Exam 2022 fall solution proposal

```
rm(list = ls())
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(rddtools)
```

```
## Loading required package: AER

## Loading required package: car

## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode

## Loading required package: lmtest

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
## Loading required package: sandwich

## Loading required package: survival

## Loading required package: np

## Nonparametric Kernel Methods for Mixed Datatypes (version 0.60-11)
## [vignette("np_faq",package="np") provides answers to frequently asked questions]
## [vignette("np",package="np") an overview]
## [vignette("entropy_np",package="np") an overview of entropy-based methods]

##
## Please consider citing R and rddtools,
## citation()
## citation("rddtools")
```

```
library(magrittr)
library(haven)
library(rddensity)
library(rdrobust)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --

## v tibble  3.1.7     v purrr   0.3.4
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x tidyr::extract()   masks magrittr::extract()
## x dplyr::filter()    masks stats::filter()
## x dplyr::lag()       masks stats::lag()
## x car::recode()      masks dplyr::recode()
## x purrr::set_names() masks magrittr::set_names()
## x purrr::some()      masks car::some()
```

```
library(dplyr)
library(broom)
library(MatchIt)
library(huxtable)
```

```
##
## Attaching package: 'huxtable'
##
## The following object is masked from 'package:ggplot2':
##
##     theme_grey
##
## The following object is masked from 'package:dplyr':
##
##     add_rownames
```

```
library(cobalt)
```

```
##  cobalt (Version 4.3.2, Build Date: 2022-01-19)
##
## Attaching package: 'cobalt'
##
## The following object is masked from 'package:MatchIt':
##
##     lalonde
```

Exercise 1 (40%)

```
alcohol <- read_csv('C:/Users/yuazh/OneDrive - Universitetet i Oslo/Desktop/Teaching/Exam 2022 fall/alc
```

```
## New names:
## Rows: 50 Columns: 4
## -- Column specification
## -------------------------------------------------------- Delimiter: "," dbl
## (4): ...1, age, death, threshold
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

1.1 Describe the data (how many variables and observations do you have?) What are the names of your variables? What are the means of the most relevant variable?

```
names(alcohol)
```

```
## [1] "...1"      "age"       "death"     "threshold"
```

```
dim(alcohol)
```

```
## [1] 50  4
```

```
summary(alcohol$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   19.07   20.08   21.00   21.00   21.92   22.93
```

```
summary(alcohol$death)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   88.43   92.79   95.69   95.67   98.03  105.27       2
```

```
summary(alcohol$threshold)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0     0.0     0.5     0.5     1.0     1.0
```

There are 3 variables, named age, death and threshold (note: there is one additional variable has no name). Students that reports either 3 or 4 variables are given points.

age, death and threshold are the most relevant variables, their means are 21, 95.67 and 0.5 accordingly.

1.2 How many are below the minimum drinking age? Calculate the average morality for those below the minimum drinking age (hint: you may want to exclude missing values). Under the regression discontinuity design, what is the running variable and what is the cutoff, and what is the treatment indicator? Create a normalized running variable.

```
table(alcohol$threshold)
```

```
##
##  0  1
## 25 25
```

25 are below the minimum drinking age.

```
alcohol %>%
  group_by(threshold) %>%
  summarise(mean(death, na.rm = TRUE))
```

threshold	mean(death, na.rm = TRUE)
0	92.8
1	98.5

The average mortality for those below the minimum drinking age is 92.80270.

The running variable is the respondent's age, the cutoff is 21 years' old, and the treatment is eligibility for drinking alcohol (measured by the threshold variable).

```
alcohol <-
  alcohol %>%
  mutate(nor_age = I(age-21))
```

1.3 Regress morality on the treatment indicator. What is the interpretation of the treatment coefficient (e.g., the sign, significance, magnitude)? What are the identifying assumptions for a causal regression? Based on the identifying assumptions, can the estimated treatment coefficient be interpreted causal? Why or why not?

```
m1 <- lm(death ~ threshold, alcohol)
tidy(m1)
```

term	estimate	std.error	statistic	p.value
(Intercept)	92.8	0.516	180	3.88e-67
threshold	5.74	0.73	7.86	4.77e-10

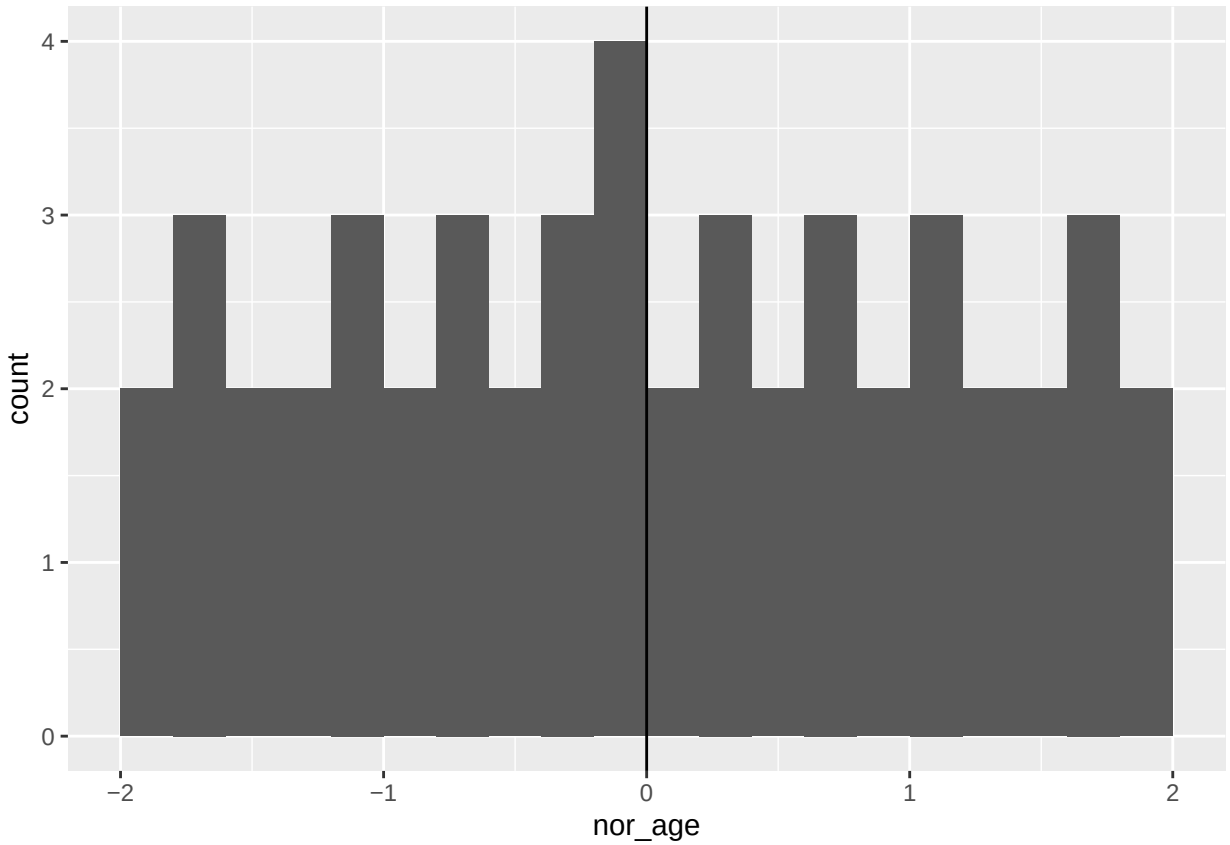
The threshold variable positive and is statistically significant at 0.1% level. It implies that people who are 21 and more have 5.74 more deaths per 100,000 person year than people who are less than 21.

The identifying assumption that regression can be used to identify causal effect is called selection on observables assumption (also known as unconfoundedness, ignorability, exogenous selection, or selection on observables). It implies that are all confounders are observed and can therefore be used as control variables.

The estimated treatment coefficient cannot be interpreted causal because do not have information on the cofounders and cannot control for them. The treatment may not be randomly assigned.

1.4 What is the identifying assumption in the RDD design? Plot a histogram of the running variable, assess graphically (use binwidth = 0.2) and formally (use a formal). Comment on whether there is any threat to the identification.

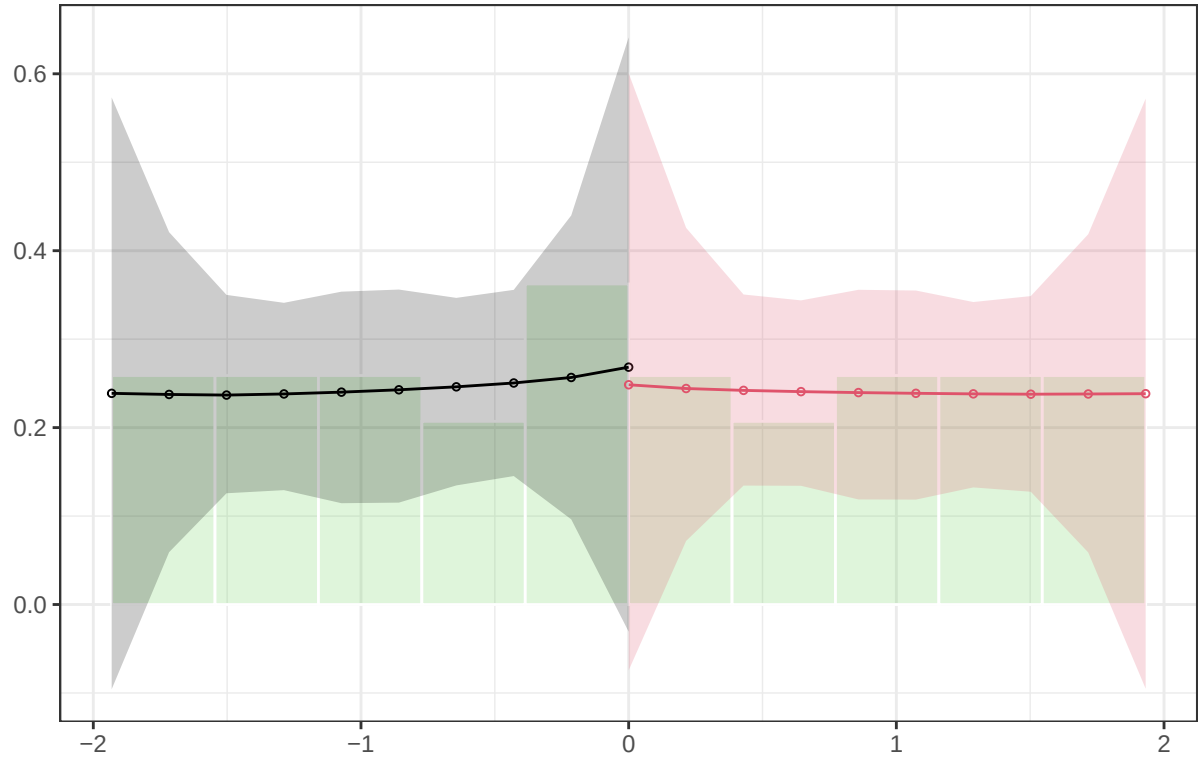
```
ggplot(alcohol, aes(x = nor_age, fill = death)) +
  geom_histogram(binwidth = 0.2, boundary = 0) +
  geom_vline(xintercept = 0)
```




