Matching Lab

INFO/STSCI/ILRST 3900: Causal Inference

16 Oct 2024

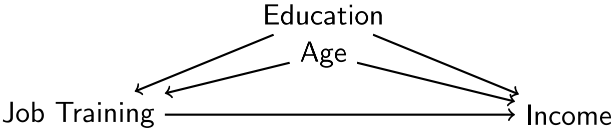
Agenda

- Review the solutions from last week's lab (5-10 minutes)
- Matching Review + Distance Metrics (10 minutes)
- Matching in R Exercise (25 minutes)
- Matching in R Discussion (5 minutes)

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?



Education

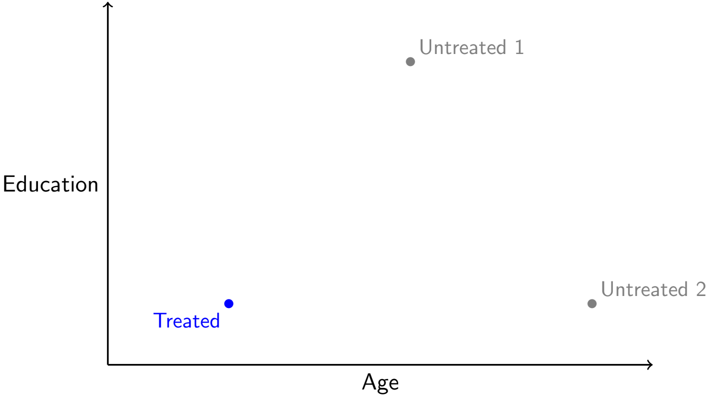
Age

Job Training

Income

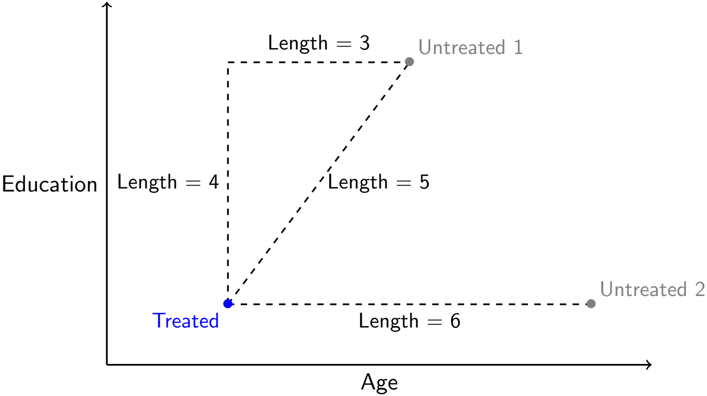
- Conditional exchangeability holds when conditioning on Age and Education!
- Matching: look for a group of untreated units which has a similar distribution of Age and Education to the treated group

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



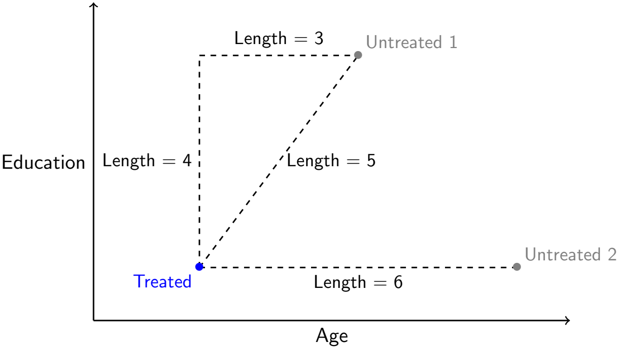
Which untreated unit should be the match?

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



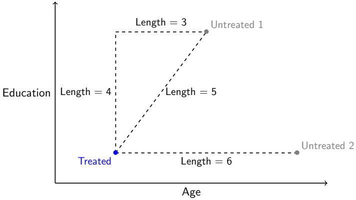
Which untreated unit should be the match?

