P8122 HW3

Sabrina Lin stl2137

11/21/2020

Part 1

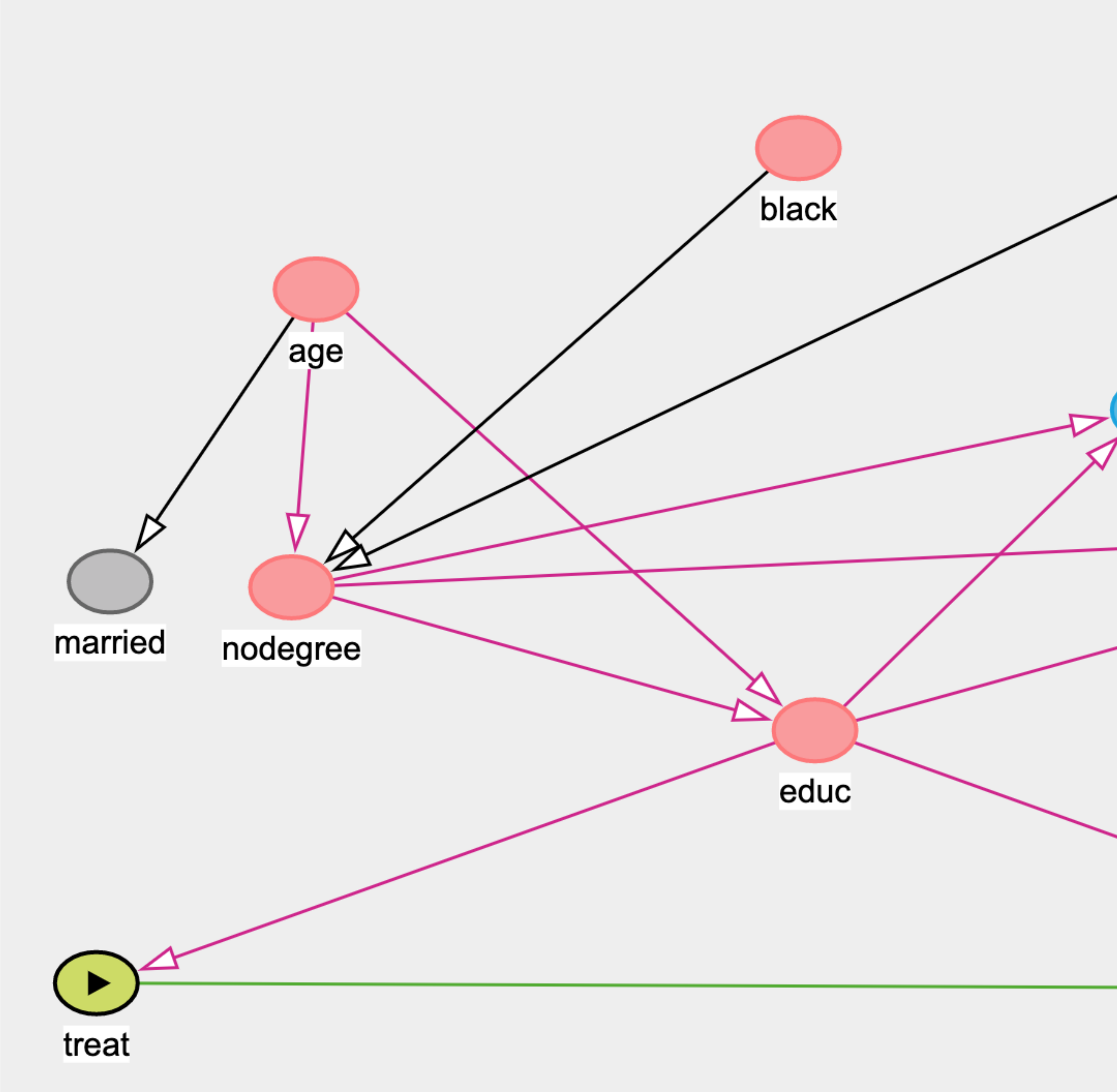
```
## Warning: Missing column names filled in: 'X1' [1]
```

```
## 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.



- Based off the DAG, `nodegree` is a collider between `black` and `hispanic`, so we will not adjust for it.
- There is no backdoor path with `married`, so we will not adjust for it.

Subpart 2

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

