Why Propensity Scores Should Not Be Used For Matching

Gary King[1]

Richard Nielsen[2]

Institute for Quantitative Social Science
Harvard University

MIT

MacMillan-CSAP Workshop on Quantitative Research Methods,
Yale University, 3/10/2016

[1]GaryKing.org
[2]www.mit.edu/∼rnielsen

The Scholarly Influence of Propensity Score Matching

The Scholarly Influence of Propensity Score Matching

- The most commonly used matching method

The Scholarly Influence of Propensity Score Matching

- The most commonly used matching method
- In 53,200 articles! (according to Google Scholar)

The Scholarly Influence of Propensity Score Matching

- The most commonly used matching method
- In 53,200 articles! (according to Google Scholar)
- Maybe even "the most developed and popular strategy for causal analysis in observational studies" (Pearl, 2010)

The Scholarly Influence of Propensity Score Matching

- The most commonly used matching method
- In 53,200 articles! (according to Google Scholar)
- Maybe even "the most developed and popular strategy for causal analysis in observational studies" (Pearl, 2010)

This paper is about:

The Scholarly Influence of Propensity Score Matching

- The most commonly used matching method
- In 53,200 articles! (according to Google Scholar)
- Maybe even "the most developed and popular strategy for causal analysis in observational studies" (Pearl, 2010)

This paper is about:

- propensity score matching

The Scholarly Influence of Propensity Score Matching

- The most commonly used matching method
- In 53,200 articles! (according to Google Scholar)
- Maybe even "the most developed and popular strategy for causal analysis in observational studies" (Pearl, 2010)

This paper is about:

- propensity score matching
- as used in practice

The Scholarly Influence of Propensity Score Matching

- The most commonly used matching method
- In 53,200 articles! (according to Google Scholar)
- Maybe even "the most developed and popular strategy for causal analysis in observational studies" (Pearl, 2010)

This paper is about:

- propensity score matching
- as used in practice

Not implicated by our results:

The Scholarly Influence of Propensity Score Matching

- The most commonly used matching method
- In 53,200 articles! (according to Google Scholar)
- Maybe even "the most developed and popular strategy for causal analysis in observational studies" (Pearl, 2010)

This paper is about:

- propensity score matching
- as used in practice

Not implicated by our results:

- Other uses of propensity scores:

The Scholarly Influence of Propensity Score Matching

- The most commonly used matching method
- In 53,200 articles! (according to Google Scholar)
- Maybe even "the most developed and popular strategy for causal analysis in observational studies" (Pearl, 2010)

This paper is about:

- propensity score matching
- as used in practice

Not implicated by our results:

- Other uses of propensity scores: E.g., regression adjustment,

The Scholarly Influence of Propensity Score Matching

- The most commonly used matching method
- In 53,200 articles! (according to Google Scholar)
- Maybe even "the most developed and popular strategy for causal analysis in observational studies" (Pearl, 2010)

This paper is about:

- propensity score matching
- as used in practice

Not implicated by our results:

- Other uses of propensity scores: E.g., regression adjustment, inverse weighting,

The Scholarly Influence of Propensity Score Matching

- The most commonly used matching method
- In 53,200 articles! (according to Google Scholar)
- Maybe even "the most developed and popular strategy for causal analysis in observational studies" (Pearl, 2010)

This paper is about:

- propensity score matching
- as used in practice

Not implicated by our results:

- Other uses of propensity scores: E.g., regression adjustment, inverse weighting, stratification,

The Scholarly Influence of Propensity Score Matching

- The most commonly used matching method
- In 53,200 articles! (according to Google Scholar)
- Maybe even "the most developed and popular strategy for causal analysis in observational studies" (Pearl, 2010)

This paper is about:

- propensity score matching
- as used in practice

Not implicated by our results:

- Other uses of propensity scores: E.g., regression adjustment, inverse weighting, stratification, pscores used in other methods

The Scholarly Influence of Propensity Score Matching

- The most commonly used matching method
- In 53,200 articles! (according to Google Scholar)
- Maybe even "the most developed and popular strategy for causal analysis in observational studies" (Pearl, 2010)

This paper is about:

- propensity score matching
- as used in practice

Not implicated by our results:

- Other uses of propensity scores: E.g., regression adjustment, inverse weighting, stratification, pscores used in other methods
- The mathematical theorems about propensity scores:

The Scholarly Influence of Propensity Score Matching

- The most commonly used matching method
- In 53,200 articles! (according to Google Scholar)
- Maybe even "the most developed and popular strategy for causal analysis in observational studies" (Pearl, 2010)

This paper is about:

- propensity score matching
- as used in practice

Not implicated by our results:

- Other uses of propensity scores: E.g., regression adjustment, inverse weighting, stratification, pscores used in other methods
- The mathematical theorems about propensity scores: Correct, but inadequate

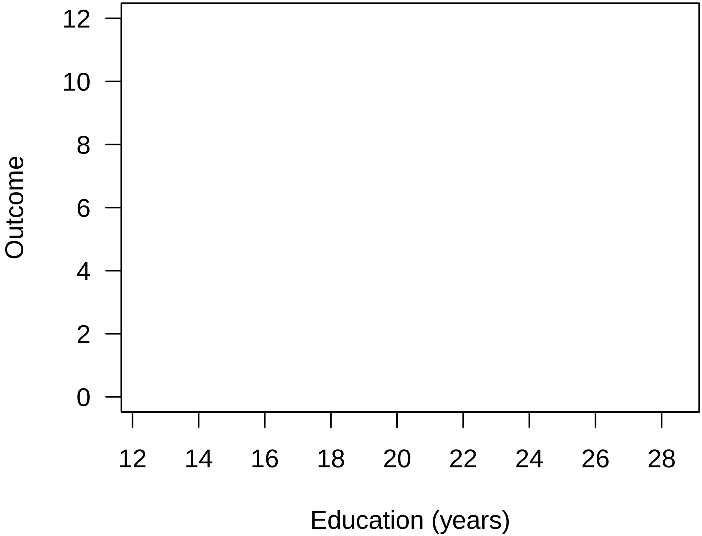
Matching to Reduce Model Dependence

Matching to Reduce Model Dependence

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

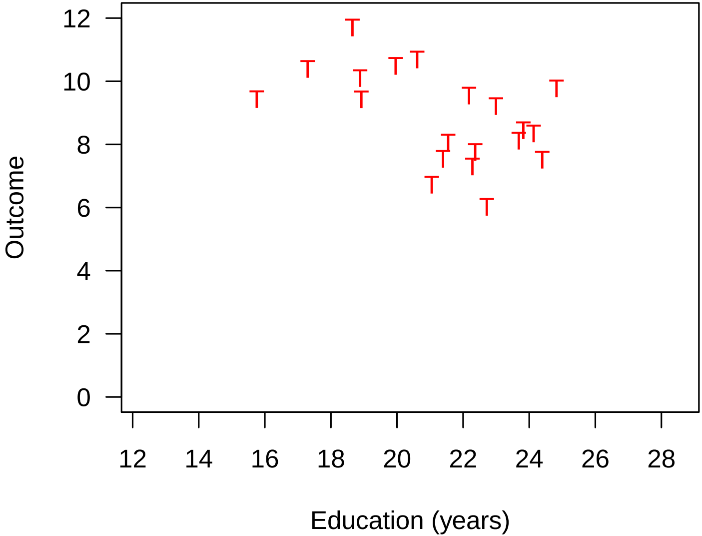
Matching to Reduce Model Dependence

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



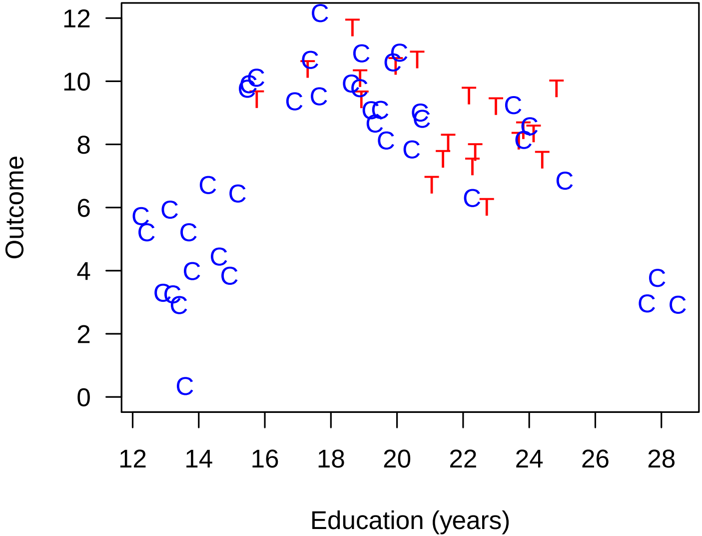
Matching to Reduce Model Dependence

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



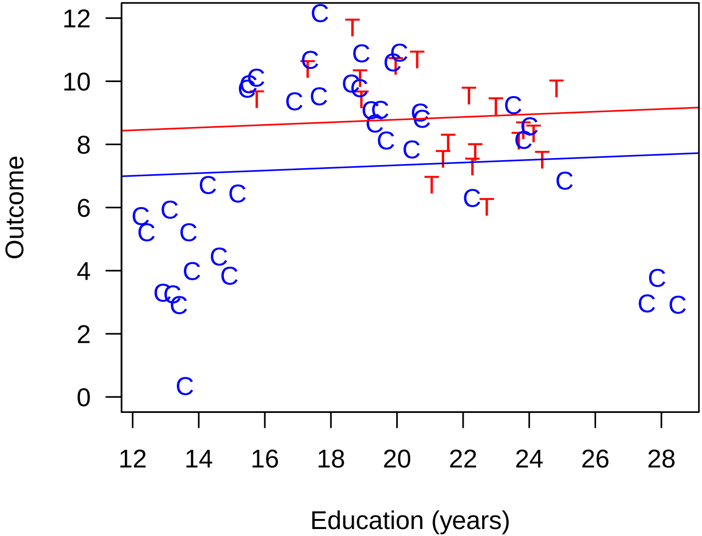
Matching to Reduce Model Dependence

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



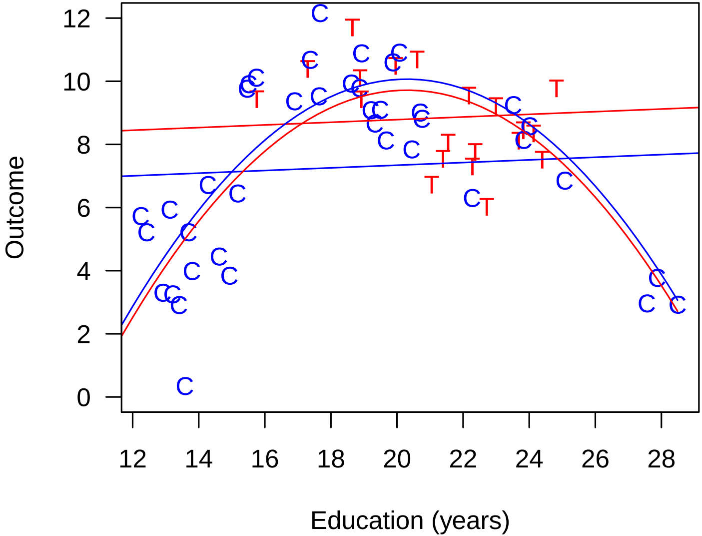
Matching to Reduce Model Dependence

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



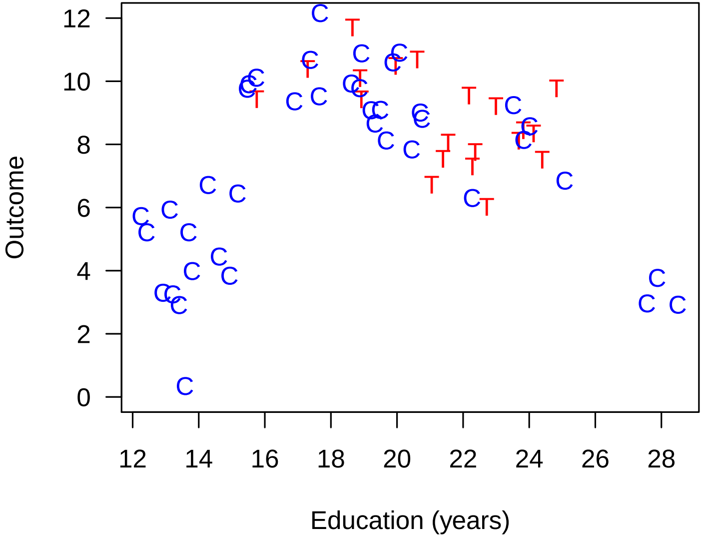
Matching to Reduce Model Dependence

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



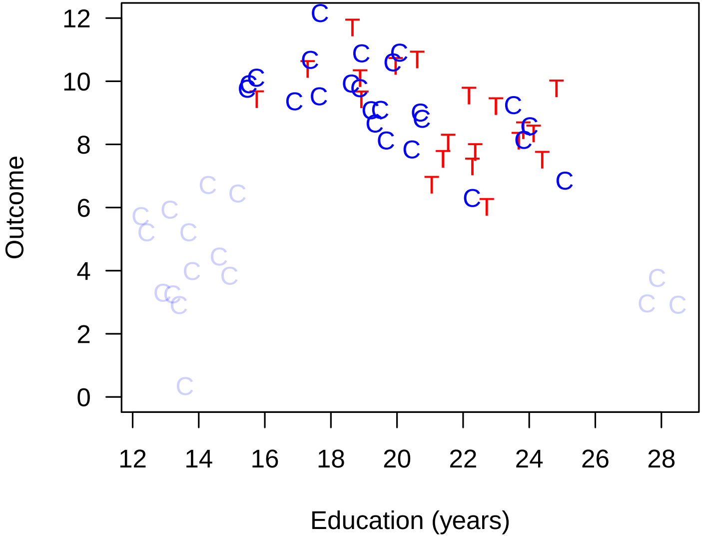
Matching to Reduce Model Dependence

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



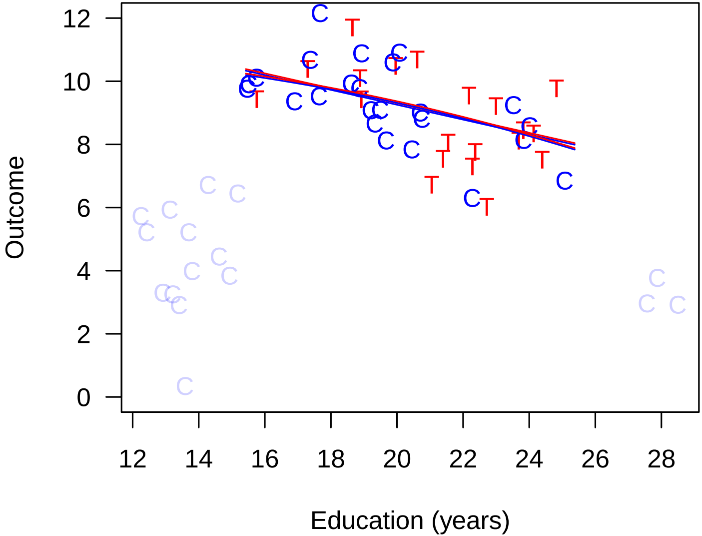
Matching to Reduce Model Dependence

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



Matching to Reduce Model Dependence

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



The Problems Matching Solves

The Problems Matching Solves

Without Matching:

