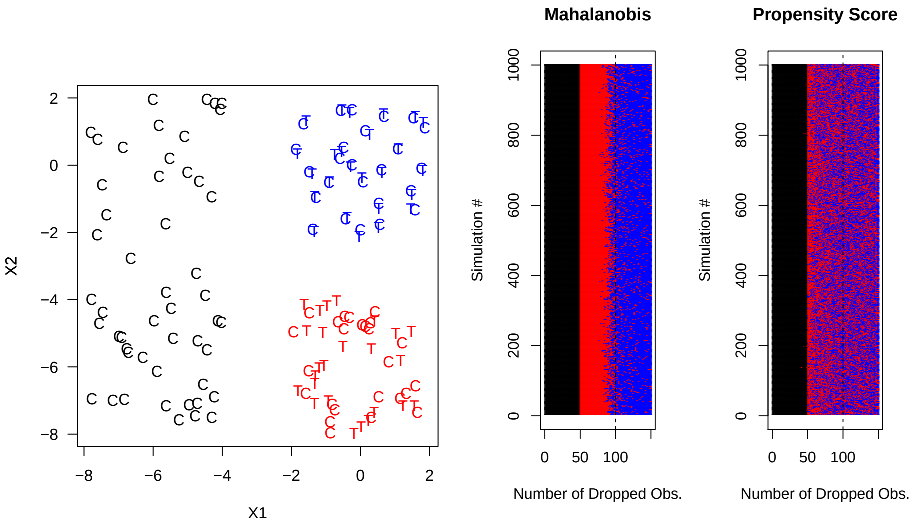
- PSM is Inefficient:
  - Efficient relative to complete randomization, but
  - Inefficient relative to full blocking (Imai, King, and Nall: up to 600% difference in SEs in experiments)
- The PSM Paradox:
  - If data are balanced to begin with, or after some pruning, $\hat{\pi} \approx 0.5$ (or constant within strata) $\rightsquigarrow$ matching is at random
  - Random matching increases imbalance!
  - Approximating complete randomization (by pruning) $\rightsquigarrow$ higher imbalance $\rightsquigarrow$ more inefficiency
  - If the data have no good matches, the paradox won't be a problem but you're cooked anyway
- PSM + Researcher Degrees of Freedom is Biased:
  - Imbalance $\rightsquigarrow$ Inefficency $\rightsquigarrow$ Model dependence $\rightsquigarrow$ Bias
- Curse of Dimensionality Problems:
  - The Promise: avoid it by balancing on π rather than X
  - The Reality: The PSM Paradox is bigger with more covariates

PSM is Blind Where Other Methods Can See

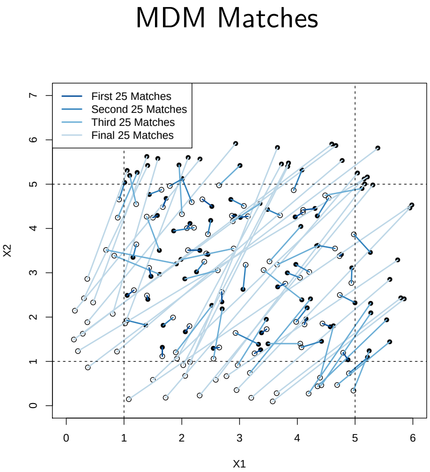
PSM is Blind Where Other Methods Can See

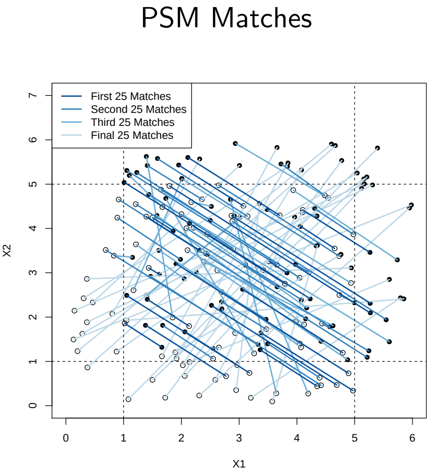


PSM is Blind Where Other Methods Can See



What Does PSM Match?



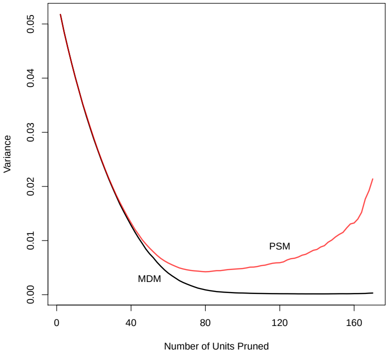


Controls: $X_1, X_2 \sim \text{Uniform}(0,5)$

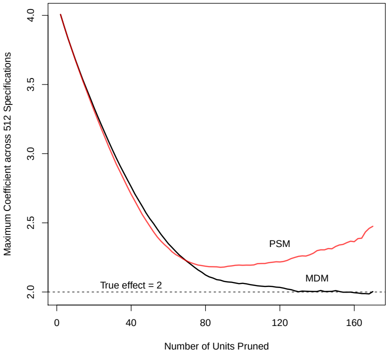
Treateds: $X_1, X_2 \sim \text{Uniform}(1,6)$

PSM Increases Model Dependence & Bias

Model Dependence

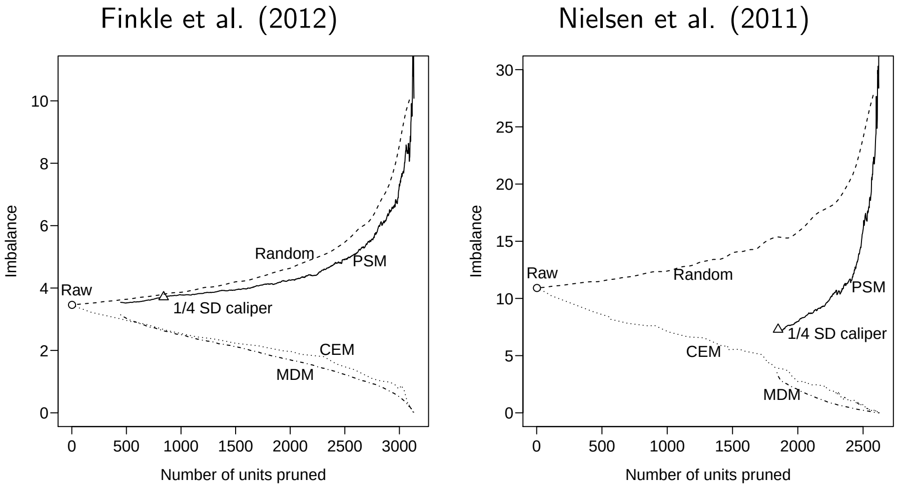


Bias



$$Y_i = 2T_i + X_{1i} + X_{2i} + \epsilon_i$$
$$\epsilon_i \sim N(0,1)$$

The Propensity Score Paradox



The Matching Frontier (King, Lucas, Nielsen 2017)

The Matching Frontier (King, Lucas, Nielsen 2017)

- Bias-Variance trade off $\rightsquigarrow$ Imbalance-*n* Trade Off
  Frontier = matched dataset with lowest imbalance for each *n*

The Matching Frontier (King, Lucas, Nielsen 2017)

- Bias-Variance trade off $\rightsquigarrow$ Imbalance-n Trade Off
  Frontier = matched dataset with lowest imbalance for each n
- To use, make 3 choices:

The Matching Frontier (King, Lucas, Nielsen 2017)

- Bias-Variance trade off $\rightsquigarrow$ Imbalance-n Trade Off
  Frontier = matched dataset with lowest imbalance for each n
- To use, make 3 choices:
  1. Imbalance metric, e.g.:

The Matching Frontier (King, Lucas, Nielsen 2017)

