Data challenge & SHS: Causal Inference & Matching

Bénédicte Colnet

February 2023

Abstract

In this tutorial, you will learn how to apply matching for the estimation of causal effects from observational data using matching. A large part of this exercice is inspired from the vignette `MatchIt`.

Contents


```
knitr::opts_chunk$set(echo = TRUE)

# Set seed for reproducibility
set.seed(123)


# Load all packages needed to execute the job
# If the packages are not installed, write
# install.packages("<name of package>")

library(ggplot2) # Plot
library(MatchIt) # Matching
```

Matching is used in the context of estimating the causal effect of a binary treatment or exposure on an outcome while controlling for measured pre-treatment variables, typically confounding variables or variables prognostic of the outcome. Here and throughout the MatchIt documentation we use the word "treatment" to refer to the focal causal variable of interest, with "treated" and "control" reflecting the names of the treatment groups. The goal of matching is to produce covariate balance, that is, for the distributions of covariates in the two groups to be approximately equal to each other, as they would be in a successful randomized experiment. The importance of covariate balance is that it allows for increased robustness to the choice of model used to estimate the treatment effect; in perfectly balanced samples, a simple difference in means can be a valid treatment effect estimate.

A matching analysis involves four primary steps:

1) planning,
2) matching,
3) assessing the quality of matches,

and 4) estimating the treatment effect and its uncertainty.

Load Matching and data set

We will use Lalonde's data on the evaluation of the National Supported Work program to demonstrate `MatchIt`'s capabilities. The Lalonde study looked at the effectiveness of a job training program (the treatment)

on the real earnings of an individual, a couple years after completion of the program. The data consists of a number of demographic variables (age, race, academic background, and previous real earnings), as well as a treatment indicator, and the real earnings in the year 1978 (the response).

> Robert Lalonde, "Evaluating the Econometric Evaluations of Training Programs", American Economic Review, Vol. 76, pp. 604-620

First, we load `MatchIt` and bring in the `lalonde` dataset. See `?lalonde` for more information on the data set.

```
library("MatchIt")
library(table1)
```

```
##
## Attaching package: 'table1'
```

```
## The following objects are masked from 'package:base':
##
##     units, units<-
```

```
data("lalonde")

lalonde$treat <- as.factor(lalonde$treat)
lalonde$nodegree <- as.factor(lalonde$nodegree)

summary(lalonde)
```

```
##  treat        age             educ           race        married       nodegree
##  0:429   Min.   :16.00   Min.   : 0.00   black :243   Min.   :0.0000   0:227
##  1:185   1st Qu.:20.00   1st Qu.: 9.00   hispan: 72   1st Qu.:0.0000   1:387
##          Median :25.00   Median :11.00   white :299   Median :0.0000
##          Mean   :27.36   Mean   :10.27                Mean   :0.4153
##          3rd Qu.:32.00   3rd Qu.:12.00                3rd Qu.:1.0000
##          Max.   :55.00   Max.   :18.00                Max.   :1.0000
##       re74            re75             re78
##  Min.   :    0   Min.   :    0.0   Min.   :    0.0
##  1st Qu.:    0   1st Qu.:    0.0   1st Qu.:  238.3
##  Median : 1042   Median :  601.5   Median : 4759.0
##  Mean   : 4558   Mean   : 2184.9   Mean   : 6792.8
##  3rd Qu.: 7888   3rd Qu.: 3249.0   3rd Qu.:10893.6
##  Max.   :35040   Max.   :25142.2   Max.   :60307.9
```