```
#fill allow us to use different colors on each side of the cutoff
#bidwidth allows us to adjust the width of each bar, boundary is a bin position specifier, boundary = 0
#geom_vline can add a vertical line on the plot, and we want the vertical line to be positioned at 0
```

```
test_density <- rddensity(alcohol$nor_age, c = 0)
#rddensity() implements manipulation testing procedures using the local polynomial density, c speicifes

rd_density_plot <- rdplotdensity(rdd = test_density, alcohol$nor_age, type = "both")
```




```
#rdplotdensity() constructs density plots. inside the function, you need to define rdd, it is the objec
#nor_age is our running variable.
#type defines point estimates are plotted, both include line and points.
```

The identifying assumption of RDD is that observations close to either side of discontinuous shift in treatment assignment are comparable.

The threat to identification is that the running variable may be manipulated. We can look at the distribution of the running variable. And check for discontinuity around cutoff and see if there is any bunching of observations around the cutoff. If there is, then it is the evidence that there is sorting. One way to do this is to plot a histogram of the running variable, and check whether there are signs of bunching around the cutoff.

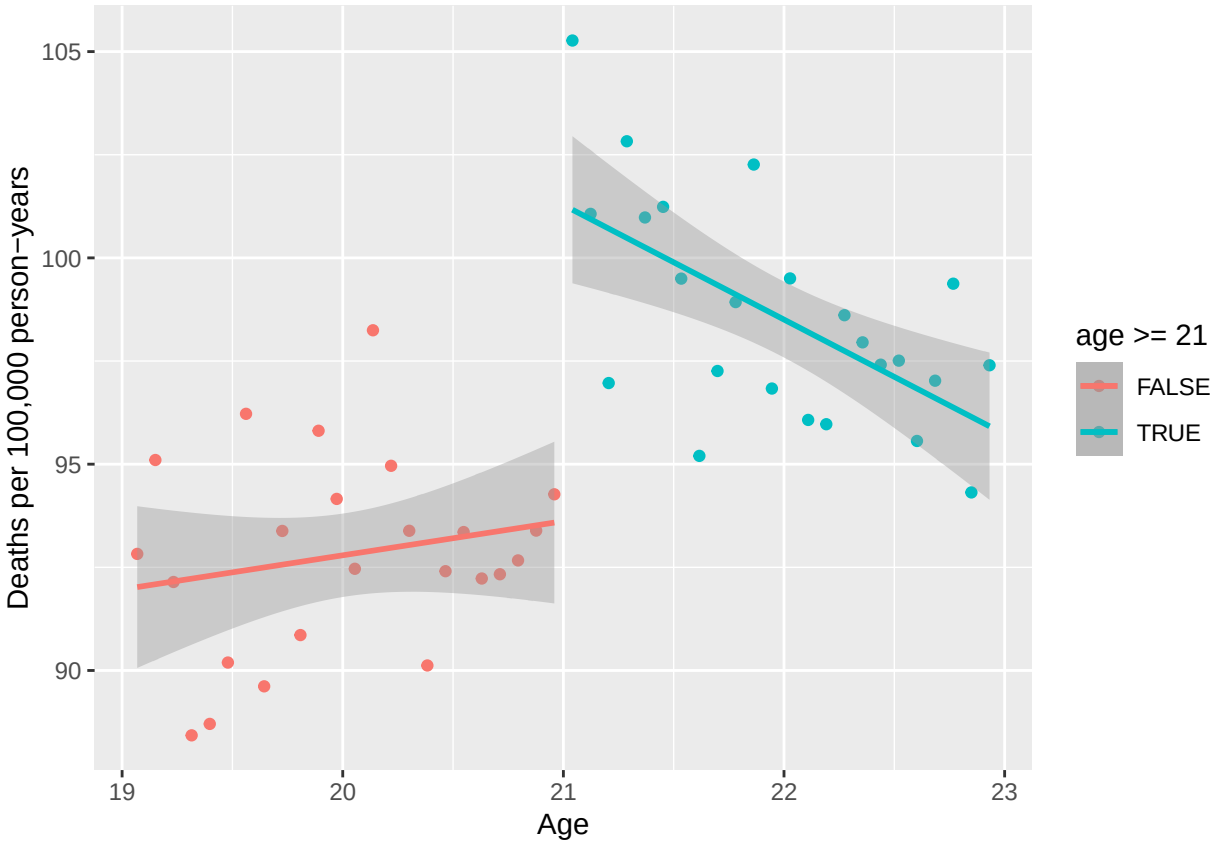
McCrary density test is a formal test for checking whether units are sorting on the running variable. The density plot shows that there is a tiny gap at the cutoff, but the gap is within the confidence interval within the both side. There is not a significant difference among the bins in the neighborhood of the cutoff. This means that the running variable is not manipulated.

1.5 Plot a scatterplot of the morality rate as a function of the running variable. Please use different colours on each side of the cut-off, include a linear fit on each side of the cut-off in your plot, and label your x-axis and y-axis accordingly. Comment on whether you can see a jump in the morality rate at the cut-off point.

```
ggplot(alcohol, aes(x = age, y = death, color = age >= 21)) +
  geom_point() +
  geom_smooth(data = filter(alcohol, age< 21), formula = y ~ x, method = "lm") +
  geom_smooth(data = filter(alcohol, age >= 21), formula = y ~ x, method = "lm") +
  ylab("Deaths per 100,000 person-years") + xlab("Age")
```

```
## Warning: Removed 1 rows containing non-finite values (stat_smooth).
## Removed 1 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```



Yes, we can see there is a discontinuous jump in the outcome variable as the running variable crosses the cutoff.

1.6 Now, run a regression of mortality on the treatment indicator and the normalized age variable, and interaction between the normalized age variable and the treatment indicator. How do you interpret the estimated treatment coefficient? Can it be interpreted as causal? Why or why not? (Relate your answer to what you found in 1.4).

```
m2 <- lm(death ~ threshold+nor_age+nor_age*threshold, data = alcohol)
summary(m2)
```

```
##
## Call:
## lm(formula = death ~ threshold + nor_age + nor_age * threshold,
##     data = alcohol)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.368 -1.787  0.117  1.108  5.341
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        93.6184     0.9325 100.399  < 2e-16 ***
## threshold           7.6627     1.3187   5.811  6.4e-07 ***
## nor_age             0.8270     0.8189   1.010  0.31809
## threshold:nor_age  -3.6034     1.1581  -3.111  0.00327 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.283 on 44 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.6677, Adjusted R-squared:  0.645
## F-statistic: 29.47 on 3 and 44 DF,  p-value: 1.325e-10
```

On average, the mortality rate per 100.000 for individuals reaching the minimum drinking age is 7.66 points higher. It is statistcally significant at 0.1% level.

The result can be interpreted as causal. This is because we do not see evidence that the running variable is manipulated in 1.4.

1.7 Include a quadratic term in the normalized running variable interacted with the treatment, and redo 1.6. Has the treatment effect changed compare to the effect in 1.6? Comment on the size and coefficient of the estimated treatment effect. What does this tell you?

```
m3 <- lm(death ~ threshold+nor_age+nor_age*threshold+I(nor_age^2)*threshold, data = alcohol)
summary(m3)
```

```
##
## Call:
## lm(formula = death ~ threshold + nor_age + nor_age * threshold +
```

```
##     I(nor_age^2) * threshold, data = alcohol)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3343 -1.3946  0.1849  1.2848  5.0817
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)             93.0729     1.4038  66.301  < 2e-16 ***
## threshold                9.5478     1.9853   4.809 1.97e-05 ***
## nor_age                 -0.8306     3.2901  -0.252    0.802
## I(nor_age^2)            -0.8403     1.6153  -0.520    0.606
## threshold:nor_age       -6.0170     4.6529  -1.293    0.203
## threshold:I(nor_age^2)   2.9042     2.2843   1.271    0.211
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.285 on 42 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.6821, Adjusted R-squared:  0.6442
## F-statistic: 18.02 on 5 and 42 DF,  p-value: 1.624e-09
```

Yes, now on average, the mortality rate per 100.000 for individuals reaching the minimum drinking age is 9.55 points higher, which is statistically significant at 0.1% level. The coefficient is bigger than that in 1.6.

