Exercise: Matching

Let's simulate some fake data and see whether we are able to recover the correct treatment effect using matching methods.

1. First, let's generate some confounder variables for 100 people.

(a) The variable 'age' should be drawn randomly from the normal distribution with mean 40 and standard deviation 7.

(b) The variable 'gender' should be drawn randomly from the binomial distribution with a 0.5 probability of being male or female.

(c) The variable 'income' should be drawn randomly from the normal distribution with mean 500 and standard deviation 50.

(d) The variable 'education' should be randomly drawn from one of four numerical categories with equal probability: 0 (None), 1 (Primary), 2 (Secondary), 3 (Tertiary). *Hint: Try using* `sample()` *(with replace=T) in R, or* `rdiscrete` *in Stata.*

```
set.seed(54321)
N <- 100
d <- tibble(age=rnorm(N,40,7),
            gender=rbinom(N,1,0.5),
            income=rnorm(N,500,50),
            education=sample(c(0,1,2,3),N,
                             prob=c(0.25,0.25,0.25,0.25), replace=T))
```

2. Our outcome is going to be attitudes to redistribution. Use the expressions below to simulate potential outcomes, with a treatment effect of 5.

$$y_0 = N(20,5) + \frac{age}{4} - 5 * gender + \frac{income}{50} - 3 * education$$

$$y_1 = y_0 + 5$$

```
set.seed(54001)
d <- d %>% mutate(y_0=rnorm(N,20,5) + age/4 - 5*gender + income/50 - education*3,
                  y_1=y_0+5)
```

3. Treatment D is receiving a government social program, but treatment is **not** randomly assigned in any way. Instead, treatment depends on age, gender, income and education. Imagine we know the treatment assignment mechanism so that binary (1/0) treatment is determined by the following expression:

$$D = \begin{cases} 1 \text{ if } (2 * gender + \frac{age}{8} + \frac{income}{50} + 2 * education + N(0,3)) > 19 \\ 0 \text{ else} \end{cases}$$

```
set.seed(54001)
d <- d %>% mutate(D=case_when(2*gender+age/8+income/50+education*2 + rnorm(N,0,3)>19~1,
                              T~0))
#summary(2*d$gender + d$age/8 + d$income/50 + d$education*2)
```

4. Calculate observed outcomes based on potential outcomes and treatment.

```
d <- d %>% mutate(y_obs=case_when(D==0~y_0,
                                  D==1~y_1))
```

5. As always, as a benchmark, let's run the 'naive' regression of the outcome on the treatment with no controls. Why is the result different from our assumed treatment effect? Be specific.

```
d %>% lm(y_obs ~ D, data=.) %>% stargazer(title="Q5")
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, May 13, 2020 - 5:20:03 PM

Table 1: Q5

	Dependent variable:
	y_obs
D	6.338***
	(1.384)
Constant	32.039***
	(0.959)
Observations	100
R^2	0.176
Adjusted R^2	0.168
Residual Std. Error	6.915 (df = 98)
F Statistic	20.964*** (df = 1; 98)
Note:	*p<0.1; **p<0.05; ***p<0.01

Gender, age, income and education are all confounders that bias our estimate.

6. Our first task is to try and do a 'manual' matching example - to try and 'match' one treated unit with one control unit so that the *only* thing that is different about them is their treatment status. Take the first treated unit in your dataset. What are its values of gender, age, income and education? Manually, by trial-and-error (not using any package or pre-prepared function), identify the most similar *control* unit. How different are your matched pair on these four variables?

```
treated_unit <- d %>% filter(D==1) %>% slice(1)
control_units <- d %>% filter(D==0 & gender==1 & education==1)
control_unit <- control_units %>% filter(age>32 & age < 36 & income>500 & income<550)

rbind(treated_unit, control_unit) %>% kable(caption="Q6")
```

Table 2: Q6

age	gender	income	education	y_0	y_1	D	y_obs
34.51176	1	539.5772	1	34.80170	39.80170	1	39.80170
33.59721	1	532.5637	1	29.67072	34.67072	0	29.67072

age	gender	income	education	y_0	y_1	D	y_obs

The treated unit is a 34.5 year-old female with income of 540 and education of level 1; the control unit is a 33.5 year-old female with income of 533 and education of level 1. These differences seem reasonably small so they are good counterfactuals for each other.

7. Compare the outcome between your matched treated unit and control unit. Is this consistent with our assumed treatment effect? Why is it similar? Why is it different?