- Bias-Variance trade off $\rightsquigarrow$ Imbalance-n Trade Off
  Frontier = matched dataset with lowest imbalance for each n
- To use, make 3 choices:
  1. Imbalance metric, e.g.:
     - Average Mahalanobis Distance (average distance from each unit to the closest in the other treatment regime)

The Matching Frontier (King, Lucas, Nielsen 2017)

- Bias-Variance trade off $\rightsquigarrow$ Imbalance-n Trade Off
  Frontier = matched dataset with lowest imbalance for each n
- To use, make 3 choices:
  1. Imbalance metric, e.g.:
     - ★ Average Mahalanobis Distance (average distance from each unit to the closest in the other treatment regime)
     - ★ Difference of multivariate histograms (L_1):

The Matching Frontier (King, Lucas, Nielsen 2017)

- Bias-Variance trade off $\rightsquigarrow$ Imbalance-n Trade Off
  Frontier = matched dataset with lowest imbalance for each n
- To use, make 3 choices:
  1. Imbalance metric, e.g.:
     - ★ Average Mahalanobis Distance (average distance from each unit to the closest in the other treatment regime)
     - ★ Difference of multivariate histograms (L_1):
  2. Quantity of interest: SATT (prune Cs only) or FSATT

The Matching Frontier (King, Lucas, Nielsen 2017)

- Bias-Variance trade off $\rightsquigarrow$ Imbalance-n Trade Off
  Frontier = matched dataset with lowest imbalance for each n
- To use, make 3 choices:
  1. Imbalance metric, e.g.:
     - Average Mahalanobis Distance (average distance from each unit to the closest in the other treatment regime)
     - Difference of multivariate histograms (L_1):
  2. Quantity of interest: SATT (prune Cs only) or FSATT
  3. Fixed- or variable-ratio matching

The Matching Frontier (King, Lucas, Nielsen 2017)

- Bias-Variance trade off $\rightsquigarrow$ Imbalance-n Trade Off
  Frontier = matched dataset with lowest imbalance for each n
- To use, make 3 choices:
  1. Imbalance metric, e.g.:
     - ★ Average Mahalanobis Distance (average distance from each unit to the closest in the other treatment regime)
     - ★ Difference of multivariate histograms (L_1):
  2. Quantity of interest: SATT (prune Cs only) or FSATT
  3. Fixed- or variable-ratio matching
- Result:

The Matching Frontier (King, Lucas, Nielsen 2017)

- Bias-Variance trade off $\rightsquigarrow$ Imbalance-n Trade Off
  Frontier = matched dataset with lowest imbalance for each n
- To use, make 3 choices:
  1. Imbalance metric, e.g.:
     - Average Mahalanobis Distance (average distance from each unit to the closest in the other treatment regime)
     - Difference of multivariate histograms (L_1):
  2. Quantity of interest: SATT (prune Cs only) or FSATT
  3. Fixed- or variable-ratio matching
- Result:
  - Simple to use

The Matching Frontier (King, Lucas, Nielsen 2017)

- Bias-Variance trade off $\rightsquigarrow$ Imbalance-n Trade Off
  Frontier = matched dataset with lowest imbalance for each n
- To use, make 3 choices:
  1. Imbalance metric, e.g.:
     - Average Mahalanobis Distance (average distance from each unit to the closest in the other treatment regime)
     - Difference of multivariate histograms (L_1):
  2. Quantity of interest: SATT (prune Cs only) or FSATT
  3. Fixed- or variable-ratio matching
- Result:
  - Simple to use
  - All solutions are optimal

The Matching Frontier (King, Lucas, Nielsen 2017)

- Bias-Variance trade off $\rightsquigarrow$ Imbalance-n Trade Off
  Frontier = matched dataset with lowest imbalance for each n
- To use, make 3 choices:
  1. Imbalance metric, e.g.:
     - Average Mahalanobis Distance (average distance from each unit to the closest in the other treatment regime)
     - Difference of multivariate histograms (L_1):
  2. Quantity of interest: SATT (prune Cs only) or FSATT
  3. Fixed- or variable-ratio matching
- Result:
  - Simple to use
  - All solutions are optimal
  - No iteration or diagnostics required

The Matching Frontier (King, Lucas, Nielsen 2017)

- Bias-Variance trade off $\rightsquigarrow$ Imbalance-n Trade Off
  Frontier = matched dataset with lowest imbalance for each n
- To use, make 3 choices:
  1. Imbalance metric, e.g.:
     - Average Mahalanobis Distance (average distance from each unit to the closest in the other treatment regime)
     - Difference of multivariate histograms (L_1):
  2. Quantity of interest: SATT (prune Cs only) or FSATT
  3. Fixed- or variable-ratio matching
- Result:
  - Simple to use
  - All solutions are optimal
  - No iteration or diagnostics required
  - No cherry picking possible

How hard is the frontier to calculate?

How hard is the frontier to calculate?

- Consider 1 point on the SATT frontier:

How hard is the frontier to calculate?

- Consider 1 point on the SATT frontier:
  - Start with matrix of N control units X_0

How hard is the frontier to calculate?

- Consider 1 point on the SATT frontier:
  - Start with matrix of N control units X_0
  - Calculate imbalance for <u>all</u> $\binom{N}{n}$ subsets of rows of X_0

How hard is the frontier to calculate?

- Consider 1 point on the SATT frontier:
  - Start with matrix of N control units X_0
  - Calculate imbalance for all $\binom{N}{n}$ subsets of rows of X_0
  - Choose subset with lowest imbalance

How hard is the frontier to calculate?

- Consider 1 point on the SATT frontier:
  - Start with matrix of N control units X_0
  - Calculate imbalance for all $\binom{N}{n}$ subsets of rows of X_0
  - Choose subset with lowest imbalance
- Evaluations needed to compute the entire frontier:

How hard is the frontier to calculate?

- Consider 1 point on the SATT frontier:
  - Start with matrix of N control units X_0
  - Calculate imbalance for <u>all</u> $\binom{N}{n}$ subsets of rows of X_0
  - Choose subset with lowest imbalance
- Evaluations needed to compute the entire frontier:
  - $\binom{N}{n}$ evaluations for <u>each</u> sample size $n = N, N-1, \ldots, 1$

How hard is the frontier to calculate?

- Consider 1 point on the SATT frontier:
  - Start with matrix of N control units X_0
  - Calculate imbalance for all $\binom{N}{n}$ subsets of rows of X_0
  - Choose subset with lowest imbalance
- Evaluations needed to compute the entire frontier:
  - $\binom{N}{n}$ evaluations for each sample size $n = N, N-1, \ldots, 1$
  - The combination is the (gargantuan) "power set"

How hard is the frontier to calculate?

- Consider 1 point on the SATT frontier:
  - Start with matrix of N control units X_0
  - Calculate imbalance for all $\binom{N}{n}$ subsets of rows of X_0
  - Choose subset with lowest imbalance
- Evaluations needed to compute the entire frontier:
  - $\binom{N}{n}$ evaluations for each sample size $n = N, N-1, \ldots, 1$
  - The combination is the (gargantuan) "power set"
  - e.g., $N > 300$ requires more imbalance evaluations than elementary particles in the universe

How hard is the frontier to calculate?

- Consider 1 point on the SATT frontier:
  - Start with matrix of N control units X_0
  - Calculate imbalance for all $\binom{N}{n}$ subsets of rows of X_0
  - Choose subset with lowest imbalance
- Evaluations needed to compute the entire frontier:
  - $\binom{N}{n}$ evaluations for each sample size $n = N, N-1, \ldots, 1$
  - The combination is the (gargantuan) "power set"
  - e.g., $N > 300$ requires more imbalance evaluations than elementary particles in the universe
  - $\rightsquigarrow$ It's **hard** to calculate!