The results change because we now use a different functional form (in this question, we use a quadratic from). We use a quadratic form to allow for the the possibility that the data-generating process was nonlinear.

1.8 Use non-parametric method to estimate the treatment effect. Is your result same or different from what you got in 1.3? Why do you think are the reasons that the results are same or different?

```
rdrobust(y = alcohol$death, x = alcohol$age, c = 21) %>% summary()
```

```
## Sharp RD estimates using local polynomial regression.
##
## Number of Obs.                   48
## BW type                       mserd
## Kernel                   Triangular
## VCE method                       NN
##
## Number of Obs.                   24           24
## Eff. Number of Obs.               6            6
## Order est. (p)                    1            1
## Order bias  (q)                   2            2
## BW est. (h)                   0.492        0.492
## BW bias (b)                   0.778        0.778
## rho (h/b)                     0.633        0.633
## Unique Obs.                      24           24
##
## ===============================================================================
##         Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
```

```
## ================================================================================
##   Conventional     9.598     3.592     2.672     0.008     [2.559 , 16.637]
##         Robust         -         -     2.206     0.027     [1.083 , 18.308]
## ================================================================================
```

```
#c specifies the RD cutoff in x; default is c = 0
```

On average, the mortality rate per 100.000 for individuals reaching the minimum drinking age is now 9.595 points higher. It is statistically significant at 1% level.

Results are different because the result from 1.8 is the causal effect (as we found the continuity assumption holds in 1.4) while that from 1.3 is not.

Exercise 2: The effect of heart treatment (60%)

2.1 Write a short report of the attached paper "The effectiveness of right heart catheterization in the initial care of critically III patients" by Connors et al (1996).

The report should include a summary of the paper, and a critical discussion of the empirical approach. The summary should identify the research questions that the paper tries to answer, how the paper answers the questions, and the results (about 1 page). The discussion of the empirical approach should give a description and critical assessment of the applied methods and its identifying assumptions.

Focus on the following questions: What are the coefficient(s) of interest(s)? What is (are) the key identifying assumption(s)? Are the identifying assumptions likely to hold? Are there data limitations, and do you have any suggestions for alternative analyses and sensitivity checks?

This question is an open question that assess the students' independent and critical thinking. Therefore the points are given for those who demonstrate their own thinking, in addition to summarize what has been done by the authors. This question is assessed based on:

1. Be able to address the research question, the method (i.e. ps matching), and the results.
2. Explain the variables of interest (the RHC itself should be key the variables of interest, in addition to other cofounders.)
3. Explain the key identifying assumption for the PS matching: i.e. the unconfoundedness assumption, and explain whether it is likely to hold (e.g. why the unobserable variable is not a threat to this identification). More points given to those who provided reasoning/rationale for why the candidates believe such assumption will hold or not.
4. Discussion on the limitation with data (e.g. a sample of severely ill patients, small sample size etc.).
5. Suggest for alternative causal method and additional sensitivity tests, higher points given for those why provide convincing rationale of how the suggested methods/tests are better or how they can overcome the problem in PS matching.

2.2 How many patients in the (right heart catheterization) RHC group died? How many patients in the non RHC group died? Report the average mortality among patients who received the treatment vs. those who did not receive the treatment (the treatment variable is called swang1 in the dataset). Which group has higher mortality? What do you think is the reason for this result? (Hint: you need numeric variables to do summary statistics).

```
rhc<- read_csv('C:/Users/yuazh/OneDrive - Universitetet i Oslo/Desktop/Teaching/Exam 2022 fall/heart.csv
```

```
## New names:
## Rows: 5735 Columns: 64
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (21): cat1, cat2, ca, death, sex, dth30, swang1, dnr1, ninsclas, resp, c... dbl
## (43): ...1, X, sadmdte, dschdte, dthdte, lstctdte, cardiohx, chfhx, deme...
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

```
rhc %>%
  group_by(swang1) %>%
  count(death)
```

```
## # A tibble: 4 x 3
## # Groups:   swang1 [2]
##   swang1 death     n
##   <chr>  <chr> <int>
## 1 No RHC No     1315
## 2 No RHC Yes    2236
## 3 RHC    No      698
## 4 RHC    Yes    1486
```

1486 patients in the RHC group died. 2236 patients in the non RHC group died.

```
#our treatment and outcome variables are string.
#convert them into numeric variables
rhc %>%
  mutate(treatment = if_else(swang1 == "RHC", 1, 0)) -> rhc

rhc %>%
  mutate(died = if_else(death == "Yes", 1, 0)) -> rhc
```

```
rhc %>%
  filter(treatment==1) %>%
  summarize(mean_treat = mean(died)) -> mean_morality_treat
mean_morality_treat
```

```
rhc %>%
  filter(treatment==0) %>%
  summarize(mean_control = mean(died)) -> mean_morality_control
mean_morality_control
```

mean_treat
0.68

mean_control
0.63

The average mortality among the RHC patients equals to 0.6804029, and the average mortality among the non-RHC patients equals to 0.6296818.

The RHC group thus has higher mortality. This is probably due to more severe patients in RHC groups.

2.3 Select some variables you believe affect both treatment and mortality and estimate the probability of receiving treatment given these variables (i.e., logistic regression). Explain why do you choose these variables. What do the coefficients from logistic regression represent? What do your results tell you?

```
rhc %>%
  mutate(men = if_else(sex == "Male", 1, 0)) -> rhc

rhc %>%
  mutate( cat_race = as.character(race)) -> rhc
```

```
logit <- glm(treatment~age+men+edu+cat_race, family="binomial", rhc)
logit %>% huxreg()
```

```
## Warning in huxreg(.): Unrecognized statistics: r.squared
## Try setting ‘statistics‘ explicitly in the call to ‘huxreg()‘
```

Exactly which variables are not that important, as long as it shows some independent thought and is justified. Bonus for discussing what are the good and bad control variables. Age should obviously be included. Also bonus if not just include only the same variables we did in class. Some discussion of trying to capture variables that indicates severity/frailty.

The coefficient represent the change in the ‘long odds’, which is uninterpretable except for its sign. A positive coefficient indicate it increase the probability of getting the treatment, and a negative coefficient decreases the probability of getting the treatment. In the example here, we see that education has a positive and significant impact on getting the treatment. Men are also more likely to receive the treatment.

2.4 Show the distribution of the propensity score (for those who get the treatment and those who do not). Discuss your finding from the graph.

```
rhc <- rhc %>%
  mutate(pscore=round(logit$fitted.values, 4),
         id=1:nrow(rhc))
list(rhc)
```

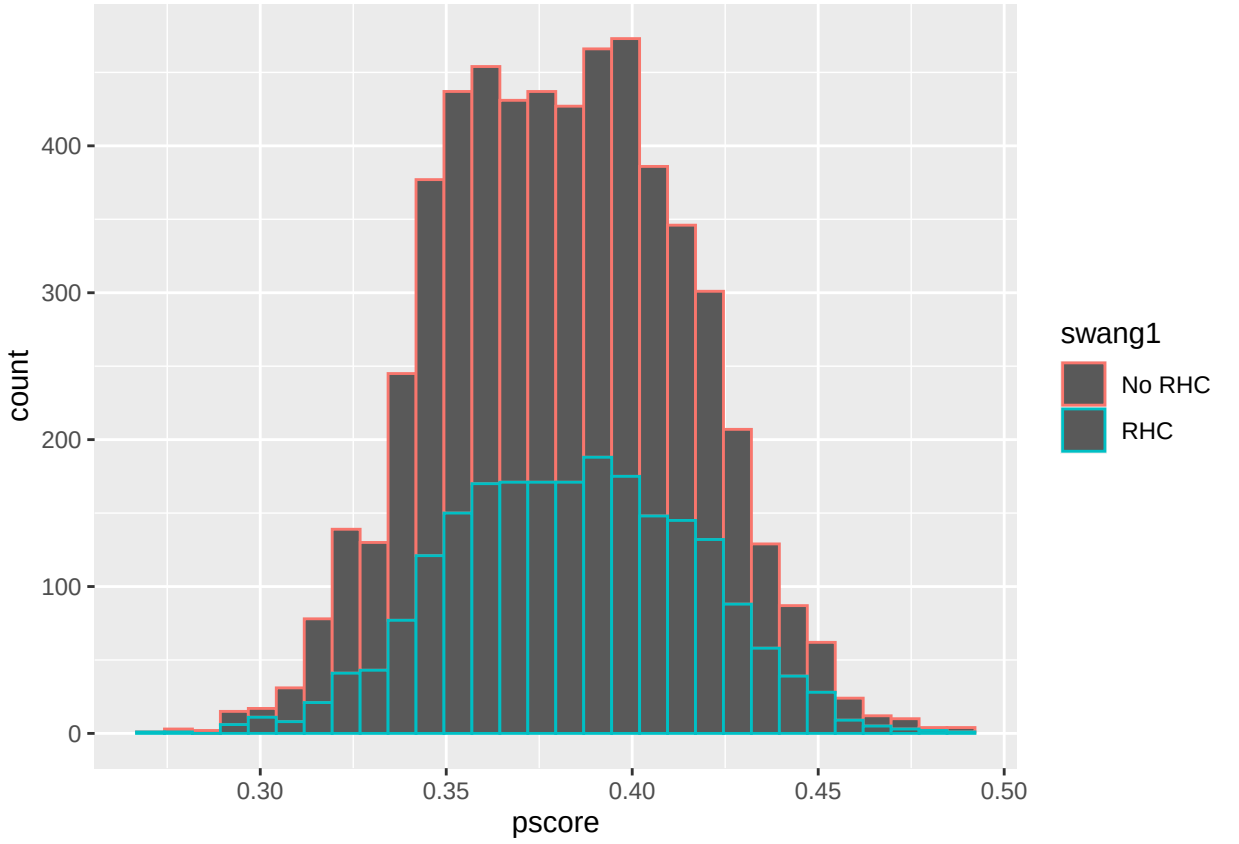
	(1)
(Intercept)	-0.771 ***
	(0.164)
age	-0.003
	(0.002)
men	0.184 ***
	(0.055)
edu	0.026 **
	(0.009)
cat_raceother	0.091
	(0.129)
cat_racewhite	0.053
	(0.077)
N	5735
logLik	-3797.441
AIC	7606.883

*** p < 0.001; ** p < 0.01; * p < 0.05.

