Homework 4

Ngoc Duong

12/5/2020

Question 1

```
##                         Stratified by practice_type
##                          0             1              2             SMD
##   n                        515           365            533
##   age (mean (SD))       14.92 (2.25)  19.46 (3.82)   21.43 (3.33)   1.429
##   age_group = 1 (%)        33 ( 6.4)    242 ( 66.3)    437 (82.0)   1.434
##   race (%)                                                           0.390
##      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 (%)                                                 0.915
##      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)   0.662
##   location (%)                                                       1.660
##      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)   1.061
```

We have 1413 subjects in the data, among whom 515 report attending a prediatric practice, 365 attending a general practice, and 533 attending an ob-gyn clinic.

Some quick observations from the descriptives above: the mean age for the pediatric clinic group is 14.92, while that for the general practice is 19.46, and 21.43 years old for the ob-gyn clinic. Insurance type 0 has the same distribution across 3 clinic types as medical assistance, suggesting women who are on medical assistance are also on insurance type 0. Finally, women living in location 2 and 3 in this sample exclusively go to ob-gyn clinic.

Question 2

- The protocol for the RCT:

(i) Control arm: patients attending pediatrics clinic and family medicine practice are grouped as one control arm

Treatment arm: patients attending ob-gyn clinic

The question of interest is whether practice type affects rate of vaccine completion, so the arms should be by practice type. There are 3 practice types in the sample, and I decide to group pediatric and general practice to preserve the sample size and power. Otherwise, I might need to exclude pediatric clinic from the analytic sample (515 subjects). As a trade-off, there could be bias if we want to compare the average completion rate between ob-gyn and general practice, but we can also propose to compare between ob-gyn practice and non-ob-gyn practice in this case.

- Cross-tab between 4 locations and treatment

```
##
##       1   2   3   4
##   0 581   0   0 299
##   1 217 165  89  62
```

- Cross-tab between age and treatment

```
##
##      11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26
##   0  29  74  81  79  95 107  98  70  61  43  34  22  26  11  21  29
##   1   1   1   2   6  15  27  30  29  44  49  49  42  54  66  65  53
```

(ii) Eligibility criteria based on levels of baseline characteristics

We want to ensure probabilistic assumption for the assignment.

We can see from the table above that none of the participants in location 2 and 3 reported going to an ob-gyn clinic, so I decided to limit the eligibility criteria to only participants in location 1 and 4, because otherwise, among the participants in location 2 and 3, the probabilistic assumption is violated.

We can also see that too few participants 11-14 years of age go to ob-gyn clinic (around 1 to 6); therefore, I decide to limit eligibility to participants at least 15 years old to increase overlap, although this will eliminate some amount of participants who go to pediatric clinic.

Although we have female patients who are above a certain age (21) do not go to pediatric clinic anymore, since we have grouped pediatric and family practice as one control arm, having patients older than 21 should not violate the probabilistic assumption.

- Some other considrations about the baseline covariates:

Age group and location type are collapsed version of age and location (which may result in loss of information and might also increase bias), so I exclude these two variables.

As noted before, medical assistance fully overlaps with insurance type 0; therefore, I don't consider this covariate.

Question 3

We have excluded 522 subjects. The remaining number of subjects is 891

Descriptive statistics of analytic sample

```
##                     Stratified by practicetype.bin
##                      0             1             SMD
##   n                    617           274
##   age (mean (SD))    18.51 (3.13)  21.74 (3.12)   1.033
##   race (%)                                        0.396
##      0                 269 (43.6)    171 (62.4)
##      1                 210 (34.0)     64 (23.4)
##      2                  25 ( 4.1)      4 ( 1.5)
##      3                 113 (18.3)     35 (12.8)
```

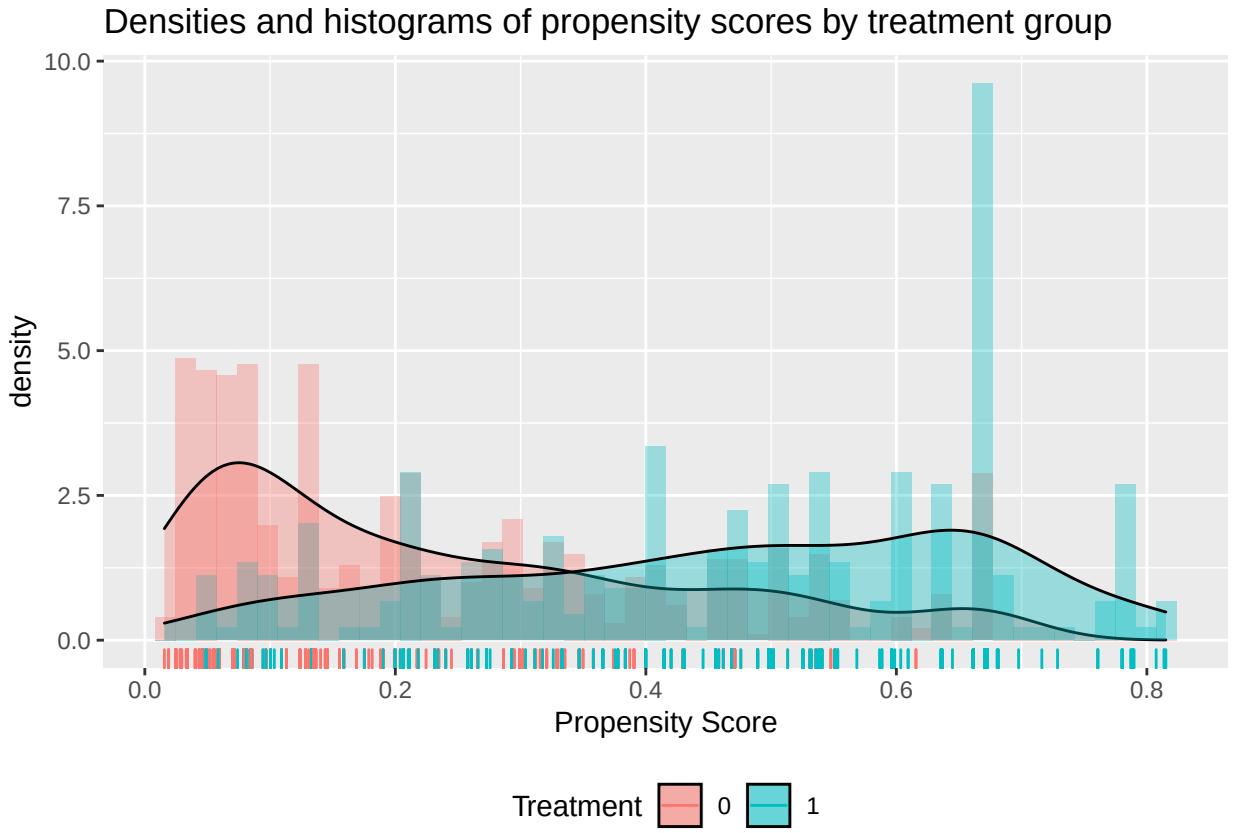
```
##   insurance_type (%)                              0.552
##      0                119 (19.3)     18 ( 6.6)
##      1                275 (44.6)    186 (67.9)
##      2                 22 ( 3.6)     14 ( 5.1)
##      3                201 (32.6)     56 (20.4)
##   location = 4 (%)    164 (26.6)     62 (22.6)   0.092
```

