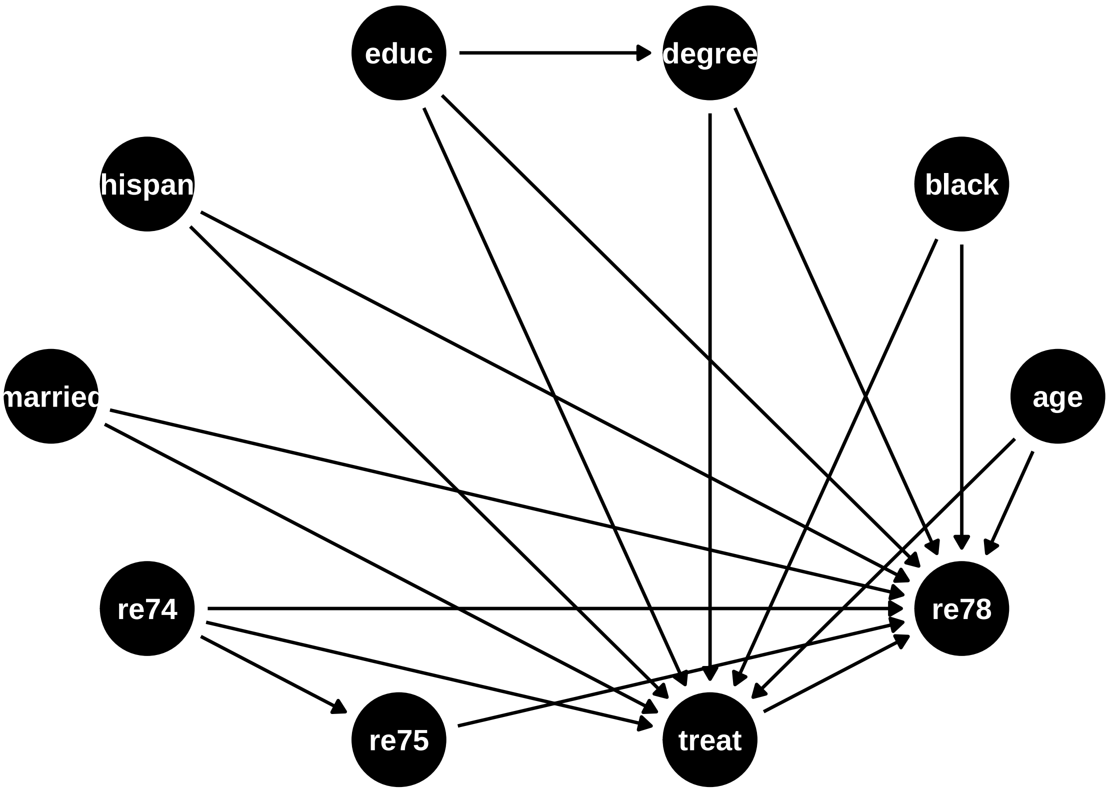
Homework 3

Ngoc Duong

11/22/2020

Question 1

1.1. Write the DAG



Describe the variables

We have the main outcome of income in the year 1978, and the main exposure as an indicator of job training (in the year 1974). We expect there is a causal relationship between additional job training in 1974 and income in 1978.

A couple other income variables are re74 (income in the year 1974) and re75 (income in the year 1975). For re74, we expect this to be associated with both job training and income in 1978. This is because incomes over the year tend to be intercorrelated (depending on profession, skills), and income in 1974 might affect whether an individual seek job training, for example to improve earning prospects in later year for people in lower income range. For re75 – income in 1975, temporally, this was measured after the job training has been implemented in 1974. Therefore, we can expect this to be on the causal pathway between “treatment” (job training) and “outcome” (re78).

We have some demographic variables such as: age (age in years), educ (years of education), married

(married indicator), nodegree (high school degree completion indicator), black (indicator of being African-American/black), hispan (indicator of being Hispanic). These variables can reasonably affect the outcome (income in year 1978) as many studies have documented links between race, age, certification, and family commitment and earning potential. Additionally, education years might also be associated with high school degree status. But it is not fully mediated by high school degree status because it might serve as proxy for other unmeasured confounders such as knowledge, skills, experience, social capital, etc. which all affected income not completly through degree/certification.

1.2. Evaluate covariate balance

```
##                   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
```

Interpret

The table above can give us a sense of the distribution of covariates across the levels of treatment. Specifically, in treatment group 0, there are 429 subjects and in treatment group 1, there are 185 subjects. The summary statistics (mean and SD for continuous variables, and count and proportion for discrete variables) were reported for each of the covariate in each treatment stratum. The standardized mean difference (SMD) computed from two treatment groups can be viewed as a measure for difference/imbalance.

Ideally, we would want the SMD to be close to 0 (sometimes the cut-off <0.2 or < 0.25 is used), but here we can see that except for education (in years) and age (in years), other variables all have larger SMD than desired, which suggests some covariate imbalance.

For example, the mean age of the no-job-training group is 28.03 years (SD = 10.79 years) and the mean age of the job-training group is 25.82 years (SD = 7.16 years). The SMD is 0.242 > 0.2, which might indicate slight imbalance.

