Matching

Ian Lundberg
Soc 114

Learning goals for today

At the end of class, you will be able to:

1. Use matching methods for causal effects
    - Select a matching algorithm
    - Define a distance metric for multivariate matching
    - Evaluate matched sets
2. Reason about choosing regression vs matching

Matching: The big idea

Matching: The big idea

Goal: Sample Average Treatment Effect on the Treated

$$\frac{1}{n_1}\sum_{i:A_i=1}\left(Y_i^1 - Y_i^0\right)$$

Matching: The big idea

Goal: Sample Average Treatment Effect on the Treated

$$\frac{1}{n_1}\sum_{i:A_i=1}\left(Y_i^1 - Y_i^0\right)$$

Problem: We don't see Y_i^0

Matching: The big idea

Goal: Sample Average Treatment Effect on the Treated

$$\frac{1}{n_1}\sum_{i:A_i=1}\left(Y_i^1 - Y_i^0\right)$$

Problem: We don't see Y_i^0

Solution: DAG + estimation

Matching: The big idea

Goal: Sample Average Treatment Effect on the Treated

$$\frac{1}{n_1}\sum_{i:A_i=1}\left(Y_i^1 - Y_i^0\right)$$

Problem: We don't see Y_i^0

Solution: DAG + estimation

$$\frac{1}{n_1}\sum_{i:A_i=1}\left(Y_i^1 - \mathsf{E}\left(Y \mid A=0, \vec{X}=\vec{x}_i\right)\right)$$

Matching: The big idea

Goal: Sample Average Treatment Effect on the Treated

$$\frac{1}{n_1} \sum_{i:A_i=1} \left(Y_i^1 - Y_i^0\right)$$

Problem: We don't see Y_i^0

Solution: DAG + estimation

$$\frac{1}{n_1} \sum_{i:A_i=1} \left(Y_i^1 - \mathrm{E}\left(Y \mid A=0, \vec{X}=\vec{x}_i\right)\right)$$

Matching: Estimate $\mathrm{E}(Y \mid A=0, \vec{X}=\vec{x}_i)$ from one or more untreated units with $\vec{X}$ "near" $\vec{x}_i$

Matching: The big idea

Goal: Sample Average Treatment Effect on the Treated

$$\frac{1}{n_1}\sum_{i:A_i=1}\left(Y_i^1 - Y_i^0\right)$$

Problem: We don't see Y_i^0

Solution: DAG + estimation

$$\frac{1}{n_1}\sum_{i:A_i=1}\left(Y_i^1 - \mathrm{E}\left(Y \mid A = 0, \vec{X} = \vec{x}_i\right)\right)$$

Matching: Estimate $\mathrm{E}(Y \mid A = 0, \vec{X} = \vec{x}_i)$ from one or more untreated units with $\vec{X}$ "near" $\vec{x}_i$

Debates: What does it mean to be "near"?

A concrete example

- In the 1970s, a research group selected disadvantaged workers

A concrete example

- In the 1970s, a research group selected disadvantaged workers
- Randomized to two treatment conditions
    - job training
    - no job training

A concrete example

- In the 1970s, a research group selected disadvantaged workers
- Randomized to two treatment conditions
  - job training
  - no job training
- Lalonde (1986) subsequently did a methodological exercise: try to recover the truth by comparing to non-randomized units with no job training

A concrete example

- In the 1970s, a research group selected disadvantaged workers
- Randomized to two treatment conditions
  - job training
  - no job training
- Lalonde (1986) subsequently did a methodological exercise: try to recover the truth by comparing to non-randomized units with no job training
- Dehejia & Wahba (1999) used this setting to illustrate matching

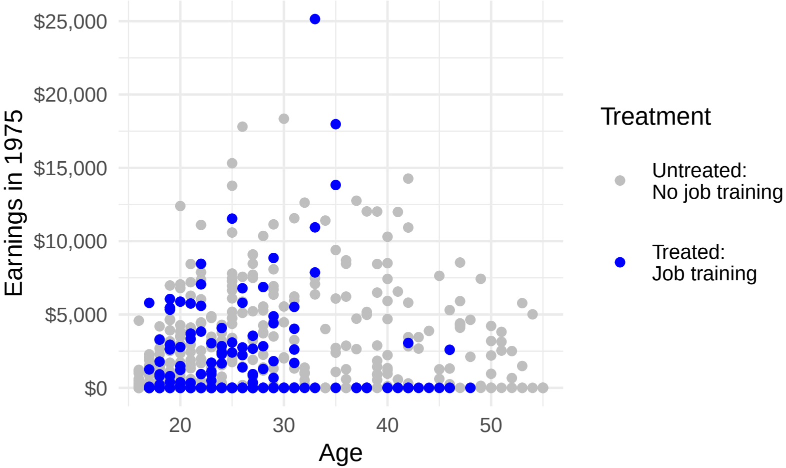
A concrete example

- In the 1970s, a research group selected disadvantaged workers
- Randomized to two treatment conditions
    - job training
    - no job training
- Lalonde (1986) subsequently did a methodological exercise: try to recover the truth by comparing to non-randomized units with no job training
- Dehejia & Wahba (1999) used this setting to illustrate matching

To get the data:

```
install.packages(“MatchIt”)
data(“lalonde”, package = “MatchIt”)
```

A concrete example



Why matching is great

Why matching is great

1. Completely transparent that Y_i^1 is observed

Why matching is great

1. Completely transparent that Y_i^1 is observed
2. Completely transparent how we estimate $\hat{Y}_i^0$ (from the matched unit)

Why matching is great

1. Completely transparent that Y_i^1 is observed
2. Completely transparent how we estimate $\hat{Y}_i^0$ (from the matched unit)
3. Easy to explain

Why matching is great

1. Completely transparent that Y_i^1 is observed
2. Completely transparent how we estimate $\hat{Y}_i^0$ (from the matched unit)
3. Easy to explain
    - We had some treated units

Why matching is great

1. Completely transparent that Y_i^1 is observed
2. Completely transparent how we estimate $\hat{Y}_i^0$ (from the matched unit)
3. Easy to explain
    - We had some treated units
    - We found comparable control units