Compared with the study sample, the analytic sample has fewer observations (617 in the control group, and 274 in the treatment group).

We have more balanced covariates between the two arms for all covariates of interest (the SMDs decreased after restricting sample by eligibility). Location seems to be quite balanced (SMD < 0.2). The proportion of categorical covariates like race and insurance type might change a bit (after combibing pediatric and general practice), and age also changes a little (somewhere in the middle of pediatric and general practice mean ages).

Question 4

```
##
## Call:
## glm(formula = practicetype.bin ~ age + I(age^2) + age + race +
##     insurance_type + location, family = binomial, data = x_new)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4944  -0.7981  -0.4124   0.8939   2.4794
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -19.28468    3.75069  -5.142 2.72e-07 ***
## age               1.50624    0.36678   4.107 4.01e-05 ***
## I(age^2)         -0.02971    0.00877  -3.388 0.000704 ***
## race1            -0.56782    0.19901  -2.853 0.004328 **
## race2            -1.17938    0.59913  -1.969 0.049010 *
## race3            -0.71241    0.24308  -2.931 0.003381 **
## insurance_type1   0.91737    0.34144   2.687 0.007215 **
## insurance_type2   1.08361    0.47945   2.260 0.023814 *
## insurance_type3   0.41065    0.38912   1.055 0.291274
## location4         0.59755    0.25695   2.325 0.020046 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1099.68  on 890  degrees of freedom
## Residual deviance:  883.98  on 881  degrees of freedom
## AIC: 903.98
##
## Number of Fisher Scoring iterations: 5
```



We observe that even after restricting eligibility criteria, there is still some subjects in treatment group that has no overlap with the control group. We later use matching to find a control with the nearest propensity score in some specified distance to these treated units. When we do this, depending on the caliper (if we decide to use it), (some of) these treated units might be excluded.

Intepret the results of PS model

Based on this propensity score model (regression output above), we see age and its quadratic term are highly associated with the propensity for being in the treated group. Specifically, the estimate coefficient for age is positive and has relatively large magnitude, which suggests the older the subjects, the more likely they go to ob-gyn clinic. Similarly, women with insurance type 1 and 2 are more likely to go to ob-gyn clinic compared to those with insurance type 0 (or who we saw were the subjects who need medical assistance). Finally, women in location 4 are more likely to be in the treatment group compared to those in location 1 as suggested by this PS model.

Question 5

Optimal matching

```
##
## Call:
## matchit(formula = practicetype.bin ~ age + I(age^2) + race +
##     insurance_type + location, data = x_new, method = "optimal",
##     distance = "logit", discard = "control")
##
## Summary of Balance for All Data:
##                  Means Treated Means Control Std. Mean Diff. Var. Ratio
## distance                0.4648        0.2377          1.1146     1.1720
```

```
## age                  21.7409       18.5105          1.0344     0.9943
## I(age^2)            482.3832      352.4327          0.9887     1.0977
## race0                 0.6241        0.4360          0.3884          .
## race1                 0.2336        0.3404         -0.2524          .
## race2                 0.0146        0.0405         -0.2161          .
## race3                 0.1277        0.1831         -0.1660          .
## insurance_type0       0.0657        0.1929         -0.5133          .
## insurance_type1       0.6788        0.4457          0.4993          .
## insurance_type2       0.0511        0.0357          0.0701          .
## insurance_type3       0.2044        0.3258         -0.3010          .
## location1             0.7737        0.7342          0.0945          .
## location4             0.2263        0.2658         -0.0945          .
##                 eCDF Mean eCDF Max
## distance           0.2886   0.4540
## age                0.2692   0.4540
## I(age^2)           0.2692   0.4540
## race0              0.1881   0.1881
## race1              0.1068   0.1068
## race2              0.0259   0.0259
## race3              0.0554   0.0554
## insurance_type0    0.1272   0.1272
## insurance_type1    0.2331   0.2331
## insurance_type2    0.0154   0.0154
## insurance_type3    0.1214   0.1214
## location1          0.0395   0.0395
## location4          0.0395   0.0395
##
##
## Summary of Balance for Matched Data:
##                 Means Treated Means Control Std. Mean Diff. Var. Ratio
## distance               0.4648        0.3966          0.3349     1.4910
## age                   21.7409       20.7810          0.3074     0.9793
## I(age^2)             482.3832      441.7737          0.3090     0.9812
## race0                 0.6241        0.5766          0.0980          .
## race1                 0.2336        0.2701         -0.0863          .
## race2                 0.0146        0.0219         -0.0609          .
## race3                 0.1277        0.1314         -0.0109          .
## insurance_type0       0.0657        0.0730         -0.0295          .
## insurance_type1       0.6788        0.6715          0.0156          .
## insurance_type2       0.0511        0.0365          0.0663          .
## insurance_type3       0.2044        0.2190         -0.0362          .
## location1             0.7737        0.7810         -0.0174          .
## location4             0.2263        0.2190          0.0174          .
##                 eCDF Mean eCDF Max Std. Pair Dist.
## distance           0.0630   0.2153          0.3350
## age                0.0818   0.1861          0.6182
## I(age^2)           0.0818   0.1861          0.6429
## race0              0.0474   0.0474          0.5802
## race1              0.0365   0.0365          0.6728
## race2              0.0073   0.0073          0.3043
## race3              0.0036   0.0036          0.6451
## insurance_type0    0.0073   0.0073          0.3241
## insurance_type1    0.0073   0.0073          0.6566
## insurance_type2    0.0146   0.0146          0.3315
```