```
##                     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
##   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
```

- 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 variable`re74` also has a relatively large SMDs (0.596 respectively), also indicating that the covariate balance for `re74` is not great.

Subpart 3

The propensity score is defined as the probability of receiving the treatment given the observed 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.

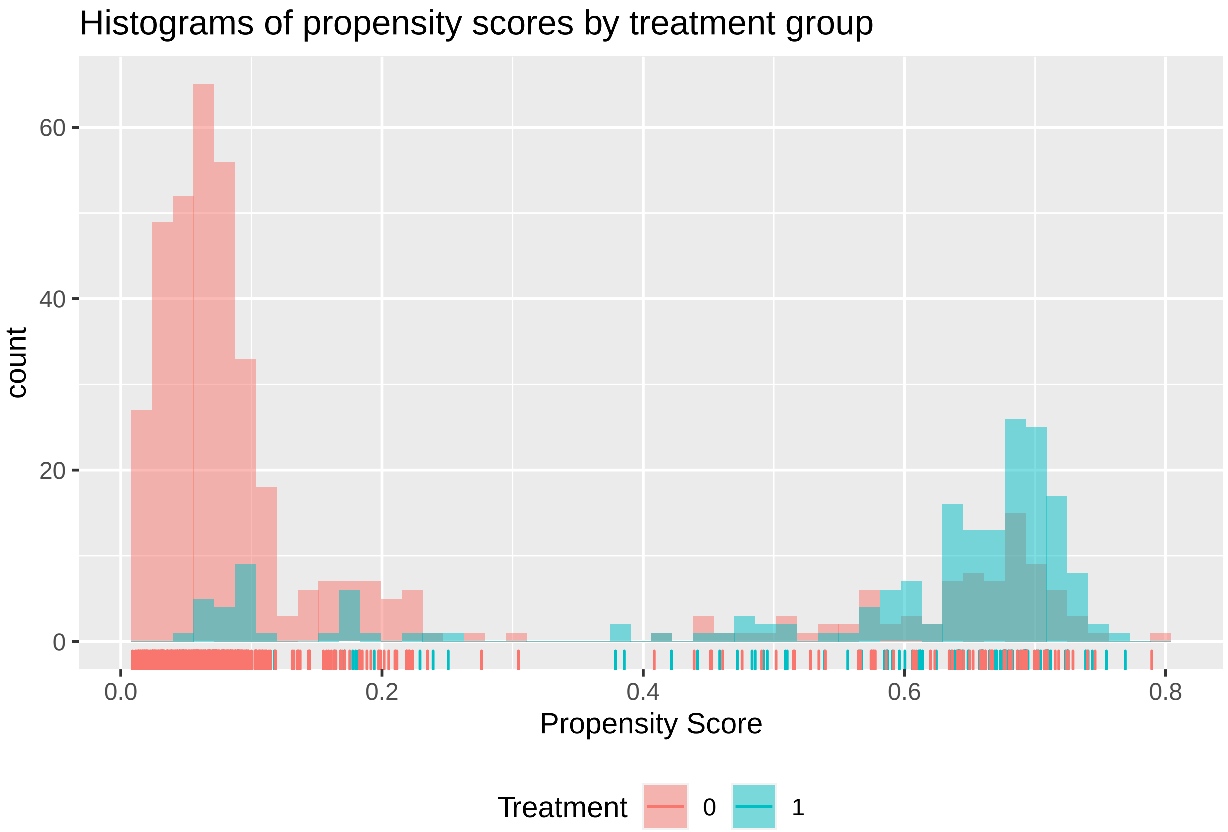
```
##
## Call:
## mlogit(formula = treat ~ 0 | age + educ + black + hispan + re74 +
##     re75, data = salary_mlogit_dat, method = "nr")
##
## Frequencies of alternatives:choice
##      0      1
## 0.6987 0.3013
##
## nr method
## 6 iterations, 0h:0m:0s
## g'(-H)^-1g = 2.24E-05
## successive function values within tolerance limits
##
## Coefficients :
##                  Estimate  Std. Error z-value  Pr(>|z|)
## (Intercept):1 -3.2323e+00  6.9382e-01 -4.6587 3.182e-06 ***
## age:1         -4.0233e-04  1.2798e-02 -0.0314  0.974920
## educ:1         7.7435e-02  4.6295e-02  1.6727  0.094396 .
## black1:1       3.2053e+00  2.8359e-01 11.3027 < 2.2e-16 ***
## hispan1:1      1.0447e+00  4.2062e-01  2.4836  0.013006 *
## re74:1        -8.3125e-05  2.8453e-05 -2.9214  0.003484 **
## re75:1         3.0206e-05  4.4258e-05  0.6825  0.494933
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Log-Likelihood: -250.39
## McFadden R^2:  0.33362
## Likelihood ratio test : chisq = 250.72 (p.value = < 2.22e-16)
```

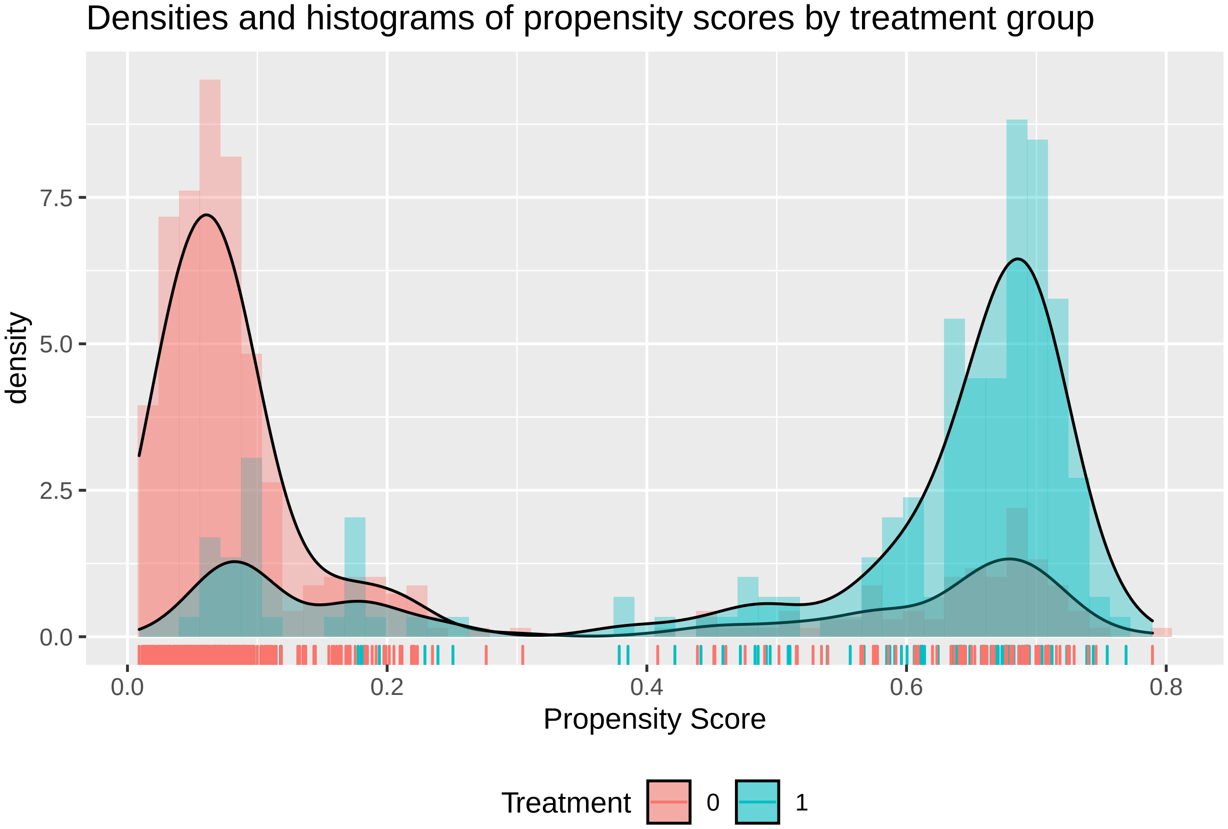
```
##
## Call:
## glm(formula = treat ~ age + educ + black + hispan + re74 + re75,
##     family = binomial, data = salary_dat)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7652  -0.4516  -0.3337   0.8392   2.4656
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.232e+00  6.938e-01  -4.659 3.18e-06 ***
## age         -4.023e-04  1.280e-02  -0.031  0.97492
## educ         7.744e-02  4.629e-02   1.673  0.09439 .
## black1       3.205e+00  2.836e-01  11.303  < 2e-16 ***
## hispan1      1.045e+00  4.206e-01   2.484  0.01301 *
## re74        -8.312e-05  2.845e-05  -2.922  0.00348 **
## re75         3.021e-05  4.426e-05   0.682  0.49493
## ---
## 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: 500.78  on 607  degrees of freedom
## AIC: 514.78
##
## Number of Fisher Scoring iterations: 5
```

Subpart 4

Given the propensity scores estimated, evaluate overlap. Trim data if necessary and evaluate the impact of trimming in your analytic sample on efficiency and generalizability.


Histograms of propensity scores by treatment group
count
60
40
20
0
0.0
0.2
0.4
0.6
0.8
Propensity Score
Treatment
0
1


Densities and histograms of propensity scores by treatment group
density
7.5
5.0
2.5
0.0
0.0
0.2
0.4
0.6
0.8
Propensity Score
Treatment
0
1

