P8122 Homework 4

Sabrina Lin stl2137

12/4/2020

Question 1

Exclude the outcome from your data. Now produce descriptive statistics of your sample by practice type.

- The following table is stratified by practice type, where 0 = pediatrics, 1 = family practice, 2 = OB-GYN

```
##                        Stratified by practice_type
##                         0             1              2              p      test
##   n                      515           365            533
##   age (mean (SD))      14.92 (2.25)  19.46 (3.82)   21.43 (3.33)  <0.001
##   age_group = 1 (%)       33 ( 6.4)    242 ( 66.3)    437 (82.0)  <0.001
##   race (%)                                                        <0.001
##      0                   232 (45.0)    169 ( 46.3)    331 (62.1)
##      1                   194 (37.7)    102 ( 27.9)    147 (27.6)
##      2                    29 ( 5.6)     10 (  2.7)     13 ( 2.4)
##      3                    60 (11.7)     84 ( 23.0)     42 ( 7.9)
##   insurance_type (%)                                              <0.001
##      0                   204 (39.6)     12 (  3.3)     59 (11.1)
##      1                   171 (33.2)    188 ( 51.5)    364 (68.3)
##      2                    25 ( 4.9)      9 (  2.5)     50 ( 9.4)
##      3                   115 (22.3)    156 ( 42.7)     60 (11.3)
##   med_assist = 1 (%)     204 (39.6)     12 (  3.3)     59 (11.1)  <0.001
##   location (%)                                                    <0.001
##      1                   216 (41.9)    365 (100.0)    217 (40.7)
##      2                     0 ( 0.0)      0 (  0.0)    165 (31.0)
##      3                     0 ( 0.0)      0 (  0.0)     89 (16.7)
##      4                   299 (58.1)      0 (  0.0)     62 (11.6)
##   location_type = 1 (%)  299 (58.1)      0 (  0.0)    151 (28.3)  <0.001
```

- There are a total of 1413 females in the study.

The levels for each variable are as follows:

- 515 females received treatment from a pediatric facility (level 0), 365 females received treatment from family practice (level 1), and 533 females received treatment from an OB-GYN (level 2).
- 701 females are in the 11 - 17 age group (level 0), and 712 females are in the 18 - 26 age group (level 1).
- There are 732 white females (level 0), 443 Black females (level 1), 52 Hispanic females (level 2), and 186 females with either other or unknown ethnicity/race (level 3).
- There are 275 females with insurance type level 0 (we are not given the insurance type/there is no metadata; even though it appears to be medical assistance, we do not know if these are the same

people for sure), 723 females with private payer insurance type (level 1), 84 females with hospital based insurance type (level 2), and 331 females with military insruance type (level 3).

- 1138 females do not have medical assistance (level 0), and 275 females have medical assistance (level 1).
- 798 females received treatment from Odenton (level 1), 165 females received treatment from White Marsh (level 2), 89 females received treatment from Johns Hopkins (level 3), and 361 females received treatment from Bayview (level 4).

Question 2

Write the protocol of the RCT you would like to conduct to address the question of interest. In particular, (i) specify control and treatment arm and (ii) specify the eligibility criteria according to levels of baseline characteristics so that the assignment to the treatment in this observational study is probabilistic. There is no single one solution. Please, explain your reasoning.

- Since we are interested in seeing whether type of practice where the Gardasil vaccines is taken affects rates of completion, the treatment and control arms should be allocated by practice type. In this case, the treatment arm will be those who received treatment through an OB-GYN, and the control arm will be those who received treatment from a family practice.
- Looking at the descriptive table and locations, we can see that only location 1 has patients that offers family practice and OB-GYN. Thus, to ensure the probabilistic assumption is not violated, we will only utilize location 1 in this RCT.
- In addition, to maintain the probabilistic assumption, we will only recruit patients who are in `age_group == 1`, or those coded as being ages 18 to 26. This is because patients who are adults cannot see a pediatrician, and most adolescent women do not see an OB-GYN.
- Making these stringent inclusion/exclusion criteria will help increase the internal validity of future matching in this assignment, but it will make our findings less generalizable to the population.

Question 3

Following the protocol, exclude subjects that are ineligible. Now conduct descriptive statis- tics of your sample by treatment group in your analytic sample. Compare the characteristics of the study sample with your analytic sample.

- The following table is the descriptive statistics table following the inclusion/exclusion criteria listed above. Because `location` and `location_type` are now only one level due to us only keeping the Odenton location, they also have been excluded from the table. `age_group` is also excluded from the table due to adults only being recruited.

```
##                     Stratified by practice_type
##                      1               2           p      test
##   n                    242           184
##   age (mean (SD))    21.61 (2.58)  22.60 (2.32)  <0.001
##   race (%)                                        0.045
##      0                  115 (47.5)    111 (60.3)
##      1                   70 (28.9)     42 (22.8)
##      2                    8 ( 3.3)      2 ( 1.1)
##      3                   49 (20.2)     29 (15.8)
##   insurance_type (%)                              0.017
##      0                    5 ( 2.1)      2 ( 1.1)
##      1                  140 (57.9)    133 (72.3)
##      2                     5 ( 2.1)      1 ( 0.5)
##      3                    92 (38.0)     48 (26.1)
##   med_assist = 1 (%)       5 ( 2.1)      2 ( 1.1)   0.687
```