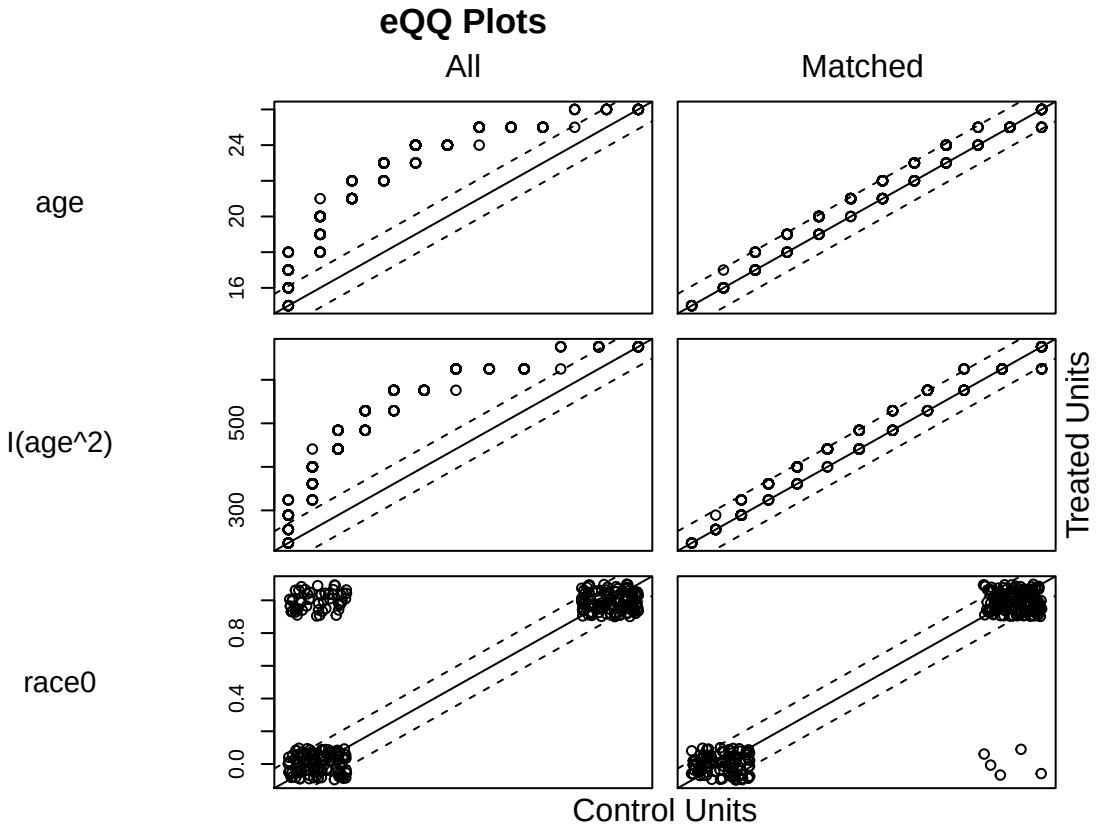
```
## insurance_type3    0.0146   0.0146          0.6516
## location1          0.0073   0.0073          0.7501
## location4          0.0073   0.0073          0.7501
##
## Percent Balance Improvement:
##                 Std. Mean Diff. Var. Ratio eCDF Mean eCDF Max
## distance                 70.0     -151.6      78.2     52.6
## age                      70.3     -267.7      69.6     59.0
## I(age^2)                 68.8       79.7      69.6     59.0
## race0                    74.8          .      74.8     74.8
## race1                    65.8          .      65.8     65.8
## race2                    71.8          .      71.8     71.8
## race3                    93.4          .      93.4     93.4
## insurance_type0          94.3          .      94.3     94.3
## insurance_type1          96.9          .      96.9     96.9
## insurance_type2           5.4          .       5.4      5.4
## insurance_type3          88.0          .      88.0     88.0
## location1                81.5          .      81.5     81.5
## location4                81.5          .      81.5     81.5
##
## Sample Sizes:
##           Control Treated
## All           617     274
## Matched       274     274
## Unmatched     280       0
## Discarded      63       0
```