How hard is the frontier to calculate?

- Consider 1 point on the SATT frontier:
  - Start with matrix of N control units X_0
  - Calculate imbalance for all $\binom{N}{n}$ subsets of rows of X_0
  - Choose subset with lowest imbalance
- Evaluations needed to compute the entire frontier:
  - $\binom{N}{n}$ evaluations for each sample size $n = N, N-1, \ldots, 1$
  - The combination is the (gargantuan) "power set"
  - e.g., $N > 300$ requires more imbalance evaluations than elementary particles in the universe
  - $\rightsquigarrow$ It's **hard** to calculate!
- King, Lucas and Nielsen develop algorithms for the (optimal) frontier which:

How hard is the frontier to calculate?

- Consider 1 point on the SATT frontier:
  - Start with matrix of N control units X_0
  - Calculate imbalance for all $\binom{N}{n}$ subsets of rows of X_0
  - Choose subset with lowest imbalance
- Evaluations needed to compute the entire frontier:
  - $\binom{N}{n}$ evaluations for each sample size $n = N, N-1, \ldots, 1$
  - The combination is the (gargantuan) "power set"
  - e.g., $N > 300$ requires more imbalance evaluations than elementary particles in the universe
  - $\rightsquigarrow$ It's **hard** to calculate!
- King, Lucas and Nielsen develop algorithms for the (optimal) frontier which:
  - runs very fast

How hard is the frontier to calculate?

- Consider 1 point on the SATT frontier:
  - Start with matrix of N control units X_0
  - Calculate imbalance for all $\binom{N}{n}$ subsets of rows of X_0
  - Choose subset with lowest imbalance
- Evaluations needed to compute the entire frontier:
  - $\binom{N}{n}$ evaluations for each sample size $n = N, N-1, \dots, 1$
  - The combination is the (gargantuan) "power set"
  - e.g., $N > 300$ requires more imbalance evaluations than elementary particles in the universe
  - $\rightsquigarrow$ It's **hard** to calculate!
- King, Lucas and Nielsen develop algorithms for the (optimal) frontier which:
  - runs very fast
  - operate as "greedy" but they prove are optimal

How hard is the frontier to calculate?

- Consider 1 point on the SATT frontier:
  - Start with matrix of N control units X_0
  - Calculate imbalance for all $\binom{N}{n}$ subsets of rows of X_0
  - Choose subset with lowest imbalance
- Evaluations needed to compute the entire frontier:
  - $\binom{N}{n}$ evaluations for each sample size $n = N, N-1, \ldots, 1$
  - The combination is the (gargantuan) "power set"
  - e.g., $N > 300$ requires more imbalance evaluations than elementary particles in the universe
  - $\rightsquigarrow$ It's **hard** to calculate!
- King, Lucas and Nielsen develop algorithms for the (optimal) frontier which:
  - runs very fast
  - operate as "greedy" but they prove are optimal
  - do not require evaluating every subset

How hard is the frontier to calculate?

- Consider 1 point on the SATT frontier:
  - Start with matrix of N control units X_0
  - Calculate imbalance for all $\binom{N}{n}$ subsets of rows of X_0
  - Choose subset with lowest imbalance
- Evaluations needed to compute the entire frontier:
  - $\binom{N}{n}$ evaluations for each sample size $n = N, N-1, \ldots, 1$
  - The combination is the (gargantuan) "power set"
  - e.g., $N > 300$ requires more imbalance evaluations than elementary particles in the universe
  - $\rightsquigarrow$ It's **hard** to calculate!
- King, Lucas and Nielsen develop algorithms for the (optimal) frontier which:
  - runs very fast
  - operate as "greedy" but they prove are optimal
  - do not require evaluating every subset
  - work with very large data sets

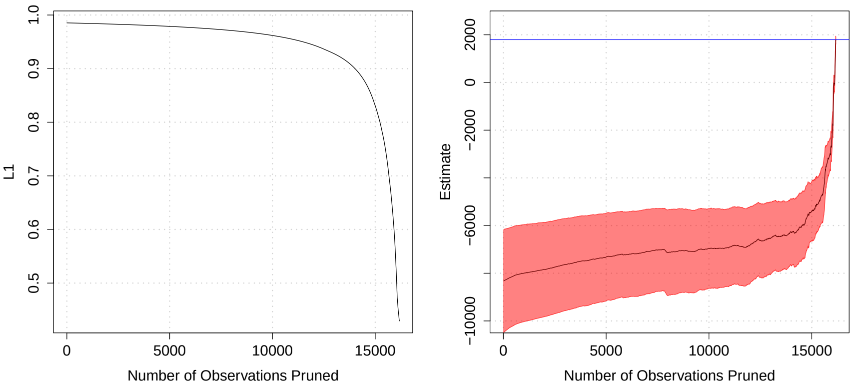
How hard is the frontier to calculate?

- Consider 1 point on the SATT frontier:
  - Start with matrix of N control units X_0
  - Calculate imbalance for all $\binom{N}{n}$ subsets of rows of X_0
  - Choose subset with lowest imbalance
- Evaluations needed to compute the entire frontier:
  - $\binom{N}{n}$ evaluations for each sample size $n = N, N-1, \ldots, 1$
  - The combination is the (gargantuan) "power set"
  - e.g., $N > 300$ requires more imbalance evaluations than elementary particles in the universe
  - $\rightsquigarrow$ It's **hard** to calculate!
- King, Lucas and Nielsen develop algorithms for the (optimal) frontier which:
  - runs very fast
  - operate as "greedy" but they prove are optimal
  - do not require evaluating every subset
  - work with very large data sets
  - is the exact frontier (no approximation or estimation)

How hard is the frontier to calculate?

- Consider 1 point on the SATT frontier:
  - Start with matrix of N control units X_0
  - Calculate imbalance for all $\binom{N}{n}$ subsets of rows of X_0
  - Choose subset with lowest imbalance
- Evaluations needed to compute the entire frontier:
  - $\binom{N}{n}$ evaluations for each sample size $n = N, N-1, \ldots, 1$
  - The combination is the (gargantuan) "power set"
  - e.g., $N > 300$ requires more imbalance evaluations than elementary particles in the universe
  - $\rightsquigarrow$ It's **hard** to calculate!
- King, Lucas and Nielsen develop algorithms for the (optimal) frontier which:
  - runs very fast
  - operate as "greedy" but they prove are optimal
  - do not require evaluating every subset
  - work with very large data sets
  - is the exact frontier (no approximation or estimation)
  - $\rightsquigarrow$ It's **easy** to calculate!

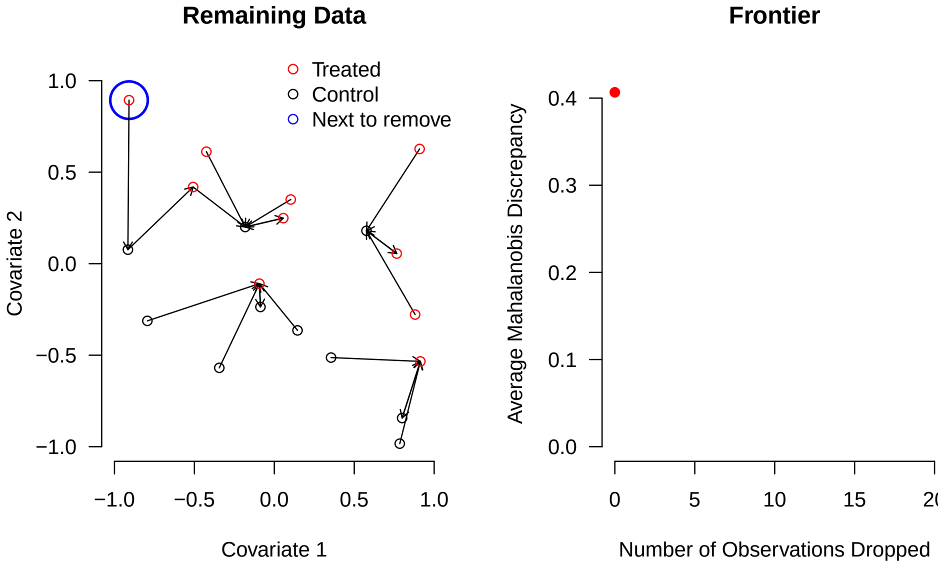
Job Training Data: Frontier and Causal Estimates



