Causal Analysis

Steph Low

2024-10-17

Goal: evaluate the impact of National Supported Work (NSW) Demonstration, which is a labor training program on post-intervention income levels. The interest is in estimating the causal effect of this training program on income.

The Lalonde data set has 614 subjects and 10 variables:

- `age` : age in years
- `educ` : years of schooling
- `black` : indicator variable for blacks
- `hispan` : indicator variable for Hispanics
- `married` : indicator variable for marital status
- `nodegree` : indicator variable for high school diploma
- `re74` : real earnings in 1974
- `re75` : real earnings in 1975
- `re78` : real earnings in 1978 (the outcome variable; post-intervention income)
- `treat` : indicator variable for treatment status

Potential confounding variables are `age` , `educ` , `black` , `hispan` , `married` , `nodegree` , `re74` , and `re75` .

Data Preparation

First I need to prepare the data.

```
# load data
data(lalonde)

# view data
# View(lalonde)

# convert indicator variables to numeric and prep data
age <- lalonde$age
educ <- lalonde$educ
re74 <- lalonde$re74
re75 <- lalonde$re75
married <- lalonde$married
nodegree <- lalonde$nodegree
black <- as.numeric(lalonde$race=="black") # 1 for black, 0 otherwise
hispan <- as.numeric(lalonde$race=="hispan") # 1 for hispanic, 0 otherwise
treatment <- lalonde$treat
outcome <- lalonde$re78

# create a dataset with these variables for simplicity
data <- cbind(age, educ, re74, re75, married, nodegree, black, hispan, treatment, outcome)
data <- data.frame(data)

# get covariates
xvars <- c("age", "educ", "black", "hispan", "married", "nodegree", "re74", "re75")
```

Unmatched Data

Now that the data is prepared, I want to find the standardized mean differences (SMD) for all the confounding variables prior to matching.

```
# create Table1, pre-matching
table1 <- CreateTableOne(vars=xvars            # covariates to be summarized
                         , strata="treatment"   # stratifying the groups by treatment
                         , data=data
                         , test=FALSE           # don't do groupwise comparisons
                         )

# include the standardized mean difference (SMD)
print(table1, smd=TRUE)
```

```
##                       Stratified by treatment
##                        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
```

Next I want to find the raw mean of real earnings in 1978 (outcome) for treated subjects minus the mean of real earnings in 1978 for untreated subjects.

```
# calculate means by treatment group
raw_means <- aggregate(x=data$outcome
                       , by = list(data$treatment)
                       , FUN = mean)

# calculate the difference between real earnings of the treated subjects minus the untreated subjects
raw_means$x[raw_means$Group.1==1] - raw_means$x[raw_means$Group.1==0]
```

```
## [1] -635.0262
```

Propensity Score Estimation

Next I want to fit a propensity score model. I'll estimate the propensity score using logistic regression where the outcome is treatment, including the 8 confounding variables in the model as predictors, with no interaction terms or non-linear terms (such as squared terms).

The propensity score for each subject is the probability of receiving treatment given the covariates.

```
# fit a propensity score model using logistic regression
psmodel <- glm(treatment~age+educ+re74+re75+married+nodegree+black+hispan
               , family=binomial()
               , data=data
               )

# show model summary
summary(psmodel)
```

```
## 
## Call:
## glm(formula = treatment ~ age + educ + re74 + re75 + married + 
##     nodegree + black + hispan, family = binomial(), data = data)
## 
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)    
## (Intercept) -4.729e+00  1.017e+00  -4.649 3.33e-06 ***
## age          1.578e-02  1.358e-02   1.162  0.24521    
## educ         1.613e-01  6.513e-02   2.477  0.01325 *  
## re74        -7.178e-05  2.875e-05  -2.497  0.01253 *  
## re75         5.345e-05  4.635e-05   1.153  0.24884    
## married     -8.321e-01  2.903e-01  -2.866  0.00415 ** 
## nodegree     7.073e-01  3.377e-01   2.095  0.03620 *  
## black        3.065e+00  2.865e-01  10.699  < 2e-16 ***
## hispan       9.836e-01  4.257e-01   2.311  0.02084 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 751.49  on 613  degrees of freedom
## Residual deviance: 487.84  on 605  degrees of freedom
## AIC: 505.84
## 
## Number of Fisher Scoring iterations: 5
```

```
# estimate the propensity scores
pscore <- psmodel$fitted.values

# find the min and max estimated propensity scores
min(pscore)
```

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

```
max(pscore)
```

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

Propensity Score Matching

Next I want to match subjects on their propensity scores.

In this case I will pair match:

- without replacement
- without a maximum distance tolerated for matching (no caliper)
- on the propensity score instead of the logit of the propensity score.