The Problems Matching Solves

Without Matching:

Imbalance

The Problems Matching Solves

Without Matching:

Imbalance $\rightsquigarrow$ Model Dependence

The Problems Matching Solves

Without Matching:

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

The Problems Matching Solves

Without Matching:

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

The Problems Matching Solves

Without Matching:

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

- Qualitative choice from unbiased estimates $=$ biased estimator

The Problems Matching Solves

Without Matching:

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

- Qualitative choice from unbiased estimates = biased estimator
  - e.g., Choosing from *results* of 50 randomized experiments

The Problems Matching Solves

Without Matching:

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

- Qualitative choice from unbiased estimates = biased estimator
  - e.g., Choosing from *results* of 50 randomized experiments
  - Choosing based on "plausibility" is probably worse[eff]

The Problems Matching Solves

Without Matching:

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

- Qualitative choice from unbiased estimates $=$ biased estimator
  - e.g., Choosing from *results* of 50 randomized experiments
  - Choosing based on "plausibility" is probably worse[eff]
- conscientious effort doesn't avoid biases (Banaji 2013)[acc]

The Problems Matching Solves

Without Matching:

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

- Qualitative choice from unbiased estimates = biased estimator
  - e.g., Choosing from *results* of 50 randomized experiments
  - Choosing based on "plausibility" is probably worse[eff]
- conscientious effort doesn't avoid biases (Banaji 2013)[acc]
- People do not have easy access to their own mental processes or feedback to avoid the problem (Wilson and Brekke 1994)[exprt]

The Problems Matching Solves

Without Matching:

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

- Qualitative choice from unbiased estimates = biased estimator
  - e.g., Choosing from *results* of 50 randomized experiments
  - Choosing based on "plausibility" is probably worse [eff]
- conscientious effort doesn't avoid biases (Banaji 2013) [acc]
- People do not have easy access to their own mental processes or feedback to avoid the problem (Wilson and Brekke 1994) [exprt]
- Experts overestimate their ability to control personal biases more than nonexperts, and more prominent experts are the most overconfident (Tetlock 2005) [tch]

The Problems Matching Solves

Without Matching:

Imbalance ⇝ Model Dependence ⇝ Researcher discretion ⇝ Bias

- Qualitative choice from unbiased estimates = biased estimator
  - e.g., Choosing from *results* of 50 randomized experiments
  - Choosing based on "plausibility" is probably worse[eff]
- conscientious effort doesn't avoid biases (Banaji 2013)[acc]
- People do not have easy access to their own mental processes or feedback to avoid the problem (Wilson and Brekke 1994)[exprt]
- Experts overestimate their ability to control personal biases more than nonexperts, and more prominent experts are the most overconfident (Tetlock 2005)[tch]
- "Teaching psychology is mostly a waste of time" (Kahneman 2011)

The Problems Matching Solves

Without Matching: ("out" is struck through)

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

The Problems Matching Solves

Without Matching:

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

The Problems Matching Solves

Without Matching:

Imbalance ⇝ Model Dependence ⇝ Researcher discretion ⇝ Bias

The Problems Matching Solves

With~~out~~ Matching:

~~Im~~balance ⇝ ~~Model Dependence~~ ⇝ ~~Researcher discretion~~ ⇝ Bias

The Problems Matching Solves

With~~out~~ Matching:

~~Im~~balance ⇝ ~~Model Dependence~~ ⇝ ~~Researcher discretion~~ ⇝ ~~Bias~~

The Problems Matching Solves

With~~out~~ Matching:

~~Im~~balance ⇝ ~~Model Dependence~~ ⇝ ~~Researcher discretion~~ ⇝ ~~Bias~~

A central project of statistics: Automating away human discretion

What's Matching?

What's Matching?

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

What's Matching?

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

What's Matching?

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

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

What's Matching?

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

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

What's Matching?

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

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

What's Matching?

- 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 ${\color{red}Y_i(0)}$ with Y_j with a matched ($X_i \approx X_j$) control

What's Matching?

- 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 with a matched ($X_i \approx X_j$) control
- Quantities of Interest:

What's Matching?

- 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 with a matched ($X_i \approx X_j$) control
- Quantities of Interest:
  1. SATT: Sample Average Treatment effect on the Treated:

$$\text{SATT} = \underset{i \in \{T_i = 1\}}{\text{Mean}} (\text{TE}_i)$$

What's Matching?

- 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 with a matched ($X_i \approx X_j$) control
- Quantities of Interest:
  1. SATT: Sample Average Treatment effect on the Treated:

$$\text{SATT} = \underset{i \in \{T_i = 1\}}{\text{Mean}} (\text{TE}_i)$$

  2. FSATT: Feasible SATT (prune badly matched treateds too)

What's Matching?

- 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 with a matched ($X_i \approx X_j$) control
- Quantities of Interest:
  1. SATT: Sample Average Treatment effect on the Treated:

$$\text{SATT} = \underset{i \in \{T_i = 1\}}{\text{Mean}} (\text{TE}_i)$$

  2. FSATT: Feasible SATT (prune badly matched treateds too)
- Big convenience: Follow preprocessing with whatever statistical method you'd have used without matching

What's Matching?

- 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 with a matched ($X_i \approx X_j$) control
- Quantities of Interest:
  1. SATT: Sample Average Treatment effect on the Treated:

$$\text{SATT} = \underset{i \in \{T_i = 1\}}{\text{Mean}} (\text{TE}_i)$$

  2. FSATT: Feasible SATT (prune badly matched treateds too)
- Big convenience: Follow preprocessing with whatever statistical method you'd have used without matching
- Pruning nonmatches makes control vars matter less: reduces imbalance, model dependence, researcher discretion, & bias

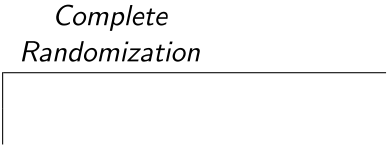
Matching: Finding Hidden Randomized Experiments

Matching: Finding Hidden Randomized Experiments

Types of Experiments

Matching: Finding Hidden Randomized Experiments

Types of Experiments



Types of Experiments

	Complete Randomization	*Fully Blocked*

Matching: Finding Hidden Randomized Experiments

Types of Experiments

Balance Covariates:	*Complete Randomization*	*Fully Blocked*
Observed		
Unobserved		

Matching: Finding Hidden Randomized Experiments

Types of Experiments

Balance Covariates:	*Complete Randomization*	*Fully Blocked*
Observed	On average	
Unobserved		

Matching: Finding Hidden Randomized Experiments

Types of Experiments

Balance Covariates:	*Complete Randomization*	*Fully Blocked*
Observed	On average	
Unobserved	On average	

Matching: Finding Hidden Randomized Experiments

Types of Experiments

Balance Covariates:	*Complete Randomization*	*Fully Blocked*
Observed	On average	Exact
Unobserved	On average	

Matching: Finding Hidden Randomized Experiments

Types of Experiments

Balance Covariates:	*Complete Randomization*	*Fully Blocked*
Observed	On average	Exact
Unobserved	On average	On average

Matching: Finding Hidden Randomized Experiments

Types of Experiments

Balance Covariates:	*Complete Randomization*	*Fully Blocked*
Observed	On average	Exact
Unobserved	On average	On average

⇝ *Fully blocked* dominates *complete randomization*

Matching: Finding Hidden Randomized Experiments

Types of Experiments

Balance Covariates:	*Complete Randomization*	*Fully Blocked*
Observed	On average	Exact
Unobserved	On average	On average

⇝ *Fully blocked* dominates *complete randomization* for:

Matching: Finding Hidden Randomized Experiments

Types of Experiments

Balance Covariates:	*Complete Randomization*	*Fully Blocked*
Observed	On average	Exact
Unobserved	On average	On average

$\rightsquigarrow$ *Fully blocked* dominates *complete randomization* for:
imbalance,

Matching: Finding Hidden Randomized Experiments

Types of Experiments

Balance Covariates:	*Complete Randomization*	*Fully Blocked*
Observed	On average	Exact
Unobserved	On average	On average

⇝ *Fully blocked* dominates *complete randomization* for:
imbalance, model dependence,

Matching: Finding Hidden Randomized Experiments

Types of Experiments

Balance Covariates:	*Complete Randomization*	*Fully Blocked*
Observed	On average	Exact
Unobserved	On average	On average

⇝ *Fully blocked* dominates *complete randomization* for:
imbalance, model dependence, power,

Matching: Finding Hidden Randomized Experiments

Types of Experiments

Balance Covariates:	*Complete Randomization*	*Fully Blocked*
Observed	On average	Exact
Unobserved	On average	On average

⇝ *Fully blocked* dominates *complete randomization* for:
imbalance, model dependence, power, efficiency,

Matching: Finding Hidden Randomized Experiments

Types of Experiments

Balance Covariates:	*Complete Randomization*	*Fully Blocked*
Observed	On average	Exact
Unobserved	On average	On average

⇝ *Fully blocked* dominates *complete randomization* for: imbalance, model dependence, power, efficiency, bias,

Matching: Finding Hidden Randomized Experiments

Types of Experiments

Balance Covariates:	*Complete Randomization*	*Fully Blocked*
Observed	On average	Exact
Unobserved	On average	On average

⇝ *Fully blocked* dominates *complete randomization* for: imbalance, model dependence, power, efficiency, bias, research costs,

Matching: Finding Hidden Randomized Experiments

Types of Experiments

Balance Covariates:	*Complete Randomization*	*Fully Blocked*
Observed	On average	Exact
Unobserved	On average	On average

$\rightsquigarrow$ *Fully blocked* dominates *complete randomization* for: imbalance, model dependence, power, efficiency, bias, research costs, robustness.

Matching: Finding Hidden Randomized Experiments

Types of Experiments

Balance Covariates:	*Complete Randomization*	*Fully Blocked*
Observed	On average	Exact
Unobserved	On average	On average

$\rightsquigarrow$ *Fully blocked* dominates *complete randomization* for: imbalance, model dependence, power, efficiency, bias, research costs, robustness. E.g., Imai, King, Nall 2009: SEs 600% smaller!

Matching: Finding Hidden Randomized Experiments

Types of Experiments

Balance Covariates:	*Complete Randomization*	*Fully Blocked*
Observed	On average	Exact
Unobserved	On average	On average

$\rightsquigarrow$ *Fully blocked* dominates *complete randomization* for: imbalance, model dependence, power, efficiency, bias, research costs, robustness. E.g., Imai, King, Nall 2009: SEs 600% smaller!

Goal of Each Matching Method (in Observational Data)

Matching: Finding Hidden Randomized Experiments

Types of Experiments

Balance Covariates:	*Complete Randomization*	*Fully Blocked*
Observed	On average	Exact
Unobserved	On average	On average

⇝ *Fully blocked* dominates *complete randomization* for: imbalance, model dependence, power, efficiency, bias, research costs, robustness. E.g., Imai, King, Nall 2009: SEs 600% smaller!

Goal of Each Matching Method (in Observational Data)

- PSM: *complete randomization*

Matching: Finding Hidden Randomized Experiments

Types of Experiments

Balance Covariates:	*Complete Randomization*	*Fully Blocked*
Observed	On average	Exact
Unobserved	On average	On average

$\rightsquigarrow$ *Fully blocked* dominates *complete randomization* for: imbalance, model dependence, power, efficiency, bias, research costs, robustness. E.g., Imai, King, Nall 2009: SEs 600% smaller!

Goal of Each Matching Method (in Observational Data)

- PSM: *complete randomization*
- Other methods: *fully blocked*

Matching: Finding Hidden Randomized Experiments

Types of Experiments

Balance Covariates:	*Complete Randomization*	*Fully Blocked*
Observed	On average	Exact
Unobserved	On average	On average

⇝ *Fully blocked* dominates *complete randomization* for: imbalance, model dependence, power, efficiency, bias, research costs, robustness. E.g., Imai, King, Nall 2009: SEs 600% smaller!