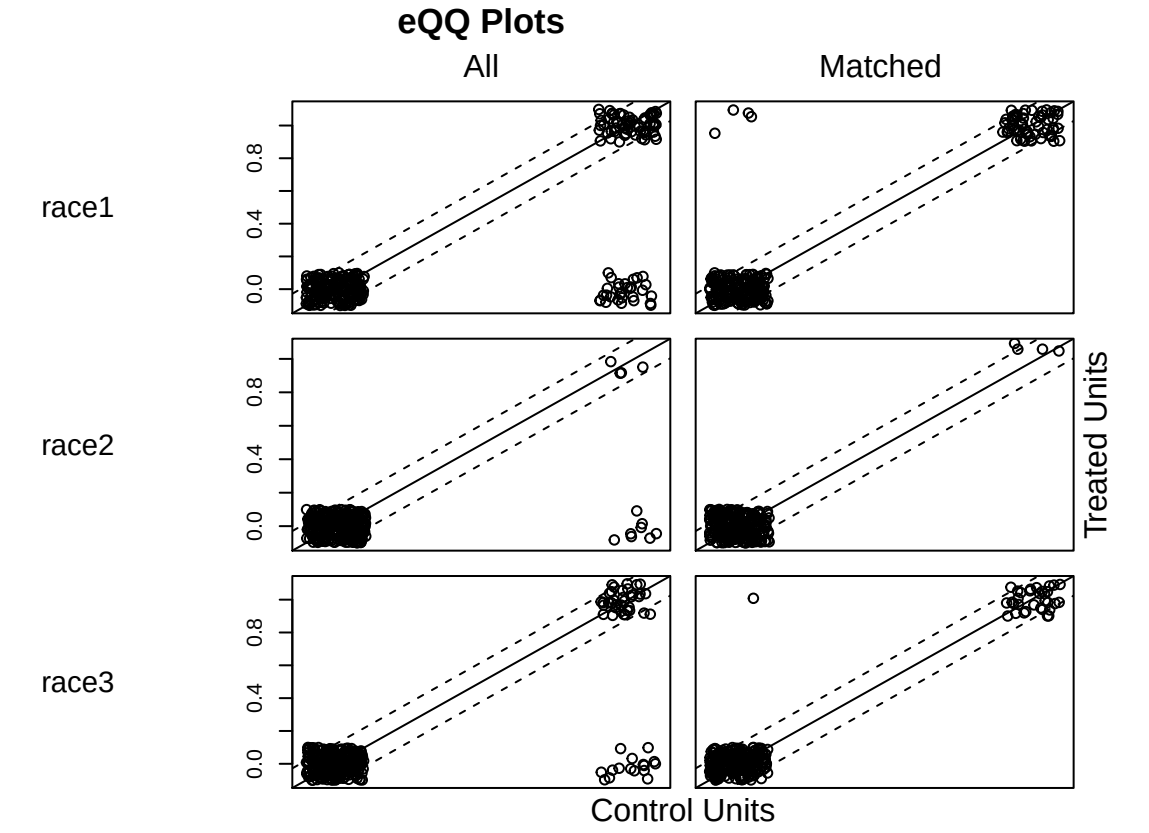
Nearest-neighbor matching with caliper

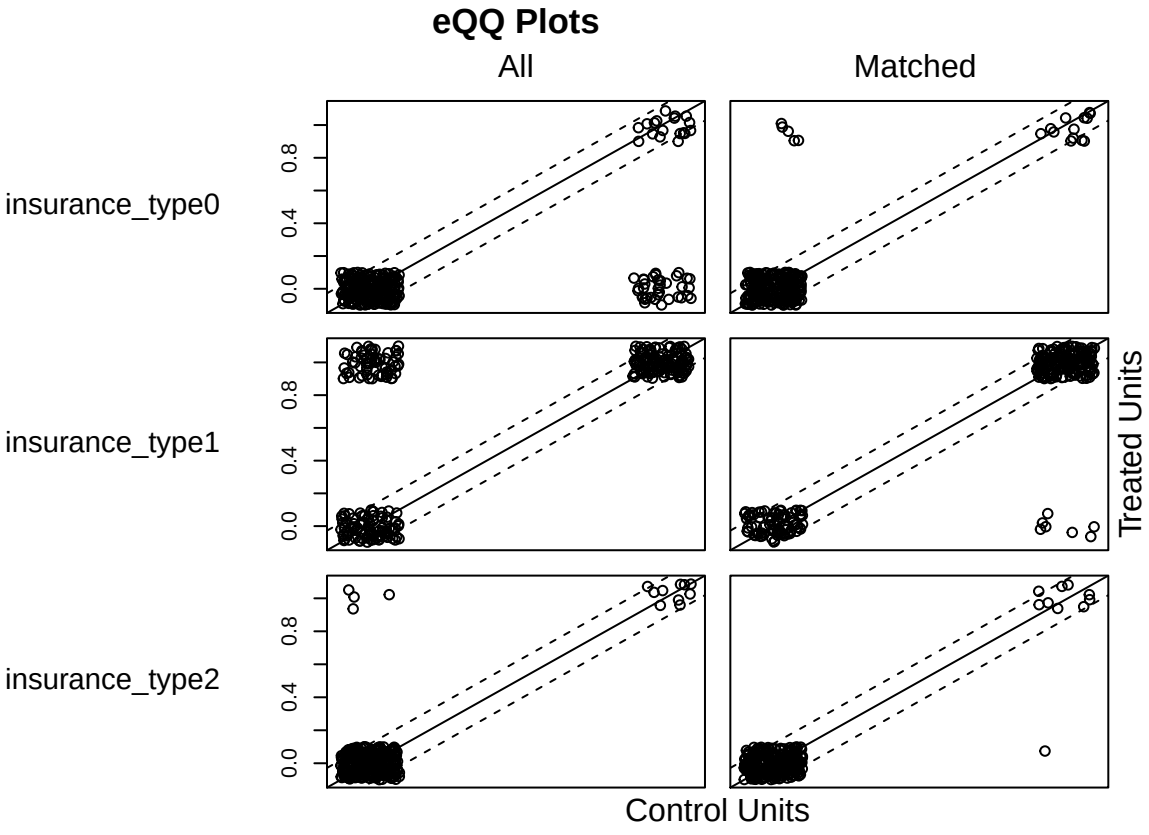
```
##
## Call:
## matchit(formula = practicetype.bin ~ age + I(age^2) + race +
##     insurance_type + location, data = x_new, method = "nearest",
##     distance = "logit", discard = "control", caliper = 0.4)
##
## Summary of Balance for All Data:
##                 Means Treated Means Control Std. Mean Diff. Var. Ratio
## distance               0.4648        0.2377          1.1146     1.1720
## age                   21.7409       18.5105          1.0344     0.9943
## I(age^2)             482.3832      352.4327          0.9887     1.0977
## race0                  0.6241        0.4360          0.3884          .
## race1                  0.2336        0.3404         -0.2524          .
## race2                  0.0146        0.0405         -0.2161          .
## race3                  0.1277        0.1831         -0.1660          .
## insurance_type0        0.0657        0.1929         -0.5133          .
## insurance_type1        0.6788        0.4457          0.4993          .
## insurance_type2        0.0511        0.0357          0.0701          .
## insurance_type3        0.2044        0.3258         -0.3010          .
## location1              0.7737        0.7342          0.0945          .
## location4              0.2263        0.2658         -0.0945          .
##                 eCDF Mean eCDF Max
## distance            0.2886   0.4540
## age                 0.2692   0.4540
## I(age^2)            0.2692   0.4540
## race0               0.1881   0.1881
```