```
treated_unit$y_obs - control_unit$y_obs
```

```
## [1] 10.13098
```

This is much larger than our assumed treatment effect, purely by chance because the y_1 of the treated unit is high and the y_0 of the control unit is low. This reflects the 'noise' in potential outcomes and not any systematic confounding, since we have already made sure the two units are balanced on these confounding variables.

8. Matching repeats this process for multiple units and then finds the average difference in outcomes between the treated and control units. Use the *matchit* package to conduct 'nearest neighbour' (the default) matching method on your dataset for the four confounder variables: gender, education, age and income. What is the result of the matching procedure - how many units were matched?

```
d <- d %>% mutate(gender=factor(gender),
                  education=factor(education))
matched_data_Q8 <- matchit(D ~ gender + education + age + income, data=d)
matched_data_Q8
```

```
##
## Call:
## matchit(formula = D ~ gender + education + age + income, data = d)
##
## Sample sizes:
##           Control Treated
## All            52      48
## Matched        48      48
## Unmatched       4       0
## Discarded       0       0
```

The result shows that all 48 treated units are matched, and 48 of the 52 control units are matched. In other words, 4 control units are thrown away because they are not useful for comparison.

9. Use *match.data* to extract the matched dataset and calculate the average difference in means between the treated and control groups. How does the result compare to the naive regression in Q5?

```
matched_data_Q8 %>% match.data() %>%
  group_by(D) %>%
  summarize(y_obs=mean(y_obs,na.rm=T)) %>%
  arrange(-D) %>%
  mutate(diff_y_obs=y_obs-lead(y_obs)) %>% kable(caption="Q9")
```

Table 3: Q9

D	y_obs	diff_y_obs
1	38.37617	6.721451
0	31.65471	NA

The matched dataset has a difference in outcomes between treatment and control of 6.6, more than our specified effect of 5 and quite similar to the naive regression in Q5.

10. To understand how matching changed our dataset, check the *summary* information about your matched data.

(a) On which variables did balance improve? Did balance deteriorate on any variables?

```
matched_data_Q8 %>% summary()
```

```
##
## Call:
## matchit(formula = D ~ gender + education + age + income, data = d)
##
## Summary of balance for all data:
##            Means Treated Means Control SD Control Mean Diff eQQ Med
## distance          0.7252        0.2537     0.2553    0.4715  0.5732
## gender0           0.3542        0.6154     0.4913   -0.2612  0.0000
## gender1           0.6458        0.3846     0.4913    0.2612  0.0000
## education1        0.1250        0.3077     0.4660   -0.1827  0.0000
## education2        0.2708        0.1731     0.3820    0.0978  0.0000
## education3        0.4583        0.1154     0.3226    0.3429  0.0000
## age              41.5227       37.3373     7.0965    4.1854  4.8056
## income          507.2430      489.9554    47.0206   17.2875 20.3831
##            eQQ Mean eQQ Max
## distance     0.4790  0.6380
## gender0      0.2500  1.0000
## gender1      0.2708  1.0000
## education1   0.1667  1.0000
## education2   0.1042  1.0000
## education3   0.3542  1.0000
## age          4.4490  6.6536
## income      20.3717 31.7572
##
##
## Summary of balance for matched data:
##            Means Treated Means Control SD Control Mean Diff eQQ Med
## distance          0.7252        0.2746     0.2548    0.4506  0.5399
## gender0           0.3542        0.5833     0.4982   -0.2292  0.0000
## gender1           0.6458        0.4167     0.4982    0.2292  0.0000
## education1        0.1250        0.2708     0.4491   -0.1458  0.0000
## education2        0.2708        0.1875     0.3944    0.0833  0.0000
## education3        0.4583        0.1250     0.3342    0.3333  0.0000
## age              41.5227       37.7301     7.2287    3.7926  3.8323
## income          507.2430      495.0532    45.1915   12.1898 13.2874
##            eQQ Mean eQQ Max
## distance     0.4506  0.5997
## gender0      0.2292  1.0000
## gender1      0.2292  1.0000
## education1   0.1458  1.0000
## education2   0.0833  1.0000
## education3   0.3333  1.0000
## age          3.7941  6.0310
## income      14.0521 28.7073
##
## Percent Balance Improvement:
```