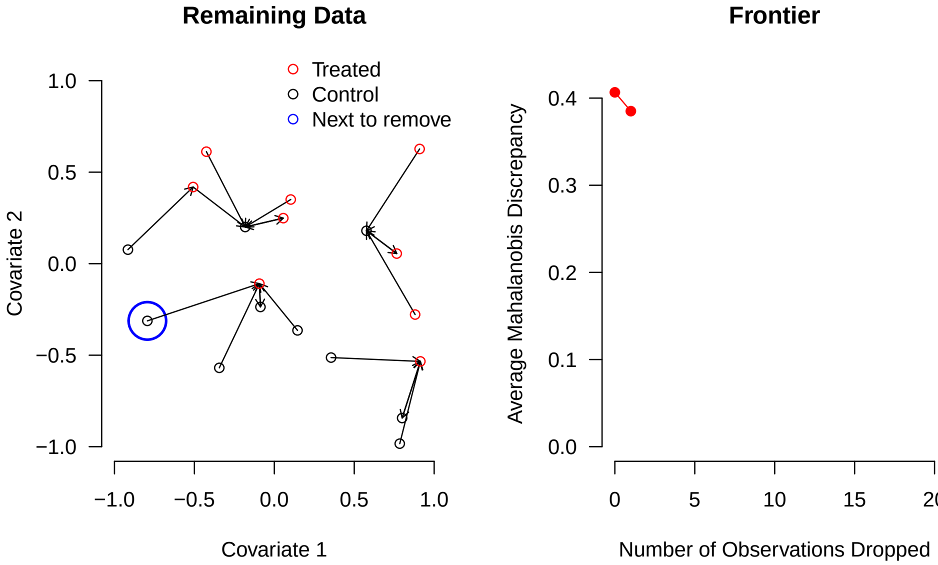
- 185 Ts; pruning most 16,252 Cs won't increase variance much
- Huge bias-variance trade-off after pruning most Cs
- Estimates converge to experiment after removing bias
- No mysteries: basis of inference clearly revealed

Constructing the FSATT Mahalanobis Frontier

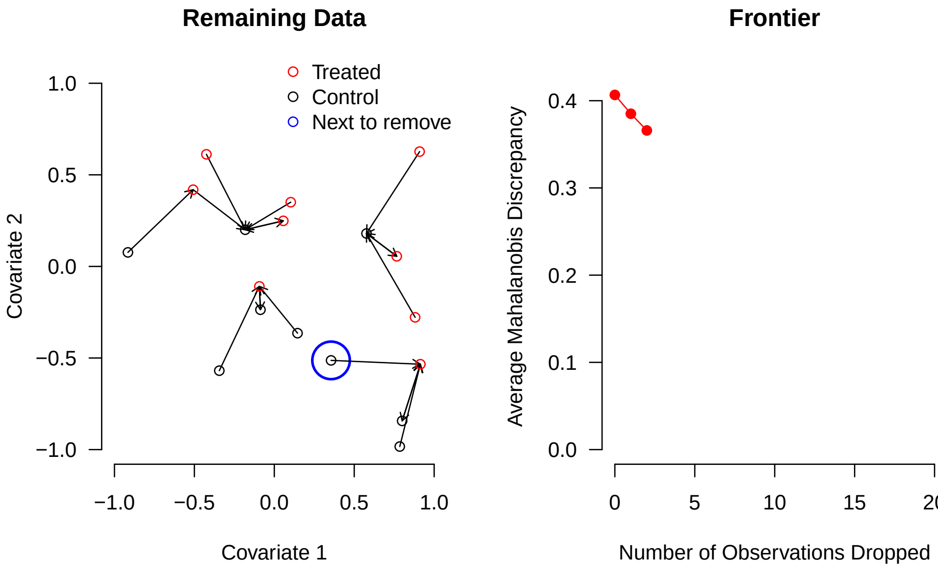
Constructing the FSATT Mahalanobis Frontier



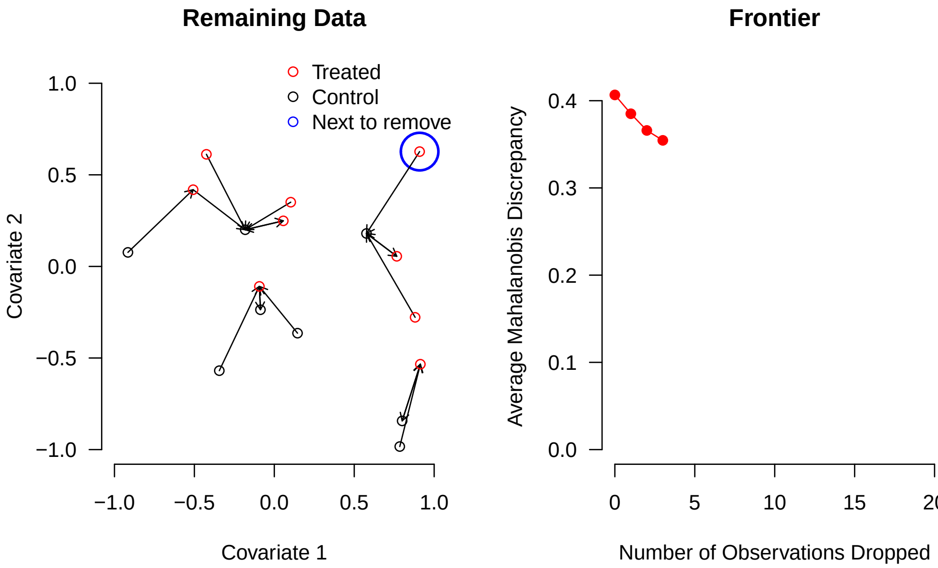
Constructing the FSATT Mahalanobis Frontier



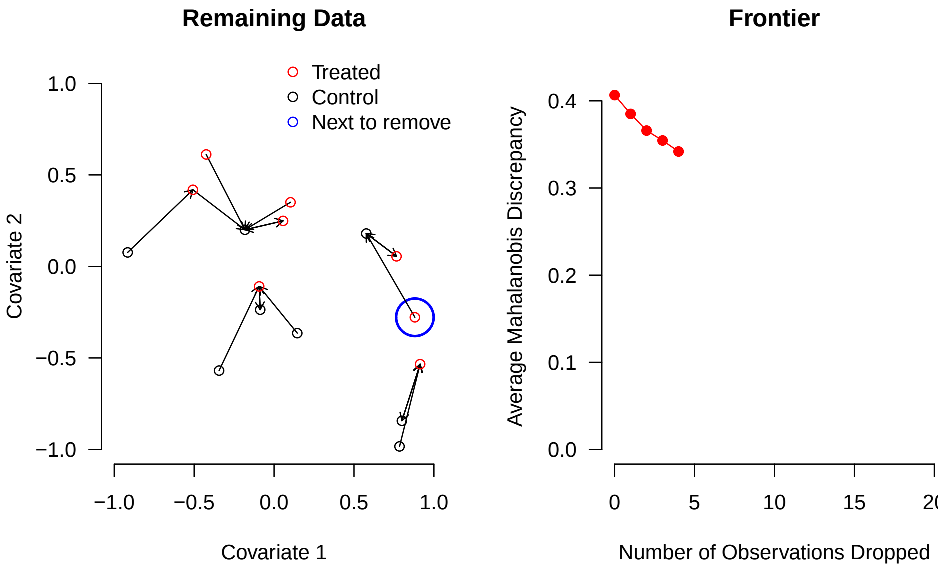
Constructing the FSATT Mahalanobis Frontier