Why matching is great

1. Completely transparent that Y_i^1 is observed
2. Completely transparent how we estimate $\hat{Y}_i^0$ (from the matched unit)
3. Easy to explain
    - We had some treated units
    - We found comparable control units
    - We took a mean difference

Why matching is great

1. Completely transparent that Y_i^1 is observed
2. Completely transparent how we estimate $\hat{Y}_i^0$ (from the matched unit)
3. Easy to explain
    - We had some treated units
    - We found comparable control units
    - We took a mean difference
4. Can assess quality of matches before we look at the outcome

Why matching is great

1. Completely transparent that Y_i^1 is observed
2. Completely transparent how we estimate $\hat{Y}_i^0$ (from the matched unit)
3. Easy to explain
   - We had some treated units
   - We found comparable control units
   - We took a mean difference
4. Can assess quality of matches before we look at the outcome
5. Model-free*

Why matching is great

1. Completely transparent that Y_i^1 is observed
2. Completely transparent how we estimate $\hat{Y}_i^0$ (from the matched unit)
3. Easy to explain
    - We had some treated units
    - We found comparable control units
    - We took a mean difference
4. Can assess quality of matches before we look at the outcome
5. Model-free*
    - * but you have to define what makes a match "good"

Matching overview

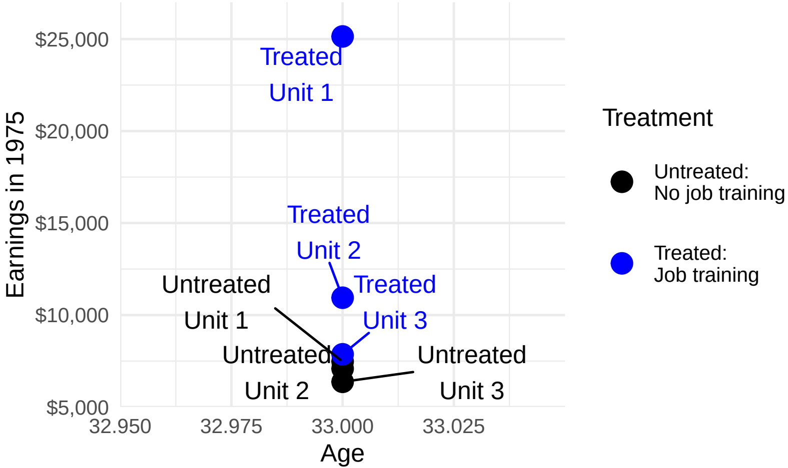
- Matching in univariate settings: Algorithms
- Matching in multivariate settings: Distance metrics

Matching in univariate settings: Algorithms

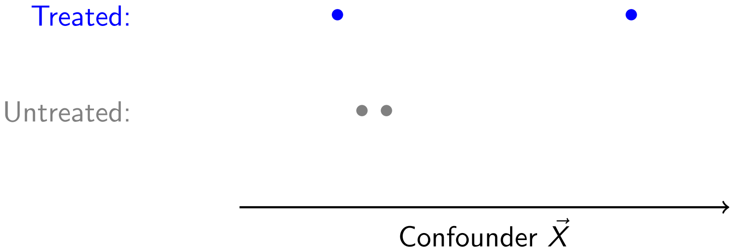
- Caliper or no caliper
- 1:1 vs k:1
- With replacement vs without replacement
- Greedy vs optimal

Caliper or no caliper matching

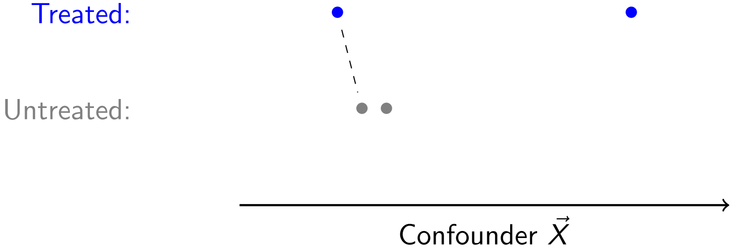
Why might I be hesitant to find a match for Treated Unit 1?



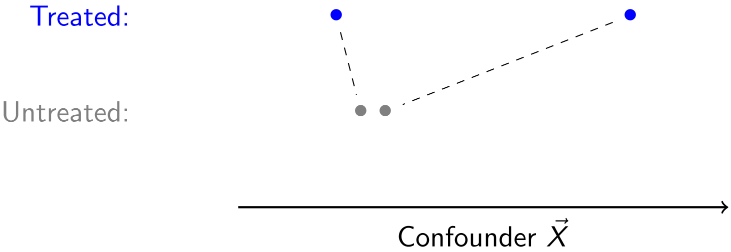
Caliper or no caliper matching



Caliper or no caliper matching



Caliper or no caliper matching

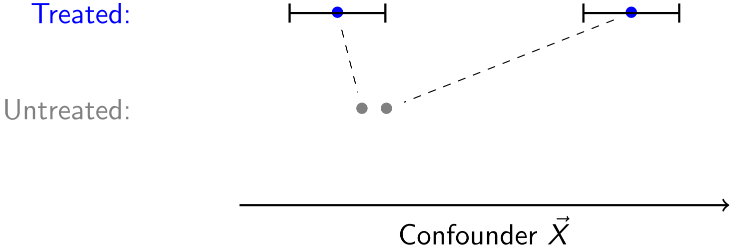


Treated:

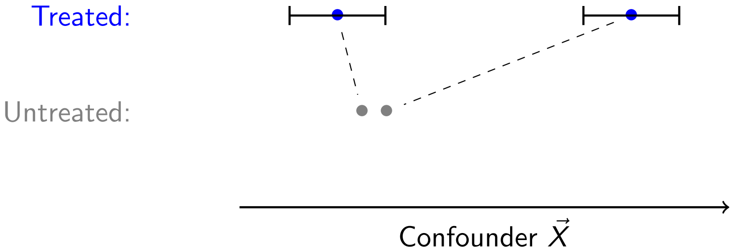
Untreated:

Confounder $\vec{X}$

Caliper or no caliper matching

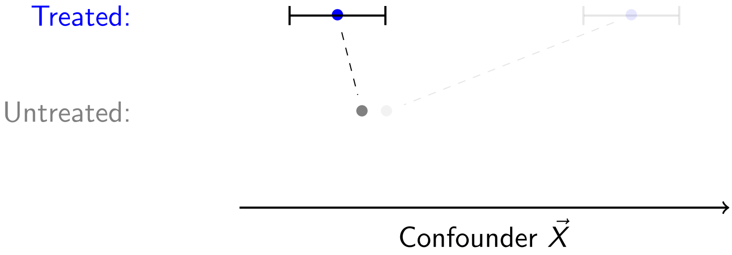


Caliper or no caliper matching



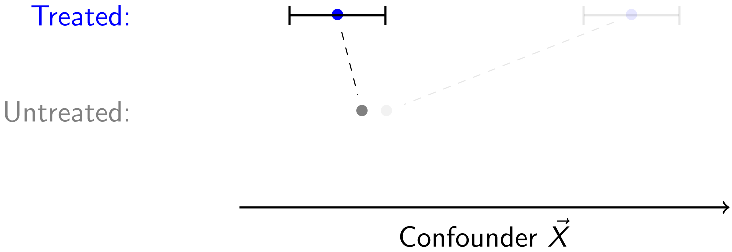
- Caliper: A radius around a treated unit such that we would rather drop the unit than make a match beyond that radius

Caliper or no caliper matching



- Caliper: A radius around a treated unit such that we would rather drop the unit than make a match beyond that radius

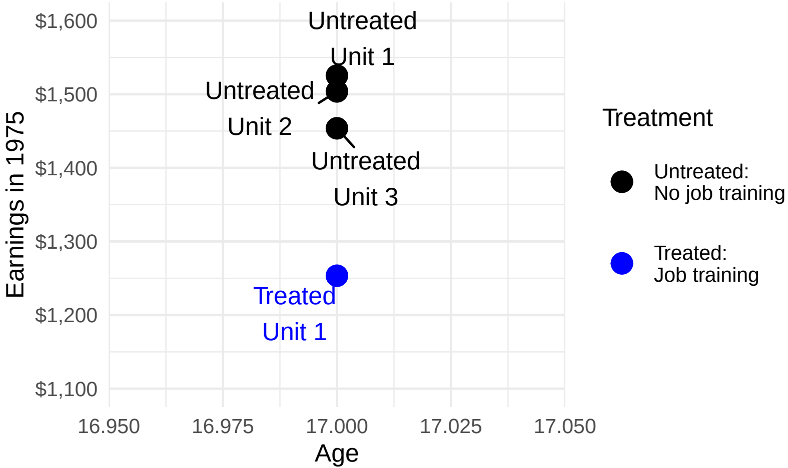
Caliper or no caliper matching



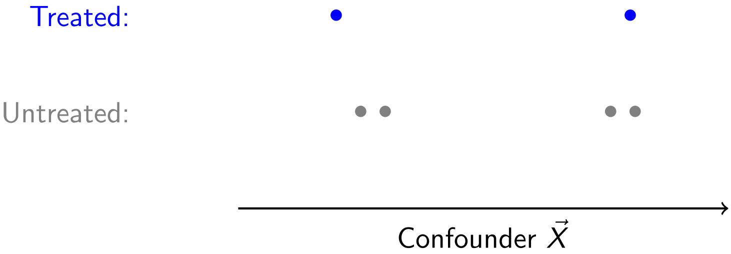
- Caliper: A radius around a treated unit such that we would rather drop the unit than make a match beyond that radius
- Feasible Sample Average Treatment Effect on the Treated (FSATT): Average among treated units for whom an acceptable match exists

1:1 vs *k*:1 matching

Can we make use of Untreated Units 1 and 2?



1:1 vs k:1 matching

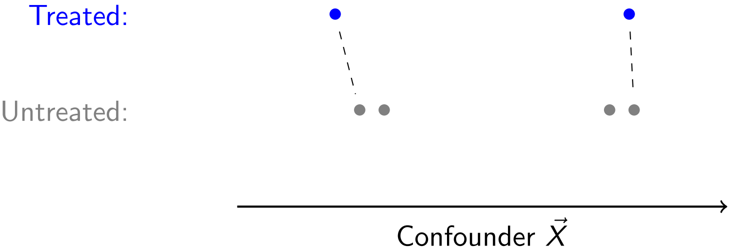


Treated:

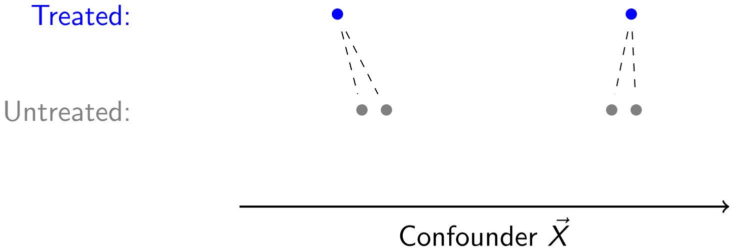
Untreated:

Confounder $\vec{X}$

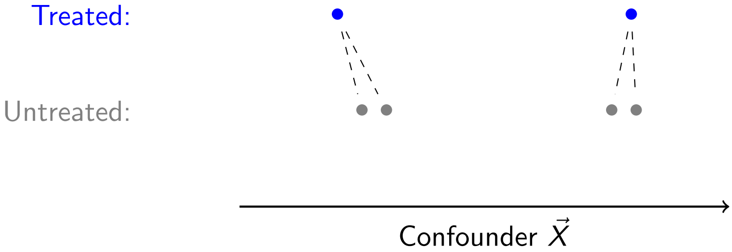
1:1 vs k:1 matching



1:1 vs k:1 matching

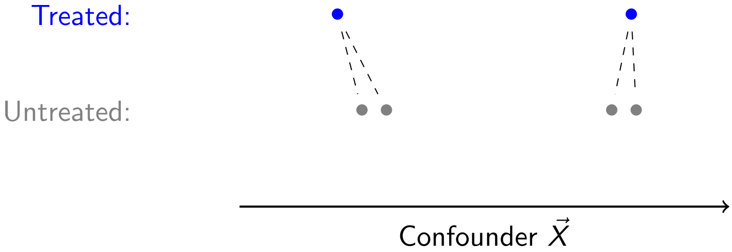


1:1 vs k:1 matching



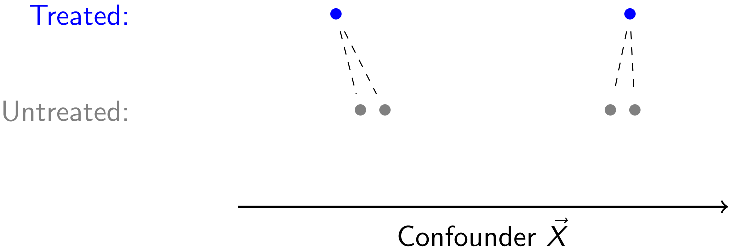
- Benefit of 2:1 matching
- Benefit of 1:1 matching

1:1 vs k:1 matching



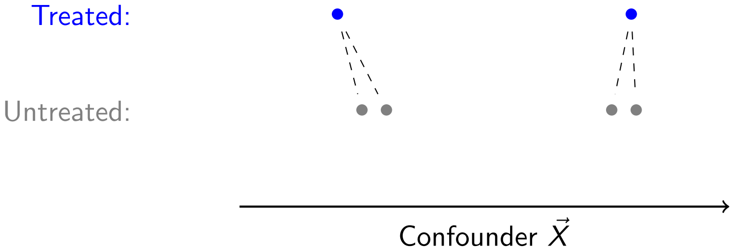
- Benefit of 2:1 matching
  - Lower variance. Averaging over more cases.
- Benefit of 1:1 matching

1:1 vs k:1 matching



- Benefit of 2:1 matching
  - Lower variance. Averaging over more cases.
- Benefit of 1:1 matching
  - Lower bias. Only the best matches.

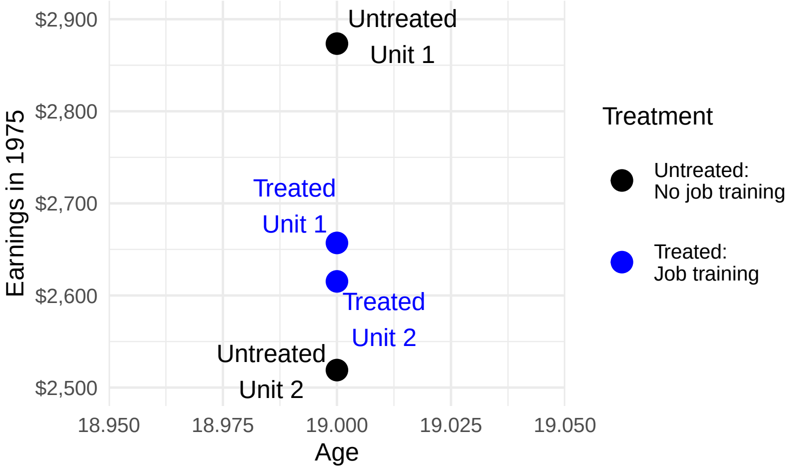
1:1 vs k:1 matching



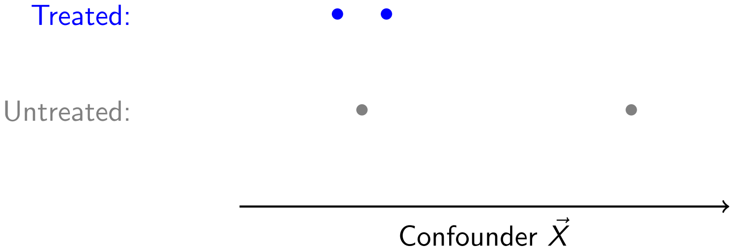
- Benefit of 2:1 matching
    - Lower variance. Averaging over more cases.
- Benefit of 1:1 matching
    - Lower bias. Only the best matches.
- Greater $k \rightarrow$ lower variance, higher bias

With replacement vs without replacement matching

We match Treated 2 to Untreated 2. Who should be the match for Treated 1?



With replacement vs without replacement matching

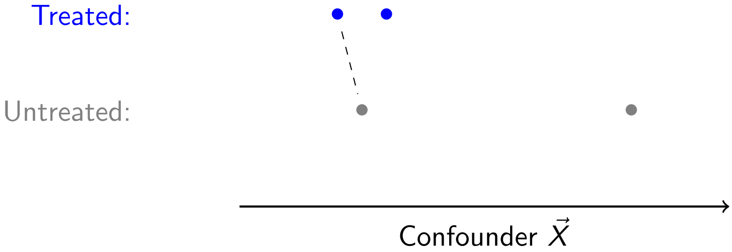


Treated:

Untreated:

Confounder $\vec{X}$

With replacement vs without replacement matching

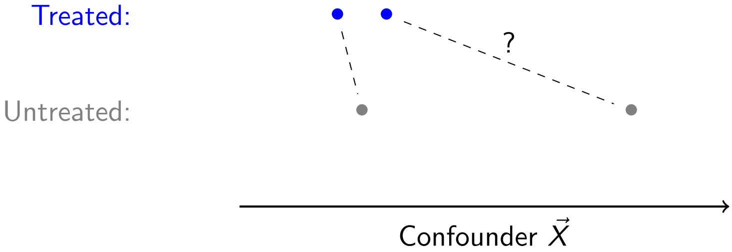


Treated:

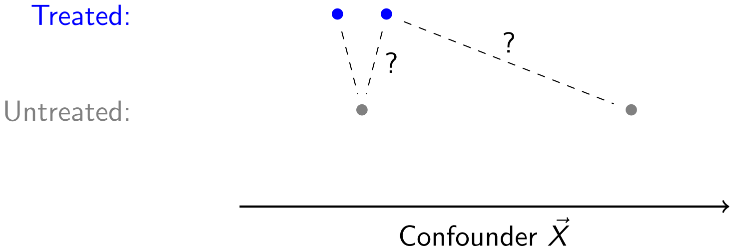
Untreated:

Confounder $\vec{X}$

With replacement vs without replacement matching

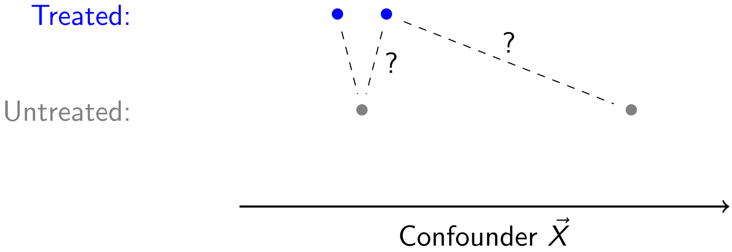


With replacement vs without replacement matching



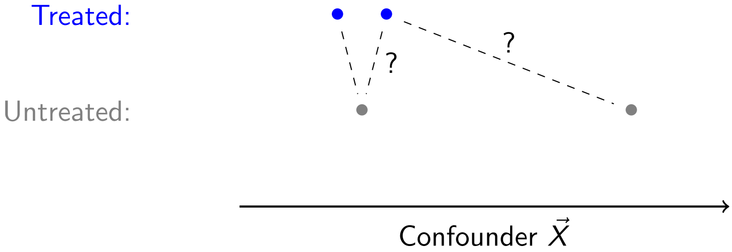
- Benefit of matching without replacement
- Benefit of matching with replacement

With replacement vs without replacement matching



- Benefit of matching without replacement
  - Lower variance. Averaging over more cases.
- Benefit of matching with replacement

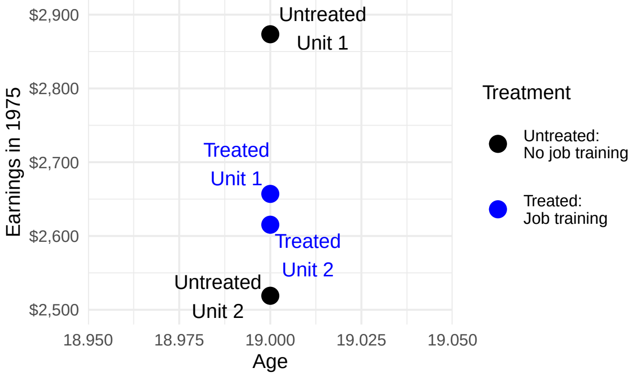
With replacement vs without replacement matching



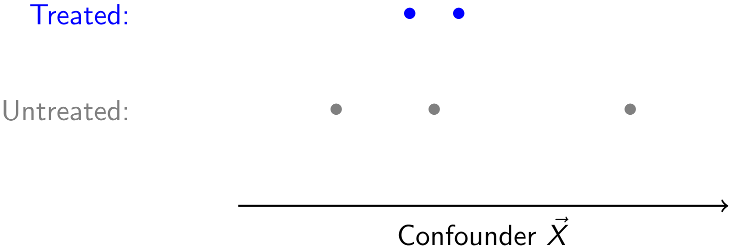
- Benefit of matching without replacement
    - Lower variance. Averaging over more cases.
- Benefit of matching with replacement
    - Lower bias. Better matches.

Greedy vs optimal matching

North side: Start with Treated 1. Find the best match.
South side: Start with Treated 2. Find the best match.
Both sides: Who is left as the match for the other treated unit?



Greedy vs optimal matching[1]



Treated:

Untreated:

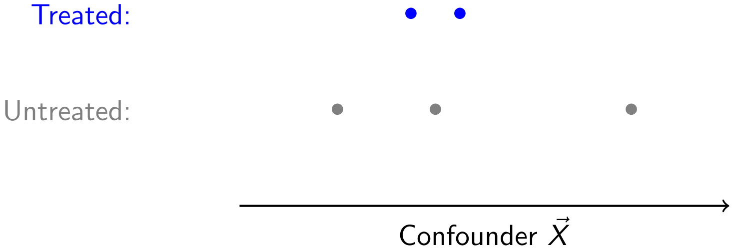
Confounder $\vec{X}$

---

[1] Gu, X. S., & Rosenbaum, P. R. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. Journal of Computational and Graphical Statistics, 2(4), 405-420.

Greedy vs optimal matching[1]

Greedy Matching:
Match sequentially



Treated:

Untreated:

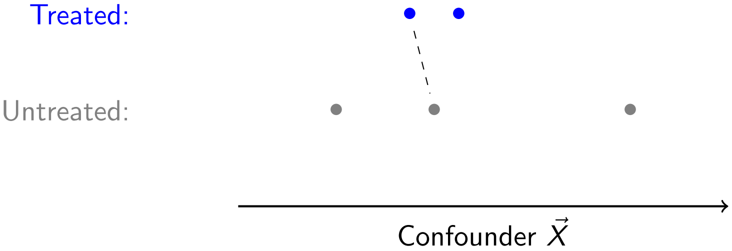
Confounder $\vec{X}$

---

[1] Gu, X. S., & Rosenbaum, P. R. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. Journal of Computational and Graphical Statistics, 2(4), 405-420.

Greedy vs optimal matching[1]

Greedy Matching:
Match sequentially

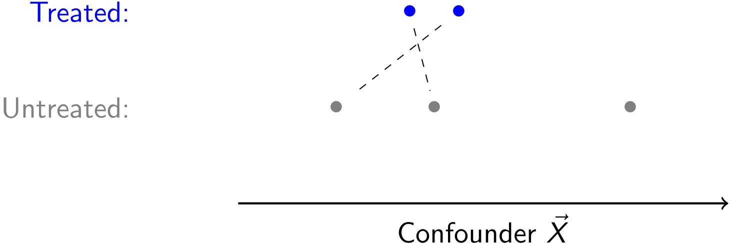


---

[1] Gu, X. S., & Rosenbaum, P. R. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. Journal of Computational and Graphical Statistics, 2(4), 405-420.