```
##            Mean Diff. eQQ Med eQQ Mean eQQ Max
## distance       4.4314  5.8121   5.9358  6.0087
## gender0       12.2699  0.0000   8.3333  0.0000
## gender1       12.2699  0.0000  15.3846  0.0000
## education1    20.1754  0.0000  12.5000  0.0000
## education2    14.7541  0.0000  20.0000  0.0000
## education3     2.8037  0.0000   5.8824  0.0000
## age            9.3851 20.2534  14.7210  9.3577
## income        29.4882 34.8117  31.0213  9.6038
##
## Sample sizes:
##           Control Treated
## All            52      48
## Matched        48      48
## Unmatched       4       0
## Discarded       0       0
```

Balance improved for gender, education, age and income.

(b) Since we still have imbalance after matching, we can try to estimate the effect of treatment using a regression *on our matched dataset*. Include all of the confounding variables as controls. Does our estimate improve?

```
matched_data_Q8 %>% match.data() %>% lm(y_obs ~ D + gender + education + age + income, data=.) %>% starg
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, May 13, 2020 - 5:20:03 PM

11. Matching *ONLY* makes a difference if we throw away some data - the data for which we cannot find good matches. The more data we throw away, the better matched/balanced is our remaining data.

(a) Conduct your nearest neighbour matching procedure again, but this time use the *exact* parameter to also require that matched treated and control units have exactly the same gender and education.

```
matched_data_Q11 <- matchit(D ~ gender + education + age + income, data=data.frame(d),exact=c("gender",
```

(b) How many units are matched now?

```
matched_data_Q11
```

```
##
## Call:
## matchit(formula = D ~ gender + education + age + income, data = data.frame(d),
##     exact = c("gender", "education"))
##
## Sample sizes:
##           Control Treated
## All            52      48
## Matched        24      24
## Unmatched      28      24
## Discarded       0       0
```

Now only 74 units are matched (37 control and 37 treated), with 15 control and 11 treated units thrown away.

(c) Has balanced improved or deteriorated on any variables?

```
matched_data_Q11 %>% summary()
```

```
##
## Call:
```

Table 4: Q10(b)