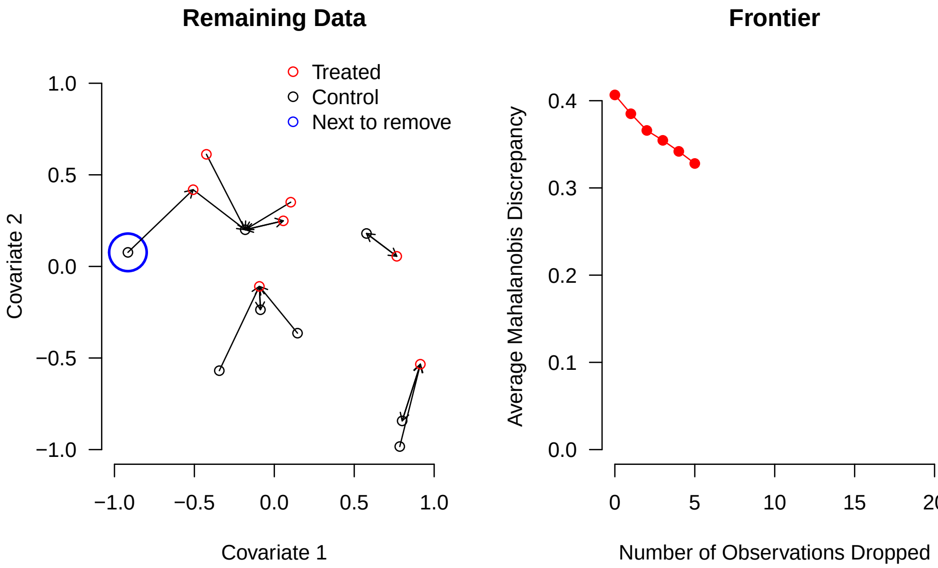
Constructing the FSATT Mahalanobis Frontier



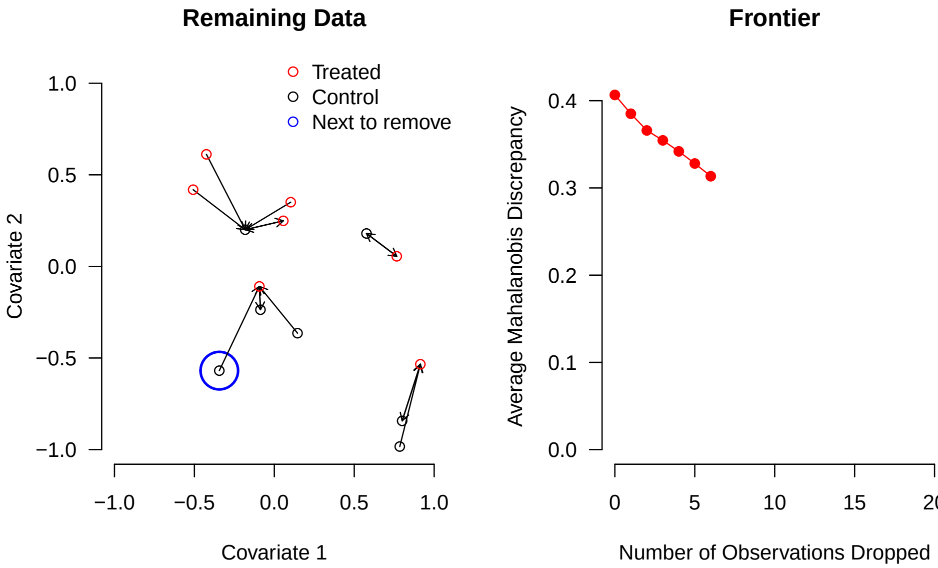
Constructing the FSATT Mahalanobis Frontier



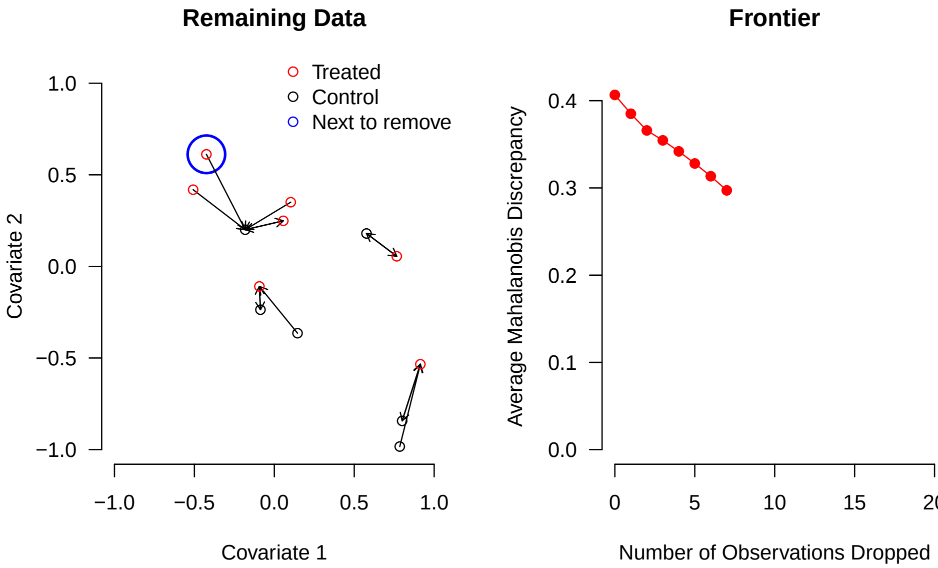
Constructing the FSATT Mahalanobis Frontier



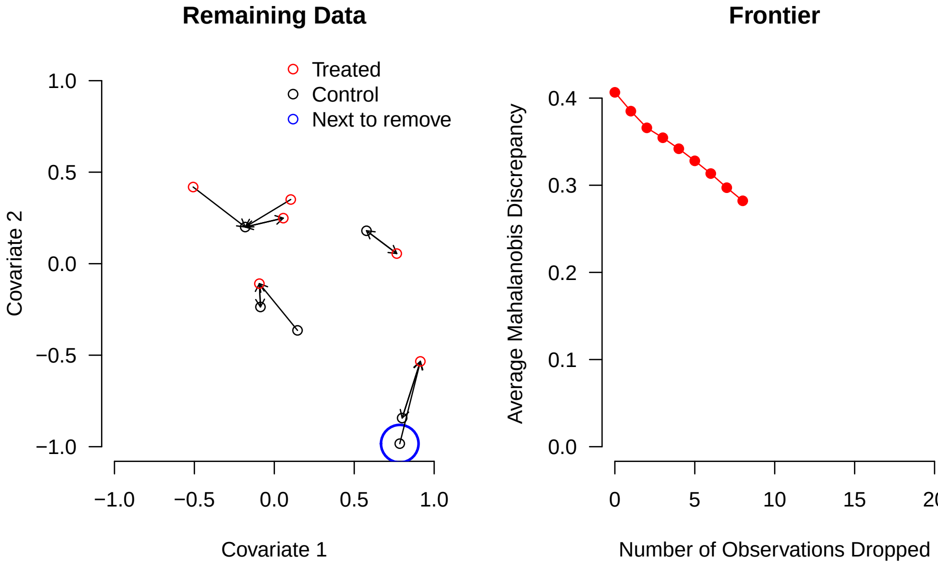
Constructing the FSATT Mahalanobis Frontier