Greedy vs optimal matching[1]

Greedy Matching:
Match sequentially



Treated:

Untreated:

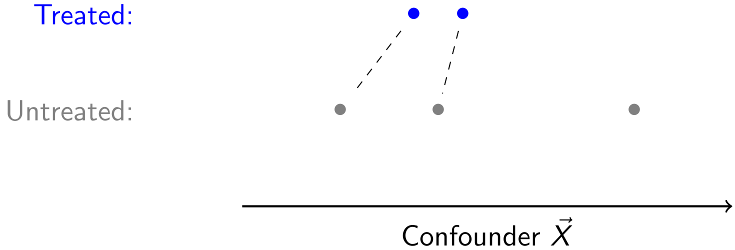
Confounder $\vec{X}$

---

[1] Gu, X. S., & Rosenbaum, P. R. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. Journal of Computational and Graphical Statistics, 2(4), 405-420.

Greedy vs optimal matching[1]

Optimal Matching:
Consider the whole set of matches

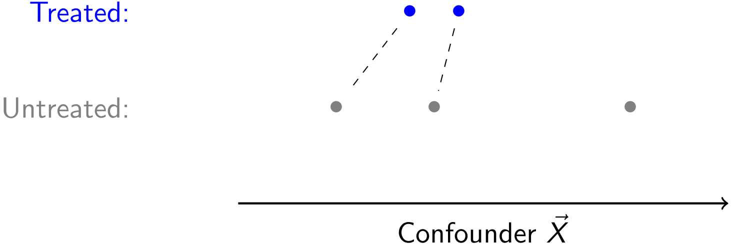


---

[1] Gu, X. S., & Rosenbaum, P. R. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. Journal of Computational and Graphical Statistics, 2(4), 405-420.

Greedy vs optimal matching[1]

Optimal Matching:
Consider the whole set of matches



- Optimal is better. Just computationally harder.

---

[1] Gu, X. S., & Rosenbaum, P. R. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. Journal of Computational and Graphical Statistics, 2(4), 405-420.

Matching in univariate settings: Algorithms (recap)

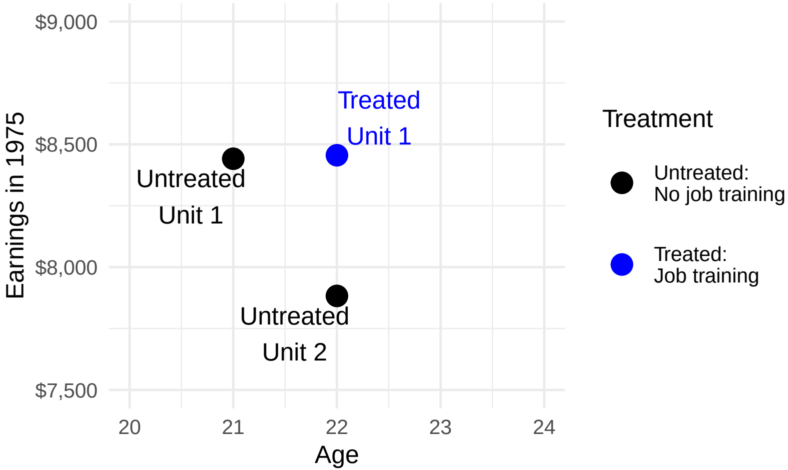
- Caliper or no caliper
- 1:1 vs k:1
- With replacement vs without replacement
- Greedy vs optimal

Matching overview
    Matching in univariate settings: Algorithms
    Matching in multivariate settings: Distance metrics

Matching overview

- Matching in univariate settings: Algorithms
- Matching in multivariate settings: Distance metrics

What if $\vec{X}$ is multivariate?



Key concept: Distance defines who is closer

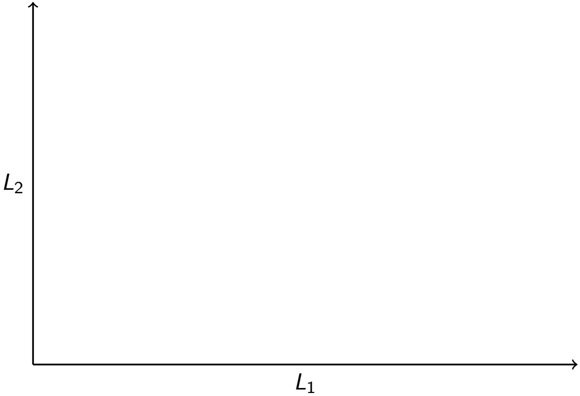
To find the closest match, we need to define what it means for unit i and j to be **close** to each other.

$$d(\vec{x}_i, \vec{x}_j) = \text{a number}$$

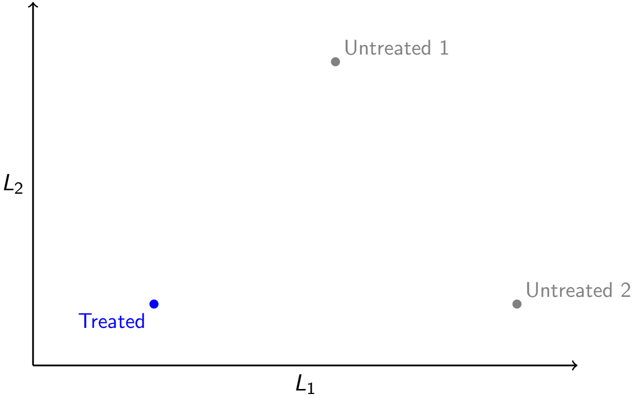
will be the **distance** between confounder vectors $\vec{x}_i$ and $\vec{x}_j$.