	Dependent variable:
	y_obs
D	12.613^{***}
	(0.942)
gender1	-7.312^{***}
	(0.808)
education1	-5.408^{***}
	(1.093)
education2	-9.956^{***}
	(1.121)
education3	-14.456^{***}
	(1.121)
age	0.176^{***}
	(0.053)
income	-0.002
	(0.009)
Constant	34.147^{***}
	(5.125)
Observations	96
R^2	0.802
Adjusted R^2	0.786
Residual Std. Error	3.567 (df = 88)
F Statistic	50.926^{***} (df = 7; 88)
Note:	*p<0.1; **p<0.05; ***p<0.01

```
## matchit(formula = D ~ gender + education + age + income, data = data.frame(d),
##     exact = c("gender", "education"))
##
## Summary of balance for all data:
##              Means Treated Means Control SD Control Mean Diff eQQ Med
## distance            0.7252        0.2537     0.2553    0.4715  0.5732
## gender0             0.3542        0.6154     0.4913   -0.2612  0.0000
## gender1             0.6458        0.3846     0.4913    0.2612  0.0000
## education1          0.1250        0.3077     0.4660   -0.1827  0.0000
## education2          0.2708        0.1731     0.3820    0.0978  0.0000
## education3          0.4583        0.1154     0.3226    0.3429  0.0000
## age                41.5227       37.3373     7.0965    4.1854  4.8056
## income            507.2430      489.9554    47.0206   17.2875 20.3831
## gender0.1           0.3542        0.6154     0.4913   -0.2612  0.0000
## gender1.1           0.6458        0.3846     0.4913    0.2612  0.0000
## education0          0.1458        0.4038     0.4955   -0.2580  0.0000
## education1.1        0.1250        0.3077     0.4660   -0.1827  0.0000
## education2.1        0.2708        0.1731     0.3820    0.0978  0.0000
## education3.1        0.4583        0.1154     0.3226    0.3429  0.0000
##              eQQ Mean eQQ Max
## distance       0.4790  0.6380
## gender0        0.2500  1.0000
## gender1        0.2708  1.0000
## education1     0.1667  1.0000
## education2     0.1042  1.0000
## education3     0.3542  1.0000
## age            4.4490  6.6536
## income        20.3717 31.7572
## gender0.1      0.2500  1.0000
## gender1.1      0.2708  1.0000
## education0     0.2500  1.0000
## education1.1   0.1667  1.0000
## education2.1   0.1042  1.0000
## education3.1   0.3542  1.0000
##
##
## Summary of balance for matched data:
##              Means Treated Means Control SD Control Mean Diff eQQ Med
## distance            0.6194        0.4560     0.2377    0.1634  0.1113
## gender0             0.5833        0.5833     0.5036    0.0000  0.0000
## gender1             0.4167        0.4167     0.5036    0.0000  0.0000
## education1          0.2500        0.2500     0.4423    0.0000  0.0000
## education2          0.2083        0.2083     0.4149    0.0000  0.0000
## education3          0.2500        0.2500     0.4423    0.0000  0.0000
## age                43.0403       39.2744     7.5337    3.7658  4.5722
## income            524.7409      513.9753    49.7142   10.7657 14.0990
## gender0.1           0.5833        0.5833     0.5036    0.0000  0.0000
## gender1.1           0.4167        0.4167     0.5036    0.0000  0.0000
## education0          0.2917        0.2917     0.4643    0.0000  0.0000
## education1.1        0.2500        0.2500     0.4423    0.0000  0.0000
## education2.1        0.2083        0.2083     0.4149    0.0000  0.0000
## education3.1        0.2500        0.2500     0.4423    0.0000  0.0000
##              eQQ Mean eQQ Max
## distance       0.1636  0.3732
```

```
## gender0        0.0000  0.0000
## gender1        0.0000  0.0000
## education1     0.0000  0.0000
## education2     0.0000  0.0000
## education3     0.0000  0.0000
## age            4.0725  7.6018
## income        15.3107 61.5963
## gender0.1      0.0000  0.0000
## gender1.1      0.0000  0.0000
## education0     0.0000  0.0000
## education1.1   0.0000  0.0000
## education2.1   0.0000  0.0000
## education3.1   0.0000  0.0000
##
## Percent Balance Improvement:
##               Mean Diff. eQQ Med eQQ Mean  eQQ Max
## distance         65.3345 80.5858  65.8448  41.5038
## gender0         100.0000  0.0000 100.0000 100.0000
## gender1         100.0000  0.0000 100.0000 100.0000
## education1      100.0000  0.0000 100.0000 100.0000
## education2      100.0000  0.0000 100.0000 100.0000
## education3      100.0000  0.0000 100.0000 100.0000
## age              10.0249  4.8575   8.4617 -14.2505
## income           37.7259 30.8297  24.8435 -93.9601
## gender0.1       100.0000  0.0000 100.0000 100.0000
## gender1.1       100.0000  0.0000 100.0000 100.0000
## education0      100.0000  0.0000 100.0000 100.0000
## education1.1    100.0000  0.0000 100.0000 100.0000
## education2.1    100.0000  0.0000 100.0000 100.0000
## education3.1    100.0000  0.0000 100.0000 100.0000
##
## Sample sizes:
##           Control Treated
## All            52      48
## Matched        24      24
## Unmatched      28      24
## Discarded       0       0
```

Balance has improved a lot on gender and education - they are now perfectly balanced - while age and income are now slightly *less* balanced.

(d) What is the average difference in mean outcomes between treated and control groups?

```
matched_data_Q11 %>% match.data() %>%
  group_by(D) %>%
  summarize(y_obs=mean(y_obs,na.rm=T)) %>%
  arrange(-D) %>%
  mutate(diff_y_obs=y_obs-lead(y_obs)) %>% kable(caption="Q611(d)")
```

Table 5: Q611(d)

D	y_obs	diff_y_obs
1	43.28320	13.88666
0	29.39654	NA

The mean difference in outcomes between treatment and control is now 8.87, higher than our specified value of 5.

12. An alternative way of limiting the number of matches is to specify a maximum distance measure beyond which paired units are dropped.

(a) Run your matching procedure again, specifying a *caliper* of 0.1 (or try other values if this doesn't work).

```
matched_data_Q12 <- matchit(D ~ gender + education + age + income, data=data.frame(d), caliper=0.1)
```

(b) How many units are matched now?

```
matched_data_Q12
```

```
##
## Call:
## matchit(formula = D ~ gender + education + age + income, data = data.frame(d),
##     caliper = 0.1)
##
## Sample sizes:
##           Control Treated
## All            52      48
## Matched        16      16
## Unmatched      36      32
## Discarded       0       0
```

58 units are matched, and 42 thrown away.