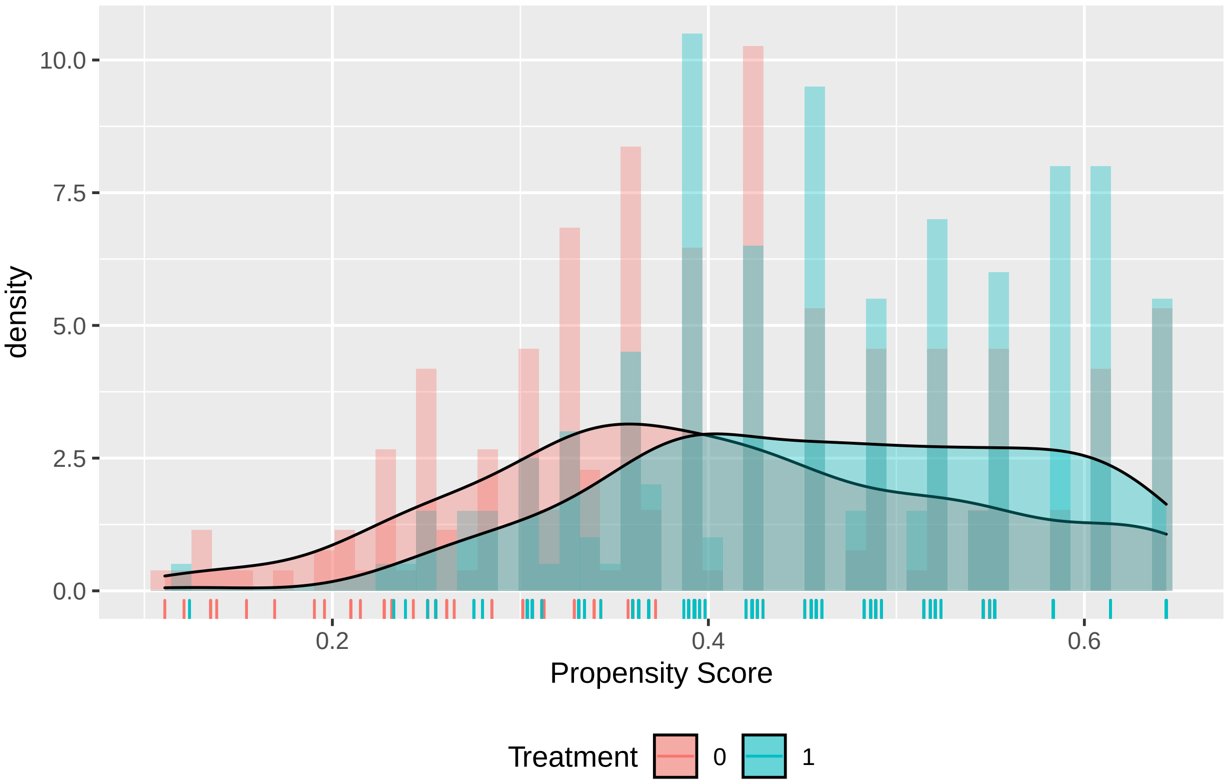
Comparing the pre-parsed and post-parsed tables, we notice the following:

- The average ages for family practice (from 19.46 to 21.61) and OB-GYN (from 21.43 to 22.60) increase post-parsed.
- Looking at practice type and race:
  - The percentage breakdowns for white, Black, and Hispanic patients are similar pre- and post-parsing for those who went to a family practice. The percentage breakdown for unknown/other ethnicity/race patients for those who went to a family practice decrease slightly (23.0 to 20.2).
  - The percentage breakdowns for white patients are similar pre- and post-parsing for those who went to an OB-GYN. The percentage breakdown for those who went to an OB-GYN decrease slightly for Black patients (27.6 to 22.8), and Hispanic patients (2.4 to 1.1). TThe percentage breakdown for those who went to an OB-GYN increase for unknown/other ethnicity/race patients (7.9 to 15.8).
- Looking at practice type and insurance type:
  - For level 0: The percentages differ pre- and post-parsing for both practice types. The pre and post difference for OB-GYN is great, going from 11.1% to 1.1%.
  - For private payer insurance type (level 1): The percentages increase slightly post-parsing for both practice types. Family practice goes from 51.5% to 57.9, and OB-GYN goes from 68.3% to 72.3%.
  - For hospital based insurance type (level 2): The percentages differ pre- and post-parsing for both practice types. The pre and post difference for OB-GYN is great, going from 9.4% to 0.5%.
  - For military insurance type (level 3): The percentages differ in different directions for each practice type post-parsing. Family practice goes from 42.7% to 38.0%, and OB-GYN goes from 11.3% to 26.1%.
- The percentages differ pre- and post-parsing for both practice types when looking at medical assistance. The pre and post difference for family practice decreases a bit, going from 3.3% to 2.1%. The pre and post difference for OB-GYN is great, going from 11.1% to 1.1%.