Goal of Each Matching Method (in Observational Data)

- PSM: *complete randomization*
- Other methods: *fully blocked*
- Other matching methods dominate PSM

Matching: Finding Hidden Randomized Experiments

Types of Experiments

Balance Covariates:	*Complete Randomization*	*Fully Blocked*
Observed	On average	Exact
Unobserved	On average	On average

⇝ *Fully blocked* dominates *complete randomization* for: imbalance, model dependence, power, efficiency, bias, research costs, robustness. E.g., Imai, King, Nall 2009: SEs 600% smaller!

Goal of Each Matching Method (in Observational Data)

- PSM: *complete randomization*
- Other methods: *fully blocked*
- Other matching methods dominate PSM (wait, it gets worse)

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. Estimation Difference in means or a model

Method 1: Mahalanobis Distance Matching

(Approximates Fully Blocked Experiment)

1. Preprocess (Matching)
   - Distance$(X_c, X_t) = \sqrt{(X_c - X_t)' S^{-1} (X_c - X_t)}$

2. Estimation Difference in means or a model

Method 1: Mahalanobis Distance Matching

(Approximates Fully Blocked Experiment)

1. Preprocess (Matching)
   - Distance$(X_c, X_t) = \sqrt{(X_c - X_t)' S^{-1} (X_c - X_t)}$
   - (Mahalanobis is for methodologists; in applications, use Euclidean!)

2. Estimation Difference in means or a model

Method 1: Mahalanobis Distance Matching

(Approximates Fully Blocked Experiment)

1. Preprocess (Matching)
   - Distance$(X_c, X_t) = \sqrt{(X_c - X_t)' S^{-1} (X_c - X_t)}$
   - (Mahalanobis is for methodologists; in applications, use Euclidean!)
   - Match each treated unit to the nearest control unit
2. Estimation Difference in means or a model

Method 1: Mahalanobis Distance Matching

(Approximates Fully Blocked Experiment)

1. Preprocess (Matching)
   - Distance$(X_c, X_t) = \sqrt{(X_c - X_t)' S^{-1} (X_c - X_t)}$
   - (Mahalanobis is for methodologists; in applications, use Euclidean!)
   - Match each treated unit to the nearest control unit
   - Control units: not reused; pruned if unused
2. Estimation Difference in means or a model

Method 1: Mahalanobis Distance Matching

(Approximates Fully Blocked Experiment)

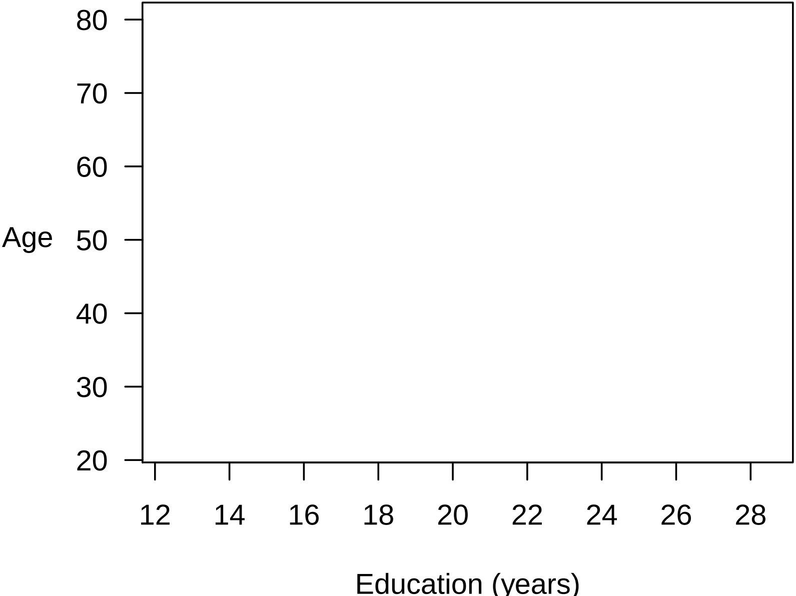
1. Preprocess (Matching)
   - Distance$(X_c, X_t) = \sqrt{(X_c - X_t)' S^{-1} (X_c - X_t)}$
   - (Mahalanobis is for methodologists; in applications, use Euclidean!)
   - Match each treated unit to the nearest control unit
   - Control units: not reused; pruned if unused
   - Prune matches if Distance$>$*caliper*
2. Estimation Difference in means or a model

Method 1: Mahalanobis Distance Matching

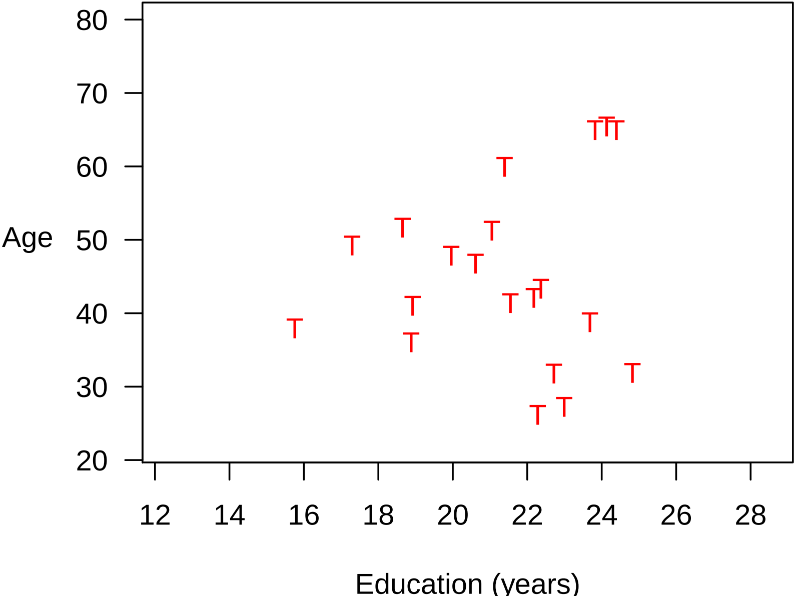
(Approximates Fully Blocked Experiment)

1. Preprocess (Matching)
   - Distance$(X_c, X_t) = \sqrt{(X_c - X_t)' S^{-1} (X_c - X_t)}$
   - (Mahalanobis is for methodologists; in applications, use Euclidean!)
   - Match each treated unit to the nearest control unit
   - Control units: not reused; pruned if unused
   - Prune matches if Distance$>$*caliper*
   - (Many adjustments available to this basic method)
2. Estimation Difference in means or a model