```
## [[1]]
## # A tibble: 5,735 x 70
##     ...1     X cat1   cat2  ca    sadmdte dschdte dthdte lstctdte death cardiohx
##    <dbl> <dbl> <chr>  <chr> <chr>   <dbl>   <dbl>  <dbl>    <dbl> <chr>    <dbl>
##  1     1     1 COPD   <NA>  Yes     11142   11151     NA    11382 No           0
##  2     2     2 MOSF ~ <NA>  No      11799   11844  11844    11844 Yes          1
##  3     3     3 MOSF ~ MOSF~ Yes     12083   12143     NA    12400 No           0
##  4     4     4 ARF    <NA>  No      11146   11183  11183    11182 Yes          0
##  5     5     5 MOSF ~ <NA>  No      12035   12037  12037    12036 Yes          0
##  6     6     6 COPD   <NA>  No      12389   12396     NA    12590 No           0
##  7     7     7 MOSF ~ <NA>  Meta~   12381   12423     NA    12616 No           0
##  8     8     8 ARF    Coma  No      11453   11487  11491    11490 Yes          0
##  9     9     9 MOSF ~ <NA>  Yes     12426   12437     NA    12560 No           0
## 10    10    10 ARF    <NA>  Yes     11381   11400     NA    11590 No           0
## # ... with 5,725 more rows, and 59 more variables: chfhx <dbl>, dementhx <dbl>,
## #   psychhx <dbl>, chrpulhx <dbl>, renalhx <dbl>, liverhx <dbl>,
## #   gibledhx <dbl>, malighx <dbl>, immunhx <dbl>, transhx <dbl>, amihx <dbl>,
## #   age <dbl>, sex <chr>, edu <dbl>, surv2md1 <dbl>, das2d3pc <dbl>,
## #   t3d30 <dbl>, dth30 <chr>, aps1 <dbl>, scoma1 <dbl>, meanbp1 <dbl>,
## #   wblc1 <dbl>, hrt1 <dbl>, resp1 <dbl>, temp1 <dbl>, pafi1 <dbl>, alb1 <dbl>,
## #   hema1 <dbl>, bili1 <dbl>, crea1 <dbl>, sod1 <dbl>, pot1 <dbl>, ...
```

```
ggplot(rhc, aes(x = pscore, color = swang1)) +
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Histograms of the propensity score distribution for two groups. The comments should discuss the degree to which there is overlap.

An good overlap indicates that the probability of getting treatment is similar between the control and the treatment group. And vice versa.

2.5 Create a matched dataset by using matching without replacement and match on the propensity score difference. Then assess the balance of the treatment and control group and plot your results.

```
m.out <- matchit(treatment~age+men+edu+cat_race, data = rhc, method = "nearest", distance = "glm", repla
#method indicates the matching method to be used. nearest means the nearest neighbor matching.
#distance indicates the distance measure to be used. glm is default for propensenity scores estimated w
#replace indicates whether matching should be done with replacement. False means no
```

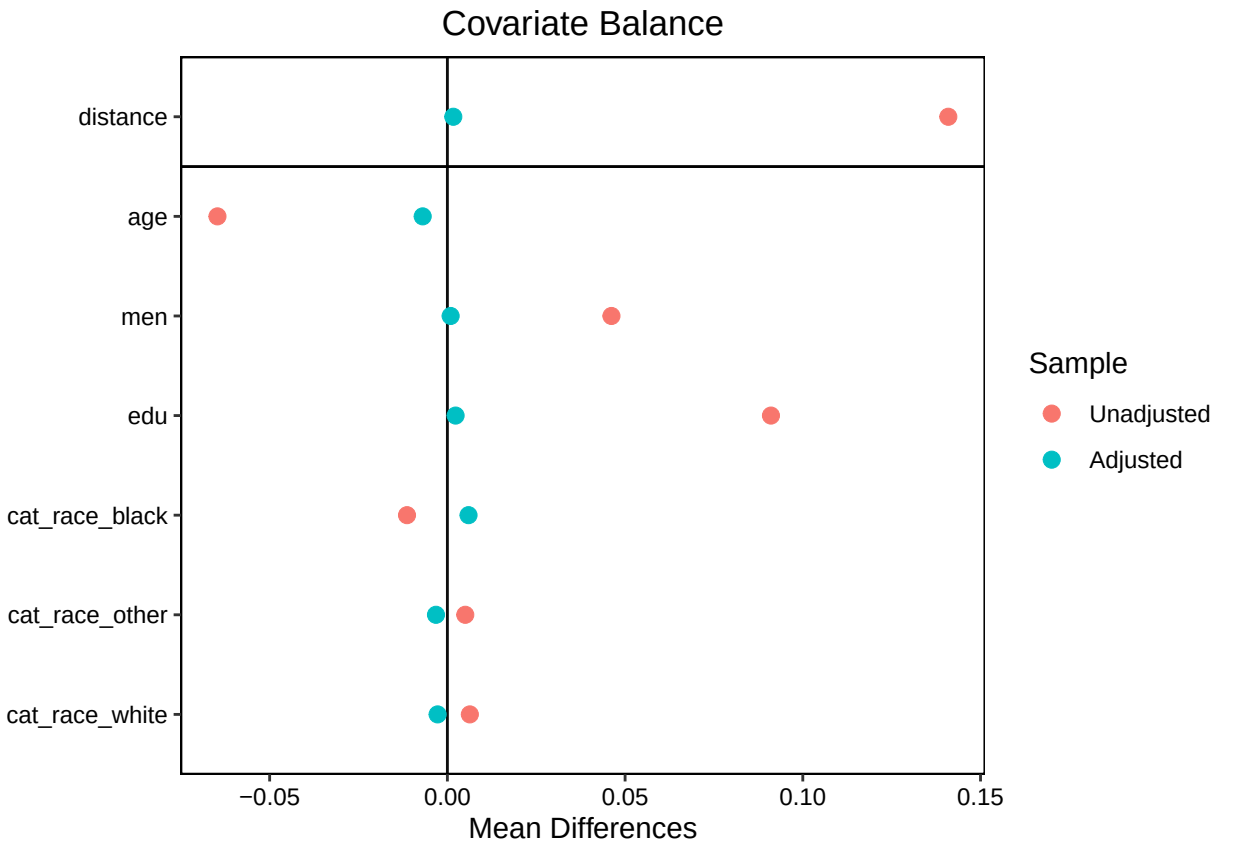
```
m_data <- match.data(m.out)
#match.data() can extract the matched dataset
```

```
summary(m.out)
```