```
## [1] 0.04785359
##   [1]  TRUE  TRUE  TRUE  TRUE FALSE FALSE  TRUE FALSE FALSE  TRUE  TRUE  TRUE
##  [13] FALSE  TRUE FALSE FALSE FALSE  TRUE FALSE  TRUE FALSE FALSE  TRUE FALSE
##  [25]  TRUE FALSE FALSE FALSE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE FALSE
##  [37]  TRUE  TRUE FALSE FALSE FALSE  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE
##  [49]  TRUE FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE
##  [61]  TRUE FALSE FALSE FALSE  TRUE FALSE  TRUE  TRUE FALSE  TRUE  TRUE FALSE
##  [73] FALSE FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE FALSE
##  [85]  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE
##  [97] FALSE FALSE FALSE FALSE  TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE
## [109] FALSE FALSE FALSE FALSE  TRUE FALSE FALSE  TRUE  TRUE FALSE FALSE  TRUE
## [121]  TRUE FALSE  TRUE FALSE  TRUE FALSE FALSE  TRUE FALSE  TRUE FALSE  TRUE
## [133]  TRUE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE
## [145] FALSE  TRUE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE
## [157] FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE  TRUE
## [169]  TRUE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE
## [181] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE
## [193]  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE  TRUE
## [205] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [217] FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE
## [229] FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [241] FALSE FALSE FALSE FALSE FALSE  TRUE FALSE  TRUE  TRUE  TRUE FALSE FALSE
## [253] FALSE  TRUE FALSE FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE  TRUE FALSE
## [265] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [277] FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE  TRUE  TRUE FALSE
## [289] FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE  TRUE FALSE
## [301]  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE
## [313]  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE
## [325] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [337] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE
## [349] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [361] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [373] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [385] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [397] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [409] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [421] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

```
## [1] 429
```

```
## [1] 0.7894206
```

```
##   [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [13] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [25] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [37] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [49] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [61] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [73] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [85] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [97] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [109] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [121] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [133] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [145] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [157] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [169] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [181] FALSE FALSE FALSE FALSE FALSE
```

```
## [1] 429
```

```
## [1] 614  12
```

```
## [1] 510  12
```

- After trimming the data, we lose 104 observations. Although trimming these 104 observations improves the internal validity since we are able to ensure comparability for the remaining units, it hurts generalizability because we are excluding certain people in the population to get to a better causal effect.