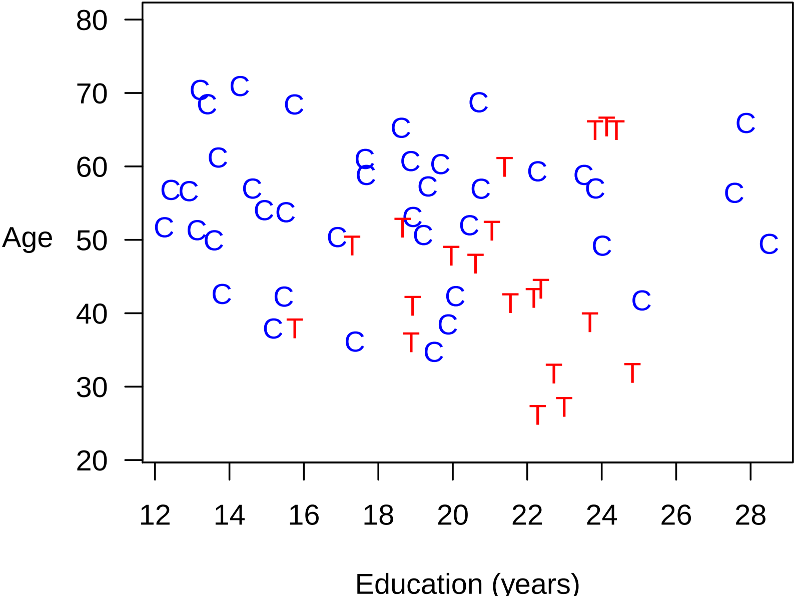
Mahalanobis Distance Matching



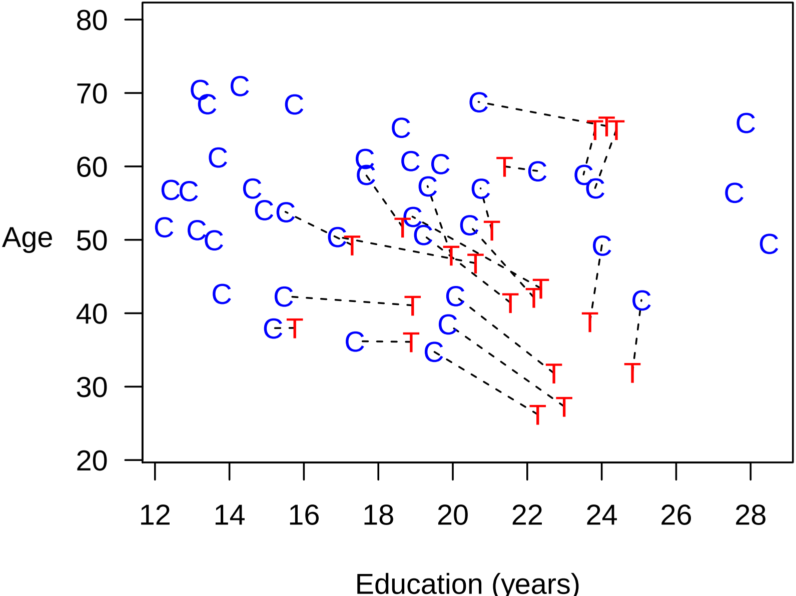
Mahalanobis Distance Matching



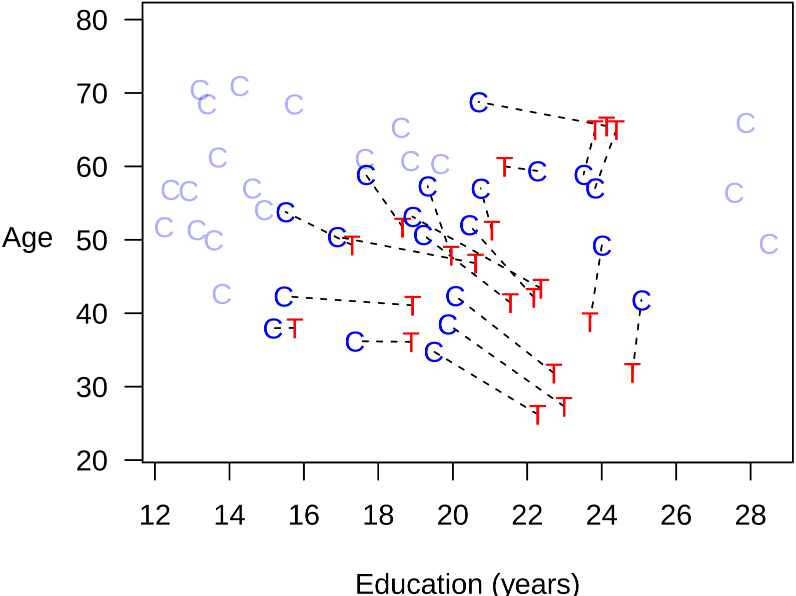
Mahalanobis Distance Matching



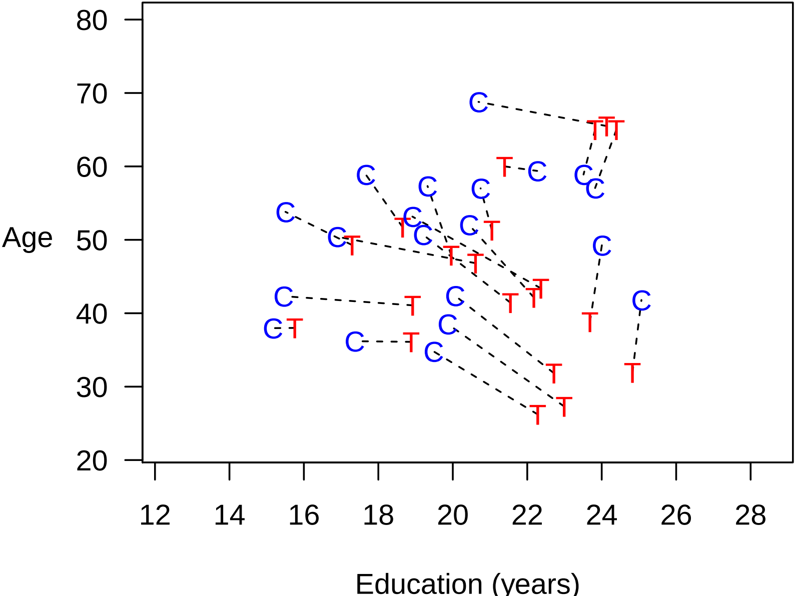
Mahalanobis Distance Matching



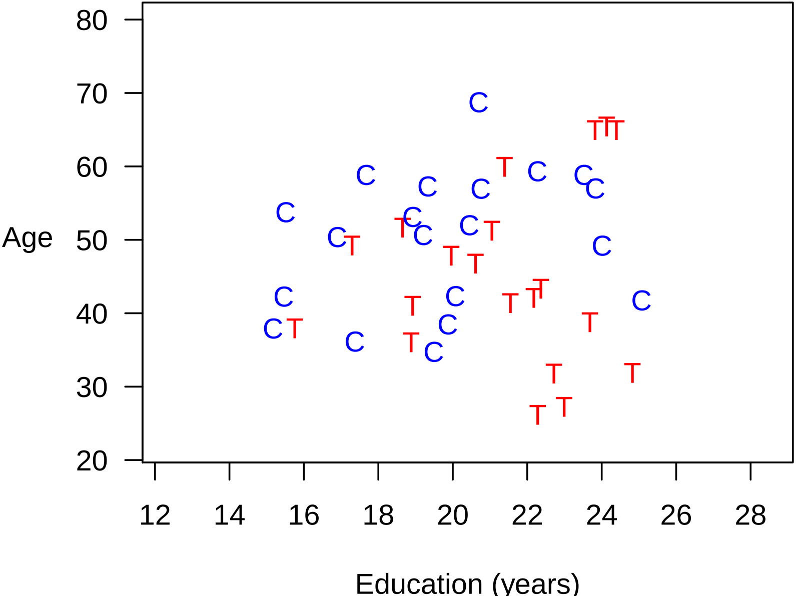
Mahalanobis Distance Matching



Mahalanobis Distance Matching

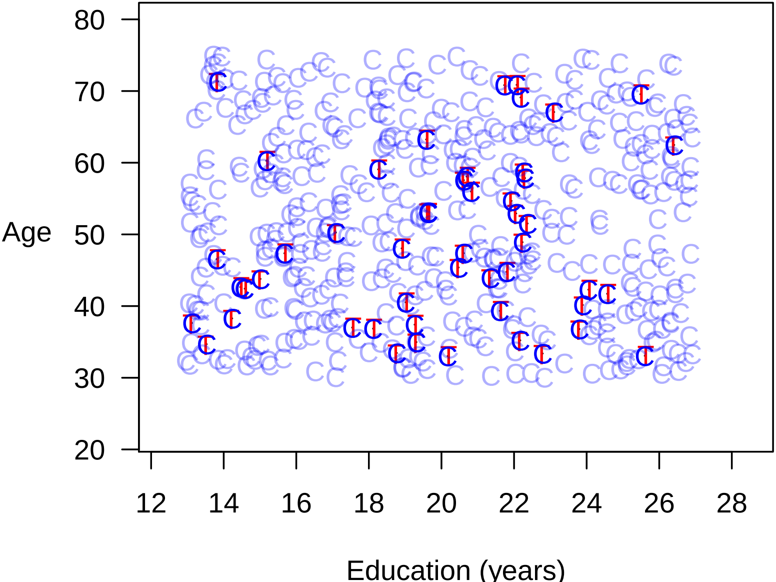


Mahalanobis Distance Matching

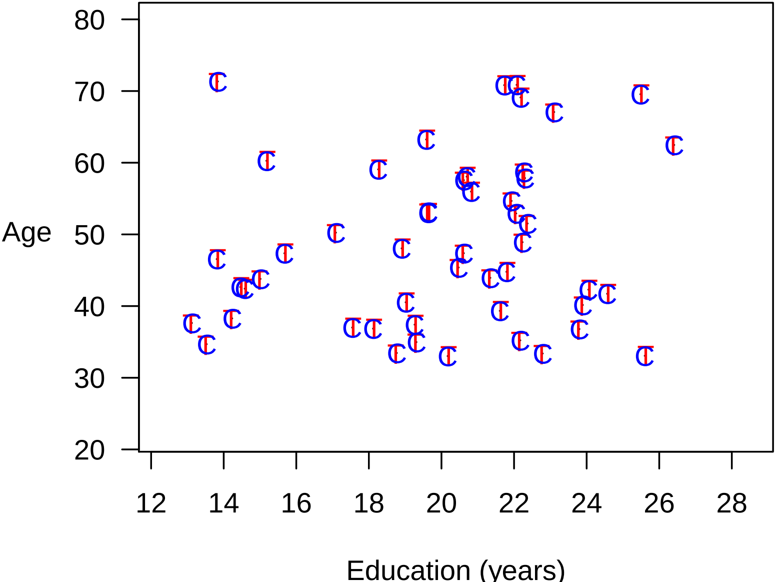


Best Case: Mahalanobis Distance Matching

Best Case: Mahalanobis Distance Matching



Best Case: 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. 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. 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. 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. 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. 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. 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. Estimation Difference in means or a model

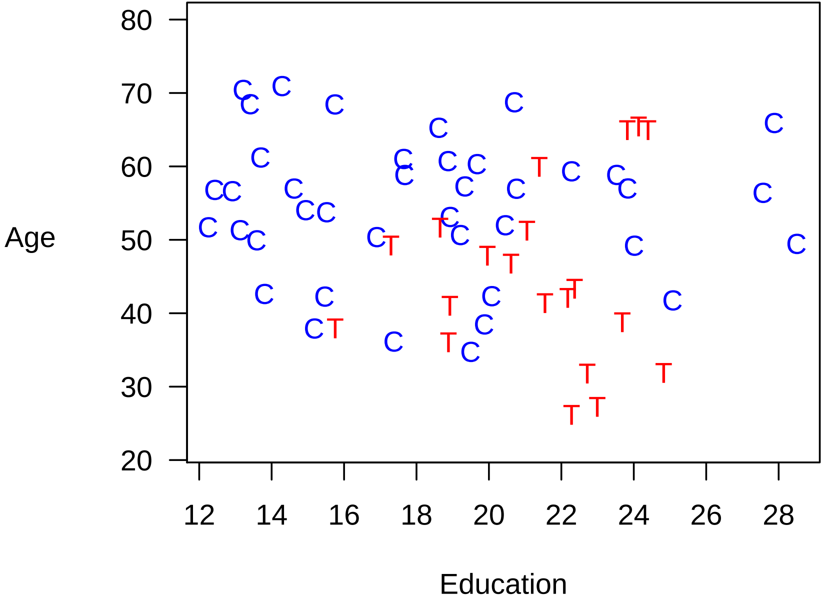
Method 2: Coarsened Exact Matching

(Approximates Fully Blocked Experiment)

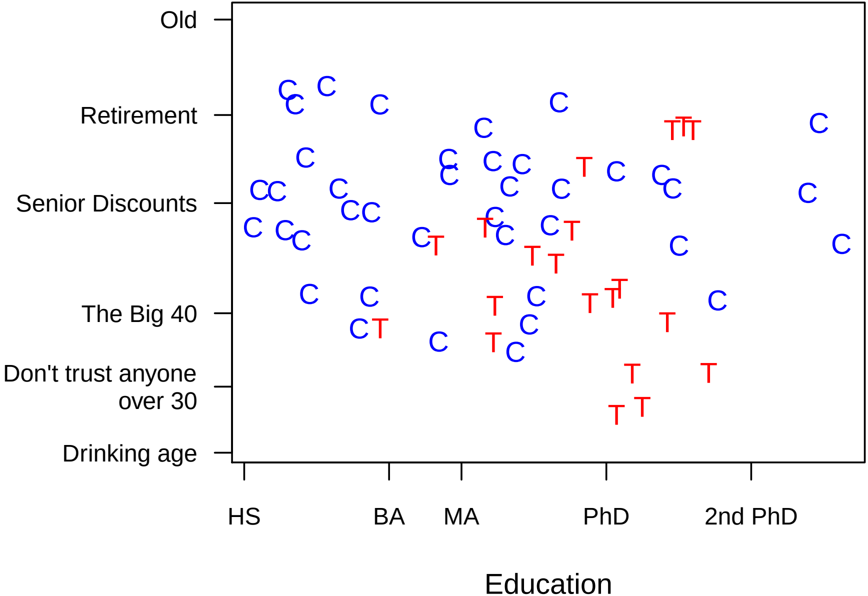
1. Preprocess (Matching)
   - Temporarily coarsen X as much as you're willing
     - e.g., Education (grade school, high school, college, graduate)
   - Apply exact matching to the coarsened X, $C(X)$
     - Sort observations into strata, each with unique values of $C(X)$
     - Prune any stratum with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

