Homework 4

Yuki Joyama

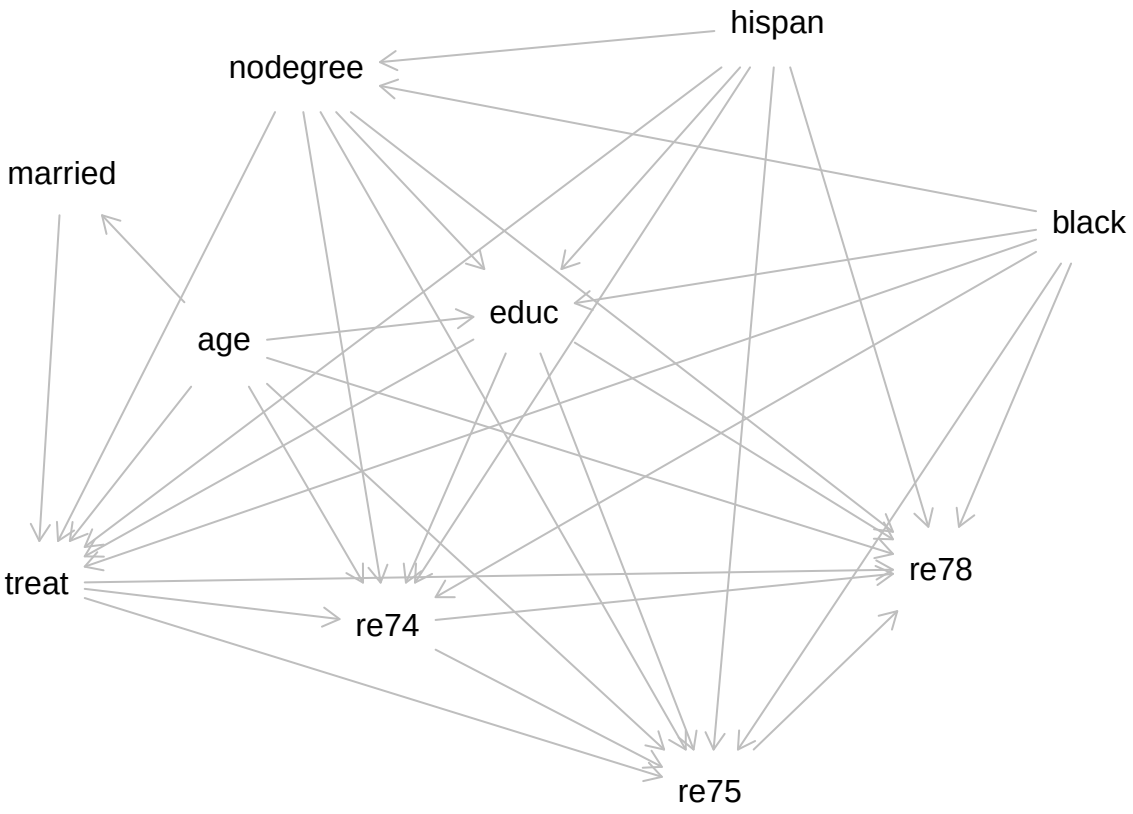
```
# libraries
library(tidyverse)
library(ggplot2)
library(dagitty)
library(tableone)
library(knitr)

# setup plot theme
theme_set(
  theme_bw() +
    theme(legend.position = "top")
  )
```

```
# import data
df = read_csv("./data/hw4_data.csv")
```

1. DAG

```
g = dagitty('dag {
age [pos="-1.842,-0.039"]
black [pos="1.458,-0.317"]
educ [pos="-0.698,-0.114"]
hispan [pos="0.269,-0.760"]
married [pos="-2.461,-0.426"]
nodegree [pos="-1.567,-0.659"]
re74 [pos="-1.219,0.594"]
re75 [pos="0.011,0.971"]
re78 [outcome,pos="0.893,0.468"]
treat [exposure,pos="-2.558,0.500"]
age -> educ
age -> married
age -> re74
age -> re75
age -> re78
age -> treat
black -> educ
black -> nodegree
black -> re74
black -> re75
black -> re78
black -> treat
educ -> re74
educ -> re75
```

```
educ -> re78
educ -> treat
hispan -> educ
hispan -> nodegree
hispan -> re74
hispan -> re75
hispan -> re78
hispan -> treat
married -> treat
nodegree -> educ
nodegree -> re74
nodegree -> re75
nodegree -> re78
nodegree -> treat
re74 -> re75
re74 -> re78
re75 -> re78
treat -> re74
treat -> re75
treat -> re78
}')
plot(g)
```