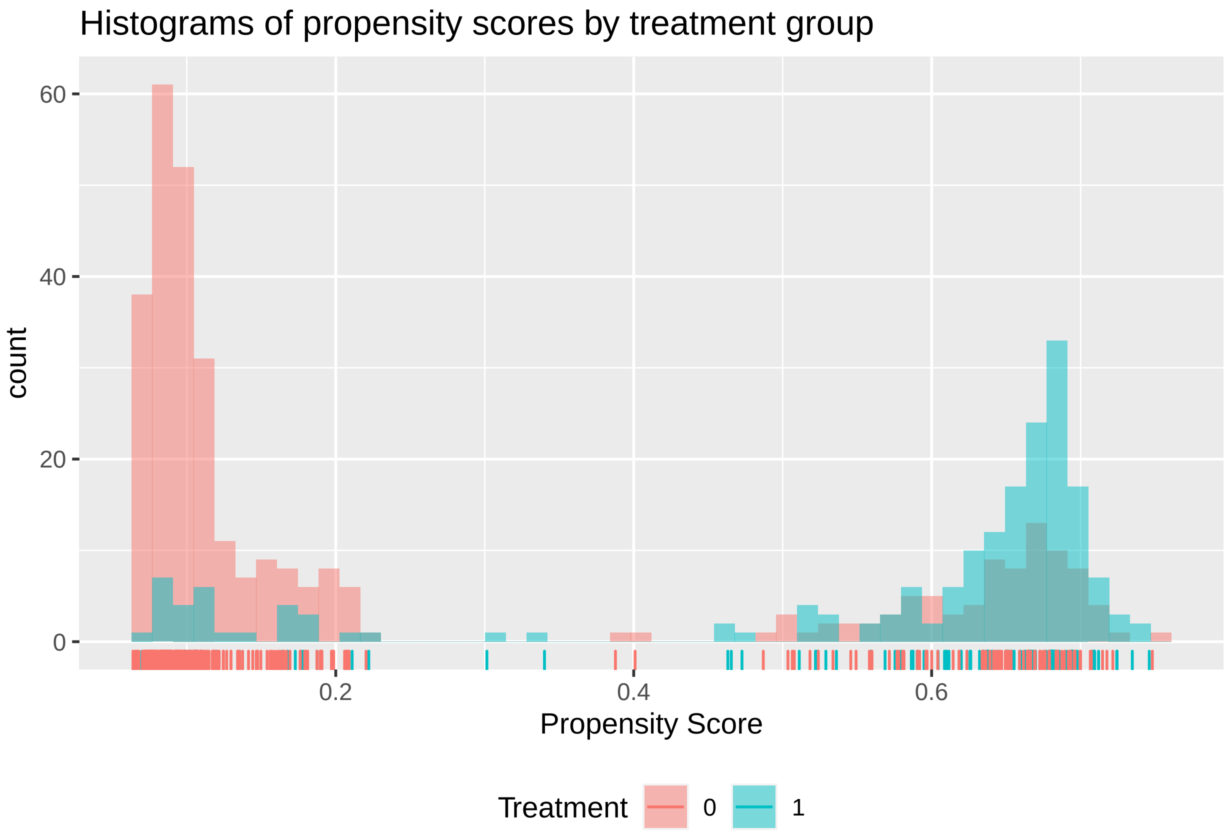
Subpart 5

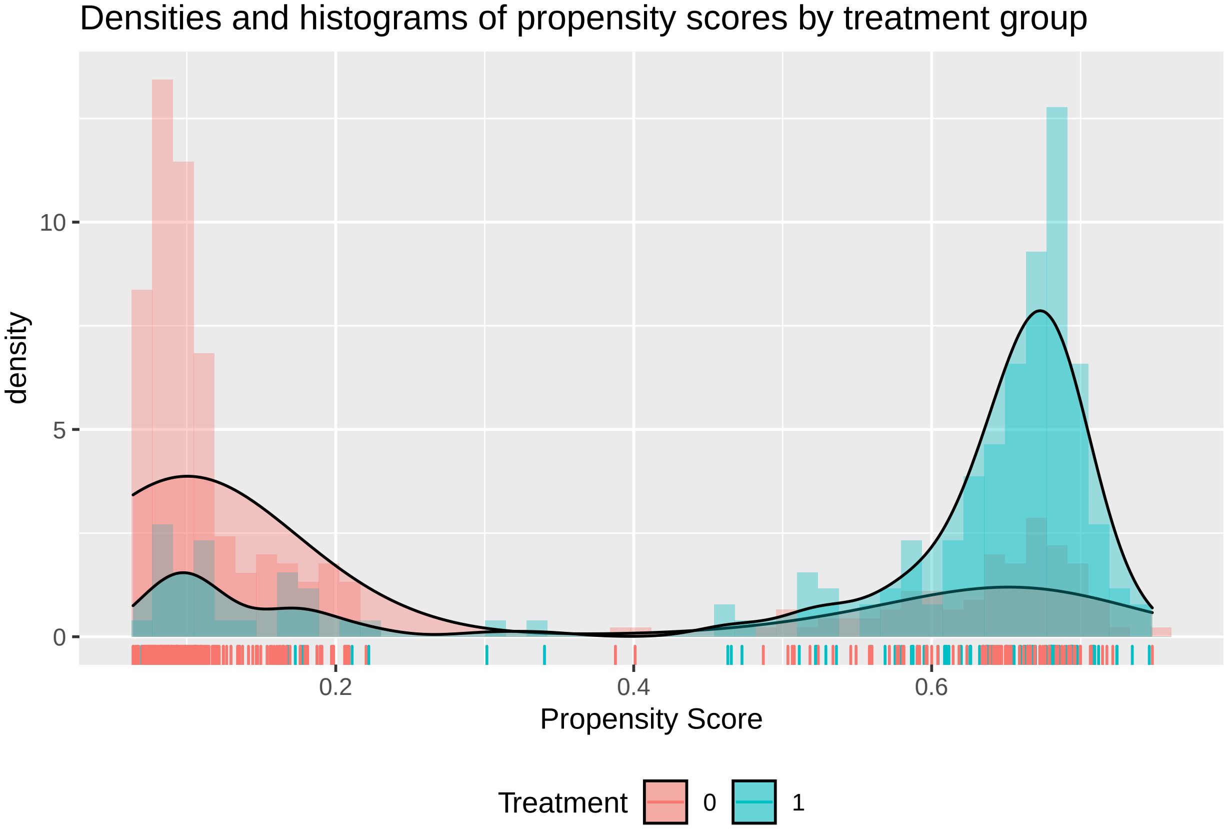
Evaluate covariate balance in the trimmed sample.

```
##                     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
##   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
```

```
##                     Stratified by treat
##                      0                 1                  SMD   
##   n                      325               185                  
##   age (mean (SD))      26.05 (10.18)     25.82 (7.16)      0.027
##   educ (mean (SD))     10.25 (2.83)      10.35 (2.01)      0.038
##   black = 1 (%)           87 (26.8)        156 (84.3)      1.421
##   hispan = 1 (%)          59 (18.2)         11 ( 5.9)      0.382
##   re74 (mean (SD)) 2811.87 (4074.79) 2095.57 (4886.62)  0.159
##   re75 (mean (SD)) 1898.54 (2740.10) 1532.06 (3219.25)  0.123
```

- Comparing the pre- and post- trimmed samples, the post-trimmed sample had lower SMDs compared to the pre-trimmed samples except for `hispan`, which will be discussed further. The SMDs that were high in the pre-trimmed sample (`black`, `re74`) are lower in the post-trimmed sample (`black` decreased from 1.671 to 1.421, and `re74` decreased from 0.596 to 0.159.) The SMDs for `age` (from 0.242 to 0.027), `educ` (0.045 to 0.038), and `re75` (0.287 to 0.123) also decreased, making them now all below the 0.2 threshold; however, the SMD for `hispan` increased (from 0.227 to 0.382), making it even higher above the 0.2 threshold.


