ECON312 Problem Set 1: question 3

Futing Chen, Hongfan Chen, Will Parker

04/15/2020

Contents


```
library(tidyverse)
library(knitr)
library(haven)
library(MatchIt)
```

Question 3

```
data <- read_dta("https://www.dropbox.com/s/dl/aw4yi13mz9z03yf/lalonde2.dta")
```

```
data %>%
  group_by(sample) %>%
  count(treated)
```

```
## # A tibble: 4 x 3
## # Groups:   sample [3]
##      sample treated     n
##   <dbl+lbl>   <dbl> <int>
## 1  1 [NSW]         0   425
## 2  1 [NSW]         1   297
## 3  2 [CPS]        NA 15992
## 4  3 [PSID]       NA  2490
```

a) Investigate whether the data is consistent with randomization of the treatment.

We assume `sample ==1` indicates the NSW sample, as that is the only sample with non-missing treatment variables

```
rct_data <- data %>%
  filter(sample == 1)
```

```
rct_data %>%
  count(treated)
```

```
## # A tibble: 2 x 2
##   treated     n
##     <dbl> <int>
## 1       0   425
## 2       1   297
```

```
rct_data %>%
  group_by(treated) %>%
  summarise( total = n(),
    mean_age = mean(age),
            mean_black = mean(black),
            mean_married = mean(married),
            mean_nodegree = mean(nodegree),
            mean_re74 = mean(re74))
```

```
## # A tibble: 2 x 7
##   treated total mean_age mean_black mean_married mean_nodegree mean_re74
##     <dbl> <int>    <dbl>      <dbl>        <dbl>         <dbl>     <dbl>
## 1       0   425     24.4      0.8          0.158         0.814     3672.
## 2       1   297     24.6      0.801        0.168         0.731     3571.
```

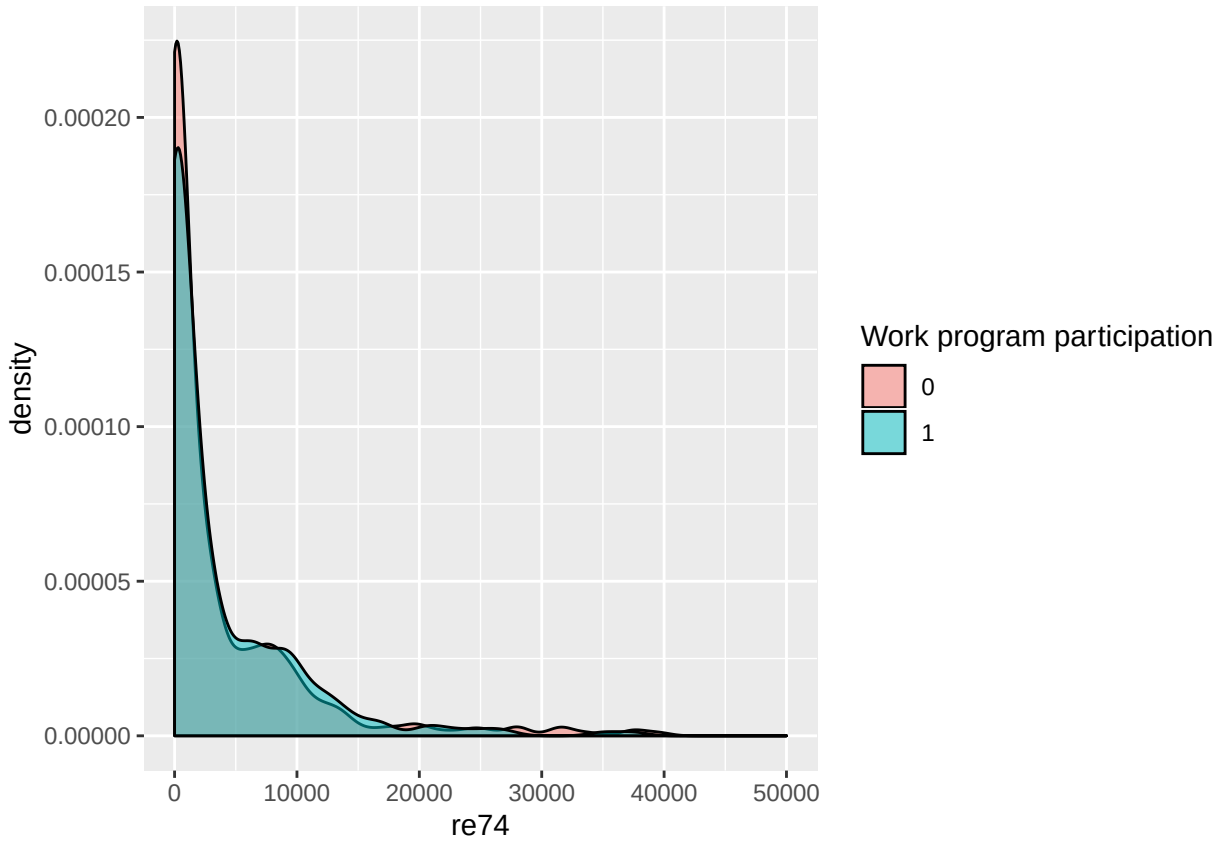
It looks like more patients were not treated than received treatment, implying that treatments were not assigned with $P(D = 1) = 0.5$. One explanation is that more patients were assigned to the control group but assignment was still probabilistic, just with $P(D = 1) = 0.4$ or so.