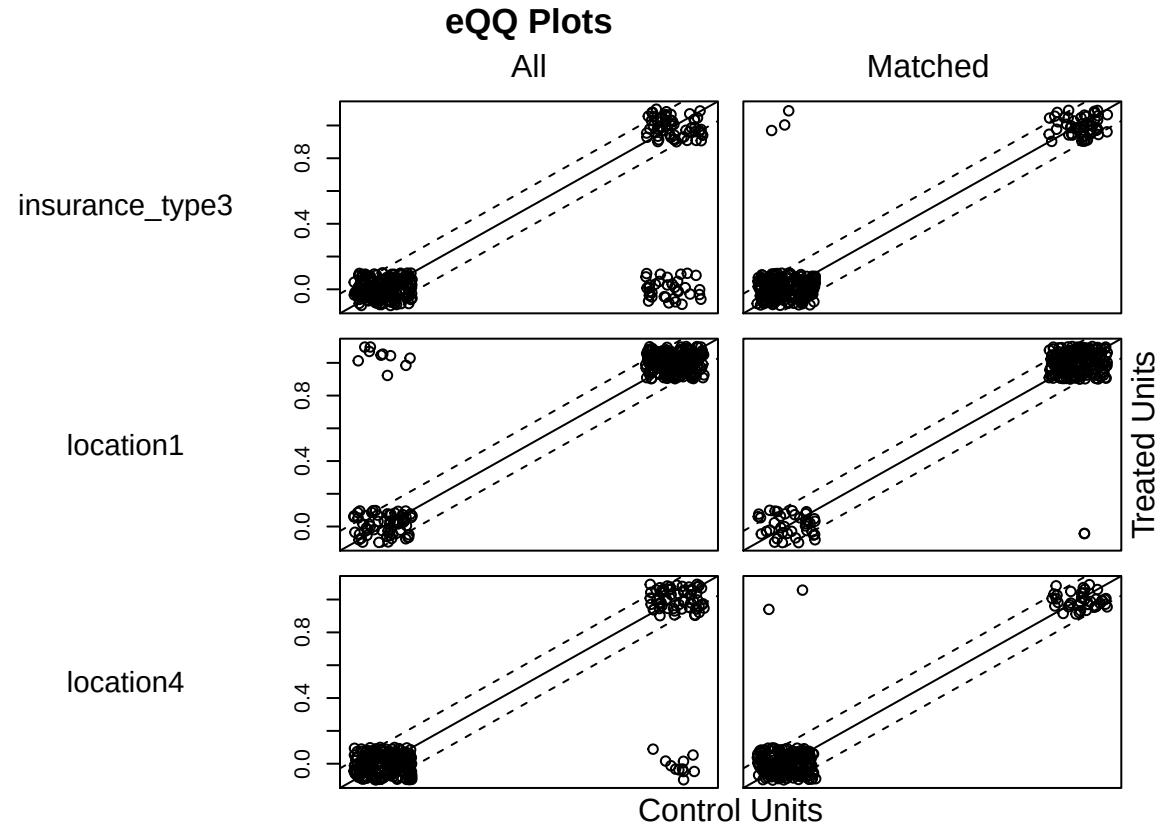
```
## race1              0.1068   0.1068
## race2              0.0259   0.0259
## race3              0.0554   0.0554
## insurance_type0    0.1272   0.1272
## insurance_type1    0.2331   0.2331
## insurance_type2    0.0154   0.0154
## insurance_type3    0.1214   0.1214
## location1          0.0395   0.0395
## location4          0.0395   0.0395
##
##
## Summary of Balance for Matched Data:
##                 Means Treated Means Control Std. Mean Diff. Var. Ratio
## distance               0.4269        0.4073          0.0961     1.1478
## age                   21.3891       20.9874          0.1286     0.9500
## I(age^2)             467.5230      451.0293          0.1255     0.9473
## race0                  0.5774        0.5983         -0.0432          .
## race1                  0.2678        0.2510          0.0396          .
## race2                  0.0167        0.0167          0.0000          .
## race3                  0.1381        0.1339          0.0125          .
## insurance_type0        0.0753        0.0544          0.0844          .
## insurance_type1        0.6527        0.6820         -0.0627          .
## insurance_type2        0.0377        0.0418         -0.0190          .
## insurance_type3        0.2343        0.2218          0.0311          .
## location1              0.8285        0.8368         -0.0200          .
## location4              0.1715        0.1632          0.0200          .
##                 eCDF Mean eCDF Max Std. Pair Dist.
## distance           0.0187   0.1004          0.0963
## age                0.0384   0.0795          0.5359
## I(age^2)           0.0384   0.0795          0.5712
## race0              0.0209   0.0209          0.6652
## race1              0.0167   0.0167          0.5736
## race2              0.0000   0.0000          0.0335
## race3              0.0042   0.0042          0.4137
## insurance_type0    0.0209   0.0209          0.2196
## insurance_type1    0.0293   0.0293          0.5287
## insurance_type2    0.0042   0.0042          0.2850
## insurance_type3    0.0126   0.0126          0.5292
## location1          0.0084   0.0084          0.4200
## location4          0.0084   0.0084          0.4200
##
## Percent Balance Improvement:
##                 Std. Mean Diff. Var. Ratio eCDF Mean eCDF Max
## distance                  91.4       13.1      93.5     77.9
## age                       87.6     -802.1      85.8     82.5
## I(age^2)                  87.3       41.9      85.8     82.5
## race0                     88.9          .      88.9     88.9
## race1                     84.3          .      84.3     84.3
## race2                    100.0          .     100.0    100.0
## race3                     92.4          .      92.4     92.4
## insurance_type0           83.5          .      83.5     83.5
## insurance_type1           87.4          .      87.4     87.4
## insurance_type2           72.9          .      72.9     72.9
## insurance_type3           89.7          .      89.7     89.7
```