Question 4

Estimate the propensity scores in the analytic sample. Interpret the results of the model.

```
## 
## Call:
## glm(formula = practice_type ~ age + race + insurance_type + med_assist, 
##     family = binomial, data = x)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.4364  -1.0481  -0.7635   1.1948   2.0432  
## 
## Coefficients: (1 not defined because of singularities)
##                 Estimate Std. Error z value Pr(>|z|)   
## (Intercept)     -3.43352    1.34569  -2.552   0.0107 * 
## age              0.12697    0.04375   2.902   0.0037 **
## race1           -0.39164    0.25152  -1.557   0.1195   
## race2           -1.25161    0.81316  -1.539   0.1238
## race3           -0.49569    0.27556  -1.799   0.0720 .
## insurance_type1  0.72322    0.86017   0.841   0.4005
## insurance_type2 -0.44255    1.40312  -0.315   0.7525
## insurance_type3  0.31772    0.88107   0.361   0.7184
## med_assist1           NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 582.64  on 425  degrees of freedom
## Residual deviance: 555.06  on 418  degrees of freedom
## AIC: 571.06
##
## Number of Fisher Scoring iterations: 4
```


Densities and histograms of propensity scores by treatment group

- Based off the density plot, we can see that the covariate balance looks decent, as a large portion of each treatment group's propensity score density overlaps each other. That being said, better covariate balance could be obtained, as certain parts of the densities do not overlap each other.