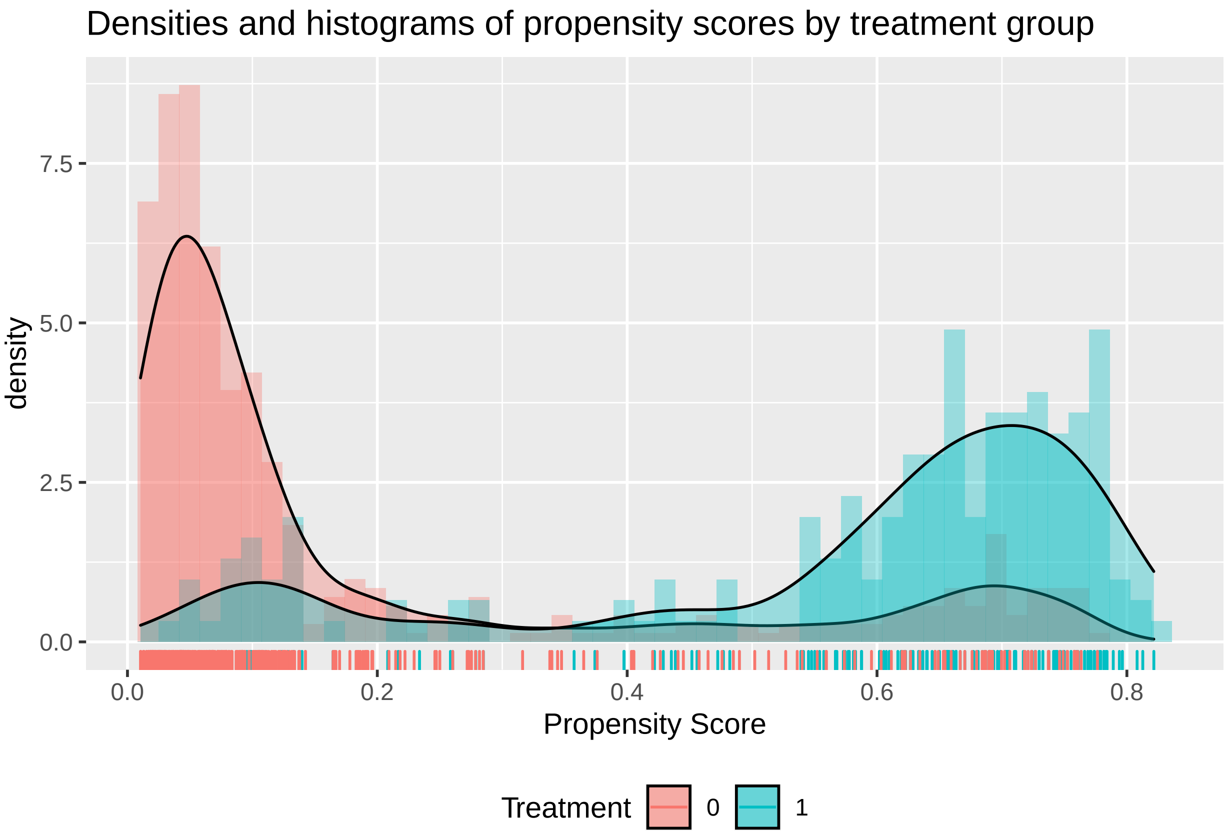
The goal of propensity score matching/subclassification is to reduce the SMD within each subclass.

1.3 Construct propensity scores

```
##
## Call:
## glm(formula = treat ~ age + educ + black + hispan + married +
##     nodegree + re74, family = binomial, data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7310  -0.4776  -0.2967   0.7480   2.6446
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.724e+00  1.020e+00  -4.632 3.61e-06 ***
## age          1.483e-02  1.353e-02   1.096  0.27311
## educ         1.628e-01  6.535e-02   2.491  0.01272 *
## black1       3.078e+00  2.864e-01  10.747  < 2e-16 ***
## hispan1      1.015e+00  4.236e-01   2.397  0.01655 *
## married1    -7.585e-01  2.819e-01  -2.690  0.00714 **
## nodegree1    7.302e-01  3.374e-01   2.164  0.03043 *
## re74        -5.336e-05  2.336e-05  -2.284  0.02234 *
## ---
## 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: 489.20  on 606  degrees of freedom
## AIC: 505.2
##
## Number of Fisher Scoring iterations: 5
```

I fitted a logistic regression with the outcome being the treatment status, and covariates being potential confounders that were identified in the DAG in part 1. I did not include re75, because given the temporal relationship between the variables, this information does not help explain the treatment – job training status in 1974, and so is also not a confounder for "re78" and "treat".

1.4. Evaluate overlap and trim data if necessary



From the Densities and histograms for propensity scores above, we can see that there are the extreme upper portion of the treatment 1 group that have no overlap with the treatment group 0. This means for these subjects in treatment group 1, we have no information about the similar subjects in treatment group 0 to recover the counterfactuals. Therefore, in order to ensure exposure assignment in random in all "blocks", we need to trim data.

In the trimming process, we can eliminate the treated subjects whose P(A=1|C) is greater that the maximum P(A=1|C) found in the untreated group, and similarly, eliminate the untreated subjects whose P(A=0|C) is lower than the minimum P(A=0|C) found in the the treated group.

```
## [1] 614  11
```

```
## [1] 518  11
```

Here, applying the above rule, I trimmed 96 subjects.

Once we have trimmed the data, we can obtain better covariate balance. However, given a highly unbalanced sample, trimming can result in small observartions at some intersection(s) of levels of covariates. Due to lowered sample size, this might lead to lower power and generalizability is also reduced. The collapsing of covariates into one score may account for potential confounders that are unmeasured, which might improve statistical efficiency.