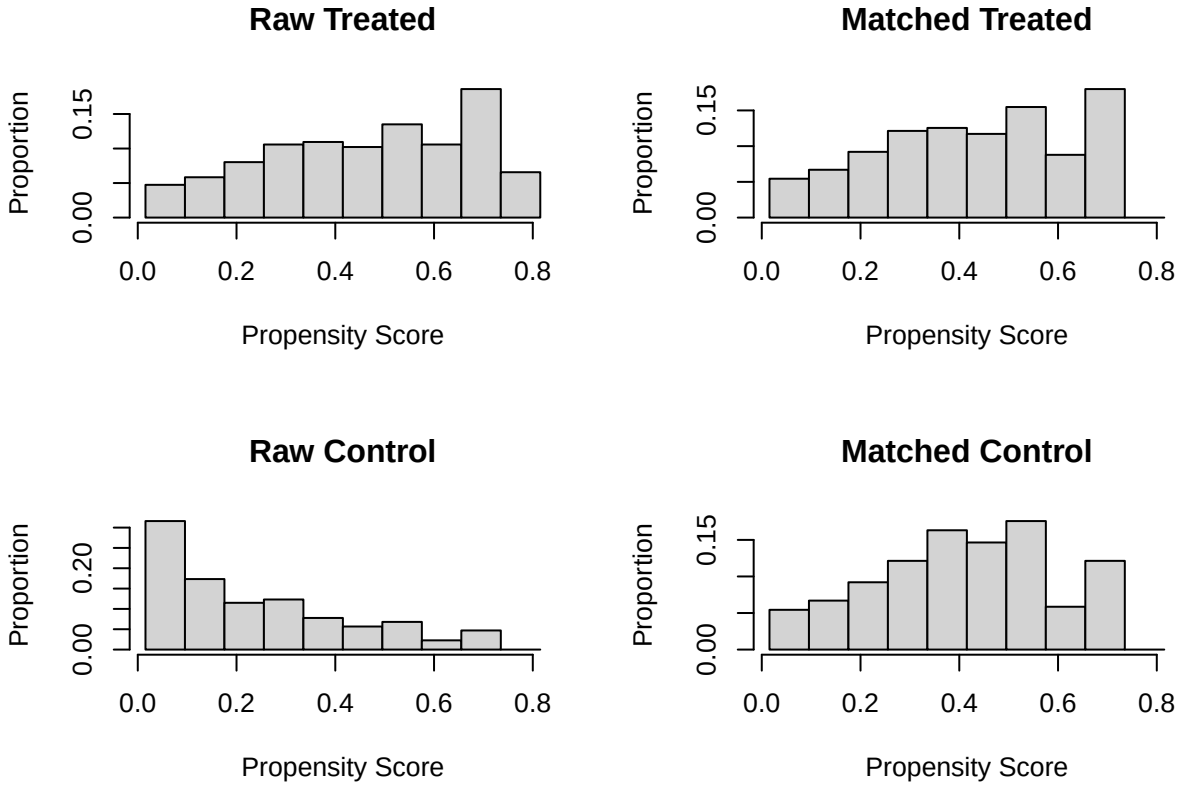
```
## location1                  78.8          .      78.8     78.8
## location4                  78.8          .      78.8     78.8
##
## Sample Sizes:
##           Control Treated
## All           617     274
## Matched       239     239
## Unmatched     315      35
## Discarded      63       0
```




eQQ Plots
All
Matched
race1
race2
race3
0.0
0.4
0.8
Treated Units
Control Units


eQQ Plots
All
Matched
insurance_type0
insurance_type1
insurance_type2
0.0
0.4
0.8
Treated Units
Control Units





- Process:

I specified the model for propensity score like in part 4. I used distance measured from the logit function of

the propensity score to account for the non-linearity in the substantive difference in the PS (easier to match for PS in the middle than at the two ends – PS around 0 or 0.8 in this case).

I originally used optimal matching to minimize overall distance, but after checking covariate balance (first table), there is still imbalance in variable age (SMD > 0.2). Therefore, I decided to use nearest neighbor matching without caliper. However, there is still some covariate imbalance. This makes sense because we saw in the PS histograms before that some subjects in the treated groups have quite high propensity scores compared to the rest. So I set caliper 0.4 to make sure the matched treated and controls are more similar based on their PS. Checking the SMDs<0.2 for all covariates helps confirm this matching strategy improves covariate balance.

However, by setting caliper to 0.4, I have also excluded some treated units (35) who don't have a nearest-neighbor match within the acceptable distance 0.4.

Since we match 1:1, we can look at the same table comparing covariates between treatment and control arm. We see the SMD reduced compared to before. The histograms also show more alignment in the distribution of propensity score between the treated units and the matched controls compared to before matching (right vs. left).