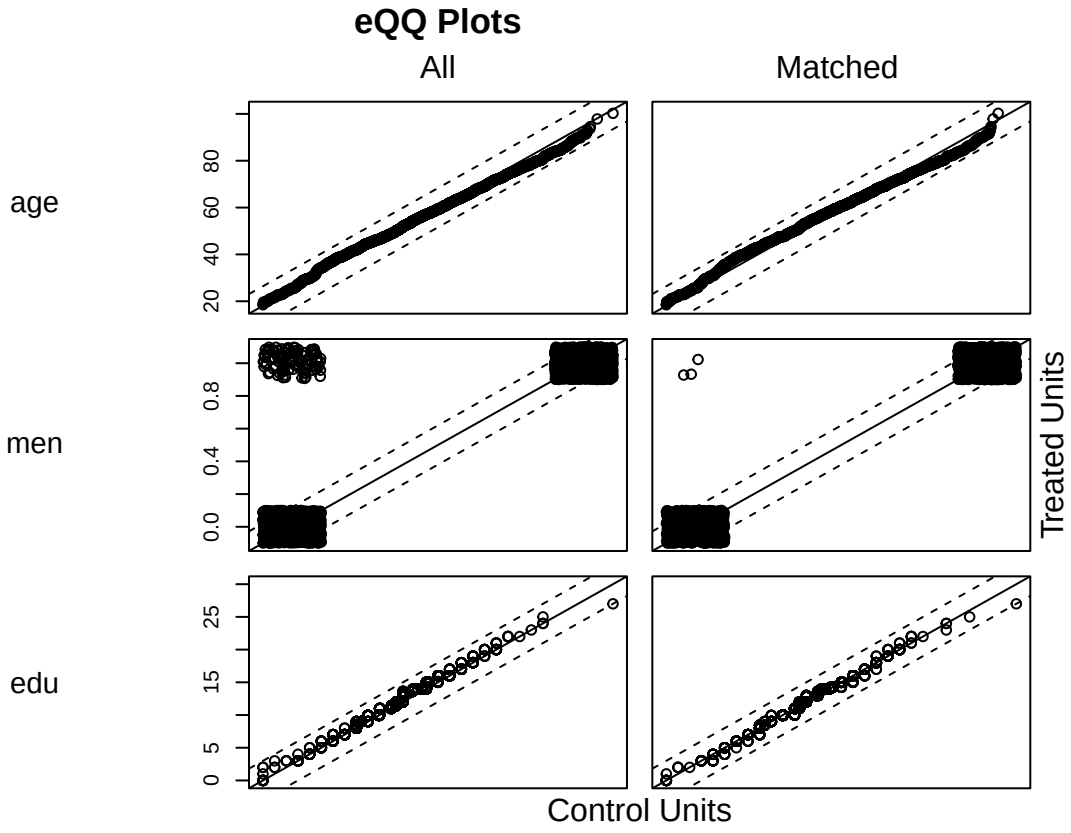
```
##
## Call:
## matchit(formula = treatment ~ age + men + edu + cat_race, data = rhc,
##     method = "nearest", distance = "glm", replace = FALSE)
##
## Summary of Balance for All Data:
##               Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean
## distance             0.3837        0.3791          0.1409     0.9779    0.0417
## age                 60.7498       61.7609         -0.0647     0.8175    0.0285
## men                  0.5852        0.5390          0.0937          .    0.0462
## edu                 11.8564       11.5690          0.0910     1.0147    0.0181
## cat_raceblack        0.1534        0.1647         -0.0315          .    0.0114
## cat_raceother        0.0650        0.0600          0.0204          .    0.0050
## cat_racewhite        0.7816        0.7753          0.0153          .    0.0063
##               eCDF Max
## distance        0.0679
## age             0.0703
## men             0.0462
## edu             0.0511
## cat_raceblack   0.0114
## cat_raceother   0.0050
## cat_racewhite   0.0063
##
##
## Summary of Balance for Matched Data:
##               Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean
## distance             0.3837        0.3836          0.0017     1.0016    0.0005
## age                 60.7498       60.8584         -0.0069     0.7904    0.0265
## men                  0.5852        0.5842          0.0019          .    0.0009
## edu                 11.8564       11.8491          0.0023     1.0144    0.0055
## cat_raceblack        0.1534        0.1474          0.0165          .    0.0060
## cat_raceother        0.0650        0.0682         -0.0130          .    0.0032
## cat_racewhite        0.7816        0.7843         -0.0066          .    0.0027
##               eCDF Max Std. Pair Dist.
## distance        0.0064          0.0029
## age             0.0604          1.0582
## men             0.0009          0.4814
## edu             0.0206          0.7313
## cat_raceblack   0.0060          0.6823
## cat_raceother   0.0032          0.4810
## cat_racewhite   0.0027          0.8156
##
## Sample Sizes:
##           Control Treated
## All          3551    2184
## Matched      2184    2184
## Unmatched    1367       0
## Discarded       0       0
```

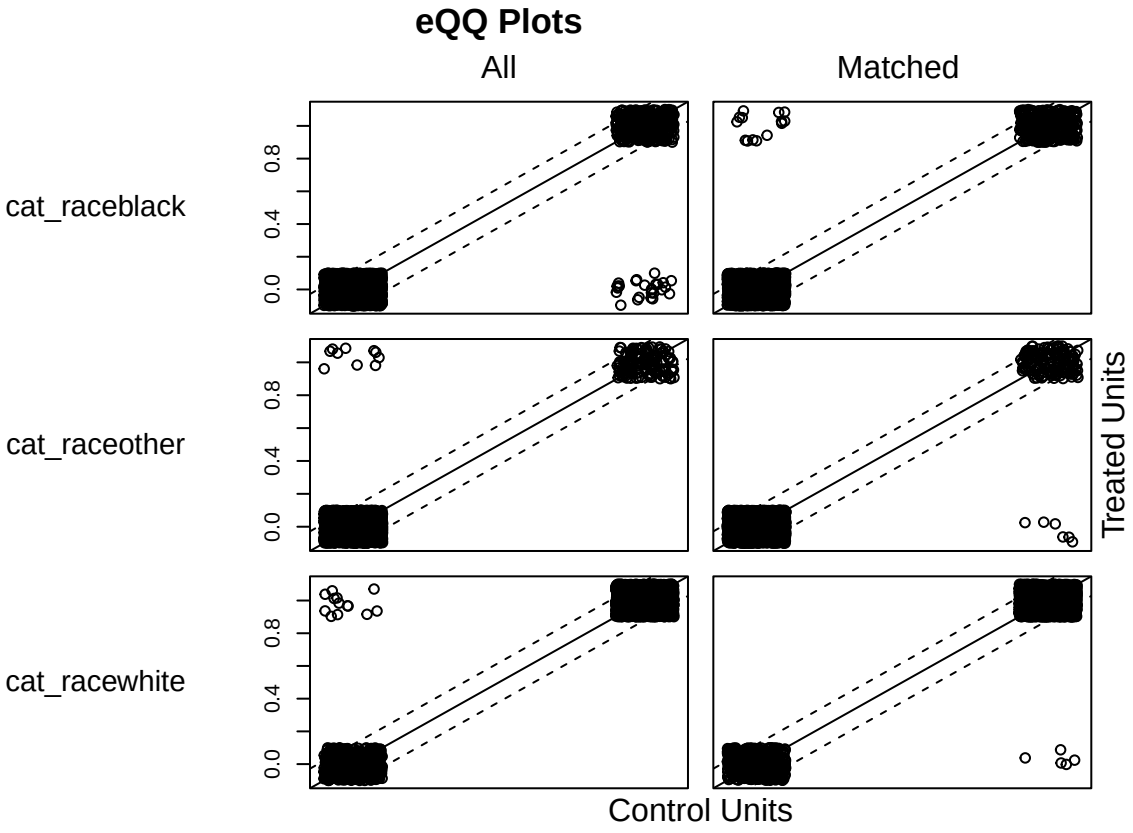
```
love.plot(m.out)
```

```
## Warning: Standardized mean differences and raw mean differences are present in the same plot.
## Use the 'stars' argument to distinguish between them and appropriately label the x-axis.
```




```
plot(m.out)
```


eQQ Plots
All
Matched
age
men
edu
20
40
60
80
0.0
0.4
0.8
0
5
15
25
Treated Units
Control Units



Bonus of showing multiple (e.g. eQQ plots, eCDF plots, or density plots of the covariates and histograms or jitter plots of the propensity score) visual assessments.

Covariate balance shows that the mean difference for the adjusted is around 0 (in blud colour), it implies the covariates have much better balance after the matching.

In eQQ plots, when values fall on the 45 degree line, the groups are balanced. Above, we can see that almost all covariates have much better balance after matching than before.

2.6 Estimate the causal effect using the matched dataset you got from 2.5. Interpret the estimated treatment coefficient.

```
lm(died ~ treatment, data = m_data) %>% huxreg()
```

The treatment effect is positive and significant at 0.1%. Thus, RHC is associated with increased morality. This is probably due to inadequate adjustment for severity.

Note that the treatment coefficient from a LPM cannot be directly interpreted as percentage: getting RHC is estimated to increase the probability of death by 0.068 (according to this regression output). If one has to interpret the coefficient as percentage, then the interpretation should be: an change in the probability of getting RHC by 0.1 is estimated to increase the probability of death by 0.68% (0.068*0.1*100).

	(1)
(Intercept)	0.613 ***
	(0.010)
treatment	0.068 ***
	(0.014)
N	4368
R2	0.005
logLik	-2963.166
AIC	5932.333

*** p < 0.001; ** p < 0.01; * p < 0.05.

2.7 Redo step 2.5 and 2.6 by using exact matching (Hint: read carefully the feature of the exact matching, you would need to do some data preparation before the matching). Evaluate and compare the results with those in 2.6.

Exact matching only matches treated and controlled individuals who have identical covariate values. It suffers from the curse of dimensinality. If one choose many variables and have continuous variables, one has to categorize them into categorical variables. Otherwise, many units without the exactly the same values will be dropped from the analysis, and with many few units remain, the estimated effects will lack prevision and cannot be generalize to the population.