Who is closer? Euclidean and Manhattan distance

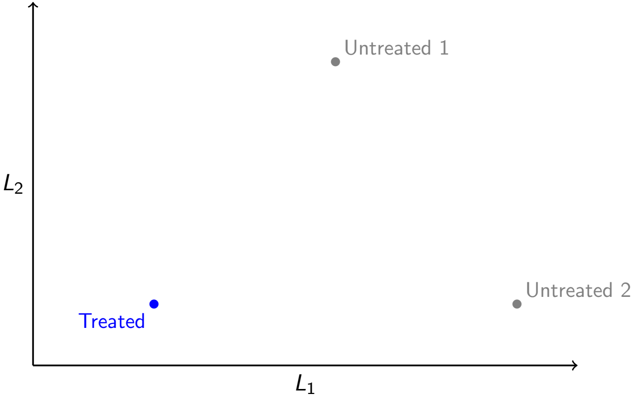
Who is closer? Euclidean and Manhattan distance



Who is closer? Euclidean and Manhattan distance

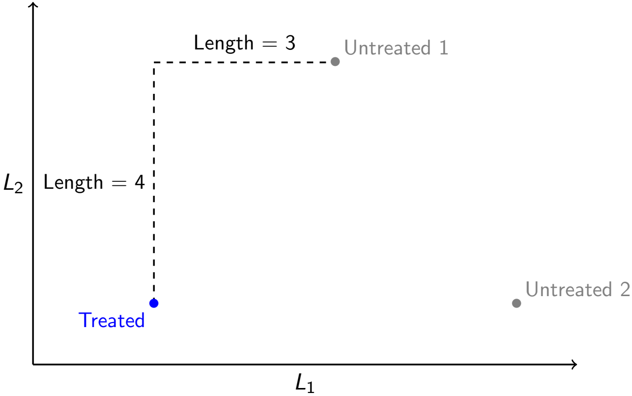


Who is closer? Euclidean and Manhattan distance



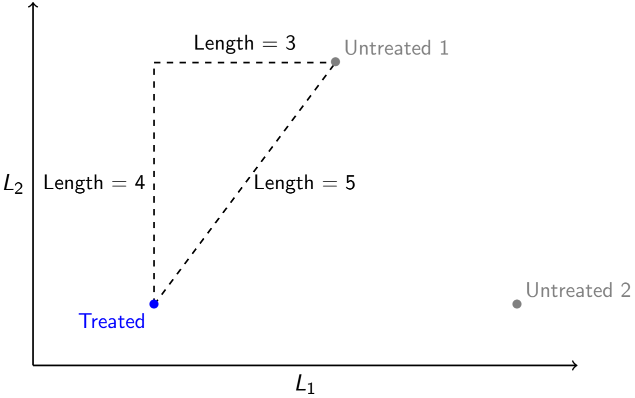
Which untreated unit should be the match?

Who is closer? Euclidean and Manhattan distance



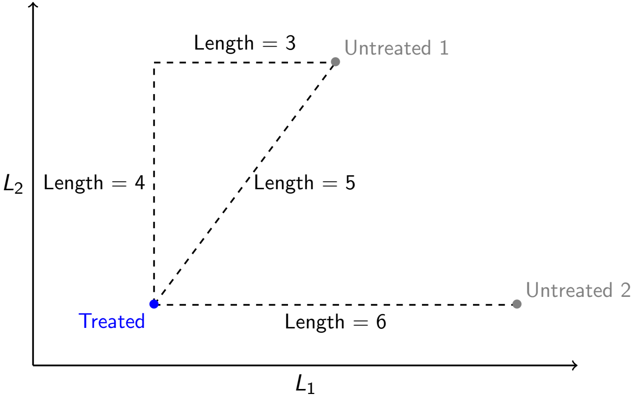
Which untreated unit should be the match?

Who is closer? Euclidean and Manhattan distance



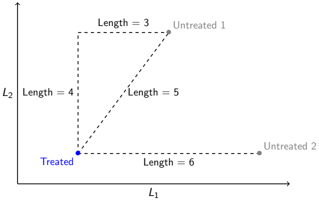
Which untreated unit should be the match?

Who is closer? Euclidean and Manhattan distance

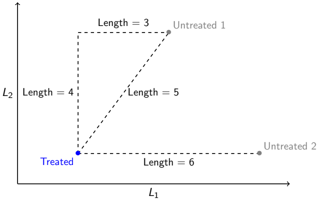


Which untreated unit should be the match?

Who is closer? Euclidean and Manhattan distance



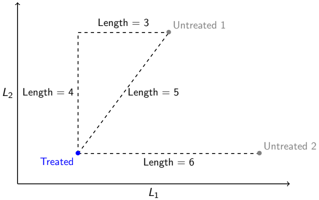
Who is closer? Euclidean and Manhattan distance



- Manhattan distance:

- Euclidean distance:

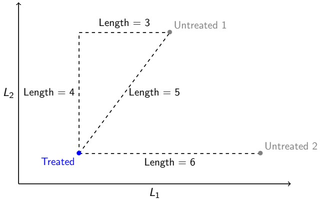
Who is closer? Euclidean and Manhattan distance



- Manhattan distance: $d(i,j) = \sum_p |L_{pi} - L_{pj}|$

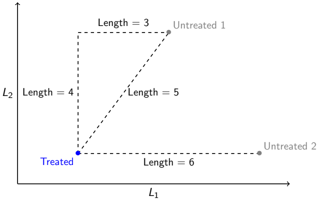
- Euclidean distance:

Who is closer? Euclidean and Manhattan distance



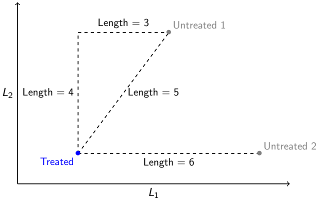
- Manhattan distance: $d(i,j) = \sum_p |L_{pi} - L_{pj}|$
  - $d(\text{Treated, Untreated 1}) = 3 + 4 = 7$
  - $d(\text{Treated, Untreated 2}) = 6 + 0 = 6$ ✓
- Euclidean distance:

Who is closer? Euclidean and Manhattan distance



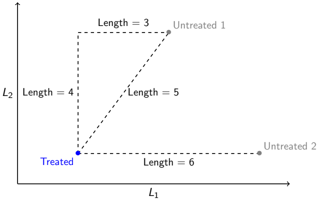
- Manhattan distance: $d(i,j) = \sum_p |L_{pi} - L_{pj}|$
  - $d(\text{Treated, Untreated 1}) = 3 + 4 = 7$
  - $d(\text{Treated, Untreated 2}) = 6 + 0 = 6$ ✓
- Euclidean distance: $d(i,j) = \sqrt{\sum_p (L_{pi} - L_{pj})^2}$