(c) Has balanced improved?

```
matched_data_Q12 %>% summary()
```

```
##
## Call:
## matchit(formula = D ~ gender + education + age + income, data = data.frame(d),
##     caliper = 0.1)
##
## Summary of balance for all data:
##            Means Treated Means Control SD Control Mean Diff eQQ Med
## distance          0.7252        0.2537     0.2553    0.4715  0.5732
## gender0           0.3542        0.6154     0.4913   -0.2612  0.0000
## gender1           0.6458        0.3846     0.4913    0.2612  0.0000
## education1        0.1250        0.3077     0.4660   -0.1827  0.0000
## education2        0.2708        0.1731     0.3820    0.0978  0.0000
## education3        0.4583        0.1154     0.3226    0.3429  0.0000
## age              41.5227       37.3373     7.0965    4.1854  4.8056
## income          507.2430      489.9554    47.0206   17.2875 20.3831
##            eQQ Mean eQQ Max
## distance     0.4790  0.6380
## gender0      0.2500  1.0000
## gender1      0.2708  1.0000
## education1   0.1667  1.0000
## education2   0.1042  1.0000
## education3   0.3542  1.0000
## age          4.4490  6.6536
## income      20.3717 31.7572
##
##
```

```
## Summary of balance for matched data:
##            Means Treated Means Control SD Control Mean Diff eQQ Med
## distance          0.5369        0.5241     0.2642    0.0128  0.0186
## gender0           0.6250        0.5625     0.5123    0.0625  0.0000
## gender1           0.3750        0.4375     0.5123   -0.0625  0.0000
## education1        0.0625        0.1875     0.4031   -0.1250  0.0000
## education2        0.2500        0.2500     0.4472    0.0000  0.0000
## education3        0.3125        0.3125     0.4787    0.0000  0.0000
## age              39.3586       39.4869     7.0048   -0.1283  2.7980
## income          521.1192      505.7814    39.3828   15.3378 17.8263
##            eQQ Mean eQQ Max
## distance     0.0207  0.0344
## gender0      0.0625  1.0000
## gender1      0.0625  1.0000
## education1   0.1250  1.0000
## education2   0.0000  0.0000
## education3   0.0000  0.0000
## age          2.6644  6.3108
## income      16.6104 39.9403
##
## Percent Balance Improvement:
##            Mean Diff. eQQ Med eQQ Mean  eQQ Max
## distance      97.2845 96.7563  95.6754  94.6087
## gender0       76.0736  0.0000  75.0000   0.0000
## gender1       76.0736  0.0000  76.9231   0.0000
## education1    31.5789  0.0000  25.0000   0.0000
## education2   100.0000  0.0000 100.0000 100.0000
## education3   100.0000  0.0000 100.0000 100.0000
## age           96.9357 41.7763  40.1115   5.1526
## income        11.2784 12.5438  18.4636 -25.7677
##
## Sample sizes:
##           Control Treated
## All            52      48
## Matched        16      16
## Unmatched      36      32
## Discarded       0       0
```

Balance has improved on all variables, and is perfect on gender and education.

(d) What is the average difference in mean outcomes between treated and control groups?

```
matched_data_Q12 %>% match.data() %>%
  group_by(D) %>%
  summarize(y_obs=mean(y_obs,na.rm=T)) %>%
  arrange(-D) %>%
  mutate(diff_y_obs=y_obs-lead(y_obs))
```

```
## # A tibble: 2 x 3
##       D y_obs diff_y_obs
##   <dbl> <dbl>      <dbl>
## 1     1  41.5       14.3
## 2     0  27.2       NA
```

The mean difference in outcomes between treatment and control is now 5.54, only slightly higher than our specified value of 5.

13. One problem with this nearest neighbour matching procedure is that it is 'dumb', matching one pair, and then another, even if the distance between all paired units would be lower if the matches were switched around.

(a) Try using the 'optimal' and 'genetic' methods of *matchit* to improve your analysis.

(b) Has balanced improved?

