Homework 7

Xinyi Lin

11/9/2019

Question 1

The treatment is practice type. As there are three kinds of practice types, we let treatment be receiving the vaccine at OB-GYN facilities and the control be receiving the vaccine at facilities other than OB-GYN to make practice types as binary. Among these variables, `Age`, `Race`, `InsuranceType`, `Location` are covariates. `Shots` contains some information about outcome, information in `AgeGroup` and `LocationType` overlap with `Age` and `Location`. There are NAs when using `MedAssist` to calculate propensity scores. So `Shots`, `MedAssist`, `AgeGroup` and `LocationType` are not regarded as covariates.

```
##       Age        AgeGroup Race    Shots     Completed      InsuranceType
##  Min.   :11.00   0:701    0:732   1:440   Min.   :0.0000   0:275
##  1st Qu.:15.00   1:712    1:443   2:436   1st Qu.:0.0000   1:723
##  Median :18.00            2: 52   3:537   Median :0.0000   2: 84
##  Mean   :18.55            3:186           Mean   :0.3319   3:331
##  3rd Qu.:22.00                            3rd Qu.:1.0000
##  Max.   :26.00                            Max.   :1.0000
##  MedAssist Location LocationType PracticeType_bin
##  0:1138    1:798    0:963        Min.   :0.0000
##  1: 275    2:165    1:450        1st Qu.:0.0000
##            3: 89                 Median :0.0000
##            4:361                 Mean   :0.3772
##                                  3rd Qu.:1.0000
##                                  Max.   :1.0000
```

The table reflects covariate balance for the original data is shown below

```
##                     Stratified by PracticeType_bin
##                      0             1             SMD
##   n                    880           533
##   Age (mean (SD))    16.80 (3.74)  21.43 (3.33)   1.305
##   Race (%)                                        0.372
##      0                 401 (45.6)    331 (62.1)
##      1                 296 (33.6)    147 (27.6)
##      2                  39 ( 4.4)     13 ( 2.4)
##      3                 144 (16.4)     42 ( 7.9)
##   InsuranceType (%)                               0.728
##      0                 216 (24.5)     59 (11.1)
##      1                 359 (40.8)    364 (68.3)
##      2                  34 ( 3.9)     50 ( 9.4)
##      3                 271 (30.8)     60 (11.3)
##   Location (%)                                    1.382
##      1                 581 (66.0)    217 (40.7)
##      2                   0 ( 0.0)    165 (31.0)
##      3                   0 ( 0.0)     89 (16.7)
##      4                 299 (34.0)     62 (11.6)
```