Who is closer? Euclidean and Manhattan distance



- Manhattan distance: $d(i,j) = \sum_p |L_{pi} - L_{pj}|$
  - $d(\text{Treated, Untreated 1}) = 3 + 4 = 7$
  - $d(\text{Treated, Untreated 2}) = 6 + 0 = 6$ ✓
- Euclidean distance: $d(i,j) = \sqrt{\sum_p (L_{pi} - L_{pj})^2}$
  - $d(\text{Treated, Untreated 1}) = \sqrt{3^2 + 4^2} = 5$ ✓
  - $d(\text{Treated, Untreated 2}) = \sqrt{6^2 + 0^2} = 6$

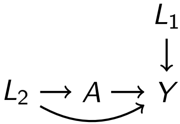
Who is closer? Euclidean and Manhattan distance



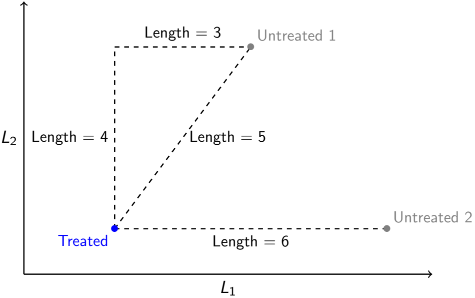
- Manhattan distance: $d(i, j) = \sum_p |L_{pi} - L_{pj}|$
  - $d(\text{Treated, Untreated 1}) = 3 + 4 = 7$
  - $d(\text{Treated, Untreated 2}) = 6 + 0 = 6$ ✓
- Euclidean distance: $d(i, j) = \sqrt{\sum_p (L_{pi} - L_{pj})^2}$
  - $d(\text{Treated, Untreated 1}) = \sqrt{3^2 + 4^2} = 5$ ✓
  - $d(\text{Treated, Untreated 2}) = \sqrt{6^2 + 0^2} = 6$
- It depends on the distance metric!

A common distance metric: Propensity scores

Now suppose only L_2 is related to treatment.

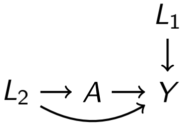


Which match do you pick?

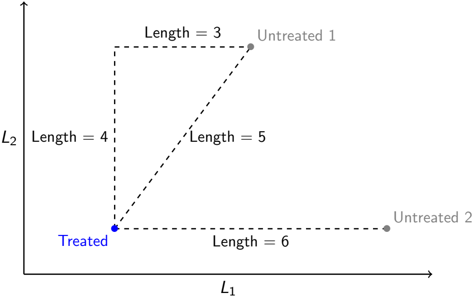


A common distance metric: Propensity scores

Now suppose only L_2 is related to treatment.



Which match do you pick? Untreated 2! Perfect match.



A common distance metric: Propensity scores

Propensity score: $\pi_i = \mathrm{P}(A = 1 \mid \vec{X} = \vec{x}_i)$

---

[2] Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1), 41-55.

A common distance metric: Propensity scores

Propensity score: $\pi_i = \mathrm{P}(A = 1 \mid \vec{X} = \vec{x}_i)$

- Univariate summary of all confounders

---

[2] Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1), 41-55.

A common distance metric: Propensity scores

Propensity score: $\pi_i = \text{P}(A = 1 \mid \vec{X} = \vec{x}_i)$

- Univariate summary of all confounders
- In expectation, a sample balanced on π is balanced on $\vec{X}$
  - Rosenbaum & Rubin theorem[2]

---

[2] Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1), 41-55.

A common distance metric: Propensity scores

Propensity scores can be nonparametric or parametric.

A common distance metric: Propensity scores

Propensity scores can be nonparametric or parametric.

- Nonparametric: $\hat{\pi}_i$ is the proportion treated in the sample stratum $\vec{X} = \vec{x}_i$.

A common distance metric: Propensity scores

Propensity scores can be nonparametric or parametric.

- Nonparametric: $\hat{\pi}_i$ is the proportion treated in the sample stratum $\vec{X} = \vec{x}_i$.
- Parametric: Often estimated as
  - Fit logistic regression

$$\text{logit}\left(\text{P}(A = 1 \mid \vec{X})\right) = \alpha + \vec{X}\vec{\beta}$$

A common distance metric: Propensity scores

Propensity scores can be nonparametric or parametric.

- Nonparametric: $\hat{\pi}_i$ is the proportion treated in the sample stratum $\vec{X} = \vec{x}_i$.
- Parametric: Often estimated as
  - Fit logistic regression

$$\text{logit}\left(\text{P}(A = 1 \mid \vec{X})\right) = \alpha + \vec{X}\vec{\beta}$$

  - Predict the probability of treatment

$$\hat{\pi}_i = \text{logit}^{-1}\left(\hat{\alpha} + \vec{x}_i\hat{\vec{\beta}}\right)$$

A common distance metric: Propensity scores

Propensity scores can be nonparametric or parametric.

- Nonparametric: $\hat{\pi}_i$ is the proportion treated in the sample stratum $\vec{X} = \vec{x}_i$.
- Parametric: Often estimated as
  - Fit logistic regression

$$\text{logit}\left(\text{P}(A = 1 \mid \vec{X})\right) = \alpha + \vec{X}\vec{\beta}$$

  - Predict the probability of treatment

$$\hat{\pi}_i = \text{logit}^{-1}\left(\hat{\alpha} + \vec{x}_i\hat{\vec{\beta}}\right)$$

Propensity score distance for matching:

$$d(i, j) = |\hat{\pi}_i - \hat{\pi}_j|$$

Workflow for matching

- Draw a DAG
- Select a sufficient adjustment set $\vec{X}$
- Define a distance: how far apart $\vec{x}_i$ and $\vec{x}_j$ are
- Choose a matching algorithm
    - 1:1 without replacement, greedy
- Conduct matching
- Estimate ATE by outcome modeling on the matched set

The **MatchIt** package makes it easy

Discuss: Why regression? Why matching?

Learning goals for today

At the end of class, you will be able to:

1. Use matching methods for causal effects
    - Select a matching algorithm
    - Define a distance metric for multivariate matching
    - Evaluate matched sets
2. Reason about choosing regression vs matching