P8122 HW3

Sabrina Lin stl2137

11/21/2020

Part 1

```
### read in data
salary_dat <- read_csv("/Users/SabrinaLin/Documents/Fall_2020_Causal_Inference/Homework/HW3/p8122_hw3_s
  mutate(
    treat = as.factor(treat),
    black = as.factor(black),
    hispan = as.factor(hispan),
    married = as.factor(married),
    nodegree = as.factor(nodegree)
  )
```

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
## Parsed with column specification:
## cols(
##   X1 = col_character(),
##   treat = col_double(),
##   age = col_double(),
##   educ = col_double(),
##   black = col_double(),
##   hispan = col_double(),
##   married = col_double(),
##   nodegree = col_double(),
##   re74 = col_double(),
##   re75 = col_double(),
##   re78 = col_double()
## )
```

- data consists of 10 variables measured for each individual:
  - an indicator of treatment assignment (job training), `treat`
  - age in years, `age`
  - education in years, `educ`
  - an indicator for African-American, `black`
  - an indicator for Hispanic, `hispan`
  - an indicator for married, `married`
  - an indicator for high school degree, `nodegree`
  - income in 1974, `re74`

  - income in 1975 `re75`
  - income in 1978, `re78`
- The variable `treat` is the treatment and the variables `re78` is the outcome.

Subpart 1

Write the DAG representing this observational study including all variables provided. Describe all the variables in the graph.

Subpart 2

Evaluate covariate balance in this observational study. Show a table or a plot. Interpret the results.

```r
## Construct a table
vars <- c("age", "educ", "black", "hispan", "married", "nodegree", "re74", "re75", "re78")

tab_presub <- CreateTableOne(vars = vars, strata = "treat", data = salary_dat, test = FALSE)

print(tab_presub, 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 = 1 (%)            87 (20.3)        156 (84.3)     1.671
##   hispan = 1 (%)           61 (14.2)         11 ( 5.9)     0.277
##   married = 1 (%)         220 (51.3)         35 (18.9)     0.721
##   nodegree = 1 (%)        256 (59.7)        131 (70.8)     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
##   re78 (mean (SD)) 6984.17 (7294.16) 6349.14 (7867.40)  0.084
```

- Given that we would like the SMD to be less than 0.2 and having seen 0.25 as a common guideline for SMD in the literature, there are several variables that surpass this rule of thumb. The SMD (standardized mean difference) for the variable `black` is very large at 1.671, indicating that the covariate balance for this variable is not good. The variables `married` and `re74` also have relatively large SMDs (0.721 and 0.596 respectively), also indicating that the covariate balance for these variables is not great.

Subpart 3

The propensity score is defined as the probability of receiving the treatment given the ob- served covariates. These scores are used to construct strata within which we assume that the exposure assignment is random. Construct propensity scores by fitting a logistic regression to the data.