Question 6

```
##
## Call:
## svyglm(formula = completed ~ practicetype.bin + age + I(age^2) +
##     race + insurance_type + location, design = svydesign(~1,
##     weights = ~weights, data = psmatch.att.data), family = binomial)
##
## Survey design:
## svydesign(~1, weights = ~weights, data = psmatch.att.data)
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -4.13216    4.87038  -0.848  0.39664
## practicetype.bin1  0.63928    0.21387   2.989  0.00295 **
## age                0.27794    0.46621   0.596  0.55134
## I(age^2)          -0.00695    0.01107  -0.628  0.53033
## race1             -0.49874    0.25943  -1.922  0.05515 .
## race2              0.36432    0.70738   0.515  0.60677
## race3             -0.41763    0.34524  -1.210  0.22702
## insurance_type1    0.33341    0.49369   0.675  0.49980
## insurance_type2    0.37992    0.63595   0.597  0.55053
## insurance_type3   -0.09699    0.56285  -0.172  0.86327
## location4          0.19309    0.36327   0.532  0.59531
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1.008159)
##
## Number of Fisher Scoring iterations: 4
```

From the regression output above, the point estimate for the average treatment effect on the treated (ATT) is 0.6393. The 95% CI for this point estimate is $\widehat{ATT} \pm 1.96 \times SE = 0.6393 \pm 1.96 \times 0.2139 = [0.22, 1.059]$. The associated p-value is $0.00295 < 0.05$, so we can reject Ho and conclude the ATT is statistically significant (effect different from 0).

Interpretation

The estimated average treatment effect in the treated (ATT) is $0.6393 > 0$ (p-value < 0.05). In the context of the question, the estimated log odds ratio of completing vaccine regimen is 0.6393 among the treated compared to if they had not been treated. We are 95% confident that the true ATT (or the log odds ratio) lies between 0.22 and 1.059 (excluding null value 0). For this analytic sample, the treatment (attending ob-gyn clinic) has a causal effect on the completion of vaccine regimen among the treated (increase the chance of vaccine regimen completion on average).

We might also want to note that some limitation of matching method applies here, specifically decreased efficiency (throwing away data) and decreased generalizability. Here by making sure all covariates achieve balance for higher internal validity, I traded unbiasedness with efficiency and power.

Question 7

```
##
## Call:
## svyglm(formula = completed ~ practicetype.bin + age + I(age^2) +
##     race + insurance_type + location, design = svydesign(~1,
##     weights = ~weights, data = psmatch.att7.data), family = binomial)
##
## Survey design:
## svydesign(~1, weights = ~weights, data = psmatch.att7.data)
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -5.292066   4.846789  -1.092 0.275380
## practicetype.bin1  0.739537   0.200489   3.689 0.000248 ***
## age                0.305623   0.459001   0.666 0.505796
## I(age^2)          -0.006766   0.010817  -0.626 0.531895
## race1             -0.513038   0.249163  -2.059 0.039971 *
## race2              0.060119   0.699434   0.086 0.931536
## race3             -0.458851   0.316262  -1.451 0.147404
## insurance_type1    0.725114   0.447177   1.622 0.105490
## insurance_type2    0.879652   0.567025   1.551 0.121408
## insurance_type3    0.499121   0.518516   0.963 0.336183
## location4          0.543454   0.277807   1.956 0.050956 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1.018205)
##
## Number of Fisher Scoring iterations: 4
```

```
##
## Call:
## svyglm(formula = completed ~ practicetype.bin + age + I(age^2) +
##     race + insurance_type + location, design = svydesign(~1,
##     weights = ~weights, data = psmatch.atc7.data), family = binomial)
##
## Survey design:
## svydesign(~1, weights = ~weights, data = psmatch.atc7.data)
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)       -0.639896   4.327359  -0.148 0.882499
## practicetype.bin1 -0.270119   0.299471  -0.902 0.367468
## age               -0.054391   0.435217  -0.125 0.900591
## I(age^2)           0.001548   0.010553   0.147 0.883455
## race1             -0.518722   0.226813  -2.287 0.022584 *
## race2             -1.126661   0.801850  -1.405 0.160577
## race3             -0.536233   0.286436  -1.872 0.061738 .
## insurance_type1    0.906706   0.373417   2.428 0.015504 *
## insurance_type2    2.186062   0.589700   3.707 0.000231 ***
## insurance_type3    0.709358   0.409652   1.732 0.083917 .
## location4          0.384508   0.299040   1.286 0.199065
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 0.999924)
##
## Number of Fisher Scoring iterations: 4
```

We have $N_1 = 274$ matched treated when estimating ATT and N_0 274 matched controls when estimating ATC, so $N = N_1 + N_0 = 548$

Combining the ATC and ATT estimates, we have the ATE as

$$\widehat{ATE} = \widehat{ATT}\frac{N_1}{N} + \widehat{ATC}\frac{N_0}{N}$$

$$= 0.739537(\frac{274}{548}) - 0.270119(\frac{274}{548}) = 0.234$$

- Interpret:

The estimated ATT is 0.739 (>0): the estimated log odds ratio of completing the vaccine regimen is 0.739 among the treated (subpopulation) compared to if they had not been treated.

The estimated ATC is -0.27 (<0): the estimated log odds ratio of completing the vaccine regimen is -0.27 among the controls (subpopulation) compared to if they had been treated.

We have the estimated ATE of 0.23 (>0). Qualitatively, for this analytic sample, the treatment (practice type, specifically ob-gyn clinic) might have a causal effect on the completion of vaccine regimen. Having access to ob-gyn clinic might increase the chance of vaccine regimen completion on average.

However, 95% CI or p-value are necessary to draw further inference on the true effect (whether it is statistically significant).

Extra: use matchit code and specify the estimand to be ATE

```
##
## Call:
## svyglm(formula = completed ~ practicetype.bin + age + I(age^2) +
##     race + insurance_type + location, design = svydesign(~1,
##     weights = ~weights, data = psmatchate.data), family = binomial)
##
## Survey design:
## svydesign(~1, weights = ~weights, data = psmatchate.data)
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -1.296228   3.996033  -0.324  0.74573
```

```
## practicetype.bin1  0.203281   0.252399   0.805  0.42081
## age                0.062425   0.401191   0.156  0.87638
## I(age^2)          -0.003295   0.009827  -0.335  0.73747
## race1             -0.565632   0.258791  -2.186  0.02910 *
## race2              0.840852   0.526393   1.597  0.11054
## race3              0.040807   0.256600   0.159  0.87368
## insurance_type1    0.395383   0.360381   1.097  0.27289
## insurance_type2    1.359671   0.460974   2.950  0.00327 **
## insurance_type3    0.603790   0.408124   1.479  0.13938
## location4         -0.209024   0.300173  -0.696  0.48640
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1.00487)
##
## Number of Fisher Scoring iterations: 4
```

We see that the ATE estimates are close (small difference might be due to different matching methods used).