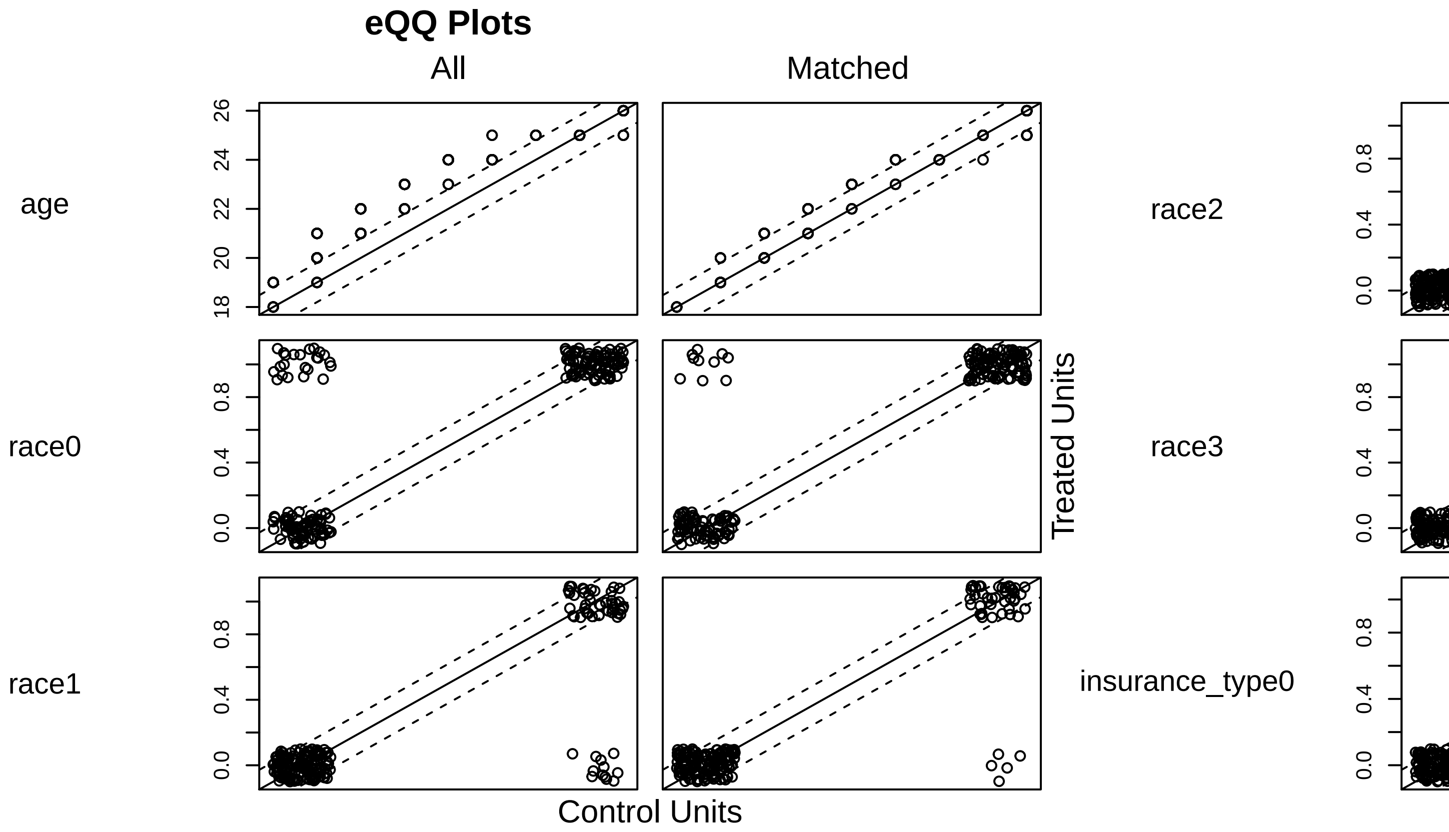
Question 5

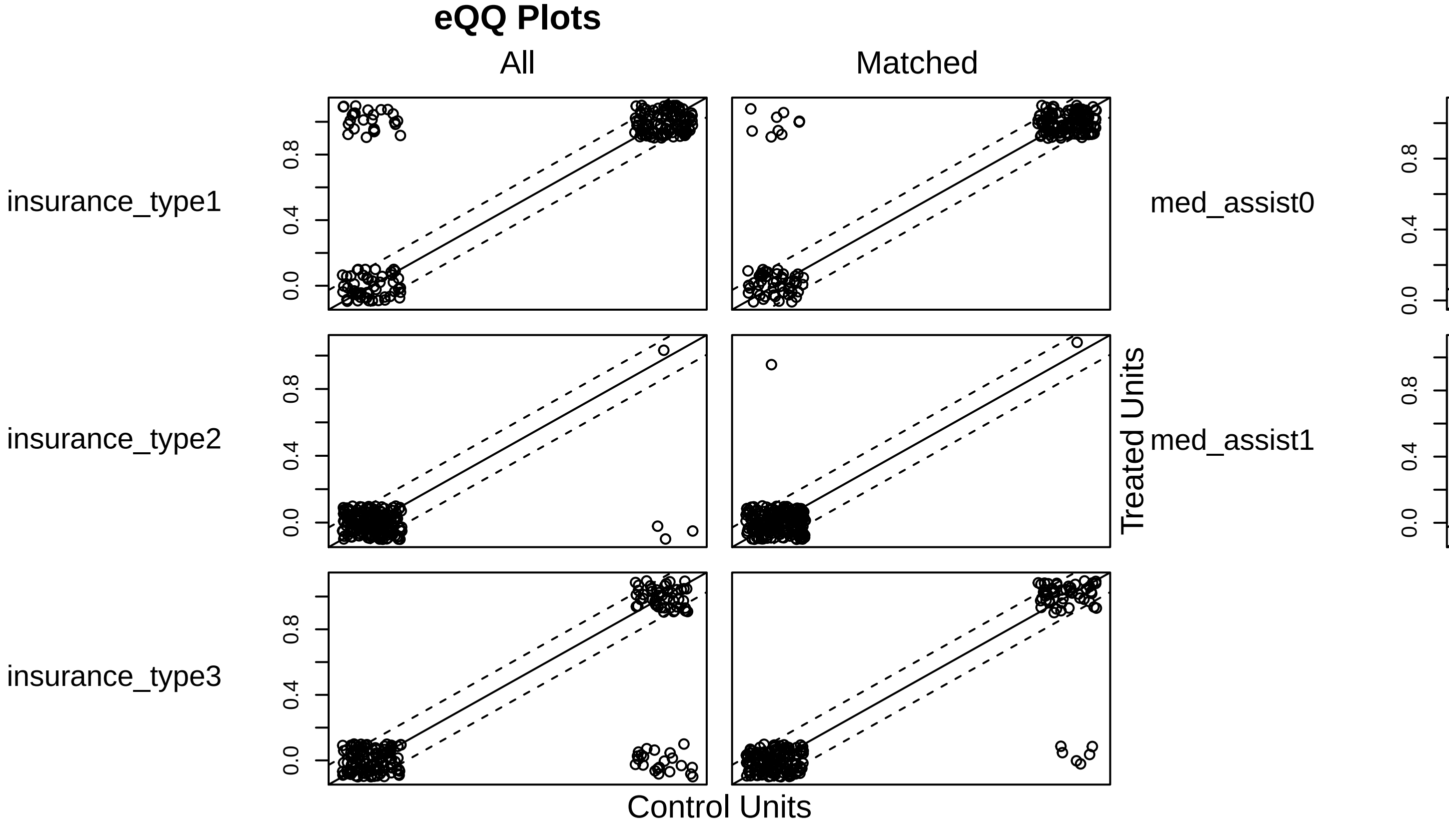
Use matching to improve covariate balance. Include your thought process, how you ultimately decide to do the matching, and a plot or table showing improvement in covariate balance.

