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:

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

```r
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.

```r
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.

```r
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 06, 2020 - 3:34:18 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?

```r
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(gender))
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	6.60123
0	31.77494	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          0.5700        0.3970     0.1825    0.1730  0.1838
## gender0           0.3542        0.6154     0.4913   -0.2612  0.0000
## gender1           0.6458        0.3846     0.4913    0.2612  0.0000
## education1        0.6458        0.3846     0.4913    0.2612  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     0.1796  0.2380
## gender0      0.2500  1.0000
## gender1      0.2708  1.0000
## education1   0.2708  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          0.5700        0.4205     0.1696    0.1495  0.1560
## gender0           0.3542        0.5833     0.4982   -0.2292  0.0000
## gender1           0.6458        0.4167     0.4982    0.2292  0.0000
## education1        0.6458        0.4167     0.4982    0.2292  0.0000
## age              41.5227       37.9400     7.0099    3.5828  3.8141
## income          507.2430      494.5461    45.7796   12.6968 14.8393
##            eQQ Mean eQQ Max
## distance     0.1498  0.2140
## gender0      0.2292  1.0000
## gender1      0.2292  1.0000
## education1   0.2292  1.0000
## age          3.5969  5.5508
## income      14.5592 28.7073
##
## Percent Balance Improvement:
##            Mean Diff. eQQ Med eQQ Mean eQQ Max
## distance      13.5801 15.1081  16.5772 10.1040
## gender0       12.2699  0.0000   8.3333  0.0000
## gender1       12.2699  0.0000  15.3846  0.0000
## education1    12.2699  0.0000  15.3846  0.0000
## age           14.3988 20.6322  19.1524 16.5747
## income        26.5551 27.1977  28.5323  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 06, 2020 - 3:34:34 PM

Table 4: Q10(b)

	Dependent variable:
	y_obs
D	6.191^{***}
	(1.328)
gender1	-4.747^{***}
	(1.279)
education1	
age	0.335^{***}
	(0.088)
income	0.024^{*}
	(0.014)
Constant	9.394
	(7.890)
Observations	96
R^2	0.413
Adjusted R^2	0.387
Residual Std. Error	6.026 (df = 91)
F Statistic	16.017^{***} (df = 4; 91)
Note:	*p<0.1; **p<0.05; ***p<0.01

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.

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

(b) How many units are matched now?

```r
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        37      37
## Unmatched      15      11
## 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?

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

```
##
## Call:
## 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            0.5700        0.3970     0.1825    0.1730  0.1838
## gender0             0.3542        0.6154     0.4913   -0.2612  0.0000
## gender1             0.6458        0.3846     0.4913    0.2612  0.0000
## education1          0.6458        0.3846     0.4913    0.2612  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          0.3542        0.6154     0.4913   -0.2612  0.0000
## education1.1        0.6458        0.3846     0.4913    0.2612  0.0000
##              eQQ Mean eQQ Max
## distance       0.1796  0.2380
## gender0        0.2500  1.0000
## gender1        0.2708  1.0000
## education1     0.2708  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   0.2708  1.0000
##
##
## Summary of balance for matched data:
##              Means Treated Means Control SD Control Mean Diff eQQ Med
```