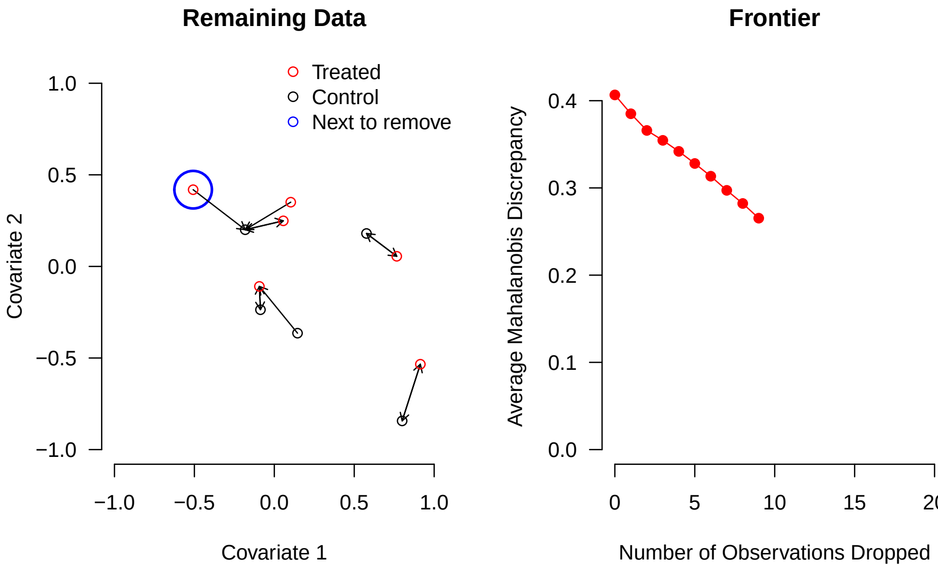
Constructing the FSATT Mahalanobis Frontier



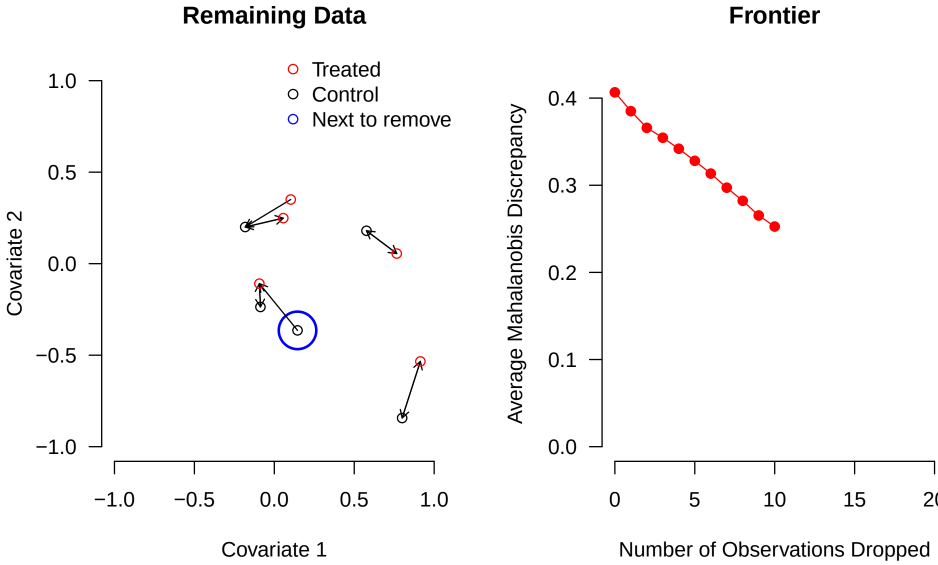
Constructing the FSATT Mahalanobis Frontier



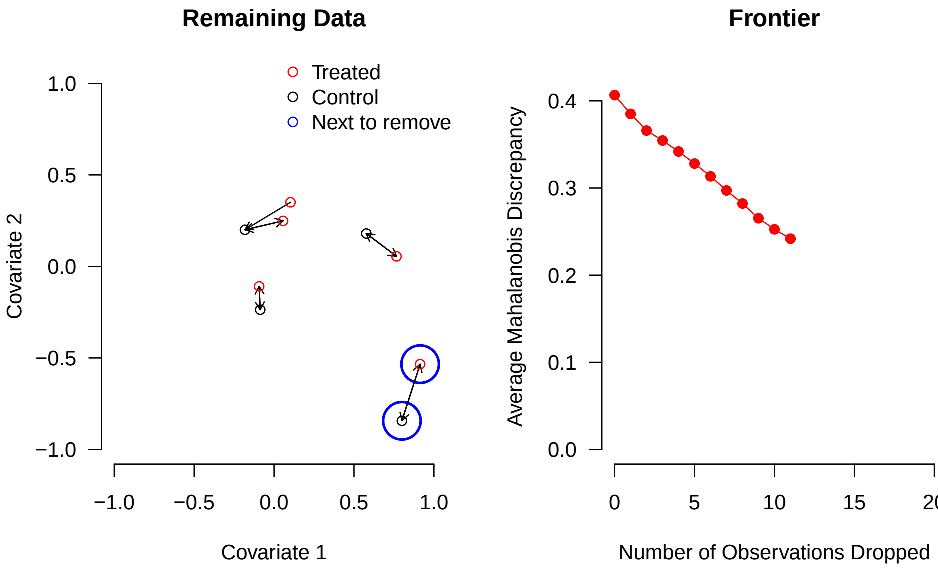
Constructing the FSATT Mahalanobis Frontier



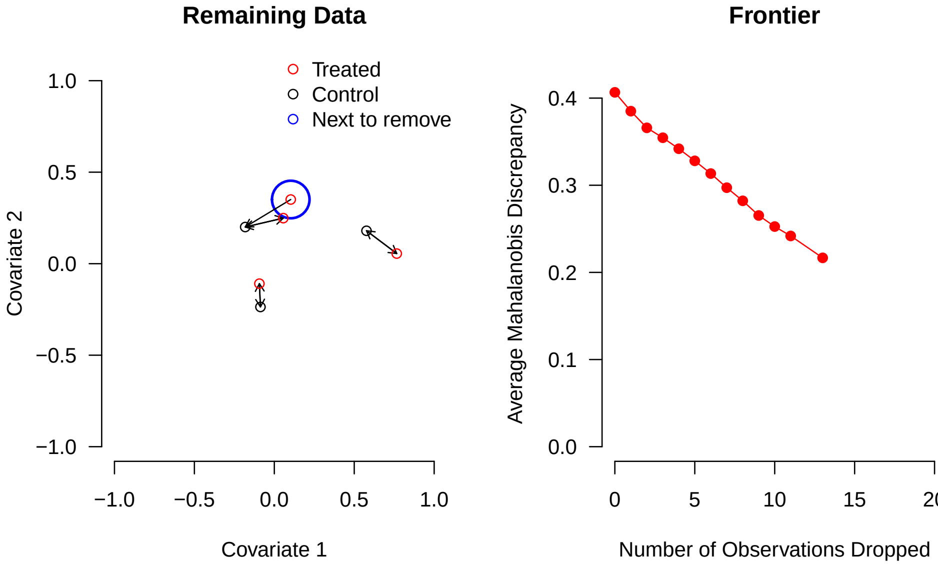
Constructing the FSATT Mahalanobis Frontier



