Matching Lab

INFO/STSCI/ILRST 3900: Causal Inference

4 Oct 2023

Agenda

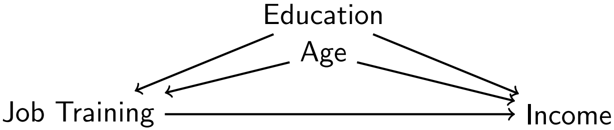
- Reminders/Announcements
- Icebreaker: Matching Lecture Review
- Matching with Multiple Covariates Overview
- R Demonstration
- Your turn (get ahead on the HW!)

Icebreaker: Matching Lecture Review

In groups of 2-4, you will be assigned one of the questions below. Your task is to explain and answer the assigned question. Have one person in your group ready to share what you discuss with the whole class.

1. What is the difference between the ATE and the ATT, and what is the challenge in estimating the ATT?
2. Explain what matching is and how we use it to estimate causal effects (like the ATT).
3. What is the difference between caliper versus no caliper matching, and what changes in the estimand when we use calipers?
4. What is $1:1$ matching versus $k:1$ matching? Explain the bias-variance trade off.
5. What is matching with replacement and without replacement? Explain the bias-variance trade off.
6. What is greedy versus optimal matching and what trade off should be considered there?

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

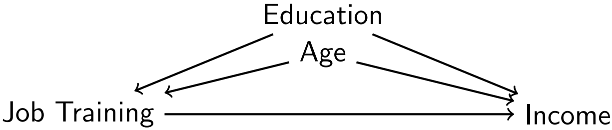


What if $\vec{L}$ is multivariate?

$$
\begin{tikzcd}
 & \text{Education} \arrow[dl] \arrow[dr] & \\
 & \text{Age} \arrow[dl] \arrow[dr] & \\
\text{Job Training} \arrow[rr] & & \text{Income}
\end{tikzcd}
$$

- Conditional exchangeability holds when conditioning on Age and Education!

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

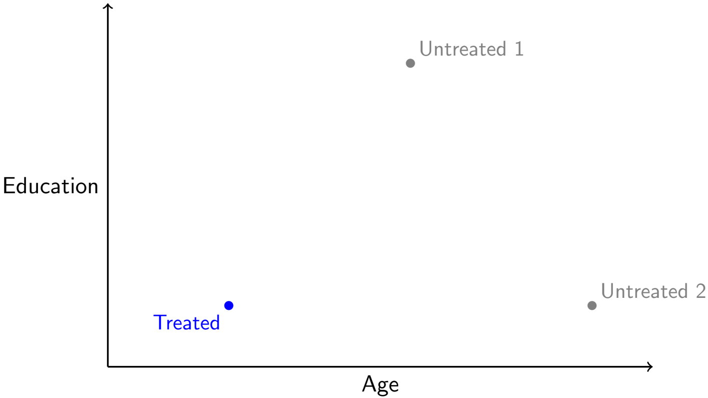


- Conditional exchangeability holds when conditioning on Age and Education!
- Estimate $E(Y^{a=0} \mid A=1)$ with a group of untreated units, $\mathcal{M}$, which has a similar distribution of Age and Education to the treated group

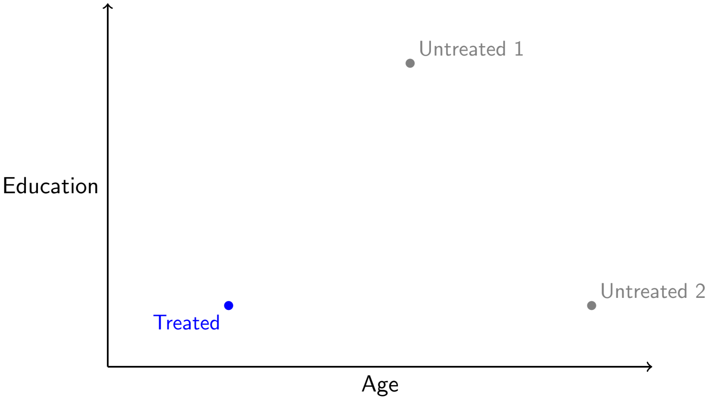
What if $\vec{L}$ is multivariate?