(c) What is the average difference in mean outcomes between treated and control groups?

```
matched_data_Q13 <- matchit(D ~ gender + education + age + income, data=data.frame(d), method="optimal")
matched_data_Q13 %>% summary()
```

```
##
## Call:
## matchit(formula = D ~ gender + education + age + income, data = data.frame(d),
##     method = "optimal")
##
## Summary of balance for all data:
##            Means Treated Means Control SD Control Mean Diff eQQ Med
## distance          0.7252        0.2537     0.2553    0.4715  0.5732
## gender0           0.3542        0.6154     0.4913   -0.2612  0.0000
## gender1           0.6458        0.3846     0.4913    0.2612  0.0000
## education1        0.1250        0.3077     0.4660   -0.1827  0.0000
## education2        0.2708        0.1731     0.3820    0.0978  0.0000
## education3        0.4583        0.1154     0.3226    0.3429  0.0000
## age              41.5227       37.3373     7.0965    4.1854  4.8056
## income          507.2430      489.9554    47.0206   17.2875 20.3831
##            eQQ Mean eQQ Max
## distance     0.4790  0.6380
## gender0      0.2500  1.0000
## gender1      0.2708  1.0000
## education1   0.1667  1.0000
## education2   0.1042  1.0000
## education3   0.3542  1.0000
## age          4.4490  6.6536
## income      20.3717 31.7572
##
##
## Summary of balance for matched data:
##            Means Treated Means Control SD Control Mean Diff eQQ Med
## distance          0.7252        0.2746     0.2548    0.4506  0.5399
## gender0           0.3542        0.5833     0.4982   -0.2292  0.0000
## gender1           0.6458        0.4167     0.4982    0.2292  0.0000
## education1        0.1250        0.2708     0.4491   -0.1458  0.0000
## education2        0.2708        0.1875     0.3944    0.0833  0.0000
## education3        0.4583        0.1250     0.3342    0.3333  0.0000
## age              41.5227       37.7301     7.2287    3.7926  3.8323
## income          507.2430      495.0532    45.1915   12.1898 13.2874
##            eQQ Mean eQQ Max
## distance     0.4506  0.5997
## gender0      0.2292  1.0000
## gender1      0.2292  1.0000
## education1   0.1458  1.0000
## education2   0.0833  1.0000
## education3   0.3333  1.0000
```

```
## age          3.7941  6.0310
## income      14.0521 28.7073
##
## Percent Balance Improvement:
##            Mean Diff. eQQ Med eQQ Mean eQQ Max
## distance       4.4314  5.8121   5.9358  6.0087
## gender0       12.2699  0.0000   8.3333  0.0000
## gender1       12.2699  0.0000  15.3846  0.0000
## education1    20.1754  0.0000  12.5000  0.0000
## education2    14.7541  0.0000  20.0000  0.0000
## education3     2.8037  0.0000   5.8824  0.0000
## age            9.3851 20.2534  14.7210  9.3577
## income        29.4882 34.8117  31.0213  9.6038
##
## Sample sizes:
##           Control Treated
## All            52      48
## Matched        48      48
## Unmatched       4       0
## Discarded       0       0
```

```
matched_data_Q13 %>% match.data() %>%
  group_by(D) %>%
  summarize(y_obs=mean(y_obs,na.rm=T)) %>%
  arrange(-D) %>%
  mutate(diff_y_obs=y_obs-lead(y_obs)) %>% kable(caption="Q13(c) Optimal Matching")
```