First, I calculate the propensity of original data

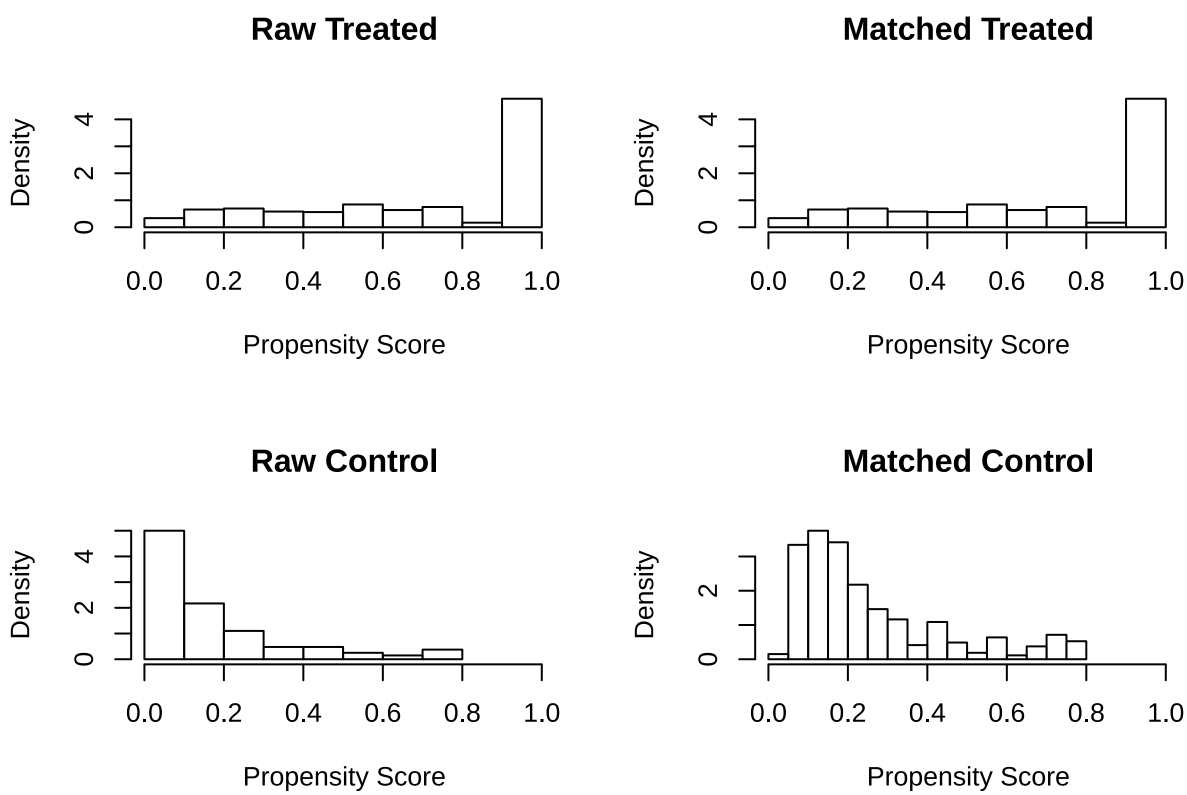
```
##
## Call:
## glm(formula = PracticeType_bin ~ Age + Race + InsuranceType +
##     Location, family = binomial, data = q1_data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.76104  -0.53890  -0.29440   0.00016   2.56680
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -7.63403    0.59059 -12.926  < 2e-16 ***
## Age              0.30458    0.02359  12.909  < 2e-16 ***
## Race1           -0.42717    0.19279  -2.216  0.02671 *
## Race2           -1.19596    0.60272  -1.984  0.04722 *
## Race3           -0.64490    0.24149  -2.670  0.00757 **
## InsuranceType1   1.02725    0.32631   3.148  0.00164 **
## InsuranceType2   1.17159    0.45553   2.572  0.01011 *
## InsuranceType3   0.59322    0.37287   1.591  0.11162
## Location2       19.33868  460.22747   0.042  0.96648
## Location3       19.76711  618.29394   0.032  0.97450
## Location4        0.43685    0.24448   1.787  0.07395 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1872.74  on 1412  degrees of freedom
## Residual deviance:  953.17  on 1402  degrees of freedom
## AIC: 975.17
##
## Number of Fisher Scoring iterations: 17
```