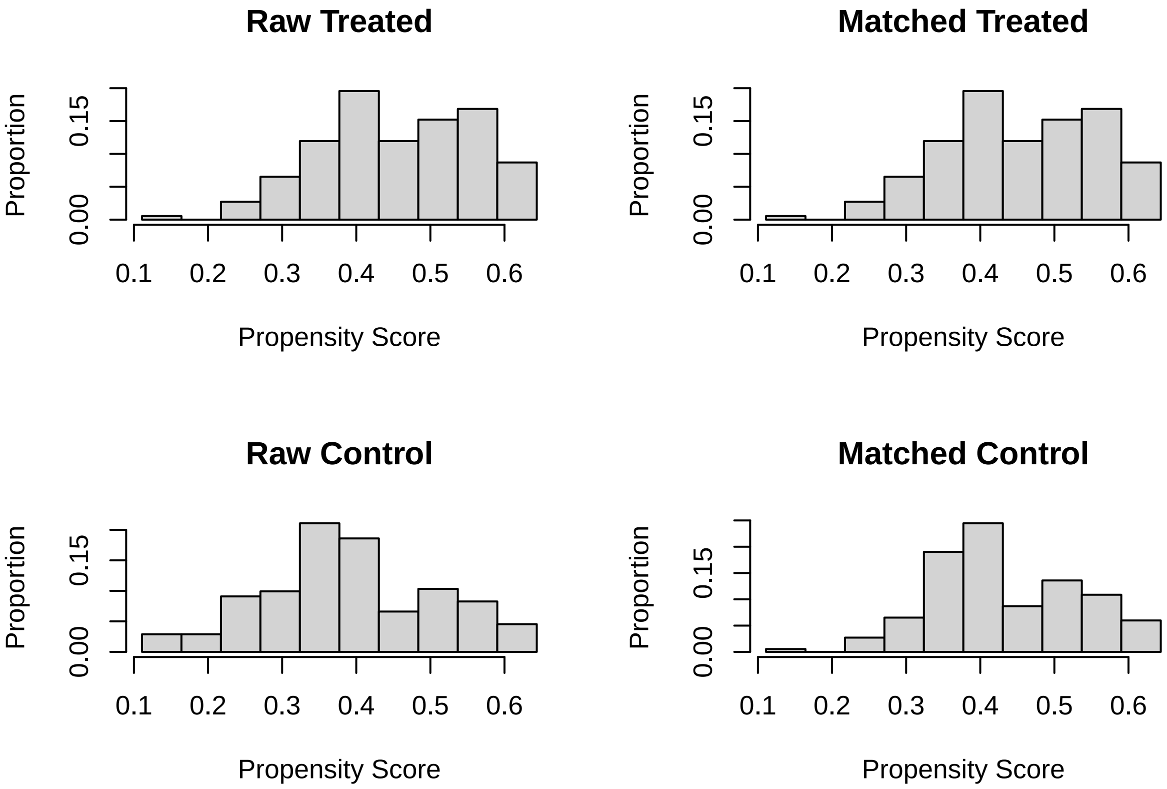
```
##
```

```
## Call:
## matchit(formula = practice_type ~ age + race + insurance_type + 
##     med_assist, data = x, method = "nearest", distance = "logit", 
##     discard = "control")
## 
## Summary of Balance for All Data:
##                 Means Treated Means Control Std. Mean Diff. Var. Ratio
## distance               0.4672        0.4051          0.5593     0.7607
## age                   22.6033       21.6074          0.4299     0.8074
## race0                  0.6033        0.4752          0.2618          .
## race1                  0.2283        0.2893         -0.1453          .
## race2                  0.0109        0.0331         -0.2140          .
## race3                  0.1576        0.2025         -0.1231          .
## insurance_type0        0.0109        0.0207         -0.0944          .
## insurance_type1        0.7228        0.5785          0.3224          .
## insurance_type2        0.0054        0.0207         -0.2071          .
## insurance_type3        0.2609        0.3802         -0.2717          .
## med_assist0            0.9891        0.9793          0.0944          .
## med_assist1            0.0109        0.0207         -0.0944          .
##                 eCDF Mean eCDF Max
## distance           0.1348   0.2465
## age                0.1143   0.1987
## race0              0.1281   0.1281
## race1              0.0610   0.0610
## race2              0.0222   0.0222
## race3              0.0449   0.0449
## insurance_type0    0.0098   0.0098
## insurance_type1    0.1443   0.1443
## insurance_type2    0.0152   0.0152
## insurance_type3    0.1193   0.1193
## med_assist0        0.0098   0.0098
## med_assist1        0.0098   0.0098
## 
## 
## Summary of Balance for Matched Data:
##                 Means Treated Means Control Std. Mean Diff. Var. Ratio
## distance               0.4672        0.4495          0.1589     1.0494
## age                   22.6033       22.2935          0.1337     0.8862
## race0                  0.6033        0.5543          0.1000          .
## race1                  0.2283        0.2609         -0.0777          .
## race2                  0.0109        0.0000          0.1048          .
## race3                  0.1576        0.1848         -0.0746          .
## insurance_type0        0.0109        0.0163         -0.0524          .
## insurance_type1        0.7228        0.6793          0.0971          .
## insurance_type2        0.0054        0.0054          0.0000          .
## insurance_type3        0.2609        0.2989         -0.0866          .
## med_assist0            0.9891        0.9837          0.0524          .
## med_assist1            0.0109        0.0163         -0.0524          .
##                 eCDF Mean eCDF Max Std. Pair Dist.
## distance           0.0344   0.1196          0.1700
## age                0.0477   0.1087          0.5513
## race0              0.0489   0.0489          0.6554
## race1              0.0326   0.0326          0.6474
## race2              0.0109   0.0109          0.1048
## race3              0.0272   0.0272          0.3729
## insurance_type0    0.0054   0.0054          0.0524
## insurance_type1    0.0435   0.0435          0.6800
## insurance_type2    0.0000   0.0000          0.0109
## insurance_type3    0.0380   0.0380          0.7055
## med_assist0        0.0054   0.0054          0.0524
## med_assist1        0.0054   0.0054          0.0524
##
## Percent Balance Improvement:
##                 Std. Mean Diff. Var. Ratio eCDF Mean eCDF Max
## distance                   71.6       82.4      74.5     51.5
## age                        68.9       43.5      58.3     45.3
## race0                      61.8          .      61.8     61.8
## race1                      46.5          .      46.5     46.5
## race2                      51.0          .      51.0     51.0
## race3                      39.4          .      39.4     39.4
## insurance_type0            44.5          .      44.5     44.5
## insurance_type1            69.9          .      69.9     69.9
## insurance_type2           100.0          .     100.0    100.0
## insurance_type3            68.1          .      68.1     68.1
## med_assist0                44.5          .      44.5     44.5
## med_assist1                44.5          .      44.5     44.5
##
## Sample Sizes:
##           Control Treated
## All           242     184
## Matched       184     184
## Unmatched      56       0
## Discarded       2       0
```