However, a more likely explanation is that the work program was made available/recommended to half of the group but compliance was with the program was somewhere around 80%. This is suggested by some minor differences in some of the baseline covariates that would be unlikely under random assigment. If this case it would be good to know initial group assignment which could be treated as an instrutmental variable.

However the 1974 income distribution looks fairly similar in between the two groups, which is good if we want to assume $D \perp\!\!\!\perp (Y_1, Y_0)$, $\hat{\beta}_{OLS}$ since clearly past and future income will be related.

1974 real-income distribution in the NSW sample

```
rct_data %>%
  ggplot(aes(x = re74, fill = factor(treated))) +
  geom_density(alpha = 0.5)+
  lims(x = c(0, 50000)) +
  labs(fill = "Work program participation")
```



b) Estimate the effect using the experimental sample.

Assuming $D \perp\!\!\!\perp (Y_1, Y_0)$, $\hat{\beta}_{OLS}$ will be a unbiased estimate of the ATE

```
model <- lm(re78 ~ treated, data = rct_data)

summary(model)
```

```
##
## Call:
## lm(formula = re78 ~ treated, data = rct_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5976  -5090  -1519   3361  54332
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5090.0      302.8  16.811   <2e-16 ***
## treated        886.3      472.1   1.877   0.0609 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6242 on 720 degrees of freedom
## Multiple R-squared:  0.004872,   Adjusted R-squared:  0.003489
## F-statistic: 3.525 on 1 and 720 DF,  p-value: 0.06086
```

```
ATE <- model$coefficients[[2]]
```

The $\hat{ATE}$ of the work experience program was +$886

c) Estimate the effect using OLS on observed data

Now use the sample consisting in the treated from the NSW sample and the comparison individuals from the CPS sample.

```
observed_controls_NSW_tx <- data %>%
  filter(treated == 1 | is.na(treated)) %>%
  mutate(treat_2 = ifelse(is.na(treated), 0, treated))

observed_model <- lm(re78 ~ treat_2, observed_controls_NSW_tx)

summary(observed_model)
```

```
##
## Call:
## lm(formula = re78 ~ treat_2, data = observed_controls_NSW_tx)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -15750  -9191   1264   9814 105423
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 15750.30      79.65  197.74   <2e-16 ***
## treat_2     -9773.95     633.36  -15.43   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10830 on 18777 degrees of freedom
## Multiple R-squared:  0.01252,    Adjusted R-squared:  0.01247
## F-statistic: 238.1 on 1 and 18777 DF,  p-value: < 2.2e-16
```

```
beta_OLS_observed <- observed_model$coefficients
```

Just comparing the observational cohort of non-treated people to the treated cohort leads a very negative and biased estimate of the ATE. There is obvious selection bias on covariates when comparing the two groups. People randomized into the NSW program had substantially lower incomes at baseline. It is quite clear that the treatment is correlated with potential outcomes in this sample

d) Investigate covariate balancing and support between the treated and the CPS sample.