Then, we can do matching. As there are 880 control group and only 533 treatment group, we have more control group than treatment group, so we can use nearest neighbor matching.

```
##
## Call:
## matchit(formula = PracticeType_bin ~ Age + Race + InsuranceType +
##     Location, data = q1_data, method = "nearest", distance = "logit",
##     discard = "control")
##
## Summary of balance for all data:
##                Means Treated Means Control SD Control Mean Diff eQQ Med
## distance              0.7123        0.1742     0.1838    0.5381  0.5825
## Age                  21.4259       16.8034     3.7433    4.6225  5.0000
## Race0                 0.6210        0.4557     0.4983    0.1653  0.0000
## Race1                 0.2758        0.3364     0.4727   -0.0606  0.0000
## Race2                 0.0244        0.0443     0.2059   -0.0199  0.0000
## Race3                 0.0788        0.1636     0.3702   -0.0848  0.0000
## InsuranceType1        0.6829        0.4080     0.4917    0.2750  0.0000
## InsuranceType2        0.0938        0.0386     0.1928    0.0552  0.0000
## InsuranceType3        0.1126        0.3080     0.4619   -0.1954  0.0000
## Location2             0.3096        0.0000     0.0000    0.3096  0.0000
```

```
## Location3             0.1670        0.0000     0.0000    0.1670  0.0000
## Location4             0.1163        0.3398     0.4739   -0.2235  0.0000
##                eQQ Mean eQQ Max
## distance         0.5384  0.8967
## Age              4.6323  6.0000
## Race0            0.1651  1.0000
## Race1            0.0600  1.0000
## Race2            0.0188  1.0000
## Race3            0.0844  1.0000
## InsuranceType1   0.2758  1.0000
## InsuranceType2   0.0563  1.0000
## InsuranceType3   0.1951  1.0000
## Location2        0.3096  1.0000
## Location3        0.1670  1.0000
## Location4        0.2233  1.0000
##
##
## Summary of balance for matched data:
##                Means Treated Means Control SD Control Mean Diff eQQ Med
## distance              0.7123        0.2600     0.1922    0.4523  0.4846
## Age                  21.4259       18.7936     3.3362    2.6323  3.0000
## Race0                 0.6210        0.5422     0.4987    0.0788  0.0000
## Race1                 0.2758        0.3077     0.4620   -0.0319  0.0000
## Race2                 0.0244        0.0169     0.1290    0.0075  0.0000
## Race3                 0.0788        0.1332     0.3401   -0.0544  0.0000
## InsuranceType1        0.6829        0.5441     0.4985    0.1388  0.0000
## InsuranceType2        0.0938        0.0525     0.2233    0.0413  0.0000
## InsuranceType3        0.1126        0.2983     0.4579   -0.1857  0.0000
## Location2             0.3096        0.0000     0.0000    0.3096  0.0000
## Location3             0.1670        0.0000     0.0000    0.1670  0.0000
## Location4             0.1163        0.2589     0.4384   -0.1426  0.0000
##                eQQ Mean eQQ Max
## distance         0.4523  0.8038
## Age              2.6323  4.0000
## Race0            0.0788  1.0000
## Race1            0.0319  1.0000
## Race2            0.0075  1.0000
## Race3            0.0544  1.0000
## InsuranceType1   0.1388  1.0000
## InsuranceType2   0.0413  1.0000
## InsuranceType3   0.1857  1.0000
## Location2        0.3096  1.0000
## Location3        0.1670  1.0000
## Location4        0.1426  1.0000
##
## Percent Balance Improvement:
##                Mean Diff. eQQ Med eQQ Mean eQQ Max
## distance          15.9367 16.8163  15.9866 10.3570
## Age               43.0550 40.0000  43.1754 33.3333
## Race0             52.3386  0.0000  52.2727  0.0000
## Race1             47.3388  0.0000  46.8750  0.0000
## Race2             62.3409  0.0000  60.0000  0.0000
## Race3             35.8665  0.0000  35.5556  0.0000
## InsuranceType1    49.5088  0.0000  49.6599  0.0000
```

```
## InsuranceType2    25.1874  0.0000  26.6667  0.0000
## InsuranceType3     4.9355  0.0000   4.8077  0.0000
## Location2          0.0000  0.0000   0.0000  0.0000
## Location3          0.0000  0.0000   0.0000  0.0000
## Location4         36.1875  0.0000  36.1345  0.0000
##
## Sample sizes:
##           Control Treated
## All           880     533
## Matched       533     533
## Unmatched     209       0
## Discarded     138       0
```



By comparing summary of balance for all data and matched data, we can find that mean difference of covariates among treatment and control groups become smaller. Based on the plot above, we can also find that the distribution of propensity score in treatment and control groups are more similar after matching. Both of these indicates matching make covariates balance better.

Question 2

```
##
## Call:
## lm(formula = Completed ~ PracticeType_bin + Age + Race + InsuranceType +
##     Location, data = match1.data)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6467 -0.3405 -0.2435  0.5600  0.9269
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.484319   0.105903   4.573 5.37e-06 ***
## PracticeType_bin  0.104326   0.036351   2.870  0.00419 **
## Age              -0.011862   0.004416  -2.686  0.00734 **
## Race1            -0.099460   0.033863  -2.937  0.00339 **
## Race2            -0.006500   0.100427  -0.065  0.94841
## Race3            -0.077372   0.048477  -1.596  0.11077
## InsuranceType1    0.076661   0.054175   1.415  0.15734
## InsuranceType2    0.176228   0.069530   2.535  0.01140 *
## InsuranceType3    0.060263   0.065688   0.917  0.35914
## Location2         0.071592   0.048812   1.467  0.14276
## Location3        -0.107679   0.063314  -1.701  0.08929 .
## Location4        -0.030463   0.045703  -0.667  0.50521
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4604 on 1054 degrees of freedom
## Multiple R-squared:  0.05103,    Adjusted R-squared:  0.04113
## F-statistic: 5.153 on 11 and 1054 DF,  p-value: 6.426e-08
```

The point estimate of the average causal effect is 0.11. The confidence interval is (0.039, 0.181).