Note of the variables in the DAG:
`treat`: treatment assignment (job training); exposure
- Suppose that the job training was completed before 1974, `treat` is likely to infludence `re74`, `re75` and

re78
age: age in years
- age may affect married, treat, educ and all the income status (re74, re75, re78)
educ: education in years
- educ may affect all the income status and treat
black, hispan: indicators for African American and hispanic
- Both of the ethnicity indicator may affect all the income status, nodegree, treat, and educ
married: indicator for married
- married could influence treat
nodegree: indicator for highschool degree
- nodegree may affect treat, educ, and all the income status
re74: income in 1974
- re74 can influence re75 and re78
re75: income in 1975
- re75 can influence re78
re78: income in 1978; outcome

Given the DAG, covariates that need to be adjusted in investigating the effect of exposure on the outcome are nodegree, hispan, black, educ and age.

2. I will evaluate the covariate balance using standardized mean differences.

```
cov = c("nodegree", "hispan", "black", "educ", "age")

# construct a table
tab = CreateTableOne(vars = cov, strata = "treat", data = df, test = FALSE)
print(tab, smd = TRUE)
```

```
##                         Stratified by treat
##                          0             1             SMD
##   n                        429           185
##   nodegree (mean (SD))  0.60 (0.49)   0.71 (0.46)  0.235
##   hispan (mean (SD))    0.14 (0.35)   0.06 (0.24)  0.277
##   black (mean (SD))     0.20 (0.40)   0.84 (0.36)  1.668
##   educ (mean (SD))     10.24 (2.86)  10.35 (2.01)  0.045
##   age (mean (SD))      28.03 (10.79) 25.82 (7.16)  0.242
```

We can see that most covariates have SMD > 0.1 except for educ, indicating the potential imbalance between the two treatment group.

3. Propensity score estimates are calculated by fitting a logistic regression.

```
# fit PS model
ps.fit <- glm(as.factor(treat) ~ as.factor(nodegree) + as.factor(hispan) + as.factor(black) + educ + ag

ps.fit |>
  broom::tidy() |>
  kable()
```

term	estimate	std.error	statistic	p.value
(Intercept)	-4.6689752	1.0063696	-4.6394238	0.0000035

term	estimate	std.error	statistic	p.value
as.factor(nodegree)1	0.7812141	0.3303690	2.3646714	0.0180461
as.factor(hispan)1	1.0760070	0.4168588	2.5812269	0.0098450
as.factor(black)1	3.3041742	0.2811462	11.7525129	0.0000000
educ	0.1523227	0.0643134	2.3684437	0.0178631
age	-0.0054560	0.0125728	-0.4339522	0.6643232

```r
# estimate PS
df.ps <- predict(ps.fit, type = 'response')
print(df.ps[1:50])
```

```
##          1          2          3          4          5          6          7
## 0.70890885 0.17352013 0.57429368 0.72003809 0.61181465 0.66089500 0.58360334
##          8          9         10         11         12         13         14
## 0.71450603 0.72158352 0.04647944 0.66455356 0.62265153 0.63106456 0.68833095
##         15         16         17         18         19         20         21
## 0.59626339 0.69761702 0.61492967 0.69299354 0.56090387 0.57962038 0.72441625
##         22         23         24         25         26         27         28
## 0.05951577 0.06156125 0.72332569 0.69876670 0.71783314 0.72223248 0.19219049
##         29         30         31         32         33         34         35
## 0.69991390 0.72332569 0.69991390 0.44122686 0.72223248 0.58757541 0.58094921
##         36         37         38         39         40         41         42
## 0.63147843 0.51040509 0.58360334 0.59484193 0.69183153 0.58625263 0.06782935
##         43         44         45         46         47         48         49
## 0.63233393 0.15559436 0.72767188 0.72223248 0.63233393 0.66698171 0.51040509
##         50
## 0.58360334
```

Listed values are the propensity scores for each observation in the dataset (only showing the first 50 observations out of 614).