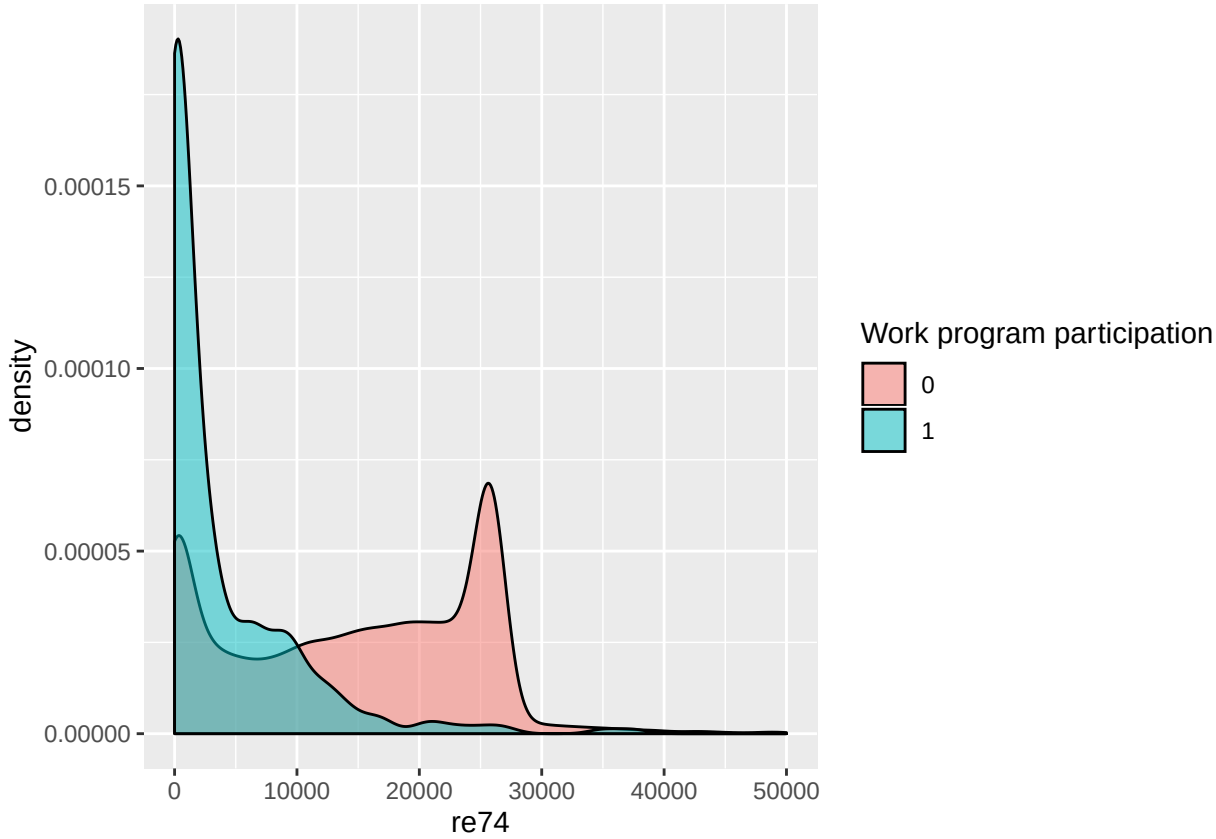
```
observed_controls_NSW_tx %>%
  group_by(treat_2) %>%
    summarise( total = n(),
    mean_age = mean(age),
            mean_black = mean(black),
            mean_married = mean(married),
```

```
            mean_nodegree = mean(nodegree),
            mean_re74 = mean(re74))
```

```
## # A tibble: 2 x 7
##   treat_2 total mean_age mean_black mean_married mean_nodegree mean_re74
##     <dbl> <int>    <dbl>      <dbl>        <dbl>          <dbl>     <dbl>
## 1       0 18482     33.4     0.0974        0.733          0.297    14746.
## 2       1   297     24.6     0.801         0.168          0.731     3571.
```

distribution of real-income in 1974 in the combined CPS and NSW treated sample

```
observed_controls_NSW_tx %>%
  ggplot(aes(x = re74, fill = factor(treat_2))) +
  geom_density(alpha = 0.5)+
  lims(x = c(0, 50000)) +
  labs(fill = "Work program participation")
```



e) Estimate the effect using 1-1 nearest neighbor propensity score matching.