Example of converting variables into categorical variables:

```
rhc$cat_age <- cut(rhc$age,
                       breaks=c(0, 20, 40, 60, 80, 100, 120),
                       labels=c('Blow 20', '20-40', '40-60', '60-80', '80-100', '120'))
```

```
rhc$cat_edu <- cut(rhc$edu,
                       breaks=c(-5, 0, 5, 10, 15, 20, 25, 30),
                       labels=c('-5-0', '0-5','5-10','10-15','15-20','20-25','25-30'))
```

```
#check if the following variables contains missing values
sum(is.na(rhc$cat_edu))
```

```
## [1] 0
```

```
sum(is.na(rhc$cat_age))
```

```
## [1] 0
```

```
sum(is.na(rhc$men))
```

```
## [1] 0
```

```r
sum(is.na(rhc$treatment))
```

```
## [1] 0
```

```r
m.out_2 <- matchit(treatment~cat_age+men+cat_edu+cat_race, data = rhc, method = "exact")
```

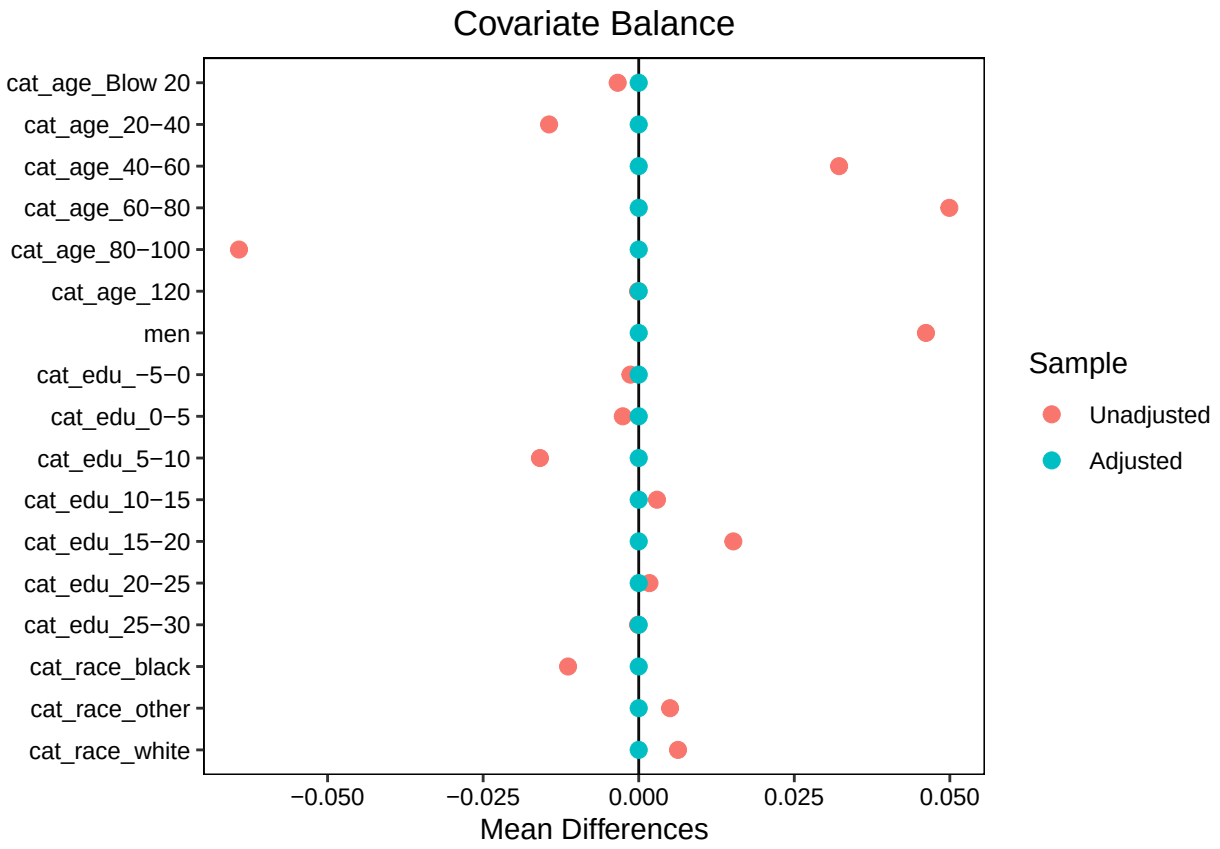
```r
m_data_2 <- match.data(m.out_2)
```

```r
summary(m.out_2)
```

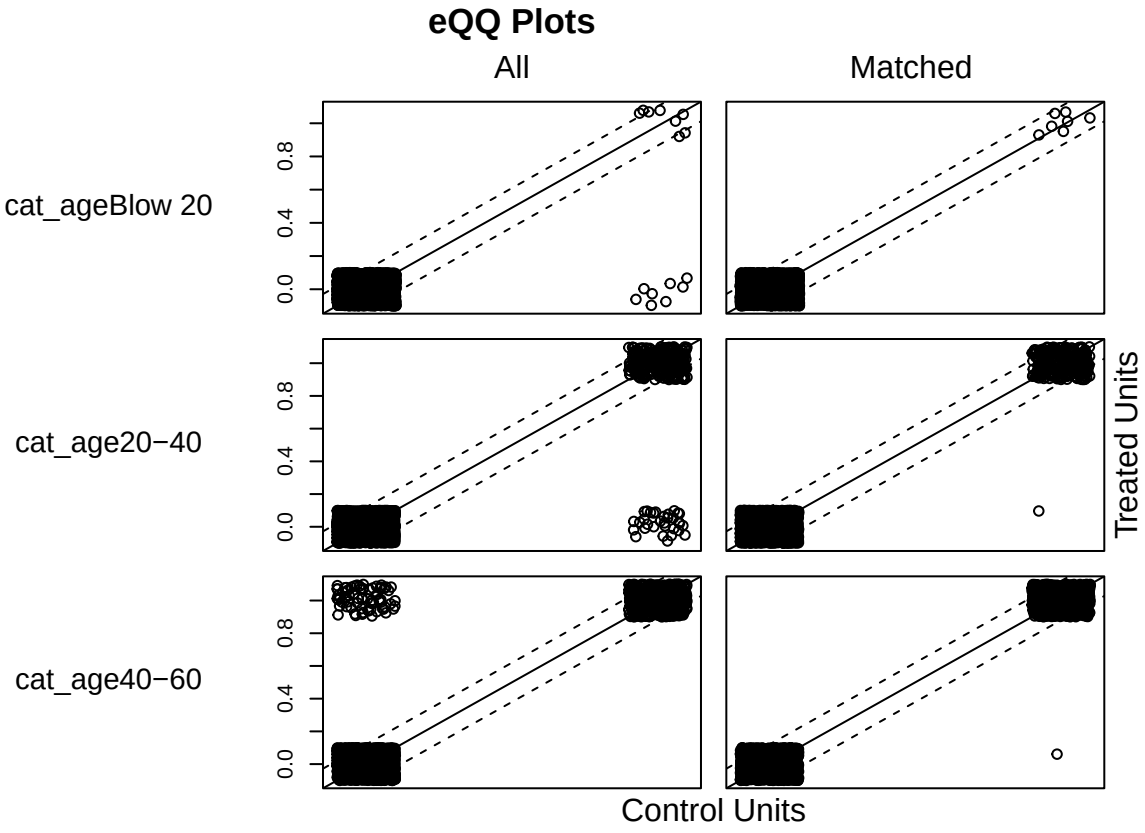
```
##
## Call:
## matchit(formula = treatment ~ cat_age + men + cat_edu + cat_race,
##     data = rhc, method = "exact")
##
## Summary of Balance for All Data:
##                Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean
## cat_ageBlow 20        0.0037        0.0070         -0.0559          .    0.0034
## cat_age20-40          0.1140        0.1284         -0.0453          .    0.0144
## cat_age40-60          0.2995        0.2672          0.0703          .    0.0322
## cat_age60-80          0.5064        0.4565          0.0998          .    0.0499
## cat_age80-100         0.0760        0.1402         -0.2424          .    0.0642
## cat_age120            0.0005        0.0006         -0.0049          .    0.0001
## men                   0.5852        0.5390          0.0937          .    0.0462
## cat_edu-5-0           0.0046        0.0059         -0.0198          .    0.0013
## cat_edu0-5            0.0234        0.0259         -0.0169          .    0.0026
## cat_edu5-10           0.2550        0.2709         -0.0364          .    0.0159
## cat_edu10-15          0.5870        0.5841          0.0060          .    0.0029
## cat_edu15-20          0.1236        0.1084          0.0462          .    0.0152
## cat_edu20-25          0.0060        0.0042          0.0225          .    0.0017
## cat_edu25-30          0.0005        0.0006         -0.0049          .    0.0001
## cat_raceblack         0.1534        0.1647         -0.0315          .    0.0114
## cat_raceother         0.0650        0.0600          0.0204          .    0.0050
## cat_racewhite         0.7816        0.7753          0.0153          .    0.0063
##                eCDF Max
## cat_ageBlow 20   0.0034
## cat_age20-40     0.0144
## cat_age40-60     0.0322
## cat_age60-80     0.0499
## cat_age80-100    0.0642
## cat_age120       0.0001
## men              0.0462
## cat_edu-5-0      0.0013
## cat_edu0-5       0.0026
## cat_edu5-10      0.0159
## cat_edu10-15     0.0029
## cat_edu15-20     0.0152
## cat_edu20-25     0.0017
## cat_edu25-30     0.0001
## cat_raceblack    0.0114
## cat_raceother    0.0050
## cat_racewhite    0.0063
```