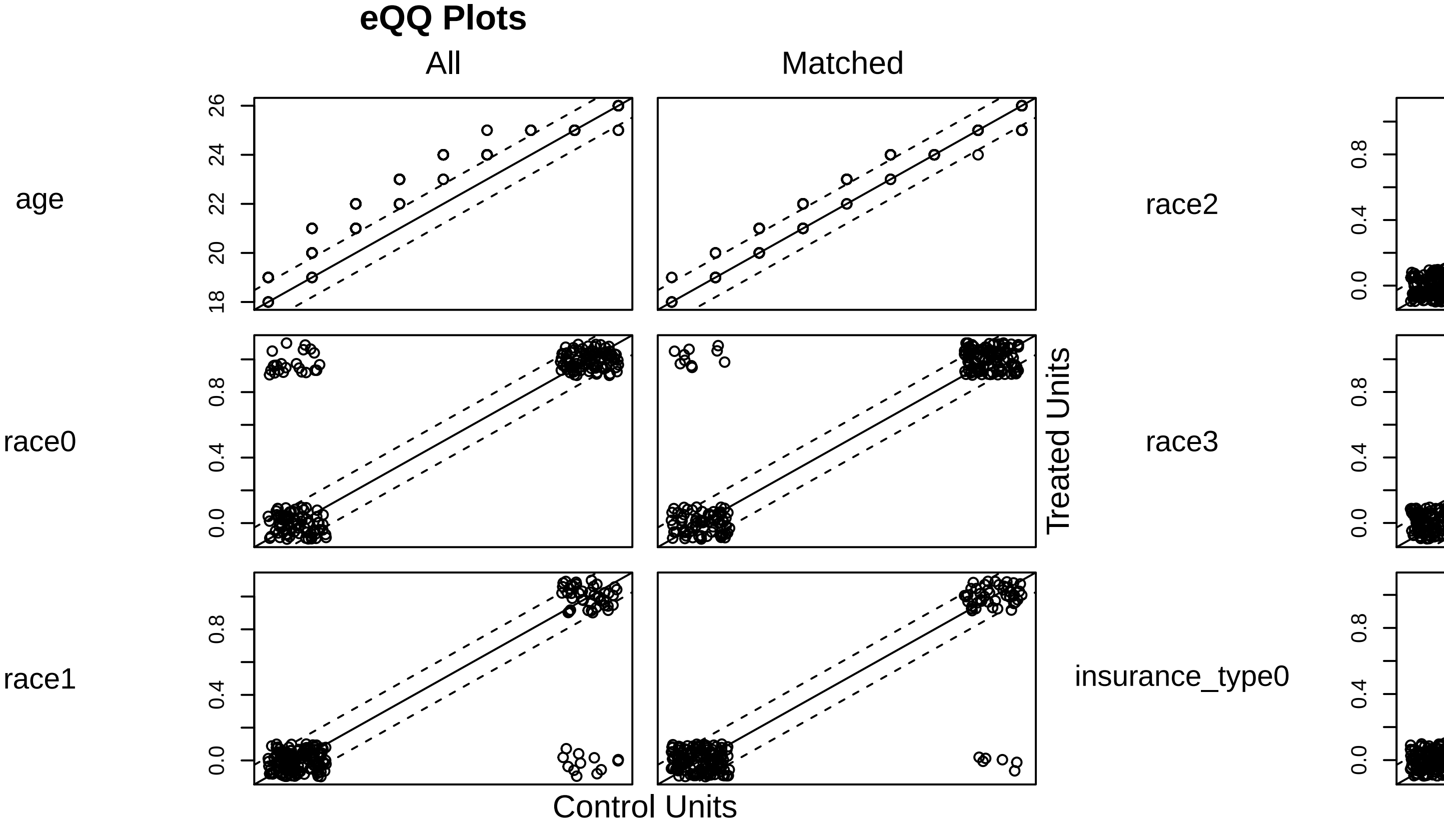
```
##                      Stratified by practice_type
##                       1             2             SMD
##   n                    242           184
##   age (mean (SD))    21.61 (2.58)  22.60 (2.32)   0.406
##   race (%)                                        0.283
##      0                 115 (47.5)    111 (60.3)
##      1                  70 (28.9)     42 (22.8)
##      2                   8 ( 3.3)      2 ( 1.1)
##      3                  49 (20.2)     29 (15.8)
##   insurance_type (%)                              0.321
##      0                   5 ( 2.1)      2 ( 1.1)
##      1                 140 (57.9)    133 (72.3)
##      2                   5 ( 2.1)      1 ( 0.5)
##      3                  92 (38.0)     48 (26.1)
##   med_assist = 1 (%)     5 ( 2.1)      2 ( 1.1)   0.079
```