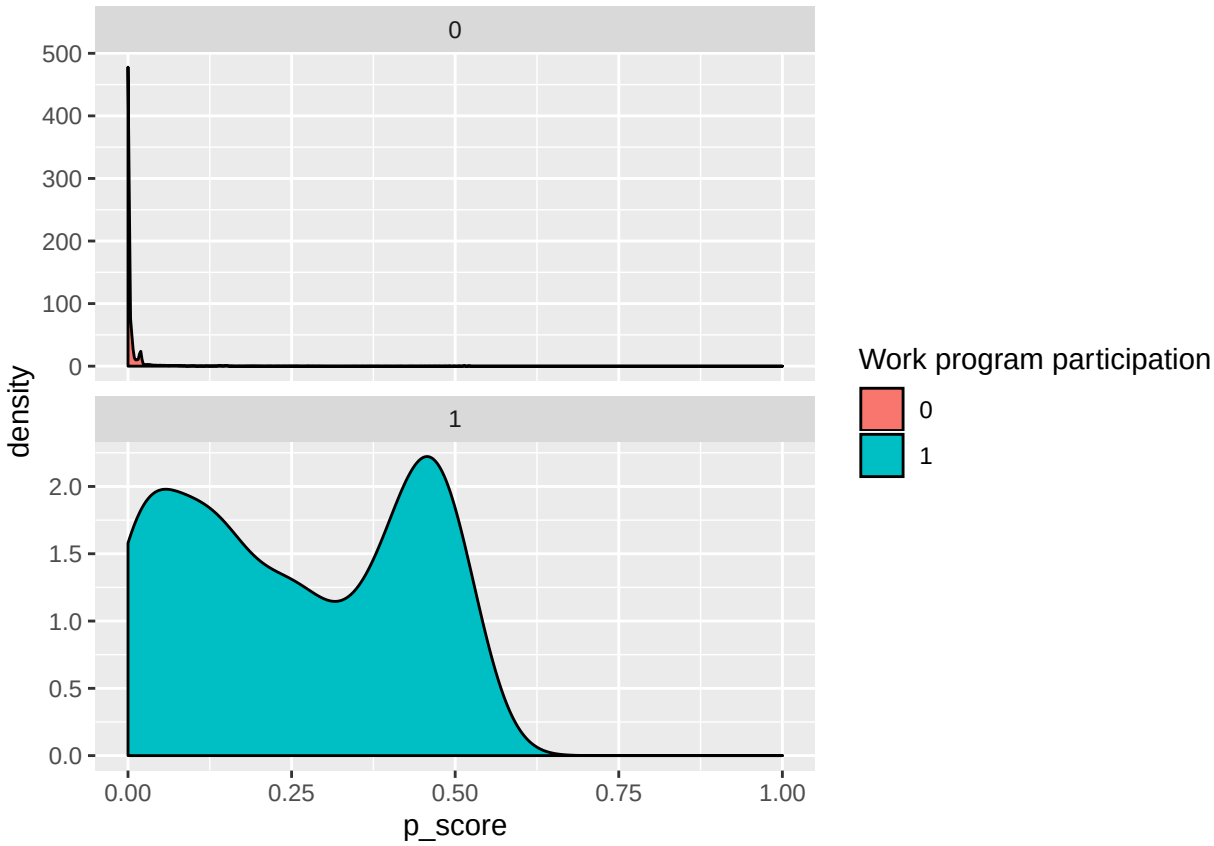
fit the propensity score

```
treat_model_formula <- formula(treat_2 ~ age +  married+ nodegree + re74 + hisp + black)

treatment_model <- glm(treat_model_formula,
```

```
                      data = observed_controls_NSW_tx,
                      family = binomial())

summary(treatment_model)
```

```
##
## Call:
## glm(formula = treat_model_formula, family = binomial(), data = observed_controls_NSW_tx)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2140  -0.0850  -0.0323  -0.0159   3.8926
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.343e+00  2.608e-01 -16.650  < 2e-16 ***
## age         -3.024e-02  8.034e-03  -3.764 0.000168 ***
## married     -1.452e+00  1.855e-01  -7.827 5.00e-15 ***
## nodegree     9.310e-01  1.468e-01   6.341 2.28e-10 ***
## re74        -1.029e-04  1.282e-05  -8.026 1.01e-15 ***
## hisp         2.169e+00  2.693e-01   8.054 8.00e-16 ***
## black        3.981e+00  1.996e-01  19.942  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3052.5  on 18778  degrees of freedom
## Residual deviance: 1665.6  on 18772  degrees of freedom
## AIC: 1679.6
##
## Number of Fisher Scoring iterations: 9
```

```
p_score <- predict(treatment_model, type = "response")

observed_controls_NSW_tx <- observed_controls_NSW_tx %>%
  cbind(p_score)

observed_controls_NSW_tx %>%
  ggplot(aes(x= p_score, fill = factor(treat_2))) +
  geom_density() +
  labs(fill = "Work program participation") +
  facet_wrap(~factor(treat_2), nrow = 2, scales = "free_y")+
  lims(x= c(0,1))
```



Most of the propensity scores for the CPS sample are around 1.