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



- Define a way to measure “distance” between two individuals as a single number
- Match individuals using that 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: $d(i,j) = \sqrt{\sum_p (L_{pi} - L_{pj})^2}$
  - $d(\text{Treated, Untreated 1}) = \sqrt{3^2 + 4^2} = 5$ ✓
  - $d(\text{Treated, Untreated 2}) = \sqrt{6^2 + 0^2} = 6$
- 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

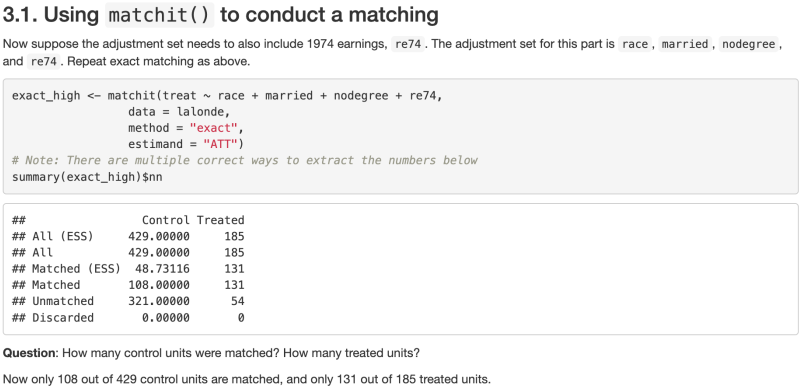
- Principle 1: Address unequal variances
  - Age might range uniformly from 18 to 80
  - Education range uniformly from 0 to 16
  - We might correct for this so age doesn't dominate the distance
- Principle 2: Address correlations
  - Suppose we included age in years, age in months, and education
  -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

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 (that will be very similar to the code from Section 2)
- If you finish early, move on to the `matching_examples.Rmd` file on the website.

3.1 Solution



3.1. Using `matchit()` to conduct a matching

Now suppose the adjustment set needs to also include 1974 earnings, `re74`. The adjustment set for this part is `race`, `married`, `nodegree`, and `re74`. Repeat exact matching as above.

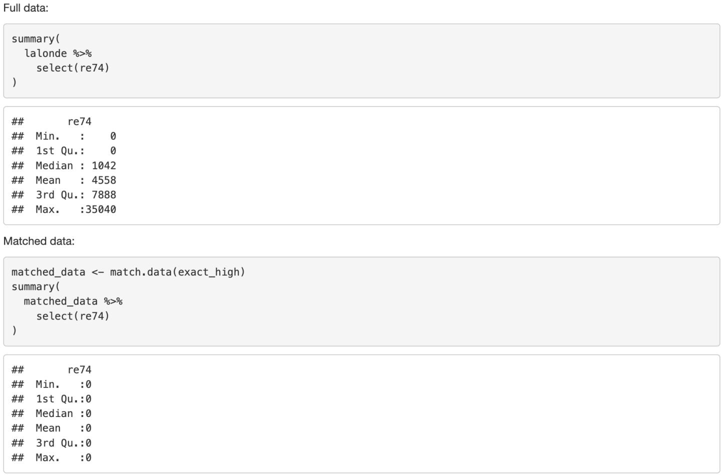
```r
exact_high <- matchit(treat ~ race + married + nodegree + re74,
                 data = lalonde,
                 method = "exact",
                 estimand = "ATT")
# Note: There are multiple correct ways to extract the numbers below
summary(exact_high)$nn
```

```
##                 Control Treated
## All (ESS)     429.00000     185
## All           429.00000     185
## Matched (ESS)  48.73116     131
## Matched       108.00000     131
## Unmatched     321.00000      54
## Discarded       0.00000       0
```

Question: How many control units were matched? How many treated units?

Now only 108 out of 429 control units are matched, and only 131 out of 185 treated units.

3.2 Solution and Discussion



Full data:

```
summary(
  lalonde %>%
    select(re74)
)
```

```
##       re74
##  Min.   :    0
##  1st Qu.:    0
##  Median : 1042
##  Mean   : 4558
##  3rd Qu.: 7888
##  Max.   :35040
```

Matched data:

```
matched_data <- match.data(exact_high)
summary(
  matched_data %>%
    select(re74)
)
```

```
##       re74
##  Min.   :0
##  1st Qu.:0
##  Median :0
##  Mean   :0
##  3rd Qu.:0
##  Max.   :0
```

What do you notice? What is different about the values of ‘re74‘ in the full data versus the matched data? Explain what happened and why it happened. Briefly interpret the result from 3.2: what is the drawback of using exact matching in this setting?