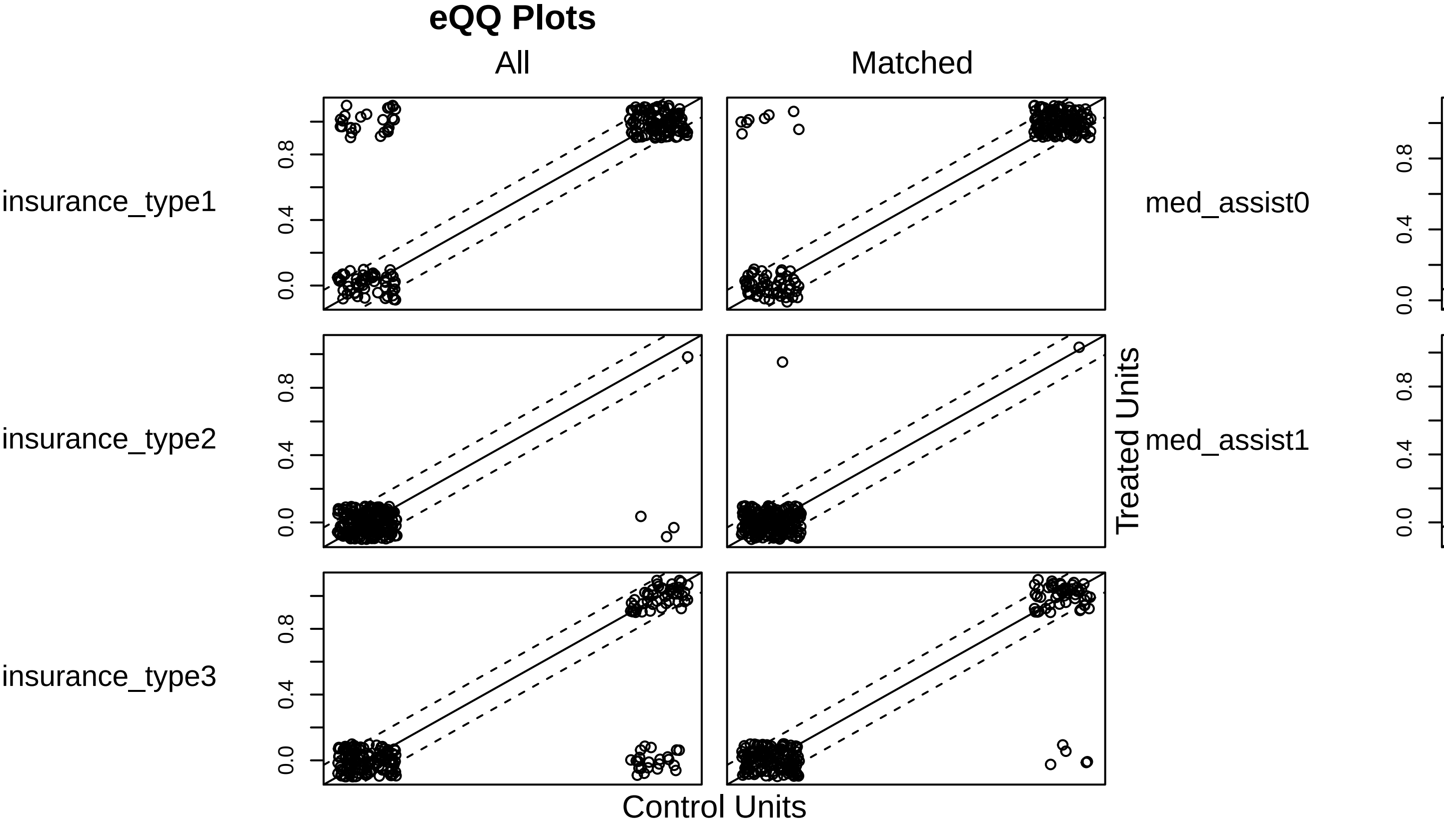

eQQ Plots
All
Matched
age
race0
race1
race2
race3
insurance_type0
18
20
22
24
26
0.0
0.4
0.8
Treated Units
Control Units