Histograms of propensity scores by treatment group


Densities and histograms of propensity scores by treatment group

- Looking at the post-trimming density plot, we can visually see that trimming has made the covariate balance better than pre-trimming; however, it is still not ideal.

Subpart 6

Using the propensity scores estimated use subclassification to balance covariates between treated and controls. Explain your process, report the breaks you decide on for your subclasses, show a plot of the propensity scores with these breaks. Inspect covariate balance for each subclass.

```
##    subclass
##       0   1   2   3
##   0  18 116 102  89
##   1   1   7  20 157
```

- Although overlap is not technically violated, the balance for the first subgroup is extremely poor. The second subgroup also does not have great balance, so we will continue trying subclassification at different percentile breaks.

```
##    subclass
##       0   1   2   3
##   0 120 107  60  38
##   1   8  20  67  90
```

- The first and second subgroup balance is still not fantastic, so we will continue trying subclassification at different percentile breaks.

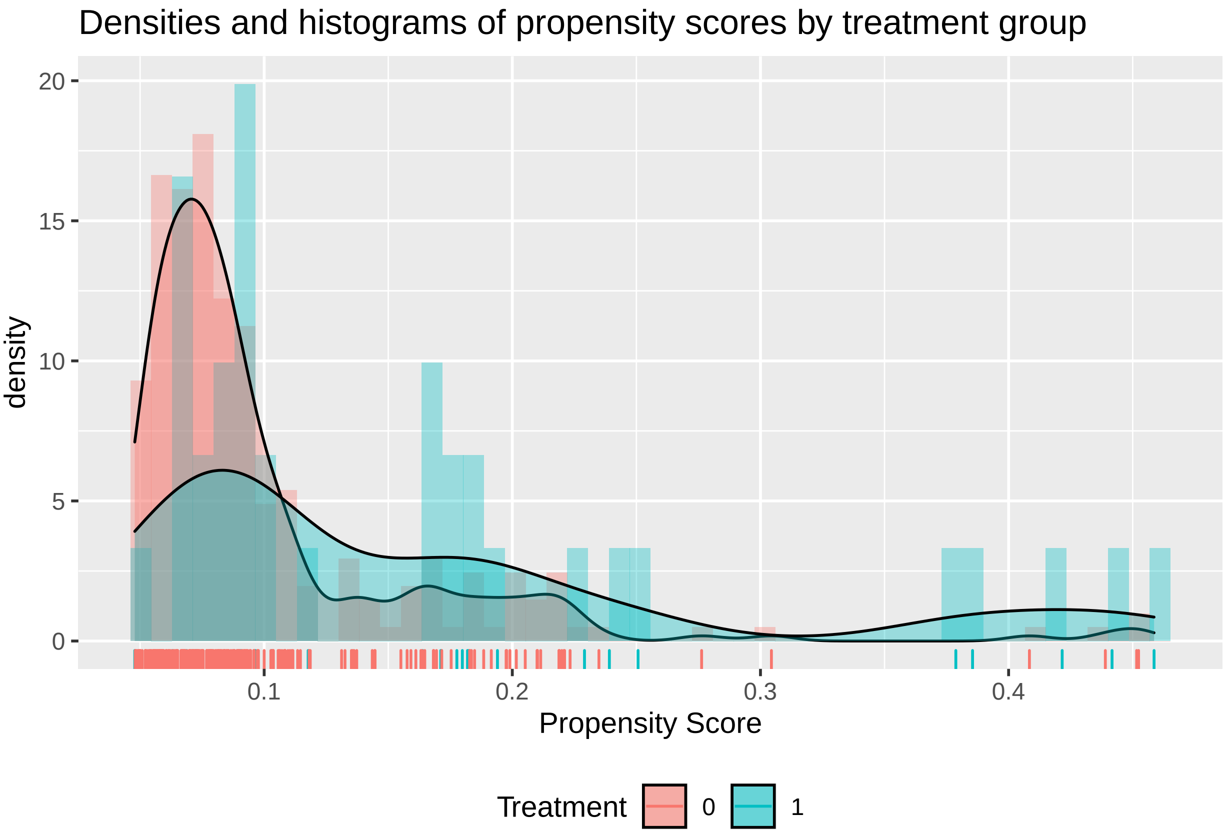
```
##    subclass
##       0   1
##   0 155 170
##   1  13 172
```

- Although the second subgroup has really good balance, the first subgroup has poor balance. We will continue trying aditional subclassification percentiles.

```
##    subclass
##       0   1   2   3
##   0 244  13  30  38
##   1  36  13  46  90
```

- After additional searching, having 4 subclasses with percentile breaks at 0.55, 0.6, 0.75 seems to lead to better balance across the subclasses. Although the first and last subclass still have sub-optimal covariate balance, the second and third subclasses have pretty good balance. We will proceed to inspect the covariate balance for each subclass more in depth.