```
## distance            0.6008        0.4619     0.1617     0.1390  0.1692
## gender0             0.4595        0.4595     0.5052     0.0000  0.0000
## gender1             0.5405        0.5405     0.5052     0.0000  0.0000
## education1          0.5405        0.5405     0.5052     0.0000  0.0000
## age                43.4863       37.9500     7.4120     5.5363  5.9792
## income            518.7072      499.0243    47.0792    19.6829 21.2294
## gender0.1           0.4595        0.4595     0.5052     0.0000  0.0000
## gender1.1           0.5405        0.5405     0.5052     0.0000  0.0000
## education0          0.4595        0.4595     0.5052     0.0000  0.0000
## education1.1        0.5405        0.5405     0.5052     0.0000  0.0000
##              eQQ Mean eQQ Max
## distance       0.1394  0.2140
## gender0        0.0000  0.0000
## gender1        0.0000  0.0000
## education1     0.0000  0.0000
## age            5.5363  8.5633
## income        22.0990 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
##
## Percent Balance Improvement:
##              Mean Diff.  eQQ Med eQQ Mean  eQQ Max
## distance        19.6808   7.9604  22.4035  10.1040
## 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
## age            -32.2765 -24.4216 -24.4390 -28.7018
## income         -13.8559  -4.1521  -8.4786 -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
##
## Sample sizes:
##           Control Treated
## All            52      48
## Matched        37      37
## Unmatched      15      11
## 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	40.39162	8.870776
0	31.52084	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        30      30
## Unmatched      22      18
## 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          0.5700        0.3970     0.1825    0.1730  0.1838
## gender0           0.3542        0.6154     0.4913   -0.2612  0.0000
## gender1           0.6458        0.3846     0.4913    0.2612  0.0000
## education1        0.6458        0.3846     0.4913    0.2612  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     0.1796  0.2380
## gender0      0.2500  1.0000
## gender1      0.2708  1.0000
## education1   0.2708  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           0.4725        0.4663     0.1679    0.0063  0.0101
## gender0            0.4667        0.4667     0.5074    0.0000  0.0000
## gender1            0.5333        0.5333     0.5074    0.0000  0.0000
## education1         0.5333        0.5333     0.5074    0.0000  0.0000
## age               38.5635       38.7122     7.0838   -0.1488  2.7648
## income           499.4717      494.5371    44.8851    4.9347  6.6626
##             eQQ Mean eQQ Max
## distance      0.0102  0.0210
## gender0       0.0000  0.0000
## gender1       0.0000  0.0000
## education1    0.0000  0.0000
## age           2.6474  5.4876
## income        8.6376 23.3280
##
## Percent Balance Improvement:
##             Mean Diff. eQQ Med eQQ Mean  eQQ Max
## distance       96.3784 94.5249  94.2936  91.1940
## 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
## age            96.4459 42.4667  40.4952  17.5239
## income         71.4553 67.3129  57.5999  26.5427
##
## Sample sizes:
##            Control Treated
## All             52      48
## Matched         30      30
## Unmatched       22      18
## 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?

```r
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  37.1       5.81
## 2     0  31.3      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?

```r
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          0.5700        0.3970     0.1825    0.1730  0.1838
## gender0           0.3542        0.6154     0.4913   -0.2612  0.0000
## gender1           0.6458        0.3846     0.4913    0.2612  0.0000
## education1        0.6458        0.3846     0.4913    0.2612  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     0.1796  0.2380
## gender0      0.2500  1.0000
## gender1      0.2708  1.0000
## education1   0.2708  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          0.5700        0.4205     0.1696    0.1495  0.1560
## gender0           0.3542        0.5833     0.4982   -0.2292  0.0000
## gender1           0.6458        0.4167     0.4982    0.2292  0.0000
## education1        0.6458        0.4167     0.4982    0.2292  0.0000
## age              41.5227       37.9400     7.0099    3.5828  3.8141
## income          507.2430      494.5461    45.7796   12.6968 14.8393
##            eQQ Mean eQQ Max
## distance     0.1498  0.2140
## gender0      0.2292  1.0000
## gender1      0.2292  1.0000
## education1   0.2292  1.0000
## age          3.5969  5.5508
## income      14.5592 28.7073
##
## Percent Balance Improvement:
##            Mean Diff. eQQ Med eQQ Mean eQQ Max
## distance      13.5801 15.1081  16.5772 10.1040
## gender0       12.2699  0.0000   8.3333  0.0000
## gender1       12.2699  0.0000  15.3846  0.0000
## education1    12.2699  0.0000  15.3846  0.0000
## age           14.3988 20.6322  19.1524 16.5747
## income        26.5551 27.1977  28.5323  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	6.60123
0	31.77494	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          0.5700        0.3970     0.1825    0.1730  0.1838
## gender0           0.3542        0.6154     0.4913   -0.2612  0.0000
## gender1           0.6458        0.3846     0.4913    0.2612  0.0000
## education1        0.6458        0.3846     0.4913    0.2612  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     0.1796  0.2380
## gender0      0.2500  1.0000
## gender1      0.2708  1.0000
## education1   0.2708  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          0.5700        0.5715     0.1980   -0.0016  0.0790
## gender0           0.3542        0.3542     0.4885    0.0000  0.0000
## gender1           0.6458        0.6458     0.4885    0.0000  0.0000
## education1        0.6458        0.6458     0.4885    0.0000  0.0000
## age              41.5227       41.4719     7.6314    0.0508  3.2627
## income          507.2430      506.7568    40.8559    0.4862  5.5924
##            eQQ Mean eQQ Max
## distance     0.0845  0.1871
```

