002 Data analysis project - analyze data in R using propensity score matching

Iuliia Allaiarova

2024-03-19

For this assignment we will use data from Lalonde (1986), that aimed to evaluate the impact of National Supported Work (NSW) Demonstration, which is a labor training program, on post-intervention income levels. Interest is in estimating the causal effect of this training program on income. The data have n=614 subjects and 10 variables.

Below is a table describing the variables in the dataset:

Variable	Description
age	Age in years.
educ	Years of schooling.
black	Indicator variable for blacks (1 if black, 0 otherwise).
hispan	Indicator variable for Hispanics (1 if Hispanic, 0 otherwise).
married	Indicator variable for marital status (1 if married, 0 otherwise).
nodegree	Indicator variable for high school diploma (1 if no diploma, 0 otherwise).
re74	Real earnings in 1974.
re75	Real earnings in 1975.
re78	Real earnings in 1978 (the outcome variable).
treat	Indicator variable for treatment status (1 if received labor training, 0 otherwise).

Q1 Find the standardized differences for all of the confounding variables (pre-matching). What is the standardized difference for married (to nearest hundredth)?

```
table1_m = CreateTableOne(vars=xvars_m,strata="treat", data=mydf, test=FALSE)
## include standardized mean difference (SMD)
print(table1_m, smd=TRUE)
```

```
##                       Stratified by treat
##                        0                1                SMD
##   n                        429              185
##   age (mean (SD))        28.03 (10.79)    25.82 (7.16)     0.242
##   educ (mean (SD))       10.24 (2.86)     10.35 (2.01)     0.045
##   black (mean (SD))       0.20 (0.40)      0.84 (0.36)     1.668
##   hispan (mean (SD))      0.14 (0.35)      0.06 (0.24)     0.277
##   married (mean (SD))     0.51 (0.50)      0.19 (0.39)     0.719
```

```
##   nodegree (mean (SD))    0.60 (0.49)       0.71 (0.46)     0.235
##   re74 (mean (SD))     5619.24 (6788.75) 2095.57 (4886.62)  0.596
##   re75 (mean (SD))     2466.48 (3292.00) 1532.06 (3219.25)  0.287
```

The standardized difference for married is 0.72.

Q2 What is the raw (unadjusted) mean of real earnings in 1978 for treated subjects minus the mean of real earnings in 1978 for untreated subjects?

```
treat_means = tapply(mydf$re78, mydf$treat, mean)
print(treat_means)
```

```
##        0        1
## 6984.170 6349.144
```

It is -635.

Fit a propensity score model. Use a logistic regression model, where the outcome is treatment. Include the 8 confounding variables in the model as predictors, with no interaction terms or non-linear terms (such as squared terms). Obtain the propensity score for each subject.

Q3 What are the minimum and maximum values of the estimated propensity score?

```
pscore<-psmodel_m$fitted.values
min(pscore)
```

```
## [1] 0.009080193
```

```
max(pscore)
```

```
## [1] 0.8531528
```

The minimum and maximum values of the estimated propensity score are 0.009 and 0.85, respectively.

Carry out propensity score matching using the Match function. Before using the Match function, first do: `set.seed(931139)`.Setting the seed will ensure that you end up with a matched data set that is the same as the one used to create this work. Use options to specify pair matching, without replacement, no caliper.

Match on the propensity score itself, not logit of the propensity score. Obtain the standardized differences for the matched data.

Q4 What is the standardized difference for married?

```
##                       Stratified by treat
##                        0                 1                 SMD
##   n                        185               185
##   age (mean (SD))        24.21 (9.55)      25.82 (7.16)     0.190
##   educ (mean (SD))       10.23 (2.37)      10.35 (2.01)     0.052
##   black (mean (SD))       0.43 (0.50)       0.84 (0.36)     0.943
##   hispan (mean (SD))      0.06 (0.24)       0.06 (0.24)    <0.001
##   married (mean (SD))     0.20 (0.40)       0.19 (0.39)     0.027
##   nodegree (mean (SD))    0.69 (0.46)       0.71 (0.46)     0.035
##   re74 (mean (SD))     2681.77 (4754.79) 2095.57 (4886.62)  0.122
##   re75 (mean (SD))     1523.69 (2810.24) 1532.06 (3219.25)  0.003
```

The standardized difference for married is 0.27.

Q5. For the propensity score matched data: Which variable has the largest standardized difference?

Variable 'black' has the largest standardized difference of 0.943.

Re-do the matching, but use a caliper this time. Set the caliper=0.1 in the options in the Match function. Again, before running the Match function, set the seed: `set.seed(931139)`.

Q6. How many matched pairs are there?

```
##                         Stratified by treat
##                          0                  1                  SMD
##   n                          111                111
##   age (mean (SD))          26.27 (11.10)      26.22 (7.18)      0.006
##   educ (mean (SD))         10.37 (2.66)       10.25 (2.31)      0.047
##   black (mean (SD))         0.72 (0.45)        0.74 (0.44)      0.040
##   hispan (mean (SD))        0.11 (0.31)        0.10 (0.30)      0.029
##   married (mean (SD))       0.24 (0.43)        0.24 (0.43)     <0.001
##   nodegree (mean (SD))      0.66 (0.48)        0.65 (0.48)      0.019
##   re74 (mean (SD))       2704.56 (4759.89) 2250.49 (5746.14)  0.086
##   re75 (mean (SD))       1969.10 (3169.08) 1222.25 (3081.19)  0.239
```

There are 111 matched pairs.

Use the matched data set (from propensity score matching with caliper=0.1) to carry out the outcome analysis. Q7. For the matched data, what is the mean of real earnings in 1978 for treated subjects minus the mean of real earnings in 1978 for untreated subjects?

```
## [1] 6151.181
```

```
## [1] 4904.375
```

It is 1246.81.

Q8. Use the matched data set (from propensity score matching with caliper=0.1) to carry out the outcome analysis. Carry out a paired t-test for the effect of treatment on earnings. What are the values of the 95% confidence interval?

```
#paired t-test
(t.test(diffy_m2))$conf.int
```

```
## [1] -420.0273 2913.6398
## attr(,"conf.level")
## [1] 0.95
```

The values of the 95% confidence interval are (-420.0273, 2913.6398).