1.5. Evaluate imbalance in trimmed sample

```
##                    Stratified by treat
##                     0                 1                 SMD
##   n                     349               169
##   age (mean (SD))     26.67 (10.54)     25.36 (7.03)     0.147
##   educ (mean (SD))    10.24 (2.78)      10.24 (2.03)     0.003
##   black = 1 (%)          87 (24.9)        140 (82.8)     1.427
##   hispan = 1 (%)         61 (17.5)         11 ( 6.5)     0.343
##   married = 1 (%)       140 (40.1)         35 (20.7)     0.432
##   nodegree = 1 (%)      222 (63.6)        117 (69.2)     0.119
##   re74 (mean (SD)) 3513.51 (4853.05) 2293.97 (5069.07)  0.246
##   re75 (mean (SD)) 1954.47 (2802.33) 1622.95 (3302.63)  0.108
```

We can see that the SMDs seem to be lower than before, which is what we desire with trimming. Now, we have age, education years, high school degree status are relatively balanced. However, black, hispan, and married variables still have relatively high SMDs. Notice that the sample size also decreases after we left out subjects with non-overlapping extreme propensitiy scores.

1.6. Use subclassification to balance covariates between treated and controls

- Process: I started out with the quintiles 0.2, 0.4, 0.6, 0.8. However, the first quintiles contains too few observations in the treated group (less than 5), which might make the subjects not exactly exchangeable within this subclass. So I combined the first two bins and chose 0.4 as the cut-off quantile for the first subclass, keeping 0.6 and 0.8 the same as by looking at the propensity score histogram and densities, these seem to be close enough.

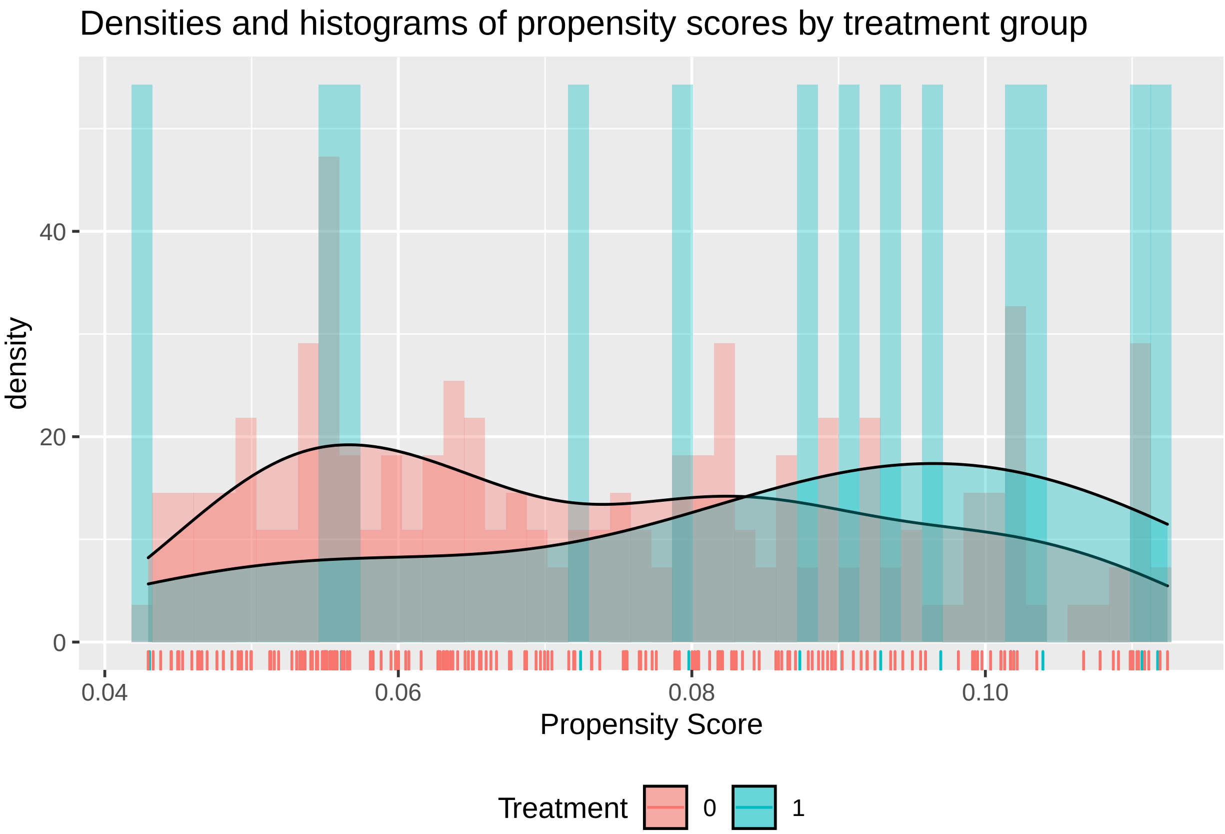
The breaks the associated propensity scores are as follows:

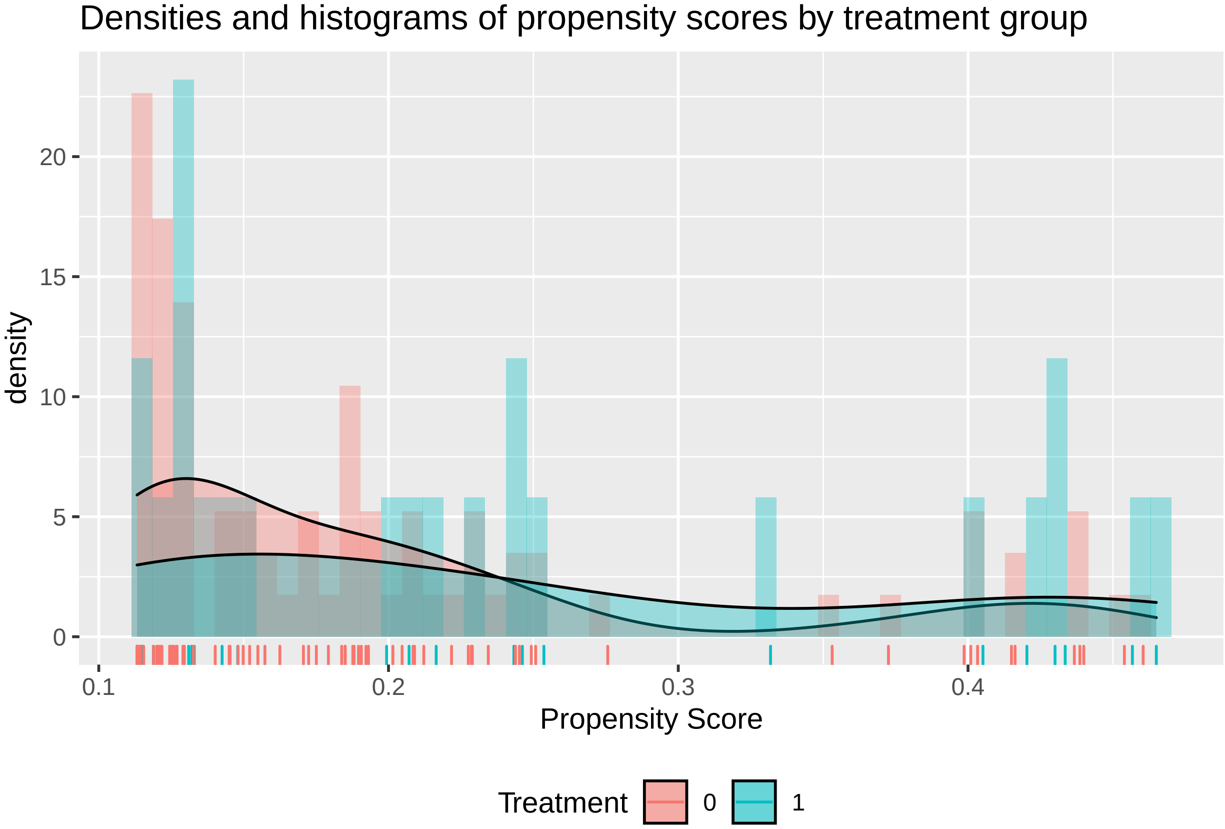
```
##       40%       60%       80% 
## 0.1130462 0.4650921 0.6447289
```

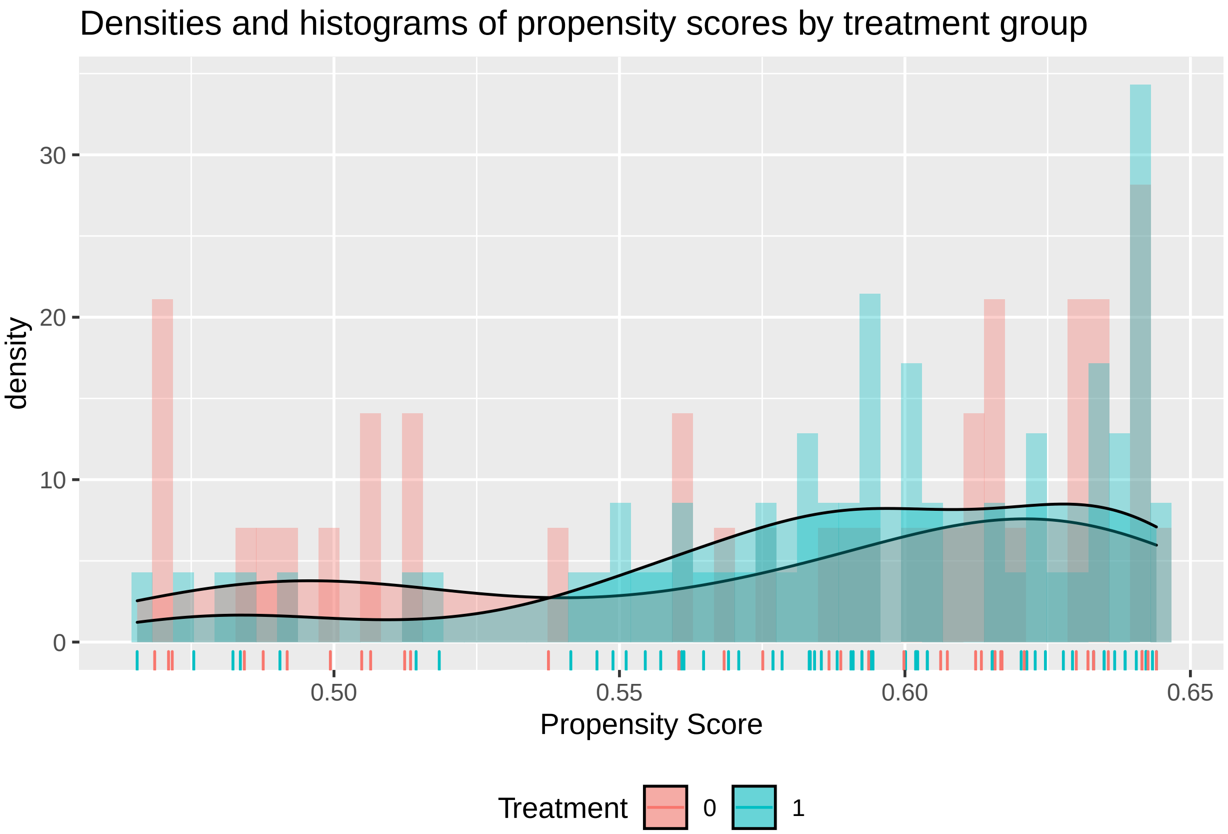
Below is the sample sizes within each subclass

