

Matching Lab

INFO/STSCI/ILRST 3900: Causal Inference

28 Oct 2026

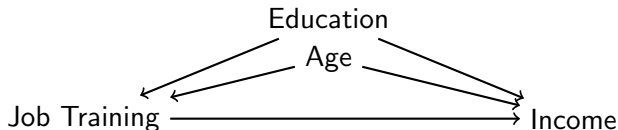
Reminders and Announcements

- ▶ In-class quiz 3 Thursday, Nov 5
- ▶ Project Part 1 due Friday, Nov 6
- ▶ Office hours:
 - ▶ Filippo: Thursday 4-5pm (CIS 325A) CIS Building
 - ▶ Shira: Monday 5-6 pm in (CIS 325A) CIS Building
 - ▶ Sam: Tuesday 3:30-4:30pm (CIS 324) CIS Building

Matching Review

- ▶ Suppose person i is in the treatment group ($A_i = 1$).
- ▶ Want to compare their outcome under treatment vs control
- ▶ Fundamental problem of causal inference: I can only observe one of these
- ▶ Matching: Find a person j in the control group ($A_j = 0$) that is *similar enough* to person i and compare their outcomes
- ▶ Reasoning: if people are *similar enough*, then maybe their potential outcomes are also *similar enough*
- ▶ How do we define *similar enough*?
- ▶ We can use covariates! $\vec{L}$

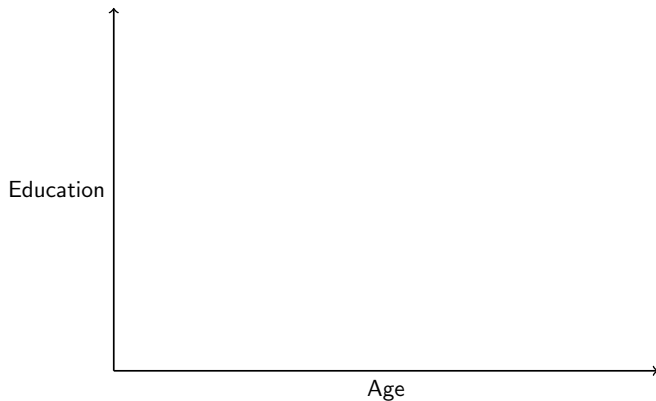
What if $\vec{L}$ is multivariate?



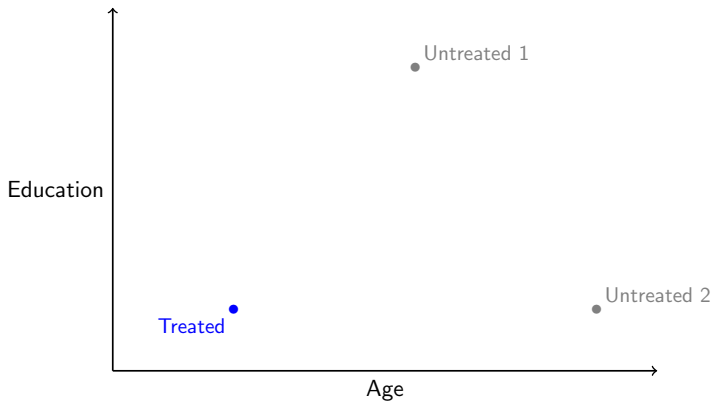
- ▶ Conditional exchangeability holds when conditioning on Age and Education!
- ▶ Matching: look for a group of untreated units which has a similar distribution of Age and Education to the treated group

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

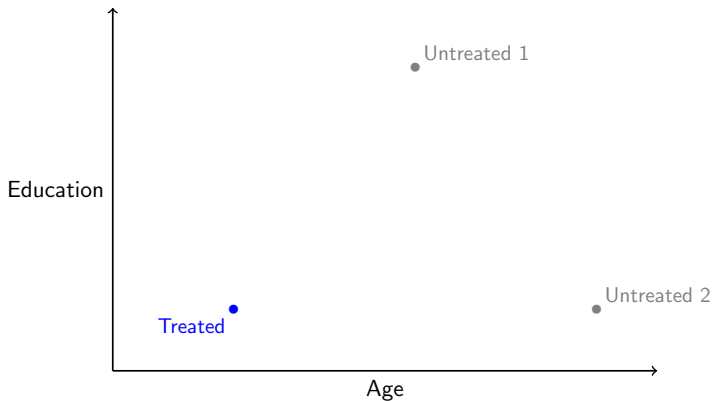
What if $\vec{L}$ is multivariate?



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

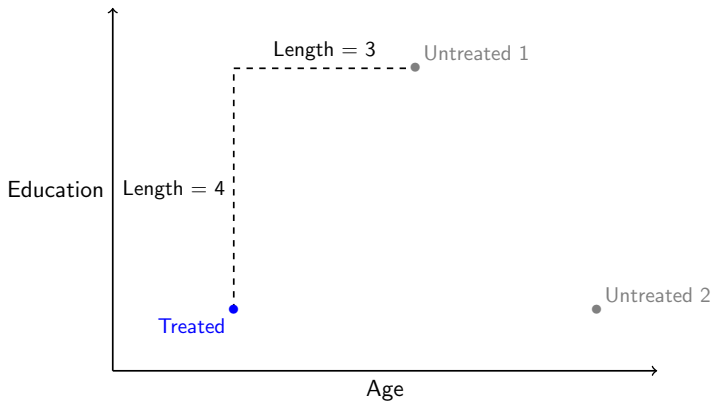


What if $\vec{L}$ is multivariate?