Table 6: Q13(c) Optimal Matching

D	y_obs	diff_y_obs
1	38.37617	6.721451
0	31.65471	NA

```
matched_data_Q13_genetic %>% summary()
```

```
##
## Call:
## matchit(formula = D ~ gender + education + age + income, data = data.frame(d),
##     method = "genetic")
##
## Summary of balance for all data:
##            Means Treated Means Control SD Control Mean Diff eQQ Med
## distance          0.7252        0.2537     0.2553    0.4715  0.5732
## gender0           0.3542        0.6154     0.4913   -0.2612  0.0000
## gender1           0.6458        0.3846     0.4913    0.2612  0.0000
## education1        0.1250        0.3077     0.4660   -0.1827  0.0000
## education2        0.2708        0.1731     0.3820    0.0978  0.0000
## education3        0.4583        0.1154     0.3226    0.3429  0.0000
## age              41.5227       37.3373     7.0965    4.1854  4.8056
## income          507.2430      489.9554    47.0206   17.2875 20.3831
##            eQQ Mean eQQ Max
## distance     0.4790  0.6380
## gender0      0.2500  1.0000
```

```
## gender1      0.2708  1.0000
## education1   0.1667  1.0000
## education2   0.1042  1.0000
## education3   0.3542  1.0000
## age          4.4490  6.6536
## income      20.3717 31.7572
##
##
## Summary of balance for matched data:
##            Means Treated Means Control SD Control Mean Diff eQQ Med
## distance          0.7252        0.6817     0.2384    0.0435  0.1750
## gender0           0.3542        0.5000     0.5154   -0.1458  0.0000
## gender1           0.6458        0.5000     0.5154    0.1458  0.0000
## education1        0.1250        0.1250     0.3409    0.0000  0.0000
## education2        0.2708        0.2083     0.4186    0.0625  0.0000
## education3        0.4583        0.3333     0.4859    0.1250  0.0000
## age              41.5227       42.0394     8.6577   -0.5167  2.7535
## income          507.2430      515.3674    28.2442   -8.1244  8.0032
##            eQQ Mean eQQ Max
## distance     0.1954  0.3742
## gender0      0.2353  1.0000
## gender1      0.2353  1.0000
## education1   0.1176  1.0000
## education2   0.1176  1.0000
## education3   0.1765  1.0000
## age          2.5132  5.1380
## income      17.2458 87.4971
##
## Percent Balance Improvement:
##            Mean Diff. eQQ Med eQQ Mean   eQQ Max
## distance      90.7742 69.4606  59.2075   41.3556
## gender0       44.1718  0.0000   5.8824    0.0000
## gender1       44.1718  0.0000  13.1222    0.0000
## education1   100.0000  0.0000  29.4118    0.0000
## education2    36.0656  0.0000 -12.9412    0.0000
## education3    63.5514  0.0000  50.1730    0.0000
## age           87.6548 42.7012  43.5109   22.7789
## income        53.0044 60.7359  15.3444 -175.5189
##
## Sample sizes:
##           Control Treated
## All            52      48
## Matched        17      48
## Unmatched      35       0
## Discarded       0       0
```

```r
matched_data_Q13_genetic %>% match.data() %>%
  group_by(D) %>%
  summarize(y_obs=mean(y_obs,na.rm=T)) %>%
  arrange(-D) %>%
  mutate(diff_y_obs=y_obs-lead(y_obs)) %>% kable(caption="Q13(c) Genetic Matching")
```

Table 7: Q13(c) Genetic Matching

D	y_obs	diff_y_obs
1	38.37617	10.28634
0	28.08982	NA

Optimal matching matches 96 units, with improvements in balance on all variables. The difference in outcomes is 6.6.

Genetic matching matches 72 units (24 control and 48 treated, some control units are reused), with improvements in balance on all variables. The difference in outcomes is 7.6.

14. Try conducting matching with the Coarsened Exact Matching (`cem`) methodology. This turns continuous variables into categorical variables and then uses exact matching. Compare balance and the outcomes for treated and control groups.

```
matched_data_Q14 <- matchit(D ~ gender + education + age + income, data=data.frame(d), method="cem")
```

```
##
## Using 'treat'='1' as baseline group
```

```
matched_data_Q14 %>% summary()
```

```
##
## Call:
## matchit(formula = D ~ gender + education + age + income, data = data.frame(d),
##     method = "cem")
##
## Summary of balance for all data:
##            Means Treated Means Control SD Control Mean Diff eQQ Med
## distance          0.7252        0.2537     0.2553    0.4715  0.5732
## gender0           0.3542        0.6154     0.4913   -0.2612  0.0000
## gender1           0.6458        0.3846     0.4913    0.2612  0.0000
## education1        0.1250        0.3077     0.4660   -0.1827  0.0000
## education2        0.2708        0.1731     0.3820    0.0978  0.0000
## education3        0.4583        0.1154     0.3226    0.3429  0.0000
## age              41.5227       37.3373     7.0965    4.1854  4.8056
## income          507.2430      489.9554    47.0206   17.2875 20.3831
##            eQQ Mean eQQ Max
## distance     0.4790  0.6380
## gender0      0.2500  1.0000
## gender1      0.2708  1.0000
## education1   0.1667  1.0000
## education2   0.1042  1.0000
## education3   0.3542  1.0000
## age          4.4490  6.6536
## income      20.3717 31.7572
##
##
## Summary of balance for matched data:
##            Means Treated Means Control SD Control Mean Diff eQQ Med
## distance          0.5116        0.4655     0.2589    0.0462  0.0689
## gender0           0.6000        0.6000     0.5477    0.0000  0.0000
## gender1           0.4000        0.4000     0.5477    0.0000  0.0000
## education1        0.4000        0.4000     0.5477    0.0000  0.0000
```

```
## education2        0.2000        0.2000     0.4472    0.0000  0.0000
## education3        0.4000        0.4000     0.5477    0.0000  0.0000
## age              37.7935       37.8157     6.1742   -0.0222  1.7662
## income          510.1941      502.2960    12.5521    7.8981  5.4186
##            eQQ Mean eQQ Max
## distance     0.0619  0.1219
## gender0      0.0000  0.0000
## gender1      0.0000  0.0000
## education1   0.0000  0.0000
## education2   0.0000  0.0000
## education3   0.0000  0.0000
## age          1.4424  2.2722
## income       8.2044 21.5821
##
## Percent Balance Improvement:
##            Mean Diff. eQQ Med eQQ Mean  eQQ Max
## distance      90.2091 87.9831  87.0742  80.8886
## gender0      100.0000  0.0000 100.0000 100.0000
## gender1      100.0000  0.0000 100.0000 100.0000
## education1   100.0000  0.0000 100.0000 100.0000
## education2   100.0000  0.0000 100.0000 100.0000
## education3   100.0000  0.0000 100.0000 100.0000
## age           99.4697 63.2476  67.5793  65.8506
## income        54.3136 73.4164  59.7266  32.0402
##
## Sample sizes:
##           Control Treated
## All            52      48
## Matched         5       5
## Unmatched      47      43
## Discarded       0       0
```

```
matched_data_Q14 %>% match.data() %>%
  group_by(D) %>%
  summarize(y_obs=mean(y_obs,na.rm=T)) %>%
  arrange(-D) %>%
  mutate(diff_y_obs=y_obs-lead(y_obs)) %>% kable(caption="Q14")
```