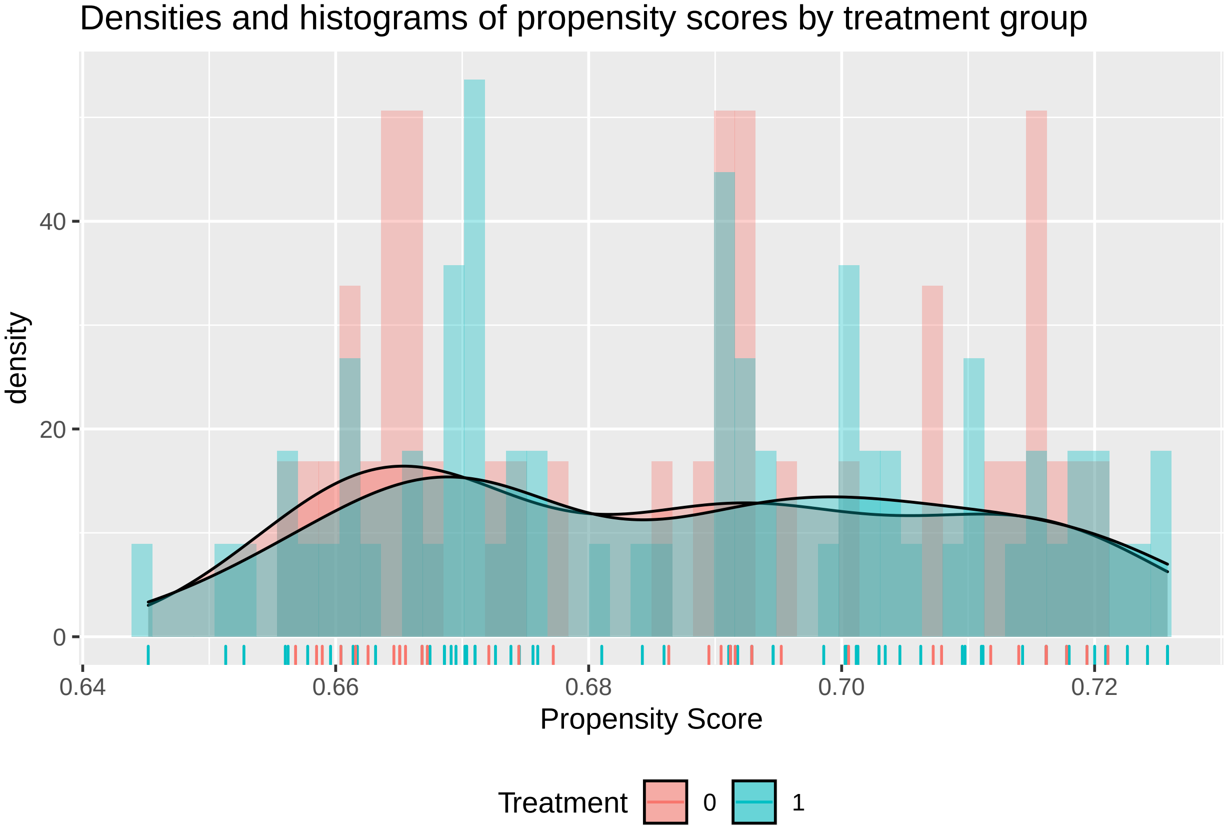
```
##    subclass
##       0   1   2   3
##   0 194  80  39  36
##   1  13  24  64  68
```

Below is the densities and histograms of propensity scores by treatment group for these 4 subclasses


Densities and histograms of propensity scores by treatment group
density
40
20
0
0.04
0.06
0.08
0.10
Propensity Score
Treatment
0
1


Densities and histograms of propensity scores by treatment group
density
20
15
10
5
0
0.1
0.2
0.3
0.4
Propensity Score
Treatment
0
1