Once the matching is done, I will find the standardized mean differences for the matched data.

```
# set seed for reproducibility
set.seed(931139)

# match on propensity score
psmatch <- Match(Tr=data$treatment     # treatment
                 , M=1                  # pair matching
                 , X=pscore             # variables to match on (estimated propensity scores)
                 , replace=FALSE        # no replacement
                 )

# extracted the matched data
matchedData <- data[unlist(psmatch[c("index.treated", "index.control")]),]

# create Table1, post-matching
matchedTable1 <- CreateTableOne(vars=xvars              # covariates to be summarized
                                , strata="treatment"   # stratifying the groups by treatment
                                , data=matchedData
                                , test=FALSE           # don't do groupwise comparisons
                                )

# get SMD
print(matchedTable1, smd=TRUE)
```

```
##                       Stratified by treatment
##                        0                 1                 SMD
##   n                        185               185
##   age (mean (SD))        25.29 (10.65)     25.82 (7.16)     0.058
##   educ (mean (SD))       10.55 (2.71)      10.35 (2.01)     0.084
##   black (mean (SD))       0.47 (0.50)       0.84 (0.36)     0.852
##   hispan (mean (SD))      0.21 (0.41)       0.06 (0.24)     0.453
##   married (mean (SD))     0.20 (0.40)       0.19 (0.39)     0.027
##   nodegree (mean (SD))    0.65 (0.48)       0.71 (0.46)     0.127
##   re74 (mean (SD))     2351.12 (4192.62) 2095.57 (4886.62)  0.056
##   re75 (mean (SD))     1605.02 (2601.68) 1532.06 (3219.25)  0.025
```

Next I'll do the same propensity matching, but this time with a caliper of 0.1 to limit the maximum tolerated distance for matching.

- a common caliper value is `caliper = 0.2 * SD(logit(propensityScore))`

```
# set seed for reproducibility
set.seed(931139)

# match on propensity score
psmatch2 <- Match(Tr=data$treatment    # treatment
                 , M=1                  # pair matching
                 , X=pscore             # variables to match on (estimated propensity scores)
                 , replace=FALSE        # no replacement
                 , caliper=0.1
                 )

# extracted the matched data
matchedData2 <- data[unlist(psmatch2[c("index.treated", "index.control")]),]

# create Table1, post-matching
matchedTable2 <- CreateTableOne(vars=xvars              # covariates to be summarized
                                , strata="treatment"   # stratifying the groups by treatment
                                , data=matchedData2
                                , test=FALSE           # don't do groupwise comparisons
                                )

# get SMD
print(matchedTable2, smd=TRUE)
```

```
##                       Stratified by treatment
##                        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
```

I'll again do some propensity matching, but this time with the logit of the propensity score and a caliper.

```
# set seed for reproducibility
set.seed(931139)

# match on propensity score
psmatch3 <- Match(Tr=data$treatment    # treatment
                 , M=1                  # pair matching
                 , X=logit(pscore)      # variables to match on (estimated propensity scores)
                 , replace=FALSE        # no replacement
                 , caliper=0.1
                 )

# extracted the matched data
matchedData3 <- data[unlist(psmatch2[c("index.treated", "index.control")]),]

# create Table1, post-matching
matchedTable3 <- CreateTableOne(vars=xvars              # covariates to be summarized
                                , strata="treatment"   # stratifying the groups by treatment
                                , data=matchedData3
                                , test=FALSE           # don't do groupwise comparisons
                                )

# get SMD
print(matchedTable3, smd=TRUE)
```

```
##                       Stratified by treatment
##                        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
```

Outcome Analysis

Next I'll do the outcome analysis.

```
# outcome analysis
y_treatment <- matchedData2$outcome[matchedData2$treatment==1]
y_control <- matchedData2$outcome[matchedData2$treatment==0]

# pairwise difference
diff <- y_treatment - y_control

# mean of real earnings in 1978 for treated subjects minus control subjects
mean(diff)
```

```
## [1] 1246.806
```

```
# paired t test
t.test(diff)
```

```
## 
##  One Sample t-test
## 
## data:  diff
## t = 1.4824, df = 110, p-value = 0.1411
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -420.0273 2913.6398
## sample estimates:
## mean of x 
##  1246.806
```

```
# let's see how this all differs with the logit-matched data

# outcome analysis
y_treatment2 <- matchedData3$outcome[matchedData3$treatment==1]
y_control2 <- matchedData3$outcome[matchedData3$treatment==0]

# pairwise difference
diff2 <- y_treatment2 - y_control2

# mean of real earnings in 1978 for treated subjects minus control subjects
mean(diff2)
```

```
## [1] 1246.806
```

```
# paired t test
t.test(diff2)
```

```
## 
##  One Sample t-test
## 
## data:  diff2
## t = 1.4824, df = 110, p-value = 0.1411
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -420.0273 2913.6398
## sample estimates:
## mean of x 
##  1246.806
```

There wasn't a difference in the outcome based off the different matched sets I did.

With a p-value of `0.1411` , I cannot reject the null hypothesis that there is no treatment effect. The NSW may not have had any impact on post-intervention income levels.

For more reading on this this topic, check out this example.