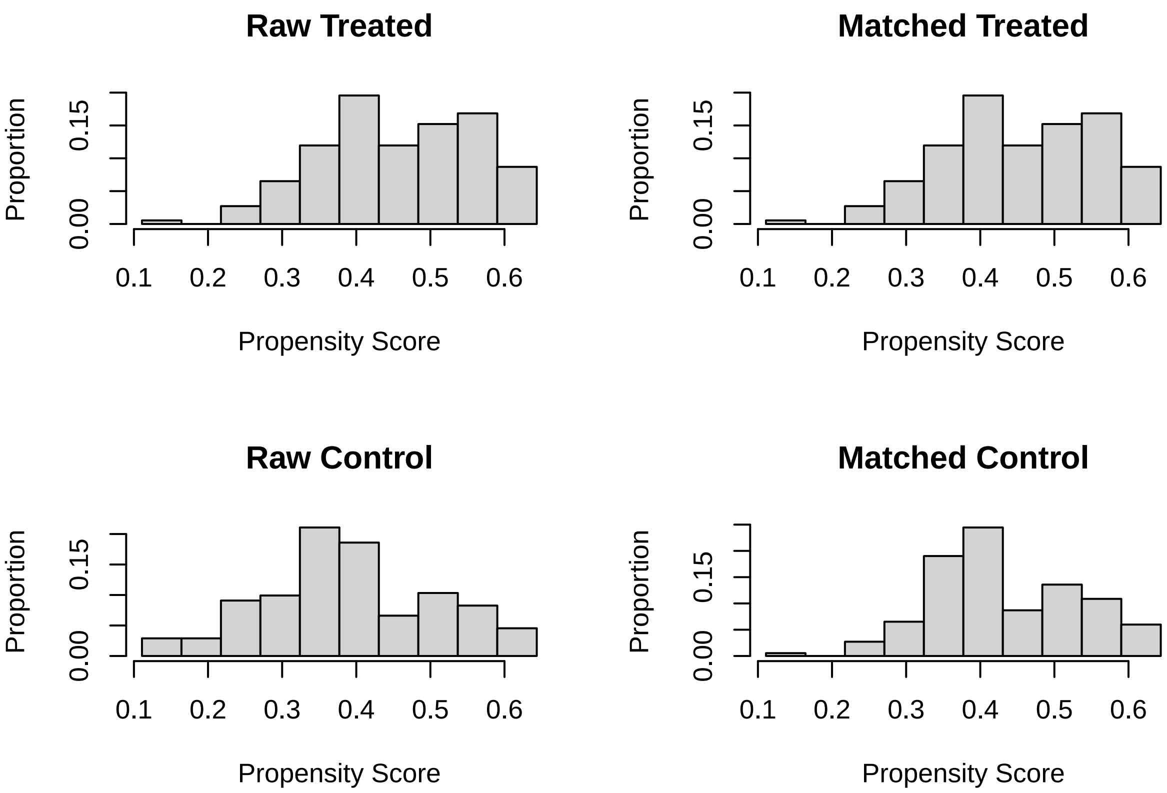

eQQ Plots
All
Matched
insurance_type1
insurance_type2
insurance_type3
med_assist0
med_assist1
0.0
0.4
0.8
Treated Units
Control Units



```
##                      Stratified by practice_type
##                       1             2             SMD
##   n                     242           184
##   age (mean (SD))    21.61 (2.58)  22.60 (2.32)   0.406
##   race (%)                                        0.283
##      0                 115 (47.5)    111 (60.3)
##      1                  70 (28.9)     42 (22.8)
##      2                   8 ( 3.3)      2 ( 1.1)
##      3                  49 (20.2)     29 (15.8)
##   insurance_type (%)                              0.321
##      0                   5 ( 2.1)      2 ( 1.1)
##      1                 140 (57.9)    133 (72.3)
##      2                   5 ( 2.1)      1 ( 0.5)
##      3                  92 (38.0)     48 (26.1)
##   med_assist = 1 (%)     5 ( 2.1)      2 ( 1.1)   0.079
```


eQQ Plots
All
Matched
age
race0
race1
race2
race3
insurance_type0
Treated Units
Control Units
18
20
22
24
26
0.0
0.4
0.8


eQQ Plots
All
Matched
insurance_type1
insurance_type2
insurance_type3
med_assist0
med_assist1
Treated Units
Control Units
0.0
0.4
0.8



- Looking at the percent balance improvement for both the greedy and optimal matching methods, we can see that greedy matching has slightly better SMD percent balance improvement. Looking at the histograms, the greedy and optimal matching are very similar. Looking into the eQQ plots, we can see that the greedy matching moves more of the matched observations inside the ideal boundaries (particularly for `race1`). Thus, we will utilize greedy matching.

Question 6

Using your matches from Question (5), estimate the average causal effect of treatment among the treated (ATT) on rates of vaccination regimen completion. Give a point estimate, a confidence interval, and a p-value for whether it had any effect, and interpret these results in context.

- The estimate for `practice_type` is 0.0748908 (-0.0138037,0.1635854), with a p-value of 0.0988063, meaning that it is not significant when looking at $\alpha = 0.05$. Thus, practice type does not have any effect on the completion of the Gardasil shots.

Question 7

Estimate the average causal effect of treatment among the controls (ATC) and the treated (ATT) now using nearest neighbour match on the PS score, one control matched with one treated (1:1), without replacement or calipers. Combine the estimates from ATC and ATT to estimate the average treatment effect on rates of vaccination regimen completion. Interpret your results.

```
## Warning: Fewer treated units than control units; not all control units will get
## a match.
```

- The ATE on rates of vaccination completion based off practice type is 0.0956504. Thus, there is a average treatment effect of vaccination appears to be better for those who go to an OB-GYN than for those who go to a family practice.