- Inspect covariate balance in each subclass.

```
##                   Stratified by treat
##                     0                 1                 SMD
##   n                    194               13
##   age (mean (SD))    27.94 (10.59)     24.85 (5.89)      0.361
##   educ (mean (SD))   10.04 (2.85)      11.08 (1.85)      0.433
##   black = 1 (%)          0 ( 0.0)          0 ( 0.0)    <0.001
##   hispan = 1 (%)        22 (11.3)          1 ( 7.7)     0.125
##   married = 1 (%)      108 (55.7)          4 (30.8)     0.519
##   nodegree = 1 (%)     120 (61.9)          4 (30.8)     0.656
##   re74 (mean (SD)) 4364.12 (4982.25) 1844.84 (3118.68)  0.606
##   re75 (mean (SD)) 2225.15 (2819.92) 1687.92 (2823.71)  0.190
```

```
##                   Stratified by treat
##                     0                 1                 SMD
##   n                     80               24
##   age (mean (SD))    24.48 (9.25)      26.92 (7.09)      0.296
##   educ (mean (SD))   10.80 (2.60)      10.08 (2.15)      0.301
##   black = 1 (%)         12 (15.0)          8 (33.3)     0.438
##   hispan = 1 (%)        39 (48.8)         10 (41.7)     0.143
##   married = 1 (%)       19 (23.8)         10 (41.7)     0.389
##   nodegree = 1 (%)      52 (65.0)         18 (75.0)     0.220
##   re74 (mean (SD)) 2661.32 (4358.51) 5581.40 (8733.30)  0.423
##   re75 (mean (SD)) 1942.03 (2984.69) 3892.77 (5238.58)  0.458
```

```
##                   Stratified by treat
##                     0                 1                 SMD
##   n                     39               64
##   age (mean (SD))    29.08 (12.53)     25.72 (6.98)      0.331
##   educ (mean (SD))    9.97 (3.43)      10.11 (2.44)      0.045
##   black = 1 (%)         39 (100.0)        64 (100.0)   <0.001
##   hispan = 1 (%)         0 (  0.0)         0 (  0.0)   <0.001
##   married = 1 (%)       13 ( 33.3)        21 ( 32.8)    0.011
##   nodegree = 1 (%)      20 ( 51.3)        38 ( 59.4)    0.163
##   re74 (mean (SD)) 3694.02 (5976.07) 3029.98 (5408.31)  0.117
##   re75 (mean (SD)) 1867.86 (3167.66) 1526.71 (3485.63)  0.102
```

```
##                   Stratified by treat
##                     0                1                SMD
##   n                    36               68
##   age (mean (SD))   22.14 (8.70)     24.56 (7.28)      0.302
##   educ (mean (SD))  10.42 (1.68)     10.25 (1.54)      0.103
##   black = 1 (%)        36 (100.0)       68 (100.0)   <0.001
##   hispan = 1 (%)        0 (  0.0)        0 (  0.0)   <0.001
##   married = 1 (%)       0 (  0.0)        0 (  0.0)   <0.001
##   nodegree = 1 (%)     30 ( 83.3)       57 ( 83.8)    0.013
##   re74 (mean (SD)) 627.89 (1170.89) 526.85 (1185.44)  0.086
##   re75 (mean (SD)) 617.26 (988.26)  899.99 (1640.48)  0.209
```

We can see that subclasses 3 and 4 have better covariate balance, with many SMDs below 0.2. Covariates in the other two subclasses have reduced SMDs from the original data but some variables still have not reached the desired cutoff. More trimming or subclassification bins might be selected to obtain better covariate balance.

1.7. Estimate marginal average causal effect of treatment on wages using 4 subclasses above

The point estimate of the marginal causal effect of job training participation on wages is 864.592. The associated p-value is 0.0940903.

The 95%CI for this point estimate is (-449.397, 2178.581)

- Interpretations:

On average, the estimated marginal causal effect of having job training in 1974 on income in 1978 is 864.592 dollars. However, job training in 1974 may not increase the income in 1978 as the estimated ACE on the difference scale is between -449.397 and 2178.581, which includes the null value of 0. Equivalently, p-value > 0.05 suggests the same conclusion.

1.8. Using direct adjustment of confounders

```
##
## Call:
## lm(formula = re78 ~ treat + age + educ + black + hispan + married +
##     nodegree + re74, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -14506  -4952  -1610   3775  54722
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.434e+02  2.439e+03   0.182   0.8558
## treat1       1.597e+03  7.835e+02   2.038   0.0420 *
## age          5.284e+00  3.241e+01   0.163   0.8705
## educ         3.922e+02  1.593e+02   2.461   0.0141 *
## black1      -1.170e+03  7.706e+02  -1.518   0.1295
## hispan1      6.558e+02  9.423e+02   0.696   0.4867
## married1     6.893e+02  6.858e+02   1.005   0.3153
## nodegree1    2.542e+02  8.502e+02   0.299   0.7650
## re74         3.590e-01  5.111e-02   7.025  5.8e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6970 on 605 degrees of freedom
## Multiple R-squared:  0.1409, Adjusted R-squared:  0.1295
## F-statistic:  12.4 on 8 and 605 DF,  p-value: < 2.2e-16
```

```
##     2.5 %    97.5 %
##   57.9656 3135.3252
```

When fitting the regression model, I used the original data. The estimated effect of job training on income in 1978 is 1596.645, and the 95%CI is 57.966, 3135.325

- Interpretations:

On average, job training in 1974 is associated with an expected 1596.645-dollar increase in income in 1978, adjusting for covariates such as income in 1974, age, race, marital status, and education. We are 95% confident that the true causal effect lies between 1596.645. Since the 95% CI does not include 0, we can conclude that job training is significantly associated with income in 1978, based on the direct adjustment model.

- Comparison:

The estimated ACE using direct adjusting of confounder is higher than the estimated ACE obatained from the propensity score subclasses. The directions of both estimates are similar (positive) which indicate some positive effect of the treatment. However, the estimated ACE of the treatment is the direct adjustment model is statistically significant at 0.05 significance level, whereas the 95%CI and p-value obtained from the subclassification approach indicate otherwise.

1.9. Discussion

Subclassification approach

- Advantages: subclasses are created to mimic blocks in block randomization, such that the propensity of receiving treatment are similar within each subclass and independent of the outcome,
- Disadvantages: searching for the optimal subclasses can be arbitrary/difficult, underpowered due to the process of trimming/reduced sample size.

Direct adjustment of confounding approach

- Advantages: straightfoward, we can easily include covariates that we want to adjust for in the regression model; more powerful since we can reserve the original sample size.
- Disadvantages: including many confounders might result in some intersection of confounders levels not having any observations/too few observations; plus, there is a potential of missing other (unmeasured) confounders, which the propensity score might be able to capture.