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

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

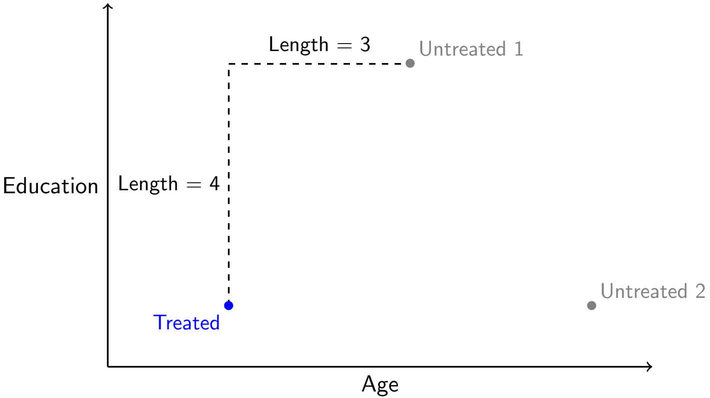


What if $\vec{L}$ is multivariate?



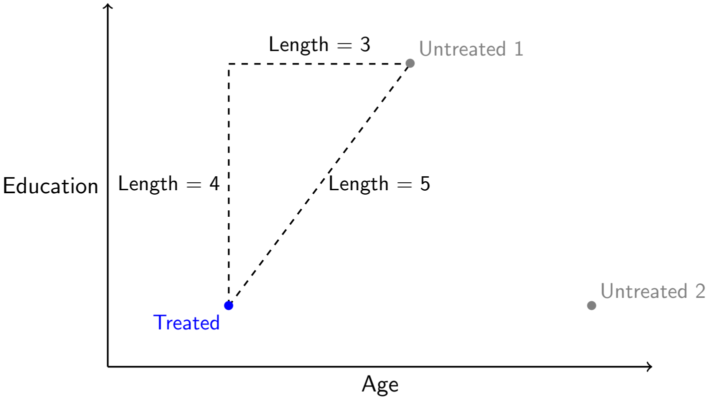
Which untreated unit should be the match?

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



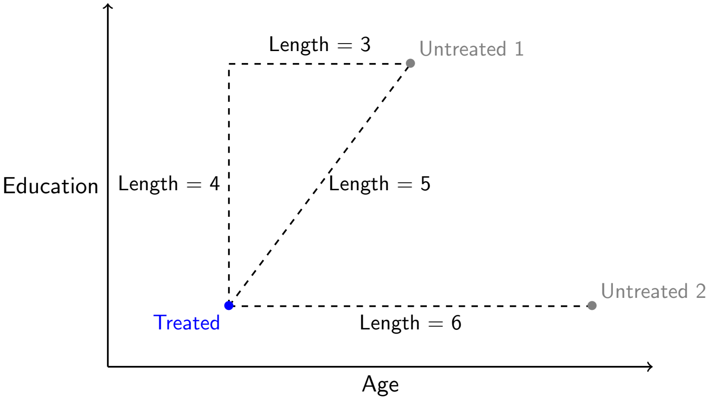
Which untreated unit should be the match?

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



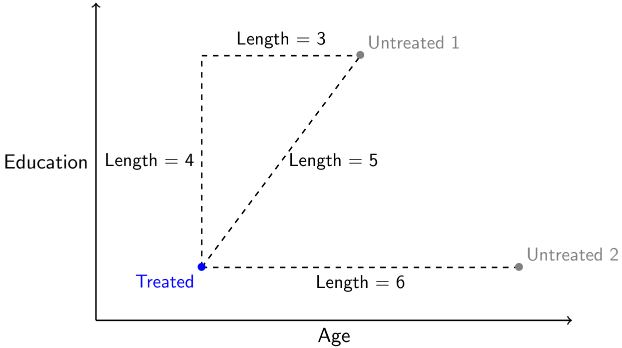
Which untreated unit should be the match?