Table 8: Q14

D	y_obs	diff_y_obs
1	38.71741	14.85253
0	23.86488	NA

Coarsened exact matching matches 42 units, with improvements in balance on all variables. The difference in outcomes is 7.6.

15. Finally, let's calculate the propensity score (the probability each unit was treated) and match treated and control units on similar values of this new propensity score.

(a) First, run a logit regression of treatment on your four confounding variables,
(b) Save the fitted values from this regression,
(c) Match on the variable for these fitted values (the probability each unit was treated) using nearest-neighbour matching and a `caliper` of 0.1 of a standard deviation.

Compare balance and the outcomes for treated and control groups.

```
d$prop_score <- d %>% glm(D ~ gender + education + age + income, data=., family="binomial") %>% fitted()

matched_data_Q15 <- matchit(D ~ prop_score, data=as.data.frame(d), caliper=0.1)

matched_data_Q15 %>% summary()
```

```
##
## Call:
## matchit(formula = D ~ prop_score, data = as.data.frame(d), caliper = 0.1)
##
## Summary of balance for all data:
##            Means Treated Means Control SD Control Mean Diff eQQ Med
## distance          0.7226        0.2561     0.2545    0.4665  0.5976
## prop_score        0.7252        0.2537     0.2553    0.4715  0.5732
##            eQQ Mean eQQ Max
## distance     0.4737  0.6996
## prop_score   0.4790  0.6380
##
##
## Summary of balance for matched data:
##            Means Treated Means Control SD Control Mean Diff eQQ Med
## distance          0.5326        0.5229     0.3007    0.0098  0.0194
## prop_score        0.5409        0.5241     0.2644    0.0168  0.0238
##            eQQ Mean eQQ Max
## distance     0.0200  0.0311
## prop_score   0.0253  0.0745
##
## Percent Balance Improvement:
##            Mean Diff. eQQ Med eQQ Mean eQQ Max
## distance      97.9081 96.7479  95.7695 95.5492
## prop_score    96.4393 95.8442  94.7160 88.3226
##
## Sample sizes:
##           Control Treated
## All            52      48
## Matched        16      16
## Unmatched      36      32
## Discarded       0       0
```

```
matched_data_Q15 %>% match.data() %>%
  group_by(D) %>%
  summarize(y_obs=mean(y_obs,na.rm=T)) %>%
  arrange(-D) %>%
  mutate(diff_y_obs=y_obs-lead(y_obs)) %>% kable(caption="Q15")
```

Table 9: Q15

D	y_obs	diff_y_obs
1	41.21112	13.15832
0	28.05280	NA

Propensity Score matching matches 58 units, with improvements in balance on the propensity score. The

difference in outcomes is 6.1.

16. The risk of using matching is that we have so many options that we can keep trying until we find a 'big' effect. So we should always be guided by a clear, measurable goal: improving balance. One possible goal is maximizing balance (ignoring considerations of sample size): Which of the matching methods you used above maximize balance on the four confounding variables?

Genetic matching seems to offer the best balance in this case.