Question 2

2.1. Write non-parametric structural equations for the DAGs

1. $Y = f_Y(A, L, \epsilon_Y), A = f_A(L, \epsilon_A), L = f_L(\epsilon_L)$
2. $Y = f_Y(A, U, \epsilon_Y), A = f_A(L, \epsilon_A), L = f_L(U, \epsilon_L), U = f_U(\epsilon_U)$
3. $Y = f_Y(U, \epsilon_Y), A = f_A(\epsilon_A), L = f_L(A, U, \epsilon_L), U = f_U(\epsilon_U)$
4. $Y = f_Y(A, L, \epsilon_Y), A = f_A(U, \epsilon_A), L = f_L(U, \epsilon_L), U = f_U(\epsilon_U)$
5. $Y = f_Y(A, U_1, \epsilon_Y), A = f_A(U_2, \epsilon_A), L = f_L(U_1, U_2, \epsilon_L), U_1 = f_{U_1}(\epsilon_{U_1}), U_2 = f_{U_2}(\epsilon_{U_2})$

2.2. Does conditioning on L properly adjust for confounding?

1. There is a backdoor path from A to Y (A - L - Y), so L is a confounder here. Conditioning on L means blocking the backdoor path which adjusts for confounding in this DAG.
2. Here, we can see that there is an arrow going from U to both A and Y (although the path is mediated by L from U to A). Regardless, there is a backdoor path from A to Y through U (A - L - U - Y), thus U is a confounder, and conditioning on U will adjust for confounding. However, since U is unmeasured, and L fully mediates the path from U to A, we can also block the backdoor path by conditioning on L.
3. We can see that there is no arrow going from A to Y, thus there is no confounding between A and Y. Since there is no backdoor path that suggests L or U is a confounder, conditioning L (or U) does not adjust for confounding.
4. Similarly to 2, we see there is an arrow going from U to A and Y, and we can have the backdoor path going from A to Y through L (A - U - L - Y). Blocking L will block this backdoor path, and so conditioning on L adjusts for confounding.
5. This DAG shows no common cause between A and Y, so there is no confounding betwen A and Y. We cannot have the backdoor path from A to Y through L, or rather, it is blocked by L being a collider on that path. Therefore, conditioning on L does not adjusting for confounding.