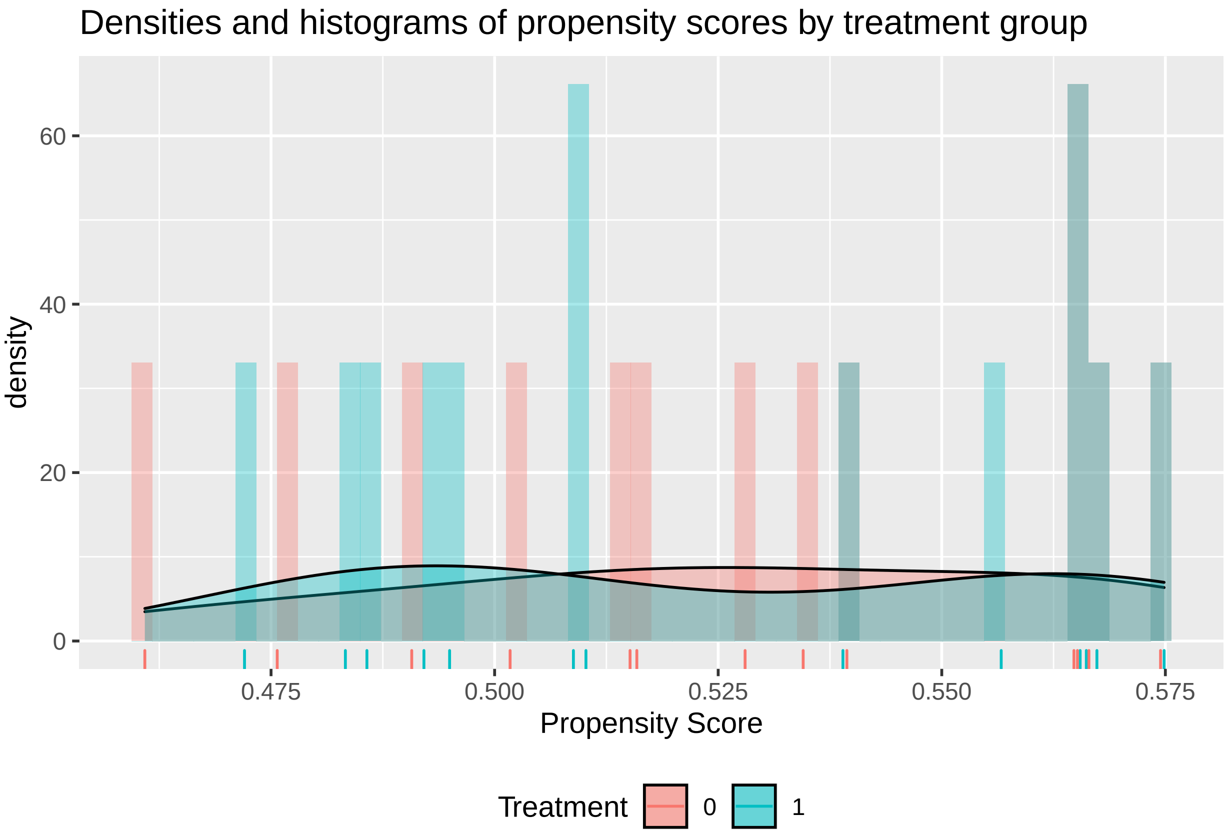
Subclass 1



```
##                      Stratified by treat
##                       0                  1                  SMD
##    n                      244                 36
##    age (mean (SD))     26.27 (10.03)     25.19 (6.00)     0.131
##    educ (mean (SD))    10.32 (2.84)      10.56 (2.05)     0.094
##    black = 1 (%)           6 ( 2.5)          7 (19.4)     0.565
##    hispan = 1 (%)         59 (24.2)         11 (30.6)     0.143
##    re74 (mean (SD)) 3040.89 (4137.02) 5160.49 (8482.26)   0.318
##    re75 (mean (SD)) 1966.78 (2582.00) 2525.13 (3429.53)   0.184
```

- The density plot for subclass 1 looks acceptable enough and is better than before subclassification, though there still is right skewedness from the no treatment group. Despite this, the covariate balance appears to be better, as the density plots overlay each other a fair amount. This skewness is not unsurprising, as the more extreme values would be placed in the first and last subgroups.
- For the first subgroup, the SMD is below 0.2 for the variables `age`, `educ`, `hispan`, `re75`; however, the SMD for the variables `black` and `re74` are above the 0.2 threshold.