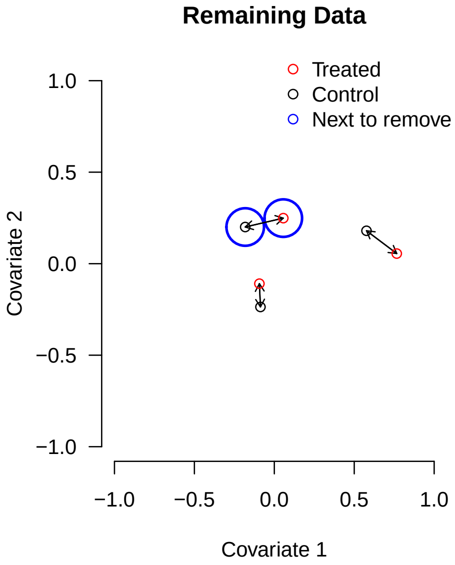
Constructing the FSATT Mahalanobis Frontier

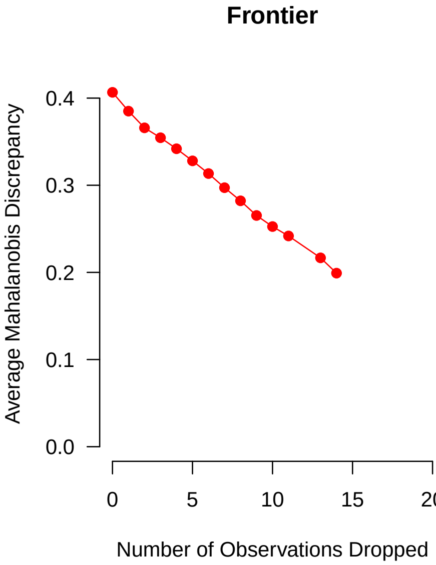


Constructing the FSATT Mahalanobis Frontier

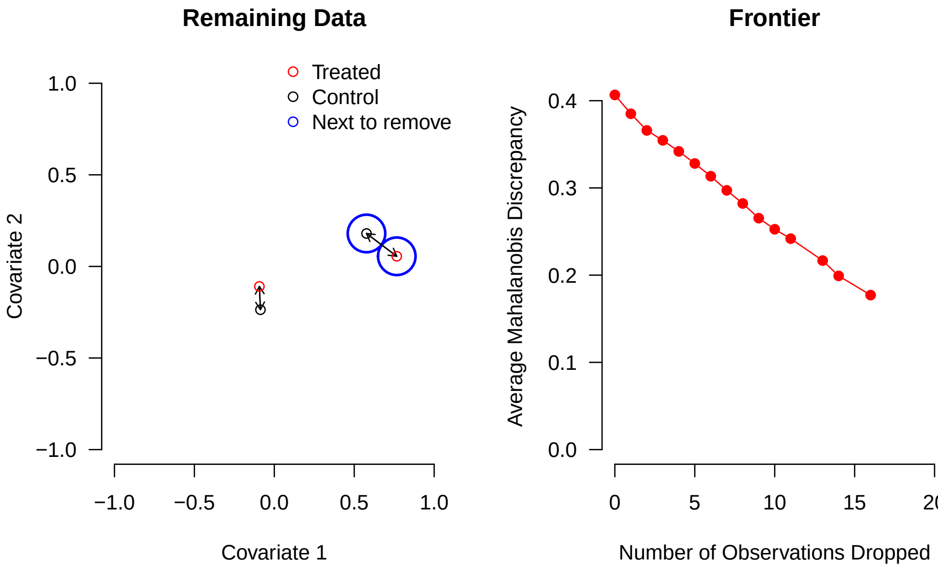


Constructing the FSATT Mahalanobis Frontier

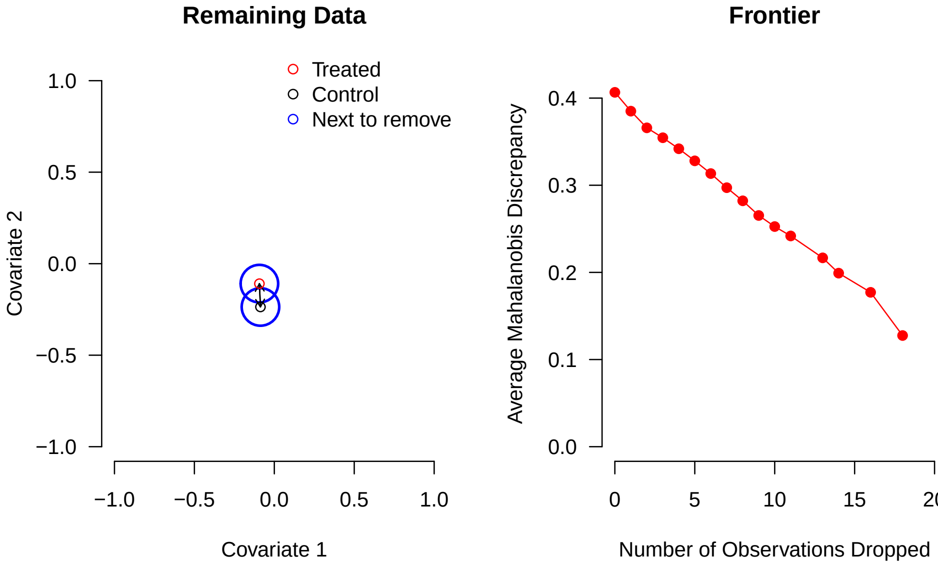




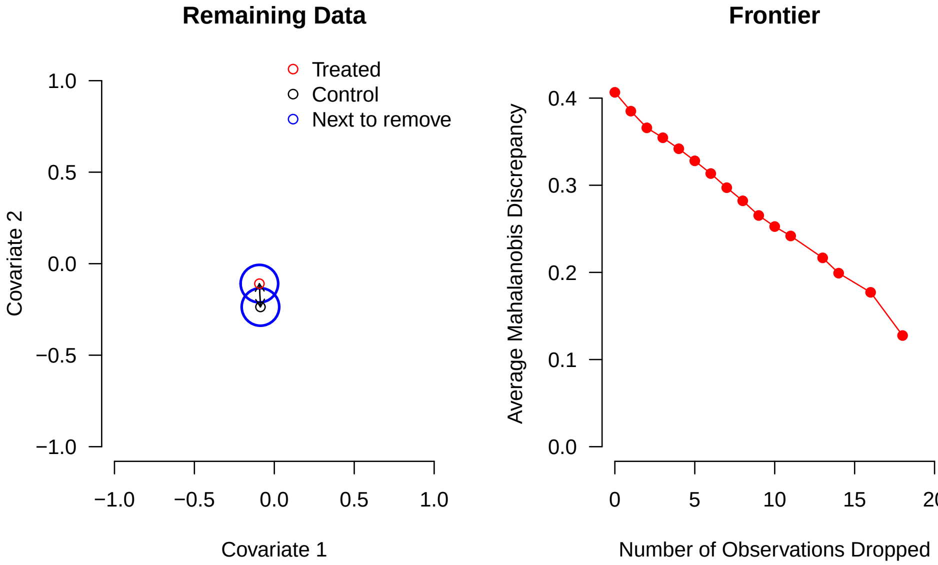
Constructing the FSATT Mahalanobis Frontier