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

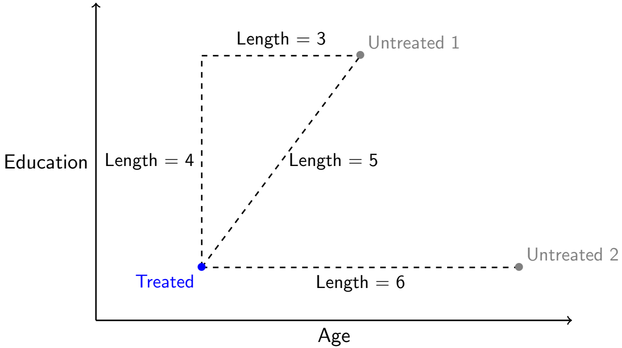


Which untreated unit should be the match?

What if $\vec{L}$ is multivariate? We need a **distance metric**

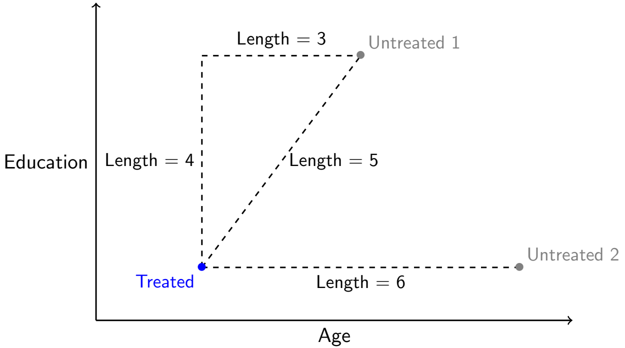


What if $\vec{L}$ is multivariate? We need a **distance metric**



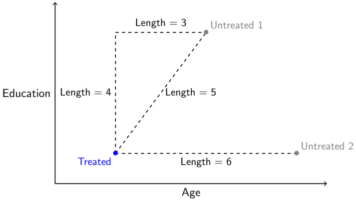
- Define a way to measure "distance" between two individuals as a single number

What if $\vec{L}$ is multivariate? We need a **distance metric**

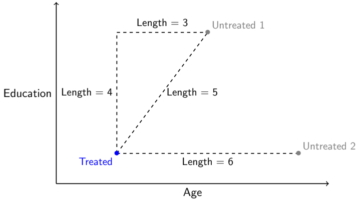


- Define a way to measure "distance" between two individuals as a single number
- Match individuals in the same way as before using that distance!

What if $\vec{L}$ is multivariate? We need a **distance metric**



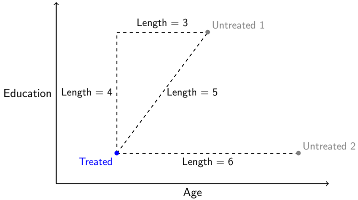
What if $\vec{L}$ is multivariate? We need a **distance metric**



- Manhattan distance:

- Euclidean distance:

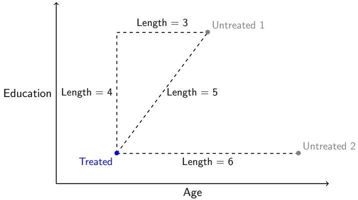
What if $\vec{L}$ is multivariate? We need a **distance metric**



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

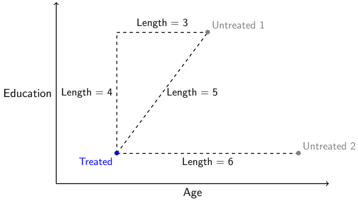
- Euclidean distance:

What if $\vec{L}$ is multivariate? We need a **distance metric**



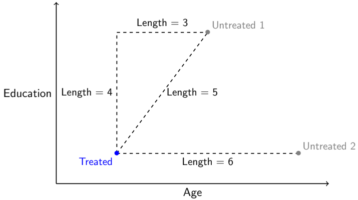
- Manhattan distance: $d(i,j) = \sum_p |L_{pi} - L_{pj}|$
  - $d(\text{Treated, Untreated 1}) = 3 + 4 = 7$
  - $d(\text{Treated, Untreated 2}) = 6 + 0 = 6$ ✓
- Euclidean distance:

What if $\vec{L}$ is multivariate? We need a **distance metric**



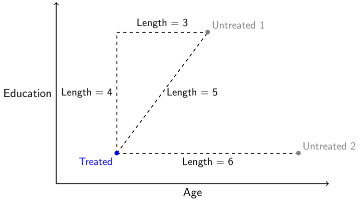
- Manhattan distance: $d(i, j) = \sum_p |L_{pi} - L_{pj}|$
  - $d(\text{Treated, Untreated 1}) = 3 + 4 = 7$
  - $d(\text{Treated, Untreated 2}) = 6 + 0 = 6$ ✓
- Euclidean distance: $d(i, j) = \sqrt{\sum_p (L_{pi} - L_{pj})^2}$

What if $\vec{L}$ is multivariate? We need a **distance metric**



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

What if $\vec{L}$ is multivariate? We need a **distance metric**



- 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: Mahalanobis distance

A common distance metric: Mahalanobis distance

Motivated by two principles

A common distance metric: Mahalanobis distance

Motivated by two principles

- Principle 1: Address unequal variances

A common distance metric: Mahalanobis distance

Motivated by two principles

- Principle 1: Address unequal variances
  - Age might range uniformly from 18 to 80

A common distance metric: Mahalanobis distance

Motivated by two principles

- Principle 1: Address unequal variances
  - Age might range uniformly from 18 to 80
  - Education range uniformly from 0 to 16

A common distance metric: Mahalanobis distance

Motivated by two principles

- Principle 1: Address unequal variances
  - Age might range uniformly from 18 to 80
  - Education range uniformly from 0 to 16
  - We might correct for this so age doesn't dominate the distance

A common distance metric: Mahalanobis distance

Motivated by two principles

- Principle 1: Address unequal variances
  - Age might range uniformly from 18 to 80
  - Education range uniformly from 0 to 16
  - We might correct for this so age doesn't dominate the distance
- Principle 2: Address correlations

A common distance metric: Mahalanobis distance

Motivated by two principles

- Principle 1: Address unequal variances
  - Age might range uniformly from 18 to 80
  - Education range uniformly from 0 to 16
  - We might correct for this so age doesn't dominate the distance
- Principle 2: Address correlations
  - Suppose we included age in years, age in months, and education

A common distance metric: Mahalanobis distance

Motivated by two principles

- Principle 1: Address unequal variances
  - Age might range uniformly from 18 to 80
  - Education range uniformly from 0 to 16
  - We might correct for this so age doesn't dominate the distance
- Principle 2: Address correlations
  - Suppose we included age in years, age in months, and education
  - Suppose we included age in years and age in months are very correlated

A common distance metric: Mahalanobis distance

Motivated by two principles

- Principle 1: Address unequal variances
  - Age might range uniformly from 18 to 80
  - Education range uniformly from 0 to 16
  - We might correct for this so age doesn't dominate the distance
- Principle 2: Address correlations
  - Suppose we included age in years, age in months, and education
  - Suppose we included age in years and age in months are very correlated
  - We should care about a correlation-corrected distance

A common distance metric: Mahalanobis distance

Motivated by two principles

- Principle 1: Address unequal variances
  - Age might range uniformly from 18 to 80
  - Education range uniformly from 0 to 16
  - We might correct for this so age doesn't dominate the distance
- Principle 2: Address correlations
  - Suppose we included age in years, age in months, and education
  - Suppose we included age in years and age in months are very correlated
  - We should care about a correlation-corrected distance

A common distance metric: Mahalanobis distance

Motivated by two principles

- Principle 1: Address unequal variances
  - Age might range uniformly from 18 to 80
  - Education range uniformly from 0 to 16
  - We might correct for this so age doesn't dominate the distance
- Principle 2: Address correlations
  - Suppose we included age in years, age in months, and education
  - Suppose we included age in years and age in months are very correlated
  - We should care about a correlation-corrected distance

$$d(i,j) = \sqrt{\left(\vec{L}_i - \vec{L}_j\right)^T \Sigma^{-1} \left(\vec{L}_i - \vec{L}_j\right)}$$

where $\Sigma = \mathrm{V}(\vec{L})$, the variance-covariance matrix of L

Code

Let's try this out in R!