Coarsened Exact Matching

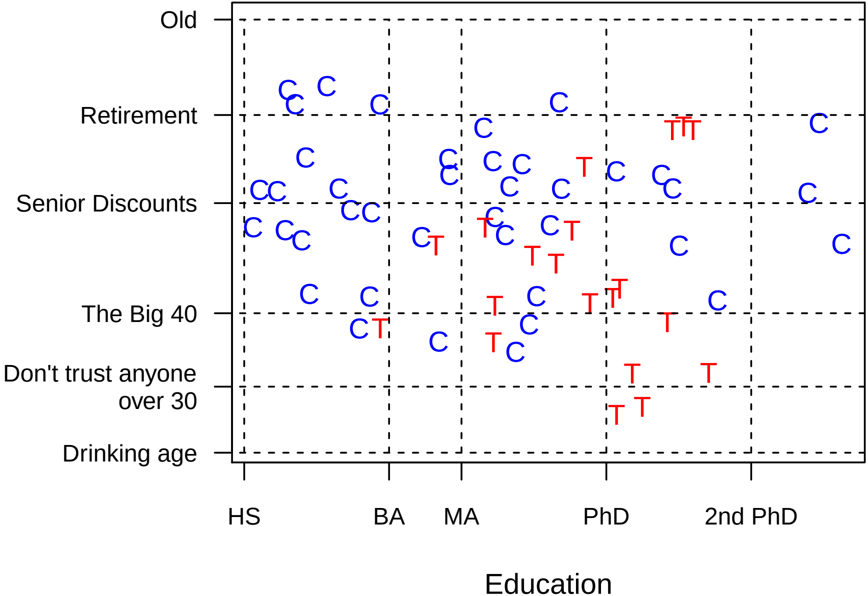
Coarsened Exact Matching



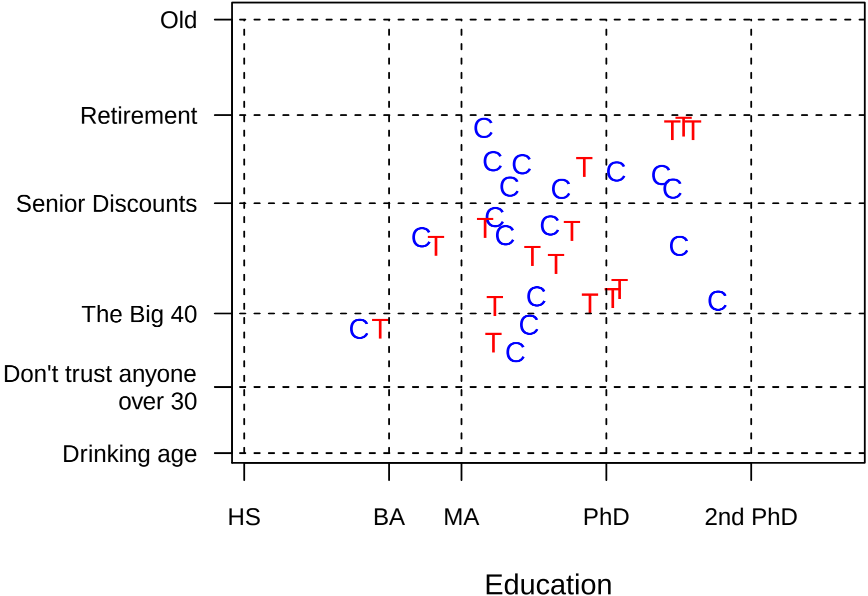
Coarsened Exact Matching



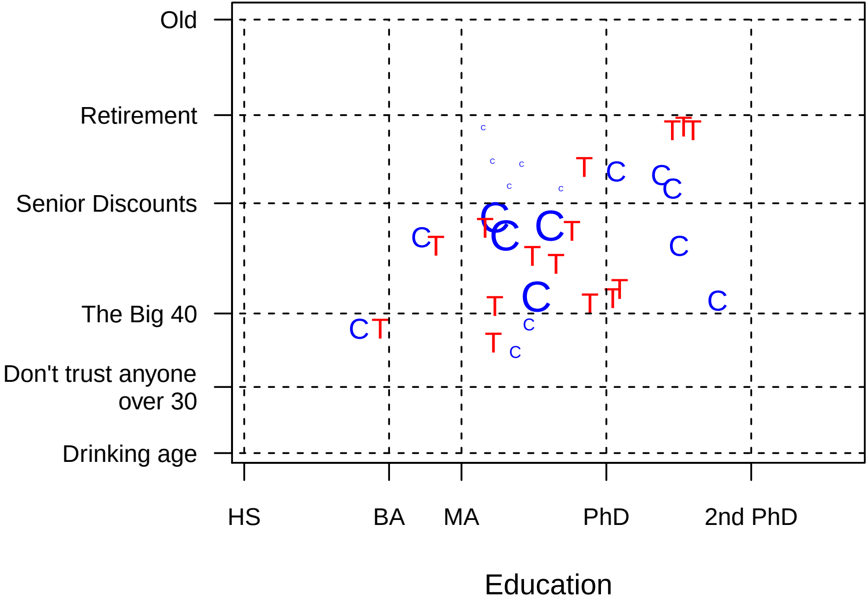
Coarsened Exact Matching



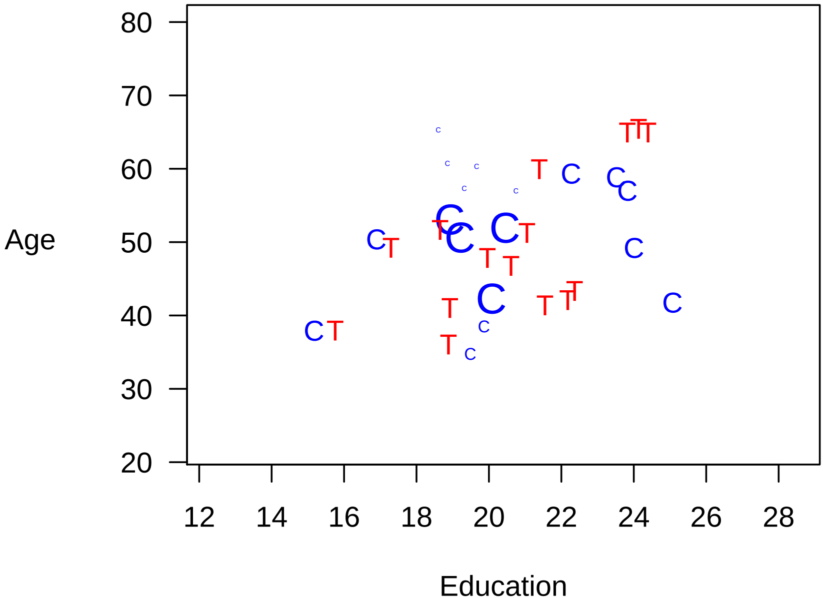
Coarsened Exact Matching



Coarsened Exact Matching

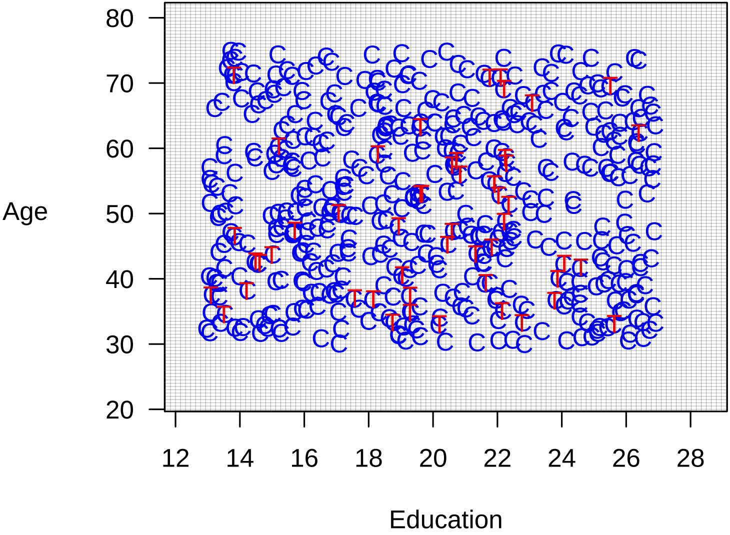


Coarsened Exact Matching

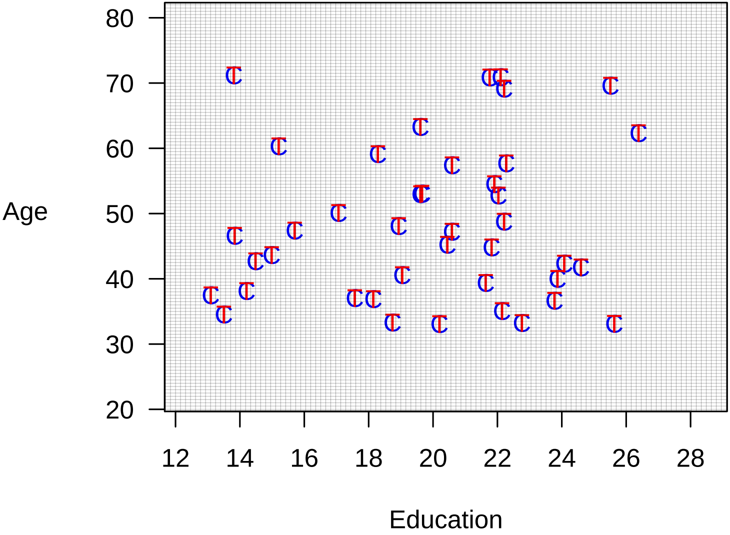


Best Case: Coarsened Exact Matching

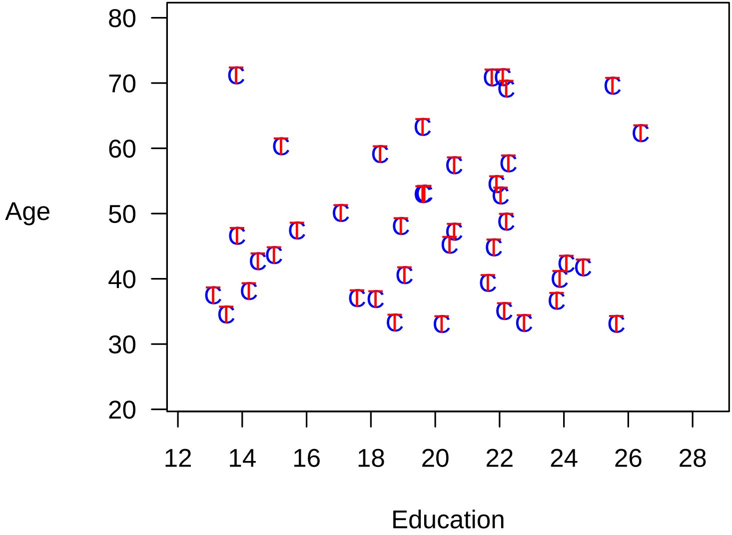
Best Case: Coarsened Exact Matching



Best Case: Coarsened Exact Matching



Best Case: 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. 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. 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 $\pi_i \equiv \Pr(T_i = 1|X) = \frac{1}{1+e^{-X_i\beta}}$
   - $\text{Distance}(X_c, X_t) = |\pi_c - \pi_t|$

2. 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 $\pi_i \equiv \Pr(T_i = 1|X) = \frac{1}{1+e^{-X_i\beta}}$
    - Distance$(X_c, X_t) = |\pi_c - \pi_t|$
    - Match each treated unit to the nearest control unit

2. 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 $\pi_i \equiv \Pr(T_i = 1|X) = \frac{1}{1+e^{-X_i\beta}}$
    - $\text{Distance}(X_c, X_t) = |\pi_c - \pi_t|$
    - Match each treated unit to the nearest control unit
    - Control units: not reused; pruned if unused

2. 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 $\pi_i \equiv \Pr(T_i = 1|X) = \frac{1}{1+e^{-X_i\beta}}$
    - $\text{Distance}(X_c, X_t) = |\pi_c - \pi_t|$
    - Match each treated unit to the nearest control unit
    - Control units: not reused; pruned if unused
    - Prune matches if Distance>*caliper*

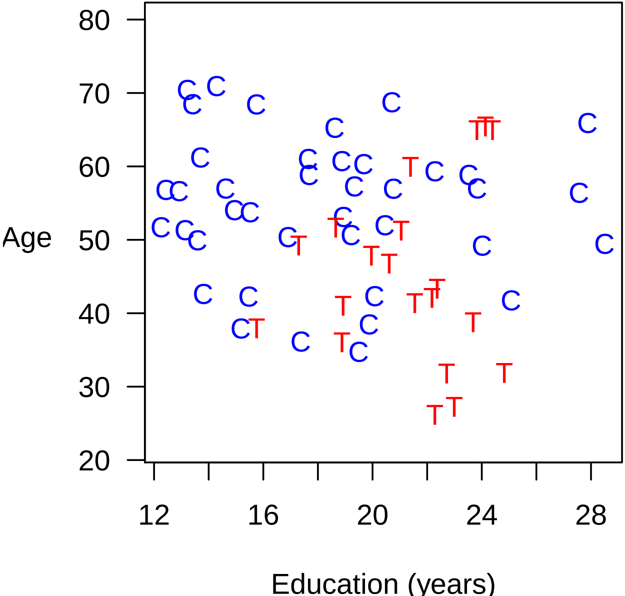
2. Estimation Difference in means or a model

Method 3: Propensity Score Matching

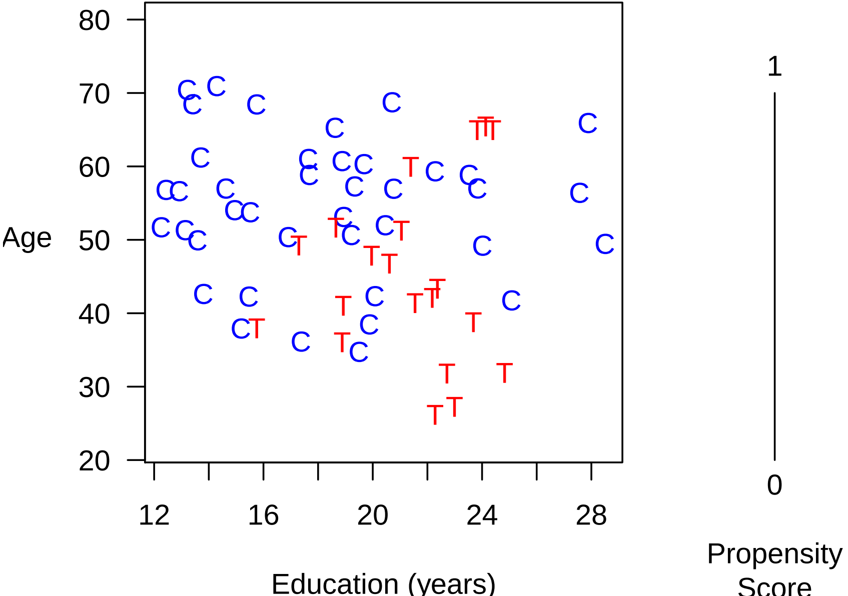
(Approximates Completely Randomized Experiment)

1. Preprocess (Matching)
   - Reduce k elements of X to scalar $\pi_i \equiv \Pr(T_i = 1|X) = \frac{1}{1+e^{-X_i\beta}}$
   - $\text{Distance}(X_c, X_t) = |\pi_c - \pi_t|$
   - Match each treated unit to the nearest control unit
   - Control units: not reused; pruned if unused
   - Prune matches if Distance>*caliper*
   - (Many adjustments available to this basic method)
2. Estimation Difference in means or a model