Interpretation:

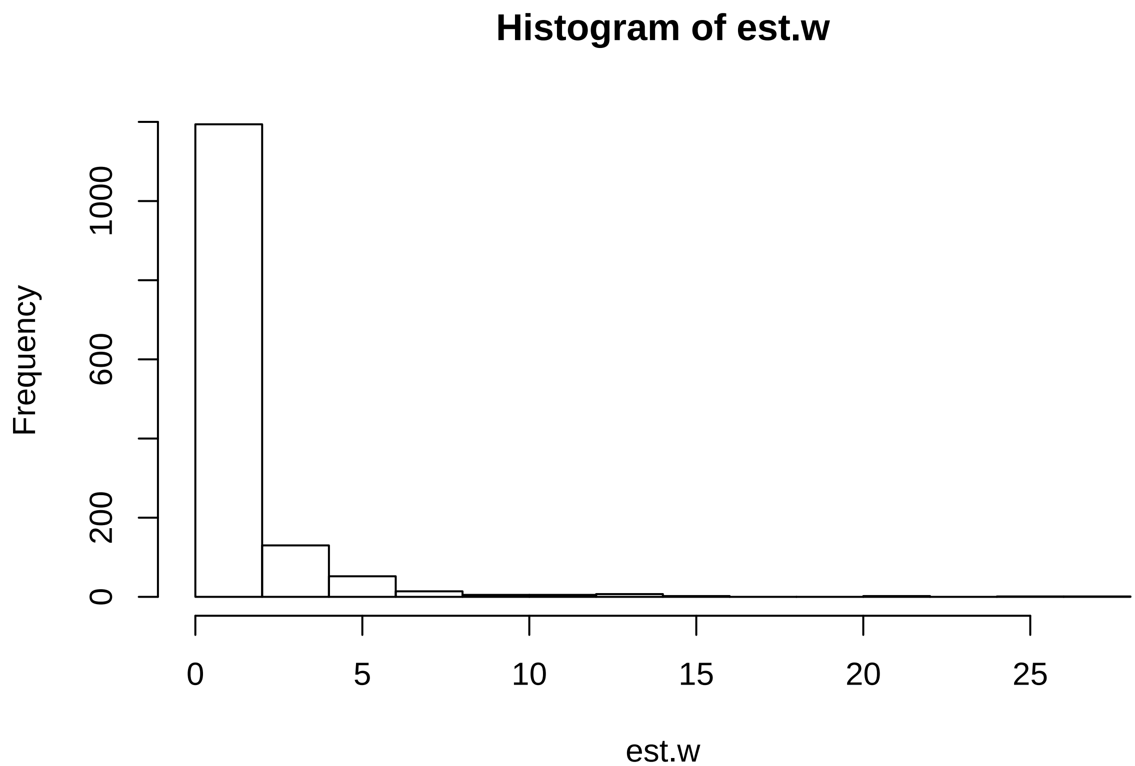
As the point estimate of the average causal effect is 0.11, the estimated true average causal effect is 0.11.

As the confidence interval is (0.039, 0.181), it means with 95% confidence, we can conclude that the true average causal effect falls between 0.039 and 0.181.

As the p-value is smaller than 0.05, we reject null hypothesis and conclude that there is average causal effect of treatment.

Question 3

The histogram of w is shown below.



Question 4

By a using marginal structual model, we can get the estimated average causal effect equals to 0.077.

Question 5

Using bootstrap to simulate distribution of estimated average causal effect and calculate confidence interval and p-value.

The point estimate of the average causal effect is 0.077. The confidence interval is (-44.61, 44.764).

Interpretation:

As the point estimate of the average causal effect is 0.077, the estimated true average causal effect is 0.077.

As the confidence interval is (-44.61, 44.764), it means with 95% confidence, we can conclude that the true average causal effect falls between -44.61 and 44.764.

As the confidence inerval covers 0 and p-value is larger than 0.05, we cannot reject null hypothesis and conclude that there is no average causal effect of treatment.

Question 6

1. When using marginal structual model, we do not exclude any observation, we just give them different weight based on propensity scores. Therefore, some individuals with specific covariates might only be

assigned to treatment group or control group. This means probabilities for them being assigned to another group is 0. This violates positivity.

2. As weights are denominators, when propensity scores are close to 1 or 0, weights can super large and blow up, extreme units can dominate.

Solutions:

1. We can set up some criterion and select observations based on this criterion to make sure subjects in control group have similar covariates as subjects in treatment group.

2. Structural: population probability is 0 and nothing we can do; Random: sample probability is 0 and need to ”borrow” information from other values of Ci to estimate e(Ci) using logistic regression modeling; Check overlap in the sample and always pay attention to the impact of trimming on the characteristics of the analytic sample.

Question 7

1. For subclassification

The point estimate of the marginal average causal effect is 0.065. The confidence interval is (-0.054, 0.184).

Interpretation:

As the point estimate of the marginal average causal effect is 0.065, the estimated true marginal average causal effect is 0.065.