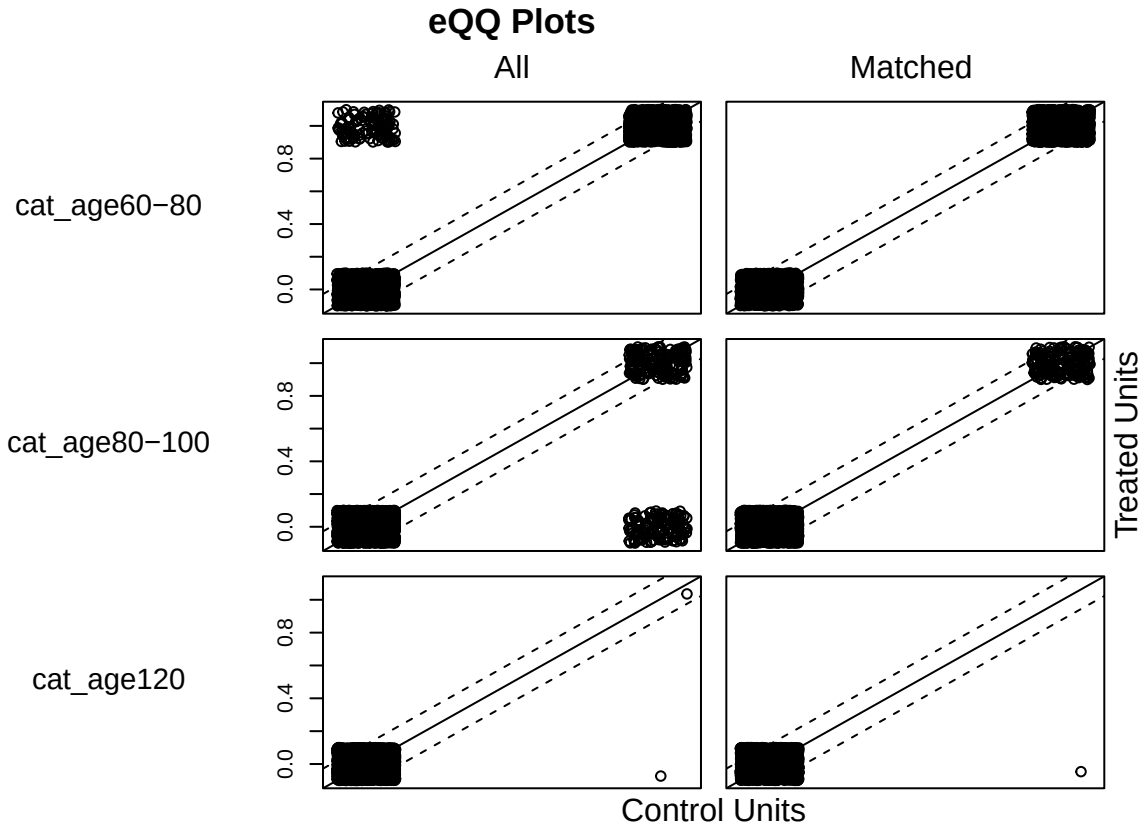
```
##
##
## Summary of Balance for Matched Data:
##                Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean
## cat_ageBlow 20        0.0032        0.0032               0          .         0
## cat_age20-40          0.1150        0.1150               0          .         0
## cat_age40-60          0.2996        0.2996               0          .         0
## cat_age60-80          0.5069        0.5069               0          .         0
## cat_age80-100         0.0753        0.0753               0          .         0
## cat_age120            0.0000        0.0000               0          .         0
## men                   0.5854        0.5854               0          .         0
## cat_edu-5-0           0.0023        0.0023               0          .         0
## cat_edu0-5            0.0212        0.0212              -0          .         0
## cat_edu5-10           0.2562        0.2562               0          .         0
## cat_edu10-15          0.5914        0.5914               0          .         0
## cat_edu15-20          0.1247        0.1247               0          .         0
## cat_edu20-25          0.0042        0.0042               0          .         0
## cat_edu25-30          0.0000        0.0000               0          .         0
## cat_raceblack         0.1519        0.1519               0          .         0
## cat_raceother         0.0619        0.0619               0          .         0
## cat_racewhite         0.7862        0.7862               0          .         0
##                eCDF Max Std. Pair Dist.
## cat_ageBlow 20        0               0
## cat_age20-40          0               0
## cat_age40-60          0               0
## cat_age60-80          0               0
## cat_age80-100         0               0
## cat_age120            0               0
## men                   0               0
## cat_edu-5-0           0               0
## cat_edu0-5            0               0
## cat_edu5-10           0               0
## cat_edu10-15          0               0
## cat_edu15-20          0               0
## cat_edu20-25          0               0
## cat_edu25-30          0               0
## cat_raceblack         0               0
## cat_raceother         0               0
## cat_racewhite         0               0
##
## Sample Sizes:
##               Control Treated
## All              3551    2184
## Matched (ESS)    3158    2166
## Matched          3461    2166
## Unmatched          90      18
## Discarded           0       0
```