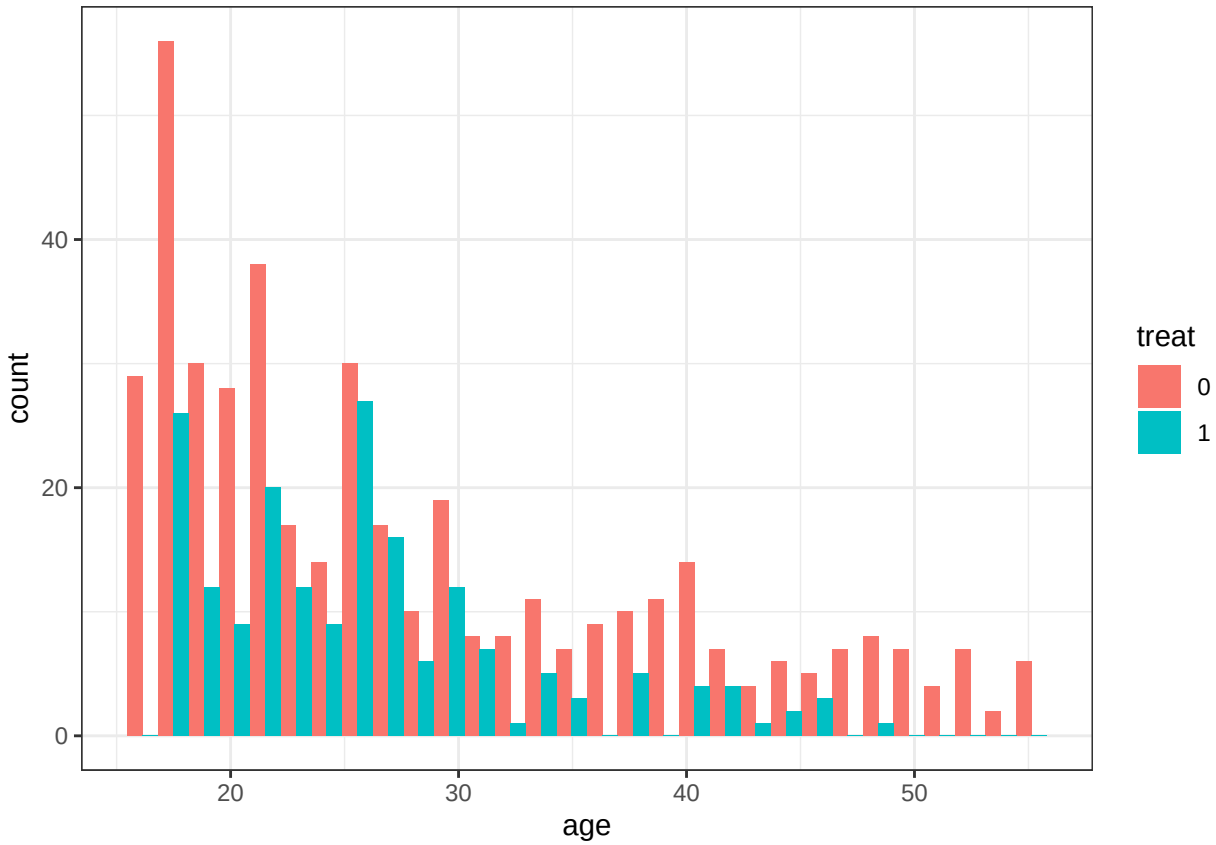
The statistical quantity of interest is the causal effect of the treatment (`treat`) on 1978 earnings (`re78`).

Planning

This is done prior to any data analysis. Ideally, draw a DAG. Great care should be taken to characterize the status of each covariates (pre-treatment, mediators, colliders). You can also check the initial imbalance.

```
ggplot(lalonde, aes(x = age, group = treat, fill = treat)) +
  geom_histogram(position = "dodge") +
  theme_bw()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
table1(~ race + age + married + educ | treat, data = lalonde, overall=F)
```

```
## Warning in !is.null(rmarkdown::metadata$output) && rmarkdown::metadata$output
## %in% : 'length(x) = 2 > 1' in coercion to 'logical(1)'
```

	0	1
	(N=429)	(N=185)
race		
black	87 (20.3%)	156 (84.3%)
hispan	61 (14.2%)	11 (5.9%)
white	281 (65.5%)	18 (9.7%)
age		
Mean (SD)	28.0 (10.8)	25.8 (7.16)
Median [Min, Max]	25.0 [16.0, 55.0]	25.0 [17.0, 48.0]
married		
Mean (SD)	0.513 (0.500)	0.189 (0.393)
Median [Min, Max]	1.00 [0, 1.00]	0 [0, 1.00]
educ		
Mean (SD)	10.2 (2.86)	10.3 (2.01)
Median [Min, Max]	11.0 [0, 18.0]	11.0 [4.00, 16.0]