1-1 nearest neighbor matching

```
treat_model_formula
```

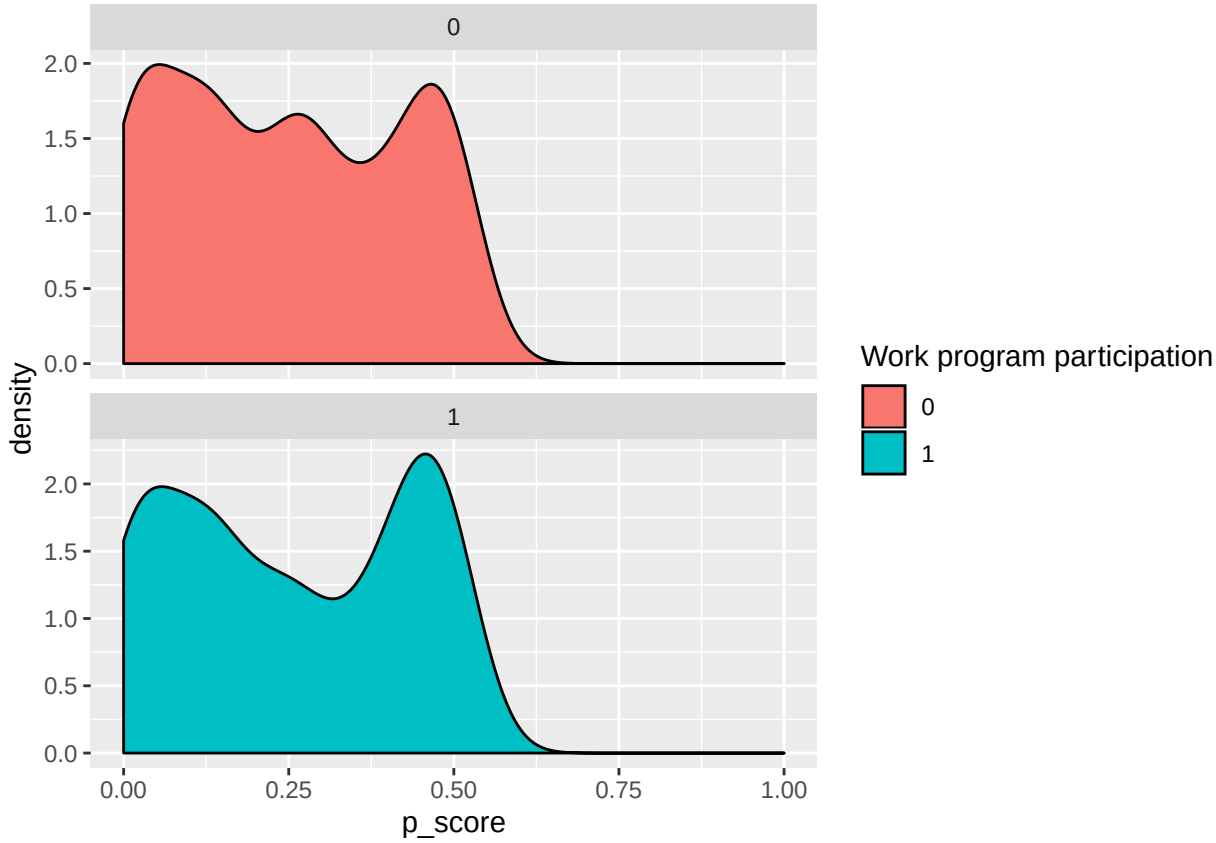
```
## treat_2 ~ age + married + nodegree + re74 + hisp + black
```

```
match <- matchit(treat_model_formula,
                        data = observed_controls_NSW_tx %>% select(treat_2, age, married,
                                                                   nodegree, re74, hisp, black),
                    method = "nearest")

p_match_sample <- observed_controls_NSW_tx %>%
  cbind(weights = match$weights) %>%
  filter(weights ==1 | treat_2 ==1)

p_match_sample %>%
  count(treat_2)
```

```
## # A tibble: 2 x 2
##   treat_2     n
##     <dbl> <int>
## 1       0   297
## 2       1   297
```