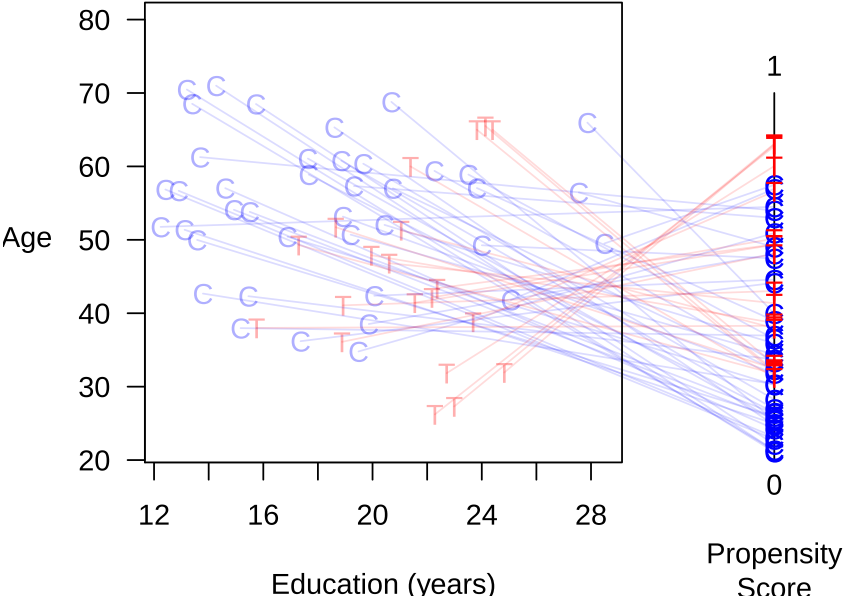
Propensity Score Matching



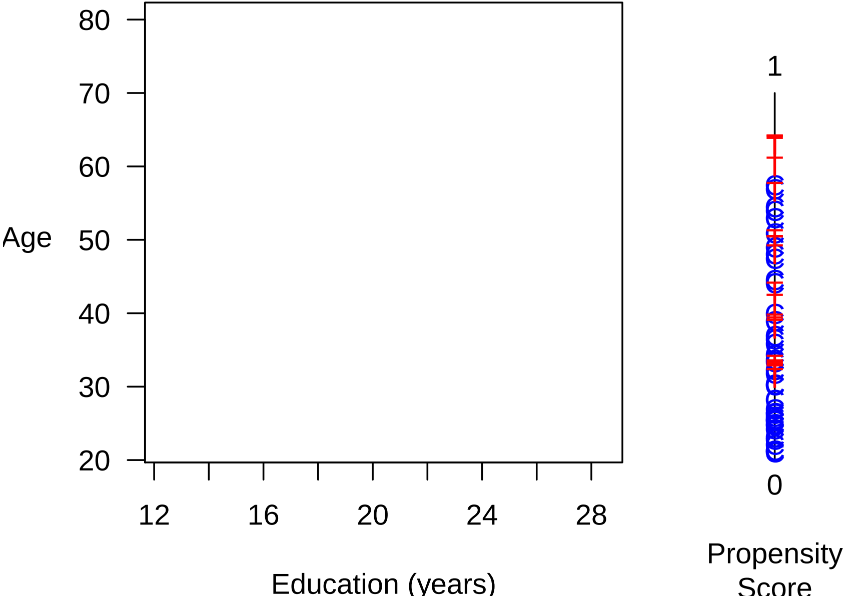
Propensity Score Matching



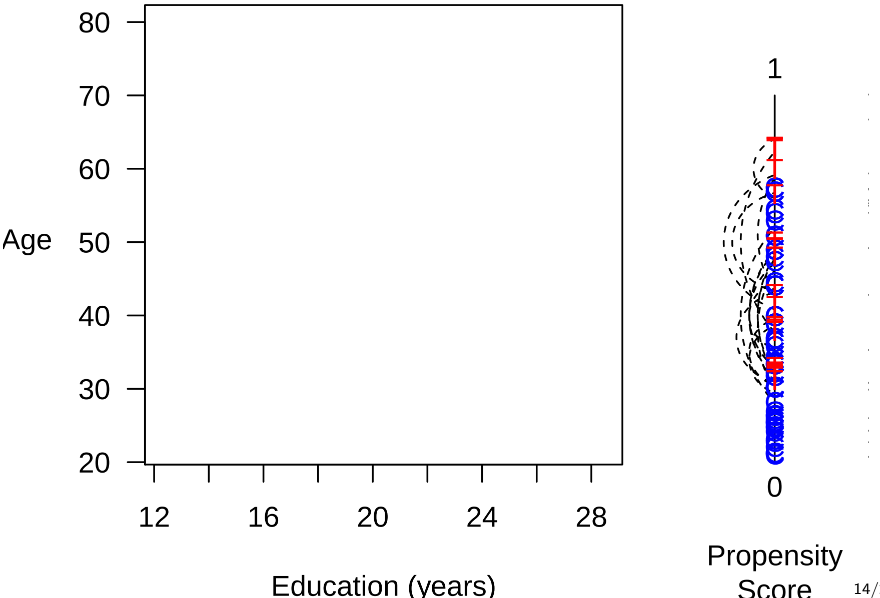
Propensity Score Matching



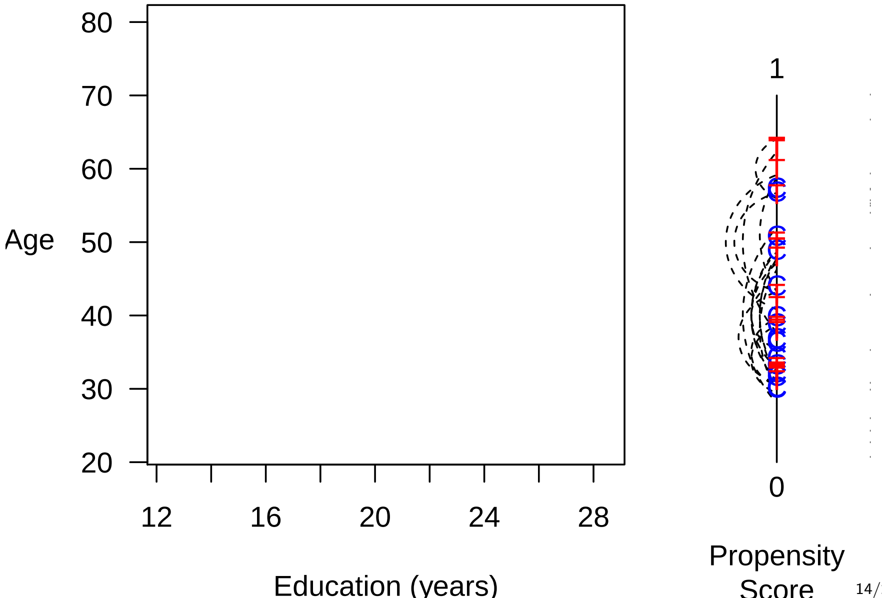
Propensity Score Matching



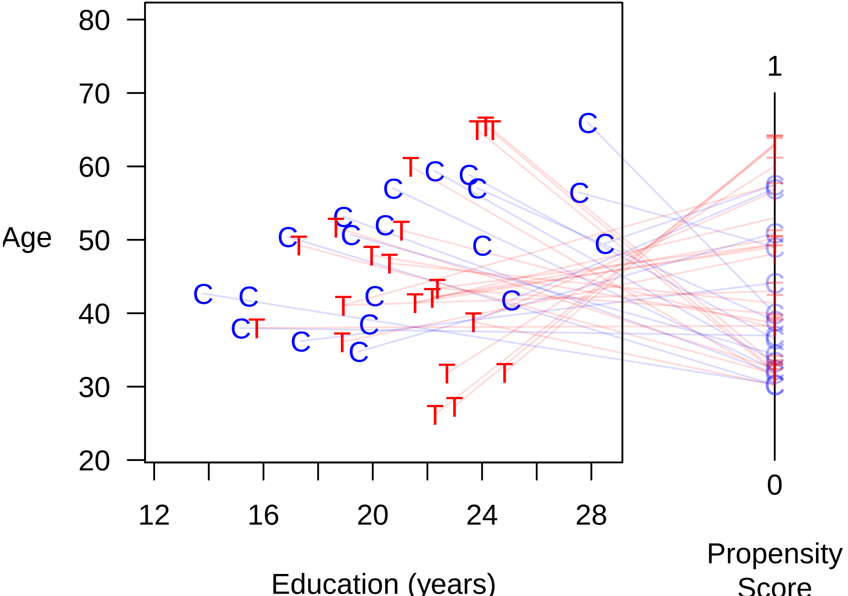
Propensity Score Matching



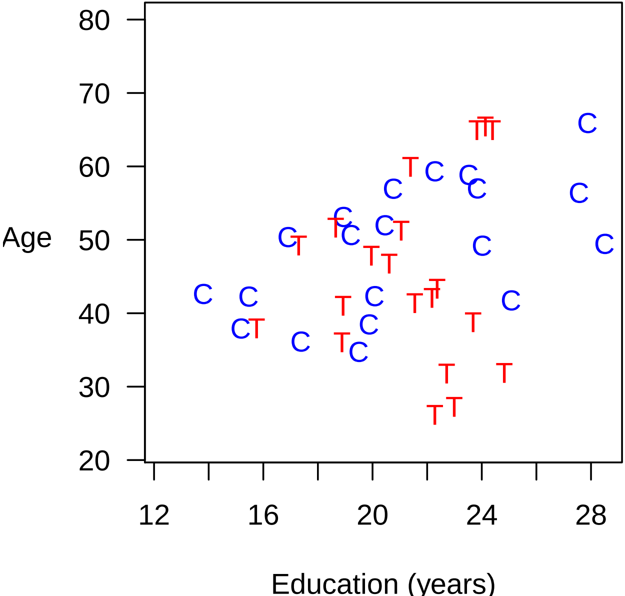
Propensity Score Matching



Propensity Score Matching

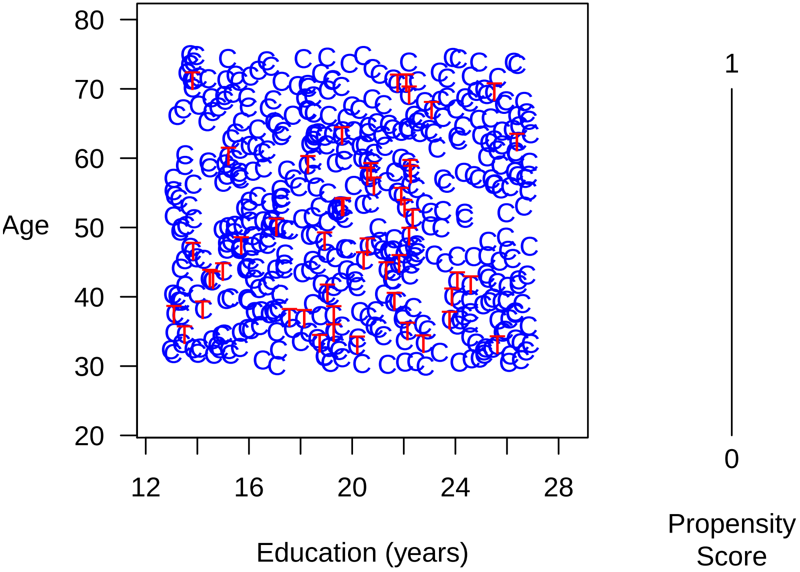


Propensity Score Matching

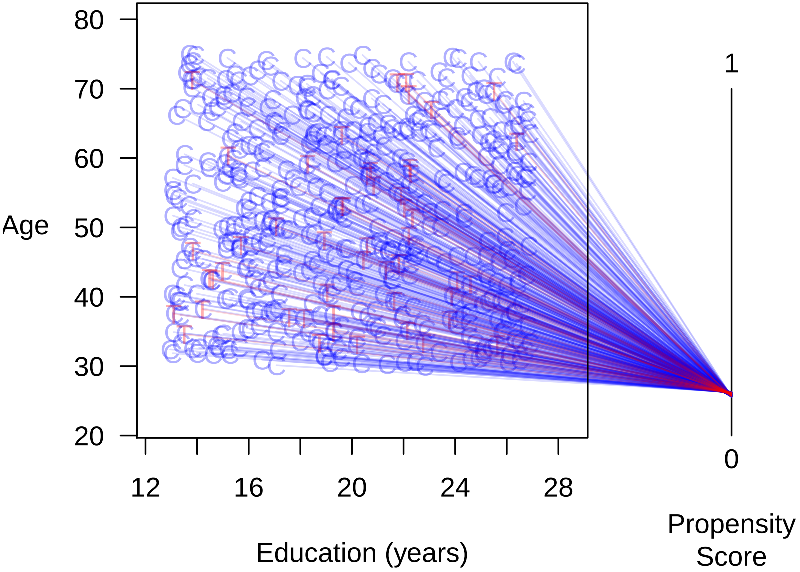


Best Case: Propensity Score Matching

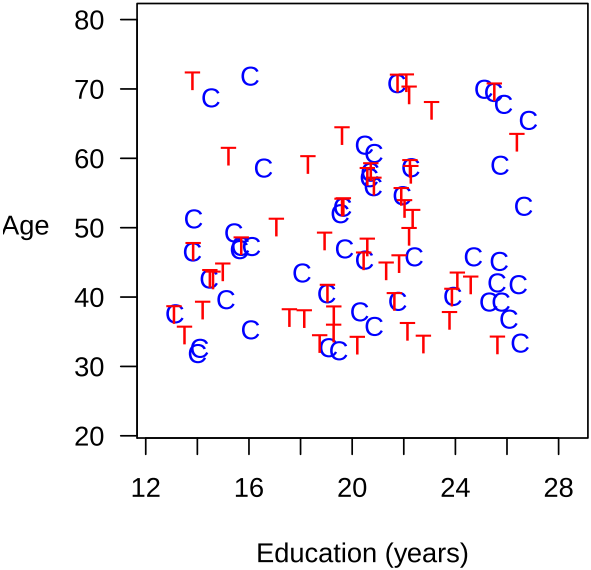
Best Case: Propensity Score Matching



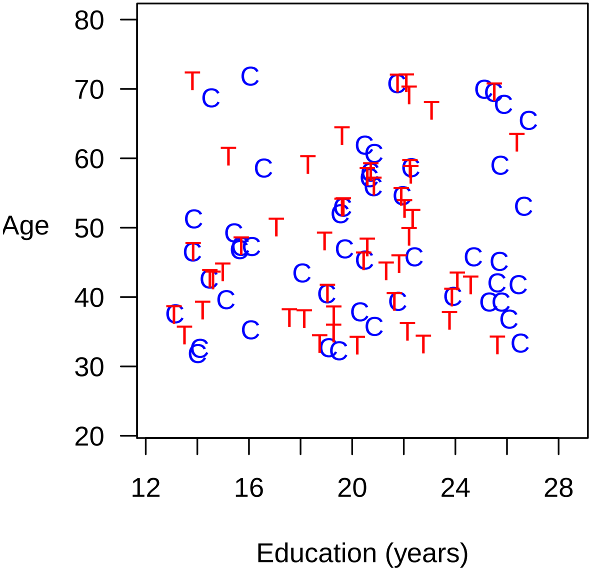
Best Case: Propensity Score Matching



Best Case: Propensity Score Matching



Best Case: Propensity Score Matching is Suboptimal



Random Pruning Increases Imbalance

Random Pruning Increases Imbalance

Deleting data only helps if you're careful!

Random Pruning Increases Imbalance

Deleting data only helps if you're careful!

- "Random pruning": pruning process is independent of X

Random Pruning Increases Imbalance

Deleting data only helps if you're careful!

- "Random pruning": pruning process is independent of X
- Discrete example

Random Pruning Increases Imbalance

Deleting data only helps if you're careful!

- "Random pruning": pruning process is independent of X
- Discrete example
  - Sex-balanced dataset: treateds M_t, F_t, controls M_c, F_c

Random Pruning Increases Imbalance

Deleting data only helps if you're careful!

- "Random pruning": pruning process is independent of X
- Discrete example
  - Sex-balanced dataset: treateds M_t, F_t, controls M_c, F_c
  - Randomly prune 1 treated & 1 control $\rightsquigarrow$ 4 possible datasets:
    2 balanced $\{M_t, M_c\}$, $\{F_t, F_c\}$
    2 imbalanced $\{M_t, F_c\}$, $\{F_t, M_c\}$

Random Pruning Increases Imbalance

Deleting data only helps if you're careful!

- "Random pruning": pruning process is independent of X
- Discrete example
  - Sex-balanced dataset: treateds M_t, F_t, controls M_c, F_c
  - Randomly prune 1 treated & 1 control $\rightsquigarrow$ 4 possible datasets:
    2 balanced $\{M_t, M_c\}$, $\{F_t, F_c\}$
    2 imbalanced $\{M_t, F_c\}$, $\{F_t, M_c\}$
  - $\implies$ random pruning increases imbalance

Random Pruning Increases Imbalance

Deleting data only helps if you're careful!

- "Random pruning": pruning process is independent of X
- Discrete example
  - Sex-balanced dataset: treateds M_t, F_t, controls M_c, F_c
  - Randomly prune 1 treated & 1 control $\rightsquigarrow$ 4 possible datasets:
    2 balanced $\{M_t, M_c\}$, $\{F_t, F_c\}$
    2 imbalanced $\{M_t, F_c\}$, $\{F_t, M_c\}$
  - $\Longrightarrow$ random pruning increases imbalance
- Continuous example

Random Pruning Increases Imbalance

Deleting data only helps if you're careful!

- "Random pruning": pruning process is independent of X
- Discrete example
  - Sex-balanced dataset: treateds M_t, F_t, controls M_c, F_c
  - Randomly prune 1 treated & 1 control $\rightsquigarrow$ 4 possible datasets:
    2 balanced $\{M_t, M_c\}$, $\{F_t, F_c\}$
    2 imbalanced $\{M_t, F_c\}$, $\{F_t, M_c\}$
  - $\Longrightarrow$ random pruning increases imbalance
- Continuous example
  - Dataset: $T \in \{0, 1\}$ randomly assigned; X any fixed variable; with n units

Random Pruning Increases Imbalance

Deleting data only helps if you're careful!