Appendix

```
knitr::opts_chunk$set(echo = TRUE)
library(tidyverse)
library(ggplot2)
library(dagitty)
library(ggdag)
library(personalized)
library(tableone)
#import data
data = read_csv("hW3 data.csv") %>%
  mutate(treat = as.factor(treat),
         black = as.factor(black),
         hispan = as.factor(hispan),
         married = as.factor(married),
         nodegree = as.factor(nodegree)) %>% select(-X1)
tidy_ggdag = dagify(
  re78 ~ treat + age + educ + degree + married + hispan + black + re74 + re75,
  treat ~ re74 + age + educ + degree + married + hispan + black,
  re75 ~ re74,
  #re74 ~ age + educ + degree + married + hispan + black,
  degree ~ educ,
  exposure = "treat",
  outcome = "re78"
) %>% tidy_dagitty(layout = "circle")

ggdag(tidy_ggdag) + theme_dag()
covar.balance = CreateTableOne(vars = c("age", "educ","black", "hispan", "married", "nodegree", "re74",
                               data = data, test = FALSE)
#print table and select to show standardized mean differences
print(covar.balance, smd = TRUE)
#fit propensity score model
ps.model = glm(treat ~ age + educ + black + hispan + married + nodegree + re74,
               data=data, family = binomial)
summary(ps.model)

#Calculate propensity score and assign it to variable "ps"
data$ps <- predict(ps.model, type="response") #gets the propensity scores for each unit, based on the m
prop.func <- function(x, trt)
{propens.model = glm(treat ~ age + educ + black + hispan + married + nodegree + re74, data=data, family
  data$ps <- predict(propens.model, type = "response")
  data$ps}

check.overlap(x = data,
              trt = data$treat,
              type = "both",
              propensity.func = prop.func)
attach(data)
#max(ps[data$treat==0]) #min(P(A=1|C))
#min(ps[data$treat==1])

data.t1 = data[ps <= max(ps[data$treat==0]) & ps >= min(ps[data$treat == 1]),]
dim(data)
dim(data.t1)
vars = c("age", "educ","black", "hispan", "married", "nodegree", "re74", "re75")
covar.balance.t1 = CreateTableOne(vars = vars, strata = "treat",
                               data = data.t1, test = FALSE)
#print table and select to show standardized mean differences
```

```
print(covar.balance.t1, smd = TRUE)
#refit propensity score on trimmed data
propens.model = glm(treat ~ age + educ + black + hispan + married + nodegree + re74,
                    data=data.t1, family = binomial)
data.t1$ps <- predict(propens.model, type = "response") #obtain propensity scores from new PS model

data.t2 = data.t1
attach(data.t2)
subclass.breaks = quantile(ps, c(0.4, 0.6, 0.8)) #0.4, 0.65
subclass.breaks
subclass = data.t2$ps
subclass = as.numeric(data.t2$ps>subclass.breaks[1])
subclass[which(data.t2$ps>subclass.breaks[1] & data.t2$ps<=subclass.breaks[2])]<- 1
subclass[which(data.t2$ps>subclass.breaks[2] & data.t2$ps<=subclass.breaks[3])]<- 2
subclass[which(data.t2$ps>subclass.breaks[3])]<- 3
#looking at sample sizes within each subclass
table(data.t2$treat, subclass)
#looking at propensity scores within subclasses
prop.func <- function(x, trt)
{data.t2$ps[which(data.t2$ps <= subclass.breaks[1])]}
check.overlap(x = data.t2[which(data.t2$ps <=subclass.breaks[1]),],
              trt = data.t2$treat[which(data.t2$ps <= subclass.breaks[1])],
              type = "both",
              propensity.func = prop.func)


prop.func <- function(x, trt)
{data.t2$ps[which(data.t2$ps>subclass.breaks[1] & data.t2$ps<=subclass.breaks[2])]}
check.overlap(x = data.t2[which(data.t2$ps>subclass.breaks[1]&data.t2$ps<=subclass.breaks[2]),],
              trt = data.t2$treat[which(data.t2$ps>subclass.breaks[1]&data.t2$ps<=subclass.breaks[2])],
              type = "both",
              propensity.func = prop.func)


prop.func <- function(x, trt)
{data.t2$ps[which(data.t2$ps>subclass.breaks[2] & data.t2$ps<=subclass.breaks[3])]}
check.overlap(x = data.t2[which(data.t2$ps>subclass.breaks[2]&data.t2$ps<=subclass.breaks[3]),],
              trt = data.t2$treat[which(data.t2$ps>subclass.breaks[2]&data.t2$ps<=subclass.breaks[3])],
              type = "both",
              propensity.func = prop.func)

prop.func <- function(x, trt)
{data.t2$ps[which(data.t2$ps>subclass.breaks[3])]}
check.overlap(x = data.t2[which(data.t2$ps>subclass.breaks[3]),],
              trt = data.t2$treat[which(data.t2$ps>subclass.breaks[3])],
              type = "both",
              propensity.func = prop.func)
tab_s0 <- CreateTableOne(vars = vars, strata = "treat", data = data.t2[which(subclass==0),], test = FALS
tab_s1 <- CreateTableOne(vars = vars, strata = "treat", data = data.t2[which(subclass==1),], test = FALS
tab_s2 <- CreateTableOne(vars = vars, strata = "treat", data = data.t2[which(subclass==2),], test = FALS
tab_s3 <- CreateTableOne(vars = vars, strata = "treat", data = data.t2[which(subclass==3),], test = FALS

## Show table with SMD
```

```
print(tab_s0, smd = TRUE)
print(tab_s1, smd = TRUE)
print(tab_s2, smd = TRUE)
print(tab_s3, smd = TRUE)
ACE0 <- mean(data.t2$re78[which(subclass==0 & data.t2$treat==1)])-mean(data.t2$re78[which(subclass==0 &
ACE1 <- mean(data.t2$re78[which(subclass==1 & data.t2$treat==1)])-mean(data.t2$re78[which(subclass==1 &
ACE2 <- mean(data.t2$re78[which(subclass==2 & data.t2$treat==1)])-mean(data.t2$re78[which(subclass==2 &
ACE3 <- mean(data.t2$re78[which(subclass==2 & data.t2$treat==1)])-mean(data.t2$re78[which(subclass==2 &

ace <- (nrow(data.t2[which(subclass==0),])/nrow(data.t2))*ACE0+ (nrow(data.t2[which(subclass==1),])/nrow
(nrow(data.t2[which(subclass==3),])/nrow(data.t2))*ACE3

v01 <- var(data.t2$re78[which(subclass==0 & data.t2$treat==1)])
v00 <- var(data.t2$re78[which(subclass==0 & data.t2$treat==0)])

v11 <- var(data.t2$re78[which(subclass==1 & data.t2$treat==1)])
v10 <- var(data.t2$re78[which(subclass==1 & data.t2$treat==0)])

v21 <- var(data.t2$re78[which(subclass==2 & data.t2$treat==1)])
v20 <- var(data.t2$re78[which(subclass==2 & data.t2$treat==0)])

v31 <- var(data.t2$re78[which(subclass==3 & data.t2$treat==1)])
v30 <- var(data.t2$re78[which(subclass==3 & data.t2$treat==0)])


n0 <- nrow(data[which(subclass==0),])
n1 <- nrow(data[which(subclass==1),])
n2 <- nrow(data[which(subclass==2),])
n3 <- nrow(data[which(subclass==3),])

n01 <- nrow(data.t2[which(subclass==0& data.t2$treat==1),])
n11 <- nrow(data.t2[which(subclass==1& data.t2$treat==1),])
n21 <- nrow(data.t2[which(subclass==2& data.t2$treat==1),])
n31 <- nrow(data.t2[which(subclass==3& data.t2$treat==1),])
n00 <- nrow(data.t2[which(subclass==0& data.t2$treat==0),])
n10 <- nrow(data.t2[which(subclass==1& data.t2$treat==0),])
n20 <- nrow(data.t2[which(subclass==2& data.t2$treat==0),])
n30 <- nrow(data.t2[which(subclass==3& data.t2$treat==0),])

varace <-(n1)^2/nrow(data)^2*((v11/n11)+(v10/n10))+ (n2)^2/nrow(data)^2*((v21/n21)+(v20/n20))+  (n0)^2/

sdace<-sqrt(varace)

CIL=ace-sdace*2
CIU=ace+sdace*2

#p-value
pval = pnorm((ace/sdace), lower.tail = FALSE)
reg.mod = lm(re78~treat + age + educ + black + hispan + married + nodegree + re74, data = data)
summary(reg.mod)

#95%CI for average treatment effect
confint(reg.mod)[2,]
```