Constructing the FSATT Mahalanobis Frontier

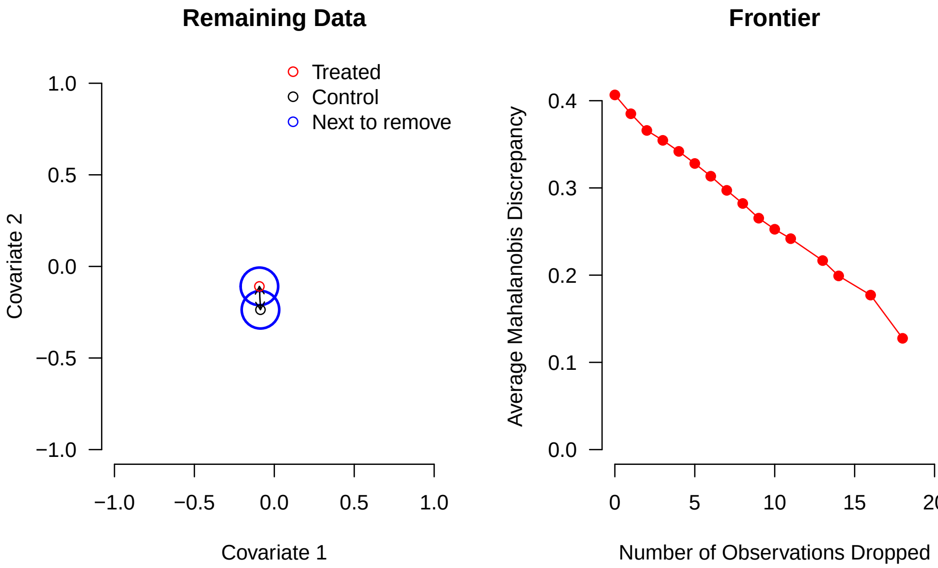


Constructing the FSATT Mahalanobis Frontier



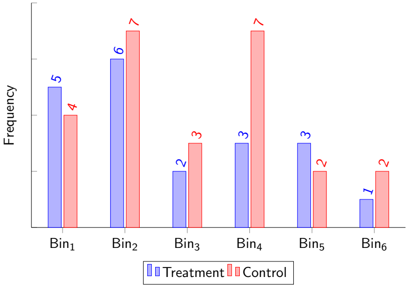
- Warning: figure omits some details!

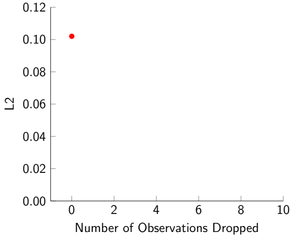
Constructing the FSATT Mahalanobis Frontier



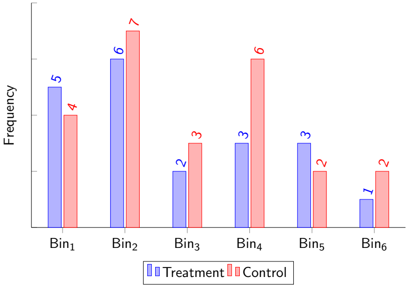
- Warning: figure omits some details!

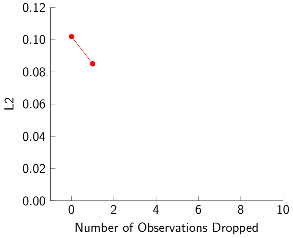
Constructing the L1/L2 SATT Frontier



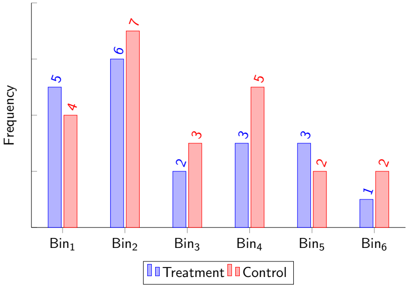


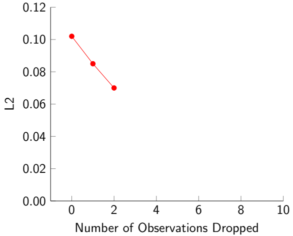
Constructing the L1/L2 SATT Frontier



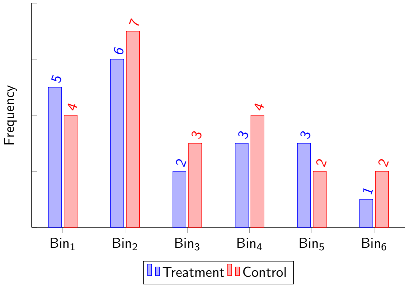


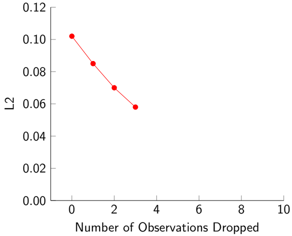
Constructing the L1/L2 SATT Frontier



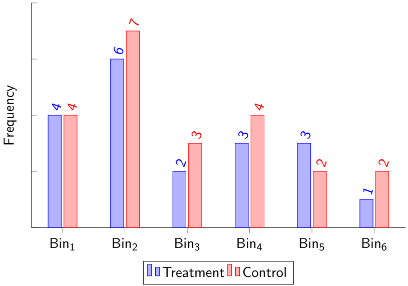


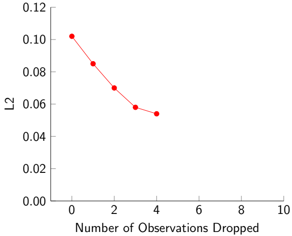
Constructing the L1/L2 SATT Frontier



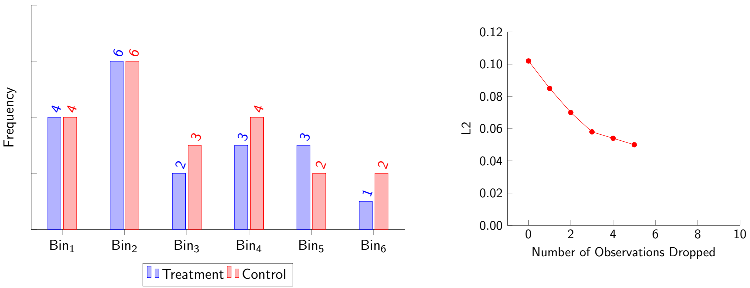


Constructing the L1/L2 SATT Frontier

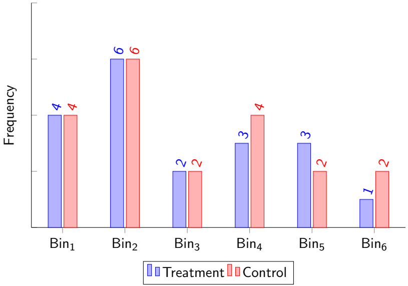


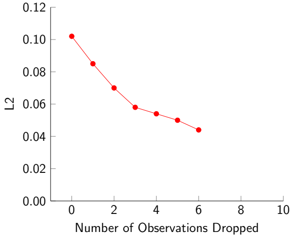


Constructing the L1/L2 SATT Frontier

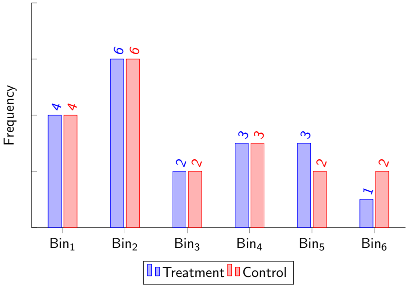


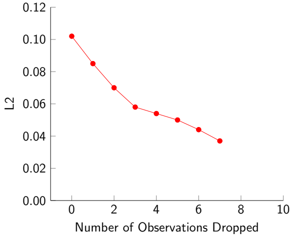
Constructing the L1/L2 SATT Frontier



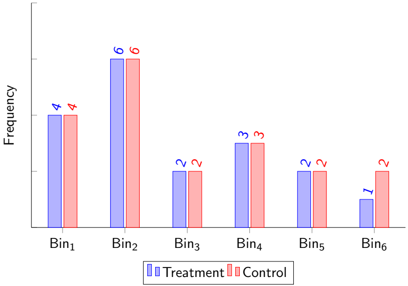


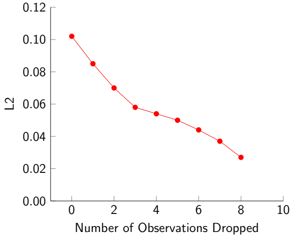
Constructing the L1/L2 SATT Frontier





Constructing the L1/L2 SATT Frontier





Constructing the L1/L2 SATT Frontier