As the confidence interval is (-0.054, 0.184), it means with 95% confidence, we can conclude that the true marginal average causal effect falls between -0.054 and 0.184.

As the confidence inerval covers 0 and p-value is larger than 0.05, we cannot reject null hypothesis and conclude that there is no average causal effect of treatment.

2. For matching:

The point estimate of the average causal effect is 0.11. The confidence interval is (0.039, 0.181).

Interpretation:

As the point estimate of the average causal effect is 0.11, the estimated true average causal effect is 0.11.

As the confidence interval is (0.039, 0.181), it means with 95% confidence, we can conclude that the true average causal effect falls between 0.039 and 0.181.

As the p-value is smaller than 0.05, we reject null hypothesis and conclude that there is average causal effect of treatment.

3. For marginal structual model:

The point estimate of the average causal effect is 0.077. The confidence interval is (-44.61, 44.764).

Interpretation:

As the point estimate of the average causal effect is 0.077, the estimated true average causal effect is 0.077.

As the confidence interval is (-44.61, 44.764), it means with 95% confidence, we can conclude that the true average causal effect falls between -44.61 and 44.764.

As the confidence inerval covers 0 and p-value is larger than 0.05, we cannot reject null hypothesis and conclude that there is no average causal effect of treatment.

We can find that the point estimates of average causal effect getting from subclassification and marginal structual model are similar. Both subclassification method and marginal structual model conclude that there is no average causal effect of treatment.

Subclassification and marginal structual model only give robust match of covariates, but matching find exact match observations in treatment group and control group. The average causal effect might not be vary significant and can only be detected by exact match samples.

```
knitr::opts_chunk$set(echo = FALSE)
library(tidyverse)
library(personalized)
library(tableone)
library(MatchIt)
library(Matching)
gard_data = read.table("./gardasil.dat",header = T)
q1_data = gard_data %>%
  mutate(AgeGroup = as.factor(AgeGroup),
         Race = as.factor(Race),
         Shots = as.factor(Shots),
         InsuranceType = as.factor(InsuranceType),
         MedAssist = as.factor(MedAssist),
         Location = as.factor(Location),
         LocationType = as.factor(LocationType)) %>%
  mutate(PracticeType_bin = ifelse(PracticeType==2,1,0)) %>%
  dplyr::select(-PracticeType)
summary(q1_data)
vars <- c("Age" , "Race", "InsuranceType" ,"Location")
## Construct a table
cov_bal <- CreateTableOne(vars = vars, strata = "PracticeType_bin", data = q1_data, test = FALSE)

## Show table with SMD
print(cov_bal, smd = TRUE)
ps.model<-glm(PracticeType_bin~Age + Race + InsuranceType + Location,data=q1_data, family = binomial)
summary(ps.model)
match1 = matchit(PracticeType_bin~Age + Race + InsuranceType + Location,distance = "logit", method = "ne
summary(match1)
plot(match1, type = "hist")
match1.data <- match.data(match1)
match1.mod <- lm(Completed ~ PracticeType_bin + Age + Race + InsuranceType + Location, data = match1.dat
summary(match1.mod)
t = qt(0.975, 1054)
CIL_match = 0.11-t*0.036
CIU_match = 0.11+t*0.036
q3.model = glm(PracticeType_bin~Age + Race + InsuranceType + Location, family = binomial, data = q1_data
#summary(q3.model)
pprobs = predict(q3.model, type = "response")
est.w = ifelse(q1_data$PracticeType_bin==1, 1/pprobs, 1/(1-pprobs))
hist(est.w)
ht.est = function(y,a,w){
  n = length(y)
  (1/n)*sum((y*a*w)-(y*(1-a)*w))
}
```

```
est_value = ht.est(q1_data$Completed, q1_data$PracticeType_bin, est.w)
boots = 1000
b.holder = rep(NA, boots)
for (i in 1:boots) {
  n = nrow(q1_data)
  S.b = sample(1:n, n, replace = TRUE)
  boot.data = q1_data[S.b,]
  boot.model = glm(PracticeType_bin~Age + Race + InsuranceType + Location, family = binomial, data = bo
  pprobs = predict(boot.model)
  est.w = ifelse(boot.data$PracticeType_bin==1, 1/pprobs, 1/(1-pprobs))
  b.holder[i] = ht.est(boot.data$Completed, boot.data$PracticeType_bin, est.w)
}
var = var(b.holder)
p.val = sum(b.holder>=0)/boots
t = quantile(b.holder, 0.975)
CIL_marginal = est_value - t*sqrt(var)
CIU_marginal = est_value + t*sqrt(var)
```