```
## gender0      0.1250  1.0000
## gender1      0.0833  1.0000
## education1   0.0833  1.0000
## age          3.2661  5.1988
## income       9.8901 43.8369
##
## Percent Balance Improvement:
##            Mean Diff. eQQ Med eQQ Mean  eQQ Max
## distance      99.1018 57.0054  52.9555  21.4143
## gender0      100.0000  0.0000  50.0000   0.0000
## gender1      100.0000  0.0000  69.2308   0.0000
## education1   100.0000  0.0000  69.2308   0.0000
## age           98.7852 32.1053  26.5876  21.8645
## income        97.1875 72.5635  51.4520 -38.0376
##
## Sample sizes:
##           Control Treated
## All            52      48
## Matched        24      48
## Unmatched      28       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	7.616294
0	30.75987	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.

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

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

```r
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           0.5700        0.3970     0.1825    0.1730  0.1838
## gender0            0.3542        0.6154     0.4913   -0.2612  0.0000
## gender1            0.6458        0.3846     0.4913    0.2612  0.0000
## education1         0.6458        0.3846     0.4913    0.2612  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      0.1796  0.2380
## gender0       0.2500  1.0000
## gender1       0.2708  1.0000
## education1    0.2708  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           0.4558        0.4497     0.1702    0.0061  0.0182
## gender0            0.4286        0.4286     0.5071    0.0000  0.0000
## gender1            0.5714        0.5714     0.5071    0.0000  0.0000
## education1         0.5714        0.5714     0.5071    0.0000  0.0000
## age               37.9352       37.5225     7.0310    0.4128  1.1920
## income           491.8847      491.5170    37.9596    0.3676  4.4257
##             eQQ Mean eQQ Max
## distance      0.0232  0.0819
## gender0       0.0952  1.0000
## gender1       0.0952  1.0000
## education1    0.0952  1.0000
## age           1.2949  3.7363
## income        6.0885 23.3280
##
## Percent Balance Improvement:
##             Mean Diff. eQQ Med eQQ Mean eQQ Max
## distance       96.4963 90.0901  87.0590 65.5879
## gender0       100.0000  0.0000  61.9048  0.0000
## gender1       100.0000  0.0000  64.8352  0.0000
## education1    100.0000  0.0000  64.8352  0.0000
## age            90.1380 75.1952  70.8938 43.8454
## income         97.8734 78.2872  70.1128 26.5427
##
## Sample sizes:
##           Control Treated
## All            52      48
## Matched        21      21
## Unmatched      31      27
## 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	37.05830	5.358786
0	31.69952	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          0.5678         0.399     0.1809    0.1688  0.1821
## prop_score        0.5700         0.397     0.1825    0.1730  0.1838
##            eQQ Mean eQQ Max
## distance     0.1751  0.2621
## prop_score   0.1796  0.2380
##
##
## Summary of balance for matched data:
##            Means Treated Means Control SD Control Mean Diff eQQ Med
## distance          0.4761        0.4705     0.1760    0.0055  0.0111
## prop_score        0.4761        0.4706     0.1727    0.0055  0.0109
##            eQQ Mean eQQ Max
## distance     0.0110  0.0197
## prop_score   0.0111  0.0282
##
## Percent Balance Improvement:
##            Mean Diff. eQQ Med eQQ Mean eQQ Max
## distance      96.7163 93.8975  93.7223 92.5008
## prop_score    96.8131 94.0857  93.8150 88.1680
##
## Sample sizes:
```

```
##           Control Treated
## All            52      48
## Matched        28      28
## Unmatched      24      20
## 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	36.89433	5.777531
0	31.11679	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.