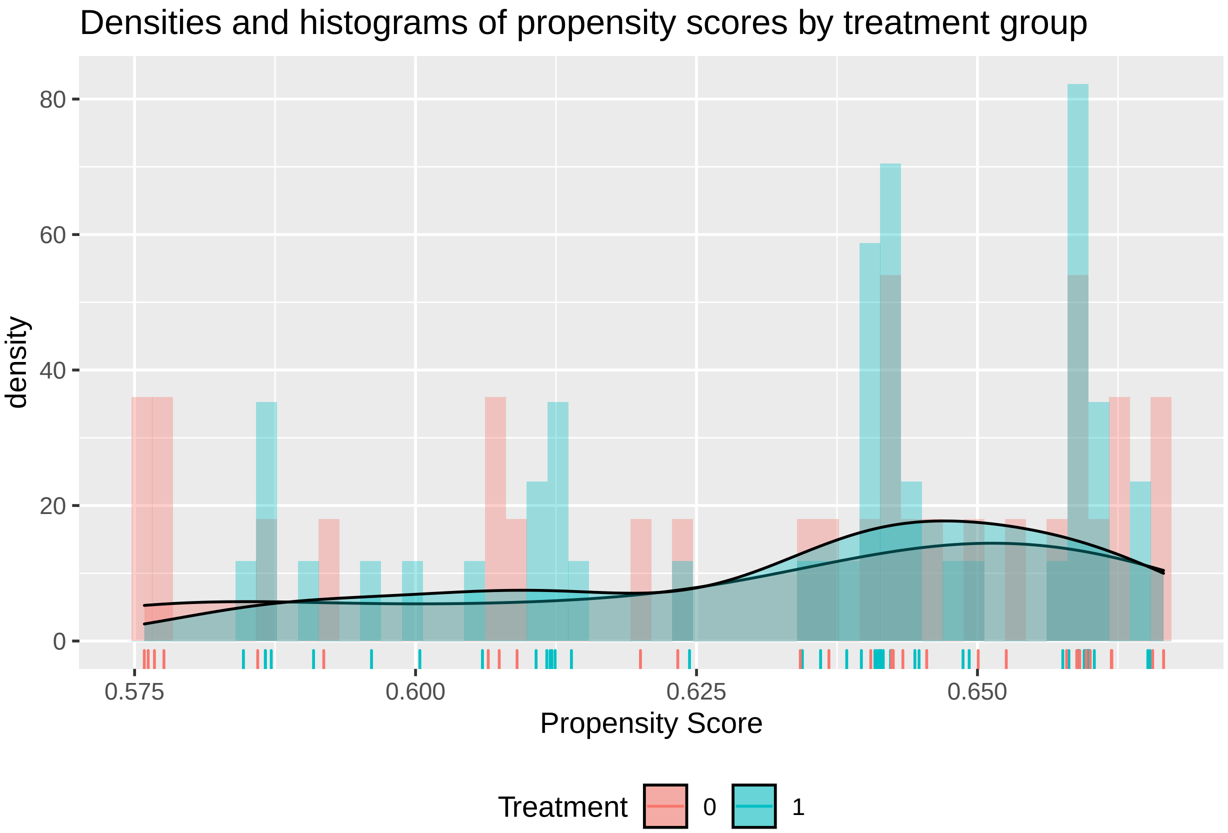
Subclass 2



```
##                     Stratified by treat
##                      0                 1                 SMD
##   n                       13                13
##   age (mean (SD))    38.00 (11.70)     30.08 (8.05)      0.789
##   educ (mean (SD))    7.62 (4.35)       8.62 (3.01)      0.267
##   black = 1 (%)         13 (100.0)        13 (100.0)   <0.001
##   hispan = 1 (%)         0 (  0.0)         0 (  0.0)   <0.001
##   re74 (mean (SD)) 6674.32 (6326.71) 8031.01 (5125.40)  0.236
##   re75 (mean (SD)) 3654.60 (5178.34) 4555.98 (5527.34)  0.168
```

- The density plot for subclass 2 looks pretty balanced, as the two density plots almost overlay each other, implying that there is decent covariate balance.
- For the second subgroup, the SMD is below 0.2 for the variables `black`, `hispan`, and `re75`; however, the SMD for the variables `age` and `educ` are above both the 0.2 threshold and the 0.25 threshold found in the literature. The variable `re74`, with a SMD of 0.236 is above the 0.2 threshold but below the 0.25 threhold.

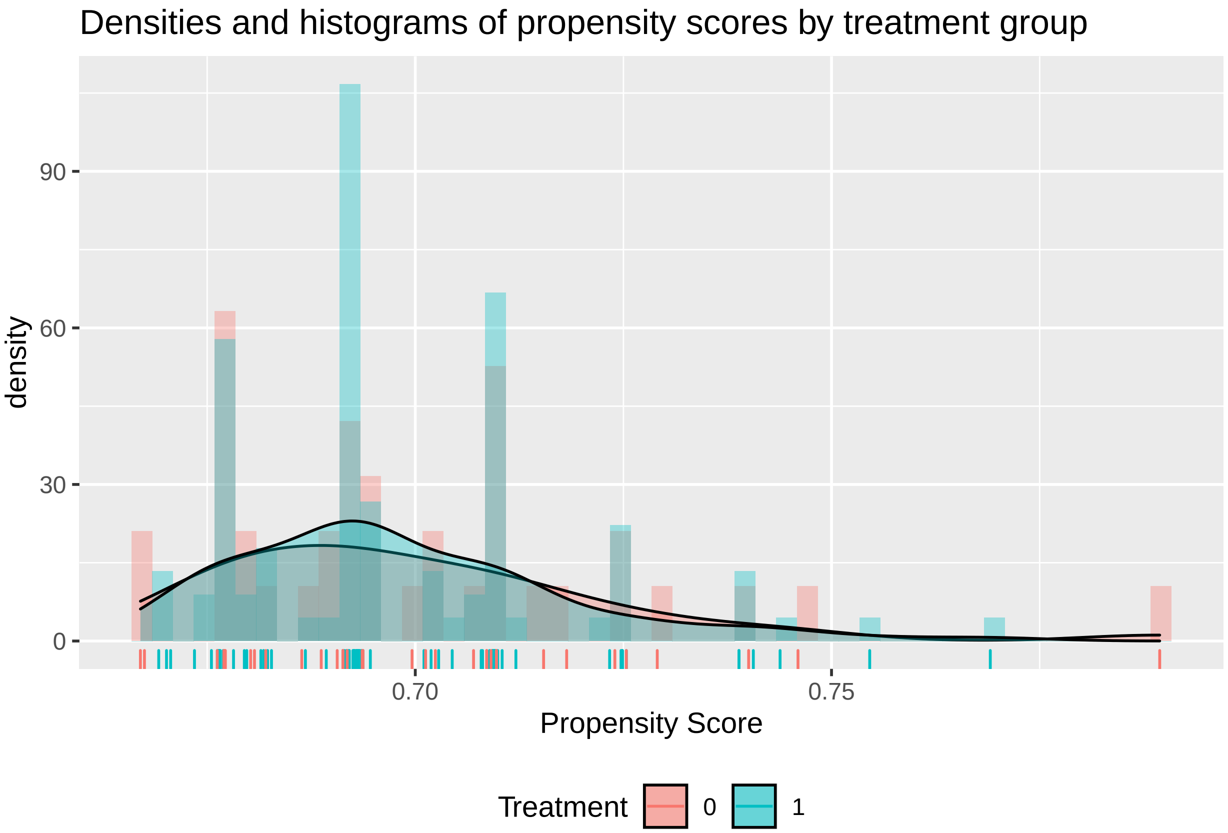
Subclass 3



```
##                    Stratified by treat
##                     0                 1                 SMD   
##   n                     30                46                  
##   age (mean (SD))    21.70 (7.99)      24.52 (8.00)      0.353
##   educ (mean (SD))    9.07 (2.18)       8.91 (1.77)      0.077
##   black = 1 (%)         30 (100.0)        46 (100.0)    <0.001
##   hispan = 1 (%)         0 (  0.0)         0 (  0.0)    <0.001
##   re74 (mean (SD)) 2165.71 (3198.56) 1793.11 (2988.62)  0.120
##   re75 (mean (SD)) 1742.16 (2843.34) 1566.32 (3896.63)  0.052
```