- "Random pruning": pruning process is independent of X
- Discrete example
  - Sex-balanced dataset: treateds M_t, F_t, controls M_c, F_c
  - Randomly prune 1 treated & 1 control $\rightsquigarrow$ 4 possible datasets:
    2 balanced $\{M_t, M_c\}$, $\{F_t, F_c\}$
    2 imbalanced $\{M_t, F_c\}$, $\{F_t, M_c\}$
  - $\Longrightarrow$ random pruning increases imbalance
- Continuous example
  - Dataset: $T \in \{0, 1\}$ randomly assigned; X any fixed variable; with n units
  - Measure of imbalance: squared difference in means d^2, where $d = \bar{X}_t - \bar{X}_c$

Random Pruning Increases Imbalance

Deleting data only helps if you're careful!

- "Random pruning": pruning process is independent of X
- Discrete example
  - Sex-balanced dataset: treateds M_t, F_t, controls M_c, F_c
  - Randomly prune 1 treated & 1 control $\rightsquigarrow$ 4 possible datasets:
    2 balanced $\{M_t, M_c\}$, $\{F_t, F_c\}$
    2 imbalanced $\{M_t, F_c\}$, $\{F_t, M_c\}$
  - $\implies$ random pruning increases imbalance
- Continuous example
  - Dataset: $T \in \{0, 1\}$ randomly assigned; X any fixed variable; with n units
  - Measure of imbalance: squared difference in means d^2, where $d = \bar{X}_t - \bar{X}_c$
  - $E(d^2) = V(d) \propto 1/n$ (note: $E(d) = 0$)

Random Pruning Increases Imbalance

Deleting data only helps if you're careful!

- "Random pruning": pruning process is independent of X
- Discrete example
  - Sex-balanced dataset: treateds M_t, F_t, controls M_c, F_c
  - Randomly prune 1 treated & 1 control $\rightsquigarrow$ 4 possible datasets:
    2 balanced $\{M_t, M_c\}$, $\{F_t, F_c\}$
    2 imbalanced $\{M_t, F_c\}$, $\{F_t, M_c\}$
  - $\Longrightarrow$ random pruning increases imbalance
- Continuous example
  - Dataset: $T \in \{0, 1\}$ randomly assigned; X any fixed variable; with n units
  - Measure of imbalance: squared difference in means d^2, where $d = \bar{X}_t - \bar{X}_c$
  - $E(d^2) = V(d) \propto 1/n$ (note: $E(d) = 0$)
  - Random pruning $\rightsquigarrow$ n declines $\rightsquigarrow$ $E(d^2)$ increases

Random Pruning Increases Imbalance

Deleting data only helps if you're careful!

- "Random pruning": pruning process is independent of X
- Discrete example
  - Sex-balanced dataset: treateds M_t, F_t, controls M_c, F_c
  - Randomly prune 1 treated & 1 control $\rightsquigarrow$ 4 possible datasets:
    2 balanced $\{M_t, M_c\}$, $\{F_t, F_c\}$
    2 imbalanced $\{M_t, F_c\}$, $\{F_t, M_c\}$
  - $\Longrightarrow$ random pruning increases imbalance
- Continuous example
  - Dataset: $T \in \{0, 1\}$ randomly assigned; X any fixed variable; with n units
  - Measure of imbalance: squared difference in means d^2, where $d = \bar{X}_t - \bar{X}_c$
  - $E(d^2) = V(d) \propto 1/n$ (note: $E(d) = 0$)
  - Random pruning $\rightsquigarrow$ n declines $\rightsquigarrow$ $E(d^2)$ increases
  - $\Longrightarrow$ random pruning increases imbalance

Random Pruning Increases Imbalance

Deleting data only helps if you're careful!

- "Random pruning": pruning process is independent of X
- Discrete example
  - Sex-balanced dataset: treateds M_t, F_t, controls M_c, F_c
  - Randomly prune 1 treated & 1 control $\rightsquigarrow$ 4 possible datasets:
    2 balanced $\{M_t, M_c\}$, $\{F_t, F_c\}$
    2 imbalanced $\{M_t, F_c\}$, $\{F_t, M_c\}$
  - $\Longrightarrow$ random pruning increases imbalance
- Continuous example
  - Dataset: $T \in \{0, 1\}$ randomly assigned; X any fixed variable; with n units
  - Measure of imbalance: squared difference in means d^2, where $d = \bar{X}_t - \bar{X}_c$
  - $E(d^2) = V(d) \propto 1/n$ (note: $E(d) = 0$)
  - Random pruning $\rightsquigarrow$ n declines $\rightsquigarrow$ $E(d^2)$ increases
  - $\Longrightarrow$ random pruning increases imbalance
- Result is completely general (see math in the paper)

PSM's Statistical Properties

PSM's Statistical Properties

1. Low Standards: Sometimes helps, never optimizes

PSM's Statistical Properties

1. Low Standards: Sometimes helps, never optimizes
   - *Efficient* relative to complete randomization, but

PSM's Statistical Properties

1. Low Standards: Sometimes helps, never optimizes
    - *Efficient* relative to complete randomization, but
    - *Inefficient* relative to (the more powerful) full blocking

PSM's Statistical Properties

1. Low Standards: Sometimes helps, never optimizes
   - *Efficient* relative to complete randomization, but
   - *Inefficient* relative to (the more powerful) full blocking
   - Other methods dominate:

PSM's Statistical Properties

1. Low Standards: Sometimes helps, never optimizes
   - *Efficient* relative to complete randomization, but
   - *Inefficient* relative to (the more powerful) full blocking
   - Other methods dominate:

     $X_c = X_t \implies \pi_c = \pi_t$

PSM's Statistical Properties

1. Low Standards: Sometimes helps, never optimizes
   - *Efficient* relative to complete randomization, but
   - *Inefficient* relative to (the more powerful) full blocking
   - Other methods dominate:

     $X_c = X_t \implies \pi_c = \pi_t$ but

     $\pi_c = \pi_t \not\Longrightarrow X_c = X_t$

PSM's Statistical Properties

1. Low Standards: Sometimes helps, never optimizes
   - *Efficient* relative to complete randomization, but
   - *Inefficient* relative to (the more powerful) full blocking
   - Other methods dominate:

     $X_c = X_t \implies \pi_c = \pi_t$ but

     $\pi_c = \pi_t \not\Longrightarrow X_c = X_t$
2. The PSM Paradox: When you do "better," you do worse

PSM's Statistical Properties

1. Low Standards: Sometimes helps, never optimizes
   - *Efficient* relative to complete randomization, but
   - *Inefficient* relative to (the more powerful) full blocking
   - Other methods dominate:
     $X_c = X_t \implies \pi_c = \pi_t$ but
     $\pi_c = \pi_t \not\Longrightarrow X_c = X_t$
2. The PSM Paradox: When you do "better," you do worse
   - When PSM approximates complete randomization (to begin with or, after some pruning)

PSM's Statistical Properties

1. Low Standards: Sometimes helps, never optimizes
   - *Efficient* relative to complete randomization, but
   - *Inefficient* relative to (the more powerful) full blocking
   - Other methods dominate:

     $X_c = X_t \implies \pi_c = \pi_t$ but

     $\pi_c = \pi_t \not\Longrightarrow X_c = X_t$
2. The PSM Paradox: When you do "better," you do worse
   - When PSM approximates complete randomization (to begin with or, after some pruning) $\rightsquigarrow$ all $\hat{\pi} \approx 0.5$ (or constant within strata)

PSM's Statistical Properties

1. <u>Low Standards:</u> Sometimes helps, never optimizes
   - *Efficient* relative to complete randomization, but
   - *Inefficient* relative to (the more powerful) full blocking
   - Other methods dominate:
     $X_c = X_t \implies \pi_c = \pi_t$ but
     $\pi_c = \pi_t \not\Longrightarrow X_c = X_t$
2. <u>The PSM Paradox:</u> When you do "better," you do worse
   - When PSM approximates complete randomization (to begin with or, after some pruning) $\rightsquigarrow$ all $\hat{\pi} \approx 0.5$ (or constant within strata) $\rightsquigarrow$ pruning at random

PSM's Statistical Properties

1. Low Standards: Sometimes helps, never optimizes
   - *Efficient* relative to complete randomization, but
   - *Inefficient* relative to (the more powerful) full blocking
   - Other methods dominate:
     $X_c = X_t \implies \pi_c = \pi_t$ but
     $\pi_c = \pi_t \not\Longrightarrow X_c = X_t$
2. The PSM Paradox: When you do "better," you do worse
   - When PSM approximates complete randomization (to begin with or, after some pruning) $\rightsquigarrow$ all $\hat{\pi} \approx 0.5$ (or constant within strata) $\rightsquigarrow$ pruning at random $\rightsquigarrow$ Imbalance

PSM's Statistical Properties

1. Low Standards: Sometimes helps, never optimizes
   - *Efficient* relative to complete randomization, but
   - *Inefficient* relative to (the more powerful) full blocking
   - Other methods dominate:
     $X_c = X_t \implies \pi_c = \pi_t$ but
     $\pi_c = \pi_t \not\Longrightarrow X_c = X_t$
2. The PSM Paradox: When you do "better," you do worse
   - When PSM approximates complete randomization (to begin with or, after some pruning) $\rightsquigarrow$ all $\hat{\pi} \approx 0.5$ (or constant within strata) $\rightsquigarrow$ pruning at random $\rightsquigarrow$ Imbalance $\rightsquigarrow$ Inefficency

PSM's Statistical Properties

1. Low Standards: Sometimes helps, never optimizes
   - *Efficient* relative to complete randomization, but
   - *Inefficient* relative to (the more powerful) full blocking
   - Other methods dominate:
     $X_c = X_t \implies \pi_c = \pi_t$ but
     $\pi_c = \pi_t \not\Longrightarrow X_c = X_t$
2. The PSM Paradox: When you do "better," you do worse
   - When PSM approximates complete randomization (to begin with or, after some pruning) $\rightsquigarrow$ all $\hat{\pi} \approx 0.5$ (or constant within strata) $\rightsquigarrow$ pruning at random $\rightsquigarrow$ Imbalance $\rightsquigarrow$ Inefficency $\rightsquigarrow$ Model dependence

PSM's Statistical Properties

1. Low Standards: Sometimes helps, never optimizes
   - *Efficient* relative to complete randomization, but
   - *Inefficient* relative to (the more powerful) full blocking
   - Other methods dominate:
     $X_c = X_t \implies \pi_c = \pi_t$ but
     $\pi_c = \pi_t \not\Longrightarrow X_c = X_t$
2. The PSM Paradox: When you do "better," you do worse
   - When PSM approximates complete randomization (to begin with or, after some pruning) $\rightsquigarrow$ all $\hat{\pi} \approx 0.5$ (or constant within strata) $\rightsquigarrow$ pruning at random $\rightsquigarrow$ Imbalance $\rightsquigarrow$ Inefficency $\rightsquigarrow$ Model dependence $\rightsquigarrow$ Bias

PSM's Statistical Properties

1. Low Standards: Sometimes helps, never optimizes
   - *Efficient* relative to complete randomization, but
   - *Inefficient* relative to (the more powerful) full blocking
   - Other methods dominate:
     $X_c = X_t \implies \pi_c = \pi_t$ but
     $\pi_c = \pi_t \not\Longrightarrow X_c = X_t$
2. The PSM Paradox: When you do "better," you do worse
   - When PSM approximates complete randomization (to begin with or, after some pruning) $\rightsquigarrow$ all $\hat{\pi} \approx 0.5$ (or constant within strata) $\rightsquigarrow$ pruning at random $\rightsquigarrow$ Imbalance $\rightsquigarrow$ Inefficency $\rightsquigarrow$ Model dependence $\rightsquigarrow$ Bias
   - If the data have no good matches, the paradox won't be a problem but you're cooked anyway.

PSM's Statistical Properties

1. Low Standards: Sometimes helps, never optimizes
   - *Efficient* relative to complete randomization, but
   - *Inefficient* relative to (the more powerful) full blocking
   - Other methods dominate:
     $X_c = X_t \implies \pi_c = \pi_t$ but
     $\pi_c = \pi_t \not\Longrightarrow X_c = X_t$
2. The PSM Paradox: When you do "better," you do worse
   - When PSM approximates complete randomization (to begin with or, after some pruning) $\rightsquigarrow$ all $\hat{\pi} \approx 0.5$ (or constant within strata) $\rightsquigarrow$ pruning at random $\rightsquigarrow$ Imbalance $\rightsquigarrow$ Inefficency $\rightsquigarrow$ Model dependence $\rightsquigarrow$ Bias
   - If the data have no good matches, the paradox won't be a problem but you're cooked anyway.
   - Doesn't PSM solve the curse of dimensionality problem?

PSM's Statistical Properties

1. Low Standards: Sometimes helps, never optimizes
   - *Efficient* relative to complete randomization, but
   - *Inefficient* relative to (the more powerful) full blocking
   - Other methods dominate:
     $X_c = X_t \implies \pi_c = \pi_t$ but
     $\pi_c = \pi_t \not\Longrightarrow X_c = X_t$
2. The PSM Paradox: When you do "better," you do worse
   - When PSM approximates complete randomization (to begin with or, after some pruning) $\rightsquigarrow$ all $\hat{\pi} \approx 0.5$ (or constant within strata) $\rightsquigarrow$ pruning at random $\rightsquigarrow$ Imbalance $\rightsquigarrow$ Inefficency $\rightsquigarrow$ Model dependence $\rightsquigarrow$ Bias
   - If the data have no good matches, the paradox won't be a problem but you're cooked anyway.
   - Doesn't PSM solve the curse of dimensionality problem? Nope.