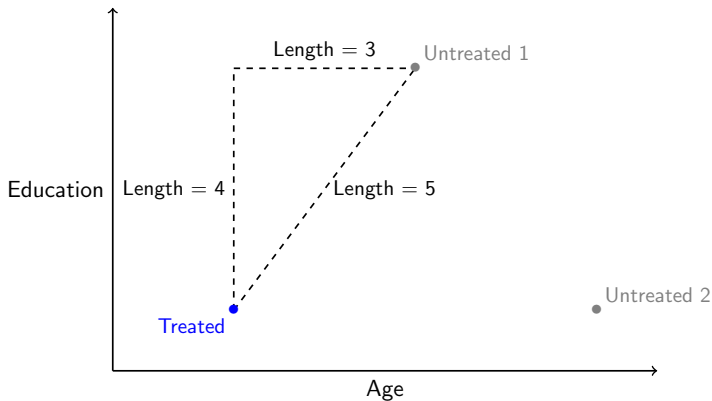
Which untreated unit should be the match?

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



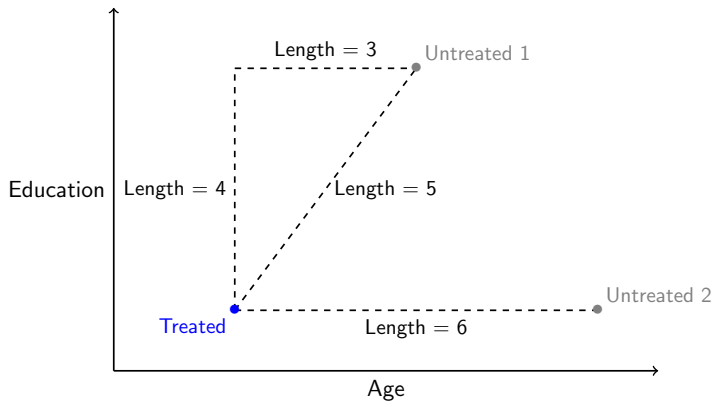
Which untreated unit should be the match?

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



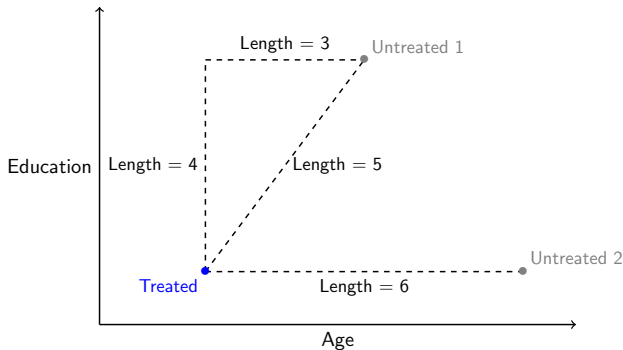
Which untreated unit should be the match?

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



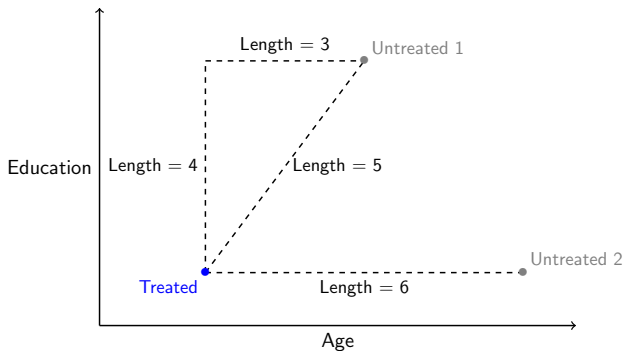
Which untreated unit should be the match?

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



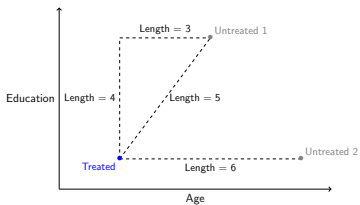
- Define a way to measure “distance” between two individuals as a single number

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

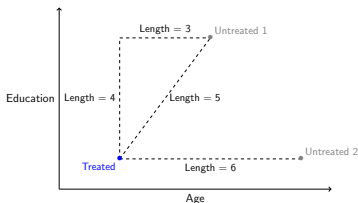


- ▶ Define a way to measure “distance” between two individuals as a single number
- ▶ Match individuals using that distance!

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



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



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

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
 - ▶ 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

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

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

Code

Let's try this out in R!

- ▶ Section 2 is worked out for you: read through, run the code blocks, and answer the questions
- ▶ Section 3 asks you to write some code (will be very similar to the code from Section 2)
- ▶ Then move on to the `matching_examples.Rmd` file on the website