- The density plot for subclass 3 looks fairly good, as the density plots overlay each other mostly. This implies that the covariate balance appears to be decent within this subgroup
- For the third subgroup, the SMD is below 0.2 for the variables `educ`, `black`, `hispan`, `re74`, and `re75`; however, the SMD for the variable `age` are above the 0.2 threshold.

Subclass 4



```
##                   Stratified by treat
##                    0                  1                  SMD
##   n                    38                 90
##   age (mean (SD))   23.97 (9.07)      26.11 (6.85)     0.266
##   educ (mean (SD))  11.63 (1.28)      11.24 (1.25)     0.306
##   black = 1 (%)        38 (100.0)        90 (100.0)   <0.001
##   hispan = 1 (%)        0 (  0.0)         0 (  0.0)   <0.001
##   re74 (mean (SD)) 530.03 (847.97)   166.86 (510.72)   0.519
##   re75 (mean (SD)) 983.08 (2221.41) 680.52 (1592.35)  0.157
```

- The density plot for subclass 4 looks pretty good for the most part, though there still is right skewedness from both treatment groups. Despite this, the covariate balance appears to be better, as the density plots overlay each other quite closely. This skewness is not unsurprising, as the more extreme values would be placed in the first and last subgroups, which is seen for the scores higher than 0.75.
- For the fourth and last subgroup, the SMD is below 0.2 for the variables `black`, `hispan`, and `re75`; however, the SMD for the variables `age`, `educ`, and `re74` are above the 0.2 threshold.

Subpart 7

Using your subclasses from Question (6), estimate the marginal average causal effect of participation in a job training on wages. Give a point estimate, a confidence interval, and a p-value for whether it had any effect, and interpret these results in context.

```
## [1] 520.5218
```

```
## Too many permutations to use exact method.
## Defaulting to approximate method.
## Increase maxiter to at least 4.33370479048693e+143 to perform exact estimation.

## [1] 0.1134
```

- The estimated marginal average causal effect of participation in a job training on wages is 1185.4095143, with a confidence interval of (-505.990081, 2876.8091096) and a p-value of 0.1134. Although 1185.4095143 is positive and thus implying that job training has a positive causal effect on wages, the wide confidence interval that crosses 0 paired with the p-value $0.1134 > 0.05$ leads us to reject the null hypothesis that there is a causal relationship between job training and wages.

Subpart 8

Estimate the marginal average causal effect between training and salary in this observa- tional study using direct adjustment of confounders. Interpret the results and compare the results with what obtained using the subclassification approach.

- Comparing the ACE using the direct adjustment of confounders vs. the ACE obtained using subclassification, the ACE using the direct adjustment of confounders is lower but still positive at 776.0137637, compared to the the ACE of 1185.4095143 obtained from subclassification. Thus the ACE found using direct adjustment would also conclude that there is a positive causual effect from training to wages but has a smaller magnitude.

Subpart 9

Discuss advantages and disadvantages of the regression based approach to confounding ad- justment and the subclassification approach.

Part 2

a) Write the non-parametric structural equation model associated with it.
b) Does conditioning on L properly adjust for confounding if we used the definition of confounder based on the backdoor criterion? Justify your answer.

DAG 1

$Y = f_Y(A, L, \epsilon_Y)$

$A = f_A(L, \epsilon_A)$

$L = f_L(\epsilon_L)$

- Conditioning on L would properly adjust for confounding, as it would block the path A-L-Y.

DAG 2

$Y = f_Y(U, A, L, \epsilon_Y)$

$A = f_A(L, \epsilon_A)$

$L = f_L(U, \epsilon_L)$

$U = f_U(\epsilon_U)$

- Conditioning on L would properly adjust for confounding, as it would block the path A-L-U-Y.

DAG 3

$Y = f_Y(U, \epsilon_Y)$

$U = f_U(\epsilon_U)$

$L = f_L(U, A, \epsilon_L)$

$A = f_A(\epsilon_A)$

- Conditioning on L would not properly adjust for confounding, as L is a collider. In this DAG, A is also not associated with Y.

DAG 4

$Y = f_Y(A, L, \epsilon_Y)$

$A = f_A(U, \epsilon_A)$

$L = f_L(U, \epsilon_L)$

$U = f_U(\epsilon_U)$

- Conditioning on L would properly adjust for confounding, as it would block the path A-U-L-Y.

DAG 5

$Y = f_Y(A, U_1, \epsilon_Y)$

$A = f_A(U_2, \epsilon_A)$

$U_2 = f_{U_2}(\epsilon_{U_2})$

$U_1 = f_{U_1}(\epsilon_{U_1})$

$L = f_L(U_1, U_2, \epsilon_L)$

- Conditioning on L would not properly adjust for confounding, as L is a collider. Thus, we do not want to condition on L.