Distribution of propensity score after 1-1 nearest neighbor matching

```
p_match_sample %>%
  ggplot(aes(x= p_score, fill = factor(treat_2))) +
  geom_density() +
  labs(fill = "Work program participation") +
  facet_wrap(~factor(treat_2), nrow = 2, scales = "free_y")+
  lims(x= c(0,1))
```




```
summary(match)
```

```
##
## Call:
## matchit(formula = treat_model_formula, data = observed_controls_NSW_tx %>%
##     select(treat_2, age, married, nodegree, re74, hisp, black),
##     method = "nearest")
##
## Summary of balance for all data:
##          Means Treated Means Control SD Control   Mean Diff    eQQ Med
## distance        0.2549        0.0120     0.0494      0.2429     0.2405
## age            24.6263       33.4442    10.9795     -8.8180     8.0000
## married         0.1684        0.7326     0.4426     -0.5642     1.0000
## nodegree        0.7306        0.2971     0.4570      0.4335     0.0000
## re74         3570.9990    14745.9287 10337.5213 -11174.9298 13446.1238
## hisp            0.0943        0.0667     0.2495      0.0276     0.0000
## black           0.8013        0.0974     0.2965      0.7040     1.0000
##            eQQ Mean    eQQ Max
## distance     0.2422     0.4788
## age          8.8687    18.0000
```

```
## married      0.5623     1.0000
## nodegree     0.4343     1.0000
## re74     11455.4296 99717.0117
## hisp         0.0269     1.0000
## black        0.7037     1.0000
##
##
## Summary of balance for matched data:
##          Means Treated Means Control SD Control Mean Diff  eQQ Med eQQ Mean
## distance        0.2549        0.2462     0.1715    0.0087   0.0002   0.0088
## age            24.6263       25.2862     8.7462   -0.6599   1.0000   1.8316
## married         0.1684        0.1616     0.3687    0.0067   0.0000   0.0067
## nodegree        0.7306        0.6869     0.4646    0.0438   0.0000   0.0438
## re74         3570.9990     3918.8766  5366.6972 -347.8777 570.1467 650.9834
## hisp            0.0943        0.1044     0.3063   -0.0101   0.0000   0.0101
## black           0.8013        0.8081     0.3945   -0.0067   0.0000   0.0067
##            eQQ Max
## distance     0.064
## age          6.000
## married      1.000
## nodegree     1.000
## re74     11569.338
## hisp         1.000
## black        1.000
##
## Percent Balance Improvement:
##          Mean Diff.  eQQ Med eQQ Mean eQQ Max
## distance    96.4379  99.9126  96.3816 86.6303
## age         92.5160  87.5000  79.3470 66.6667
## married     98.8065 100.0000  98.8024  0.0000
## nodegree    89.9038   0.0000  89.9225  0.0000
## re74        96.8870  95.7598  94.3173 88.3978
## hisp        63.3524   0.0000  62.5000  0.0000
## black       99.0434 100.0000  99.0431  0.0000
##
## Sample sizes:
##           Control Treated
## All         18482     297
## Matched       297     297
## Unmatched   18185       0
## Discarded       0       0
```

The covariate balance has improved considerably after propensity score matching

Outcome model after propensity score matching

```
p_match_model <- lm(re78 ~ treat_2, data = p_match_sample)

summary(p_match_model)
```

```
##
## Call:
## lm(formula = re78 ~ treat_2, data = p_match_sample)
##
## Residuals:
```

```
##    Min     1Q Median     3Q    Max
##  -7139  -5976  -1720   3661  54332
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7138.8      413.5  17.266   <2e-16 ***
## treat_2      -1162.4      584.7  -1.988   0.0473 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7125 on 592 degrees of freedom
## Multiple R-squared:  0.006632,   Adjusted R-squared:  0.004954
## F-statistic: 3.952 on 1 and 592 DF,  p-value: 0.04727
```

After propensity score matching the treatement effect estimate, while less "wrong", still has a negative sign. This is concerning for either 1) misspecification of the propensity score model based on the observed covariates or 2) selection on unobservables.

f) Estimate the effect using the propensity score and local linear regression.

TBD

```
guassian_kernel <- function(p_1, p_2, h){
  diff <- (p_1 - p_2)/h
  exp(-diff^2/2)
}
```