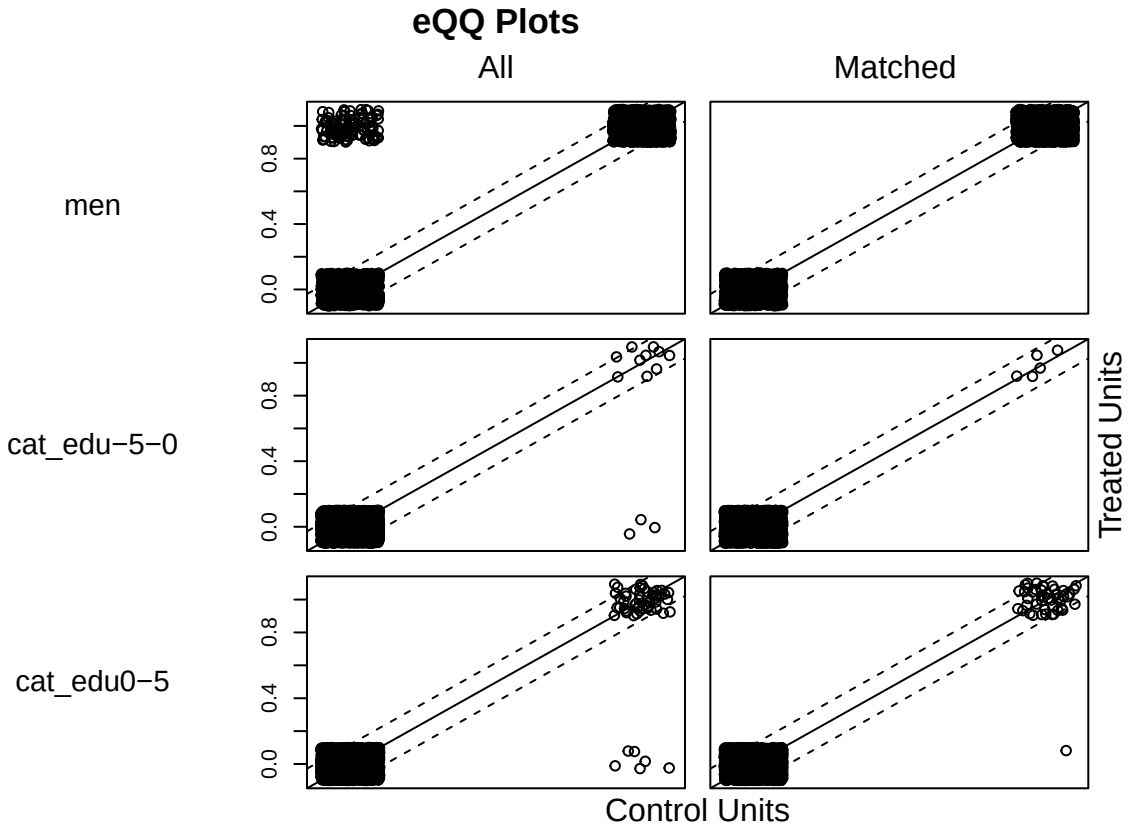
```
love.plot(m.out_2)
```

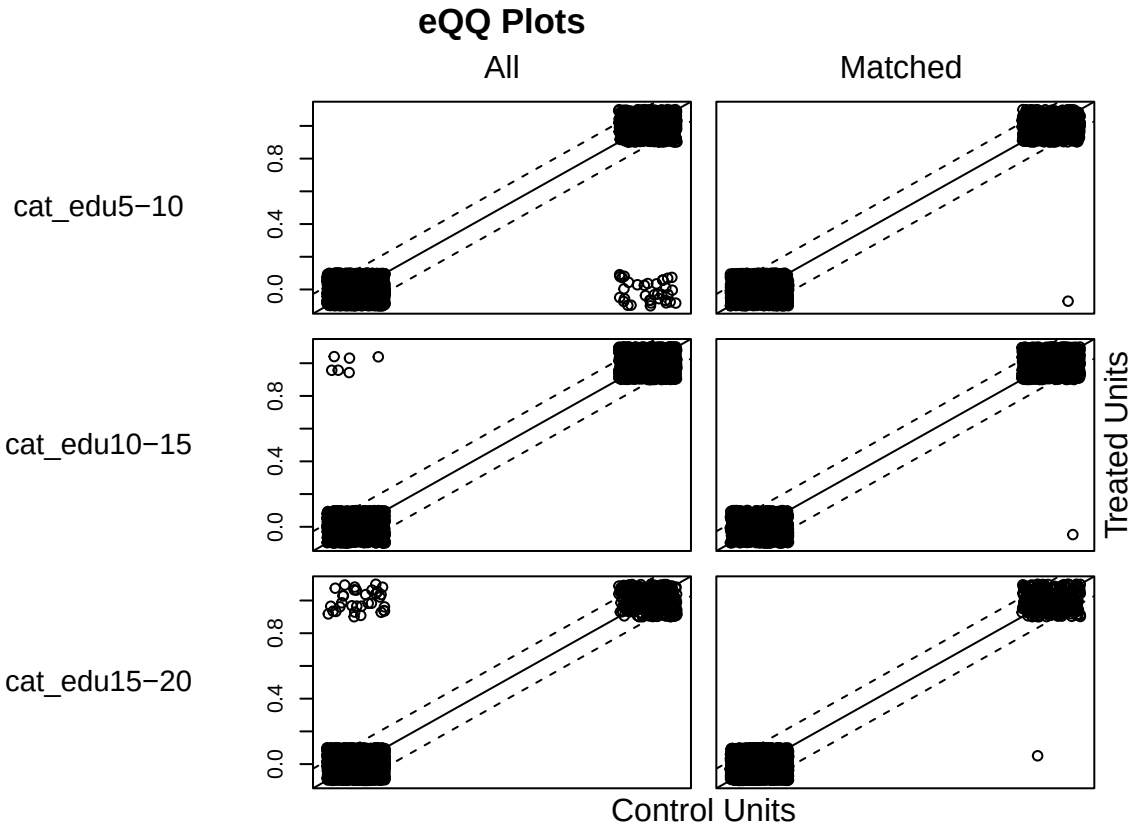

Covariate Balance
cat_age_Blow 20
cat_age_20−40
cat_age_40−60
cat_age_60−80
cat_age_80−100
cat_age_120
men
cat_edu_−5−0
cat_edu_0−5
cat_edu_5−10
cat_edu_10−15
cat_edu_15−20
cat_edu_20−25
cat_edu_25−30
cat_race_black
cat_race_other
cat_race_white
−0.050
−0.025
0.000
0.025
0.050
Mean Differences
Sample
Unadjusted
Adjusted

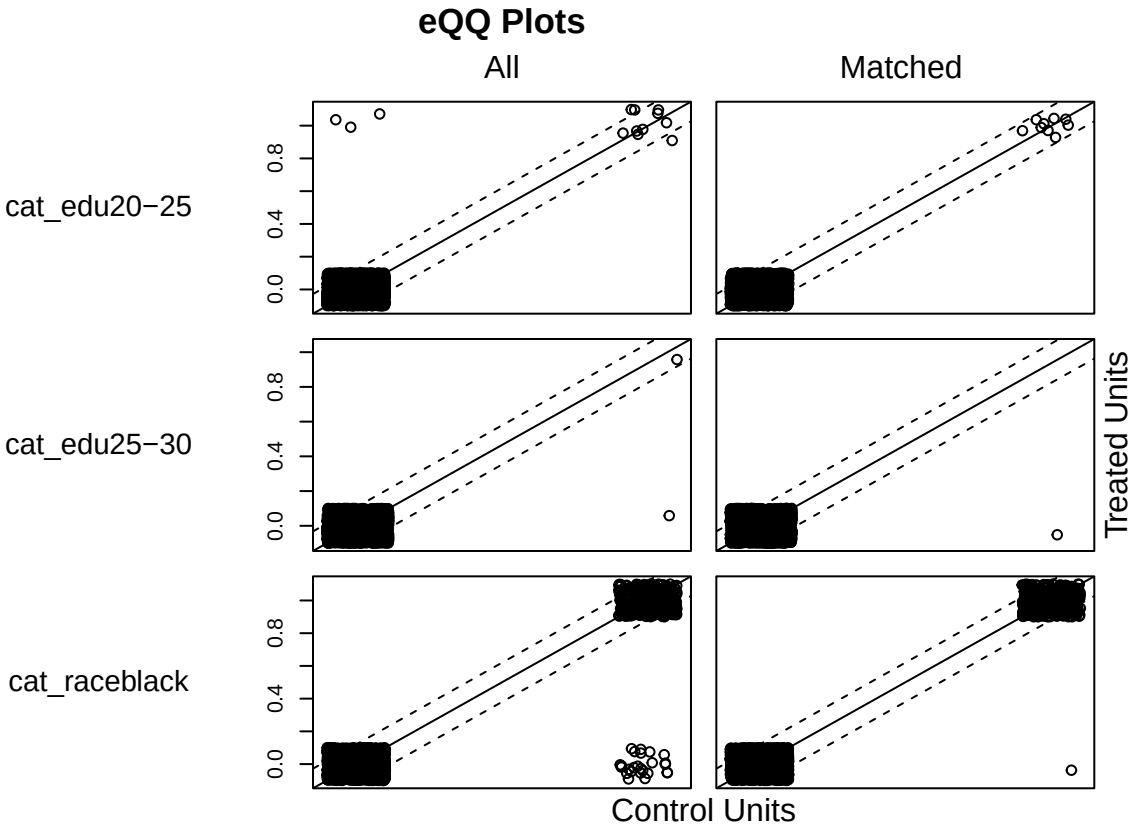
```
plot(m.out_2)
```

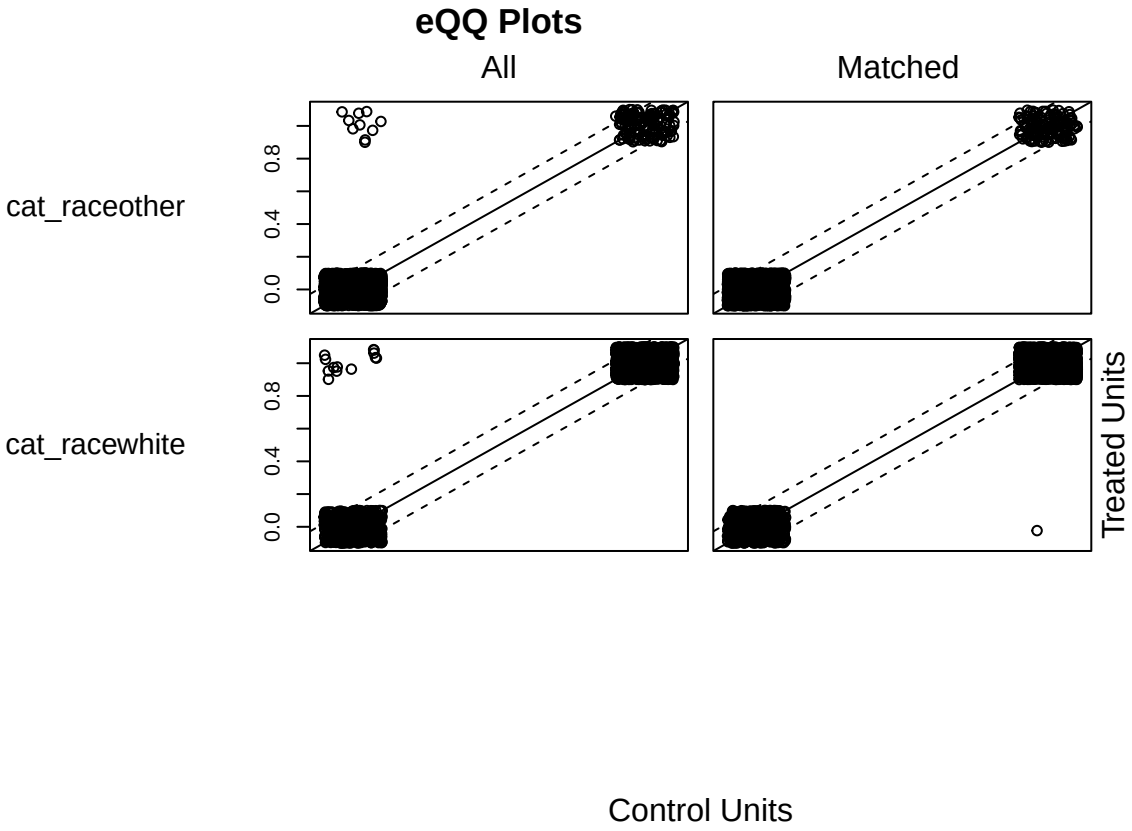

eQQ Plots
All
Matched
cat_ageBlow 20
cat_age20−40
cat_age40−60
0.0
0.4
0.8
Treated Units
Control Units


eQQ Plots
All
Matched
cat_age60−80
cat_age80−100
cat_age120
0.0
0.4
0.8
Treated Units
Control Units


eQQ Plots
All
Matched
men
cat_edu−5−0
cat_edu0−5
0.0
0.4
0.8
Treated Units
Control Units


eQQ Plots
All
Matched
cat_edu5−10
cat_edu10−15
cat_edu15−20
0.0
0.4
0.8
Treated Units
Control Units


eQQ Plots
All
Matched
cat_edu20−25
cat_edu25−30
cat_raceblack
0.0
0.4
0.8
Treated Units
Control Units



Points are given to giving explanations for table of the summary of balance, plot of covariate balance, and bonus for using multiple (e.g. eQQ plots, eCDF plots, or density plots of the covariates and histograms or jitter plots of the propensity score) visual assessments.

```
lm(died ~ treatment, data = m_data_2) %>% huxreg()
```

	(1)
(Intercept)	0.630 ***
	(0.008)
treatment	0.049 ***
	(0.013)
N	5627
R2	0.003
logLik	-3814.520
AIC	7635.040

*** p < 0.001; ** p < 0.01; * p < 0.05.

The treatment effect is positive and significant at 0.1%. According to this estimation results, getting RHC

is estimated to increase the probability of death by 0.049 (or a change in the probability of getting RHC is estimated to increase the probability of death by 0.49%).

The magnitude is smaller than that in 2.6 (where the treatment coefficient is equal to 0.068). In both 2.6 and 2.7, we just go ahead estimate the treatment effect by assuming we can't get better balance.

Causal conclusions in practice shouldn't depend on just the two approaches we've tried. We should test other methods to see if we can get better balance: for instance, trying matching with replacement, make the PS estimating function more flexible, try a totally different matching method.