Appendix

```
knitr::opts_chunk$set(echo = TRUE)
library(tidyverse)
library(ggplot2)
library(personalized)
library(tableone)
library(MatchIt)
library(optmatch)
library(texreg)
library("MatchIt")
library("lmtest")
library("boot")
library(survey)
#import data
data = read.table("gardasil.dat.txt", header = TRUE) %>% janitor::clean_names() %>%
  mutate(across(c("age_group", "race", "med_assist", "insurance_type", "practice_type", "location", "lo
#0 pediatrics #1 family practice  #2 OB-GYN

#exclude the outcome
data_x = data %>% dplyr::select(-completed, -shots)
vars <- colnames(data)[! colnames(data) %in% c('completed', 'practice_type', 'shots')]

## Construct a table
tab1 <- CreateTableOne(vars = vars, strata = "practice_type", data = data_x, test = FALSE)
print(tab1, smd = TRUE)
#group practice_type to 2 levels
data_x = data %>% mutate(practicetype.bin = as.factor(ifelse(practice_type == 0 | practice_type == 1, 0
table(data_x$practicetype.bin,data_x$location) #probabilisitc assumption violation
table(data_x$practicetype.bin,data_x$age)
#Q3 -- exclude subjects that are ineligible
x_new = data_x %>% filter(location %in% c(1,4), age >= 15) %>% mutate(location = factor(location, levels
#descriptive statistics
vars <- colnames(data)[! colnames(data) %in% c('completed', 'practicetype.bin', 'shots', 'practice_type
```

```r
## Construct a table
tab1.new<- CreateTableOne(vars = vars, strata = "practicetype.bin", data = x_new, test = FALSE)
print(tab1.new, smd = TRUE)
#Q4 -estimate propensity score in analytic sample
ps.model<-glm(practicetype.bin~ age + I(age^2) + age + race + insurance_type + location, data = x_new, f
summary(ps.model)

ps <- predict(ps.model, type="response")

prop.func <- function(x, trt)
{
  # fit propensity score model
  propens.model <- glm(trt~age + I(age^2) + age + race + insurance_type + location, data = x, family = b
  pi.x <- predict(propens.model, type = "response")
  pi.x
}

#check.overlap(x = x_new,
#              trt = x_new$practicetype.bin,
#              propensity.func = prop.func)
check.overlap(x = x_new,
              trt = x_new$practicetype.bin,
              type = "both",
              propensity.func = prop.func)
#Q5 - Use matching to improve covariate balance
#optimal
psmatch.opt <- matchit(practicetype.bin ~ age + I(age^2) + race + insurance_type + location,
                   distance="logit", method = "optimal",
                   discard = "control", data = x_new)

# 2. Check balance
summary(psmatch.opt, standardize=TRUE)
#nearest with caliper
psmatch.cal <- matchit(practicetype.bin ~ age + I(age^2) + race + insurance_type + location,
                   distance="logit", method = "nearest", caliper = 0.4,
                   discard = "control", data = x_new)

# 2. Check balance
summary(psmatch.cal, standardize=TRUE)
plot(psmatch.cal)
plot(psmatch.cal, type="hist")
#Q6 - use matches from Q5, estimate ACE of treatment among the treated (ATT)
#match ATT
psmatch.att <- matchit(practicetype.bin ~ age + I(age^2) + race + insurance_type + location,
                   distance="logit", method = "nearest", caliper = 0.4,
                   estimand = "ATT", discard = "control", data = x_new)
psmatch.att.data <- match.data(psmatch.att) #Create matched data for analysis

#estimate ATT
psmatchate.mod <- svyglm(completed ~ practicetype.bin + age + I(age^2) + race + insurance_type + locatio
summary(psmatchate.mod)
#Q7- use NN match without replacement or calipers
#ATT
```

```r
psmatch7.att <- matchit(practicetype.bin ~ age + I(age^2) + race + insurance_type + location,
                    distance="logit", method = "nearest", replace = FALSE,
                    estimand = "ATT", discard = "control", data = x_new)
psmatch.att7.data = match.data(psmatch7.att) #create dataset

#estimate ATT from model
psmatch.att7.mod <- svyglm(completed ~ practicetype.bin + age + I(age^2) + race + insurance_type + loca
                           family = binomial, design = svydesign(~ 1, weights = ~ weights,data=psmatch.a
summary(psmatch.att7.mod)

#ATC
psmatch7.atc <- matchit(practicetype.bin ~ age + I(age^2) + race + insurance_type + location,
                    distance="logit", method = "nearest", replace = FALSE,
                    estimand = "ATC", discard = "control", data = x_new)
psmatch.atc7.data = match.data(psmatch7.atc) #create matched data

#estimate ATC from model
psmatch.atc7.mod <- svyglm(completed ~ practicetype.bin + age + I(age^2) + race + insurance_type + loca
                           family = binomial, design = svydesign(~ 1, weights = ~ weights,data=psmatch.a
summary(psmatch.atc7.mod)

#combine the estimates from ATC and ATT -- calculate manually
ate = (-0.270119 + 0.739537)/2
### AVERAGE TREATMENT EFFECT
set.seed(2020)
psmatch_ate <- matchit(practicetype.bin ~ age + I(age^2) + race + insurance_type + location,
                       data = x_new,  distance = "logit", method = "full", estimand = "ATE")

psmatchate.data <- match.data(psmatch_ate) %>% mutate(completed = as.factor(completed))
psmatchate.mod <- svyglm(completed ~ practicetype.bin + age + I(age^2) + race + insurance_type + locati
                         family = binomial, design = svydesign(~ 1, weights = ~ weights,data=psmatchate
summary(psmatchate.mod)
```