PSM's Statistical Properties

1. Low Standards: Sometimes helps, never optimizes
   - *Efficient* relative to complete randomization, but
   - *Inefficient* relative to (the more powerful) full blocking
   - Other methods dominate:
     $X_c = X_t \implies \pi_c = \pi_t$ but
     $\pi_c = \pi_t \not\Longrightarrow X_c = X_t$
2. The PSM Paradox: When you do "better," you do worse
   - When PSM approximates complete randomization (to begin with or, after some pruning) $\rightsquigarrow$ all $\hat{\pi} \approx 0.5$ (or constant within strata) $\rightsquigarrow$ pruning at random $\rightsquigarrow$ Imbalance $\rightsquigarrow$ Inefficency $\rightsquigarrow$ Model dependence $\rightsquigarrow$ Bias
   - If the data have no good matches, the paradox won't be a problem but you're cooked anyway.
   - Doesn't PSM solve the curse of dimensionality problem? Nope. The PSM Paradox gets worse with more covariates

PSM's Statistical Properties

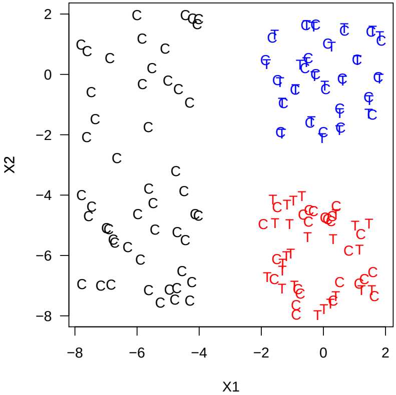
1. Low Standards: Sometimes helps, never optimizes
    - *Efficient* relative to complete randomization, but
    - *Inefficient* relative to (the more powerful) full blocking
    - Other methods dominate:
    $X_c = X_t \implies \pi_c = \pi_t$ but
    $\pi_c = \pi_t \not\Longrightarrow X_c = X_t$
2. The PSM Paradox: When you do "better," you do worse
    - When PSM approximates complete randomization (to begin with or, after some pruning) $\rightsquigarrow$ all $\hat{\pi} \approx 0.5$ (or constant within strata) $\rightsquigarrow$ pruning at random $\rightsquigarrow$ Imbalance $\rightsquigarrow$ Inefficency $\rightsquigarrow$ Model dependence $\rightsquigarrow$ Bias
    - If the data have no good matches, the paradox won't be a problem but you're cooked anyway.
    - Doesn't PSM solve the curse of dimensionality problem? Nope. The PSM Paradox gets worse with more covariates
    - What if I match on a few important covariates and then use PSM?

PSM's Statistical Properties

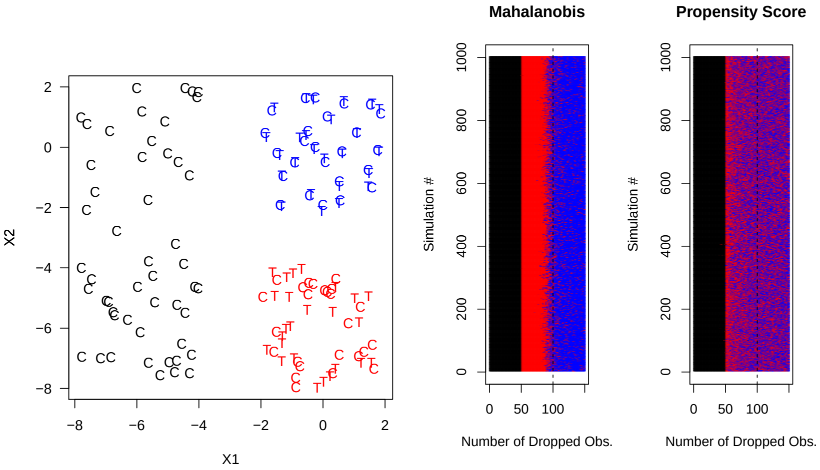
1. Low Standards: Sometimes helps, never optimizes
   - *Efficient* relative to complete randomization, but
   - *Inefficient* relative to (the more powerful) full blocking
   - Other methods dominate:
     $X_c = X_t \implies \pi_c = \pi_t$ but
     $\pi_c = \pi_t \not\Longrightarrow X_c = X_t$
2. The PSM Paradox: When you do "better," you do worse
   - When PSM approximates complete randomization (to begin with or, after some pruning) $\rightsquigarrow$ all $\hat{\pi} \approx 0.5$ (or constant within strata) $\rightsquigarrow$ pruning at random $\rightsquigarrow$ Imbalance $\rightsquigarrow$ Inefficency $\rightsquigarrow$ Model dependence $\rightsquigarrow$ Bias
   - If the data have no good matches, the paradox won't be a problem but you're cooked anyway.
   - Doesn't PSM solve the curse of dimensionality problem? Nope. The PSM Paradox gets worse with more covariates
   - What if I match on a few important covariates and then use PSM? The low standards will be raised some, but the PSM Paradox will kick in earlier

PSM is Blind Where Other Methods Can See

PSM is Blind Where Other Methods Can See

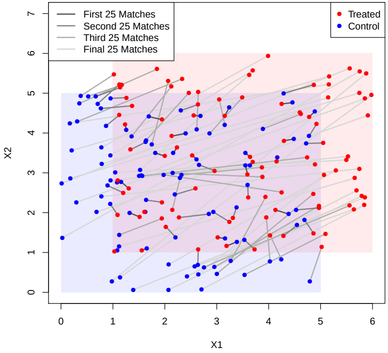


PSM is Blind Where Other Methods Can See

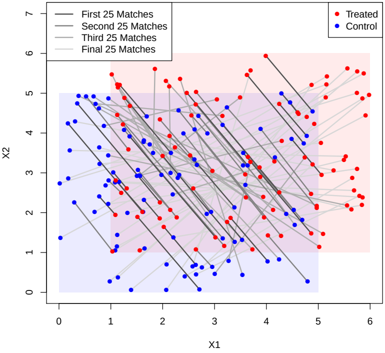


What Does PSM Match?

MDM Matches



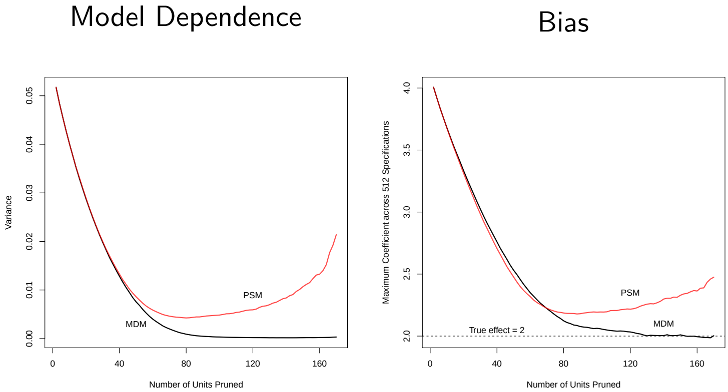
PSM Matches



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

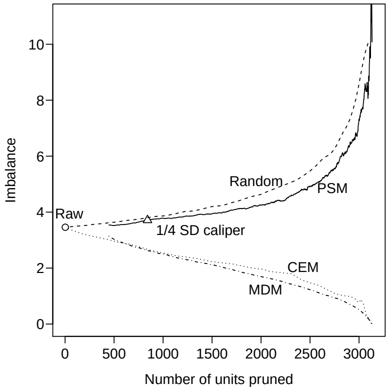


$$Y_i = 2T_i + X_{1i} + X_{2i} + \epsilon_i$$
$$\epsilon_i \sim N(0,1)$$

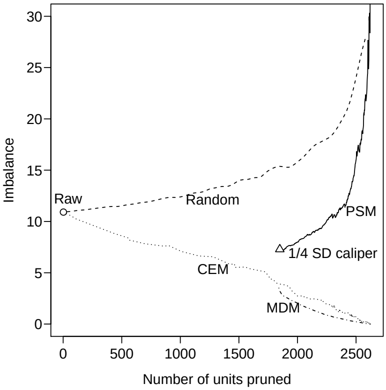
The Propensity Score Paradox in Real Data

The Propensity Score Paradox in Real Data

Finkel et al. (JOP, 2012)

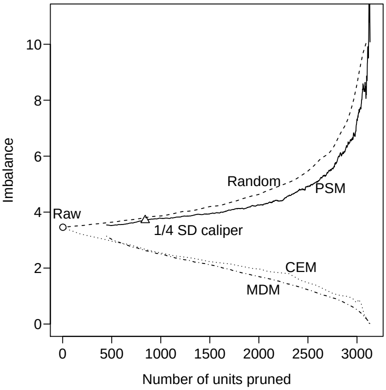


Nielsen et al. (AJPS, 2011)

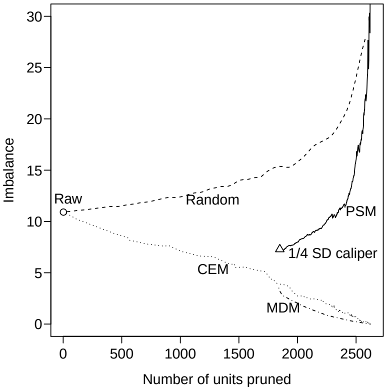


The Propensity Score Paradox in Real Data

Finkel et al. (JOP, 2012)



Nielsen et al. (AJPS, 2011)



Similar pattern for $>$ 20 other real data sets we checked

Conclusions

Conclusions

- Why propensity scores should not be used for matching

Conclusions

- Why propensity scores should not be used for matching
  - Low Standards: sometimes helps, never optimizes

Conclusions

- Why propensity scores should not be used for matching
  - Low Standards: sometimes helps, never optimizes
  - The PSM Paradox: When you do "better," you do worse

Conclusions

- Why propensity scores should not be used for matching
  - Low Standards: sometimes helps, never optimizes
  - The PSM Paradox: When you do "better," you do worse
  - Some mistakes with PSM:

Conclusions

- Why propensity scores should not be used for matching
  - Low Standards: sometimes helps, never optimizes
  - The PSM Paradox: When you do "better," you do worse
  - Some mistakes with PSM: Controlling for irrelevant covariates;

Conclusions

- Why propensity scores should not be used for matching
  - Low Standards: sometimes helps, never optimizes
  - The PSM Paradox: When you do “better,” you do worse
  - Some mistakes with PSM: Controlling for irrelevant covariates; Adjusting experimental data;

Conclusions

- Why propensity scores should not be used for matching
  - Low Standards: sometimes helps, never optimizes
  - The PSM Paradox: When you do "better," you do worse
  - Some mistakes with PSM: Controlling for irrelevant covariates; Adjusting experimental data; Reestimating propensity score after eliminating noncommon support;

Conclusions

- Why propensity scores should not be used for matching
  - Low Standards: sometimes helps, never optimizes
  - The PSM Paradox: When you do "better," you do worse
  - Some mistakes with PSM: Controlling for irrelevant covariates; Adjusting experimental data; Reestimating propensity score after eliminating noncommon support; 1/4 caliper on propensity score;

Conclusions

- Why propensity scores should not be used for matching
  - Low Standards: sometimes helps, never optimizes
  - The PSM Paradox: When you do "better," you do worse
  - Some mistakes with PSM: Controlling for irrelevant covariates; Adjusting experimental data; Reestimating propensity score after eliminating noncommon support; 1/4 caliper on propensity score; Not switching to other methods.

Conclusions

- Why propensity scores should not be used for matching
  - Low Standards: sometimes helps, never optimizes
  - The PSM Paradox: When you do "better," you do worse
  - Some mistakes with PSM: Controlling for irrelevant covariates; Adjusting experimental data; Reestimating propensity score after eliminating noncommon support; 1/4 caliper on propensity score; Not switching to other methods.
- A warning for any matching method:

Conclusions

- Why propensity scores should not be used for matching
  - Low Standards: sometimes helps, never optimizes
  - The PSM Paradox: When you do "better," you do worse
  - Some mistakes with PSM: Controlling for irrelevant covariates; Adjusting experimental data; Reestimating propensity score after eliminating noncommon support; 1/4 caliper on propensity score; Not switching to other methods.
- A warning for any matching method:
  - Pruning discards information; you must overcome this.

Conclusions

- Why propensity scores should not be used for matching
  - Low Standards: sometimes helps, never optimizes
  - The PSM Paradox: When you do "better," you do worse
  - Some mistakes with PSM: Controlling for irrelevant covariates; Adjusting experimental data; Reestimating propensity score after eliminating noncommon support; 1/4 caliper on propensity score; Not switching to other methods.
- A warning for any matching method:
  - Pruning discards information; you must overcome this.
  - Other methods can generate a "paradox" if you prune after approximating full blocking (rare, but possible)

Conclusions

- Why propensity scores should not be used for matching
  - Low Standards: sometimes helps, never optimizes
  - The PSM Paradox: When you do "better," you do worse
  - Some mistakes with PSM: Controlling for irrelevant covariates; Adjusting experimental data; Reestimating propensity score after eliminating noncommon support; 1/4 caliper on propensity score; Not switching to other methods.
- A warning for any matching method:
  - Pruning discards information; you must overcome this.
  - Other methods can generate a "paradox" if you prune after approximating full blocking (rare, but possible)
  - If you're not doing positive good, you may be hurting yourself

Conclusions

- Why propensity scores should not be used for matching
  - Low Standards: sometimes helps, never optimizes
  - The PSM Paradox: When you do "better," you do worse
  - Some mistakes with PSM: Controlling for irrelevant covariates; Adjusting experimental data; Reestimating propensity score after eliminating noncommon support; 1/4 caliper on propensity score; Not switching to other methods.
- A warning for any matching method:
  - Pruning discards information; you must overcome this.
  - Other methods can generate a "paradox" if you prune after approximating full blocking (rare, but possible)
  - If you're not doing positive good, you may be hurting yourself
- Matching methods still highly recommended; choose one with higher standards

For more information, papers, & software

GaryKing.org

www.mit.edu/∼rnielsen