CS112 Causal Inference Assignment

Akmarzhan Abylay

March 2020

1 Question 1

1.1 Part (A)

Before doing any matching, RUN A REGRESSION, with Y = nowtot, and the independent vars mentioned below: Dems, Repubs, Christian, age, srvlng, demvote. Treatment = hasgirls. Show the regression specification. Use the regression to estimate a treatment effect and confidence interval. Check the balance of this not-matched data set.

[ ]:

```
library(Matching) #loading libraries
library(rgenoud)

set.seed(2324)
foo <- read.csv(url("https://course-resources.minerva.kgi.edu/uploaded_files/mke/
 ↪00089202-1711/daughters.csv"))

#regression
lm1 <- lm(nowtot ~ hasgirls + Dems + Repubs + Christian + age + srvlng +
          demvote, data=foo)
summary(lm1)

# Call:
# lm(formula = nowtot ~ hasgirls + Dems + Repubs + Christian +
#     age + srvlng + demvote, data = foo)

# Residuals:
#     Min      1Q  Median      3Q     Max
# -56.028 -10.322  -1.517  11.208  69.642

# Coefficients:
#              Estimate Std. Error t value Pr(>|t|)
# (Intercept)  38.6991    18.6306   2.077 0.038390 *
# hasgirls     -0.4523     1.9036  -0.238 0.812322
# Dems         -8.1022    17.5861  -0.461 0.645238
# Repubs      -55.1069    17.6340  -3.125 0.001901 **
# Christian   -13.3961     3.7218  -3.599 0.000357 ***
# age           0.1260     0.1117   1.128 0.259938
# srvlng       -0.2251     0.1355  -1.662 0.097349 .
# demvote      87.5501     8.4847  10.319  < 2e-16 ***
# ---
# Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

# Residual standard error: 17.19 on 422 degrees of freedom
# Multiple R-squared:  0.7821,        Adjusted R-squared:  0.7784
# F-statistic: 216.3 on 7 and 422 DF,  p-value: < 2.2e-16

confint(lm1)[2,]
#     2.5 %    97.5 %
# -4.194060  3.289525

mean(confint(lm1)[2,])
# [1] -0.4522678

#balance check
mb1 <- MatchBalance(hasgirls ~ Dems + Repubs + Christian + age + srvlng +
                    demvote, data=foo)

# ***** (V1) Dems *****
# before matching:
# mean treatment........ 0.45833
# mean control.......... 0.50847
# std mean diff......... -10.047

# mean raw eQQ diff..... 0.050847
# med  raw eQQ diff..... 0
# max  raw eQQ diff..... 1

# mean eCDF diff........ 0.025071
# med  eCDF diff........ 0.025071
# max  eCDF diff........ 0.050141

# var ratio (Tr/Co)..... 0.98809
# T-test p-value........ 0.35571


# ***** (V2) Repubs *****
# before matching:
# mean treatment........ 0.53846
# mean control.......... 0.49153
# std mean diff......... 9.4

# mean raw eQQ diff..... 0.042373
# med  raw eQQ diff..... 0
# max  raw eQQ diff..... 1
# mean eCDF diff........ 0.023468
# med  eCDF diff........ 0.023468
# max  eCDF diff........ 0.046936

# var ratio (Tr/Co)..... 0.98911
# T-test p-value........ 0.3873


# ***** (V3) Christian *****
# before matching:
# mean treatment........ 0.9391
# mean control.......... 0.94915
# std mean diff......... -4.1958

# mean raw eQQ diff..... 0.016949
# med  raw eQQ diff..... 0
# max  raw eQQ diff..... 1

# mean eCDF diff........ 0.005025
# med  eCDF diff........ 0.005025
# max  eCDF diff........ 0.01005

# var ratio (Tr/Co)..... 1.1787
# T-test p-value........ 0.68107


# ***** (V4) age *****
# before matching:
# mean treatment........ 52.628
# mean control.......... 49.178
# std mean diff......... 38.385

# mean raw eQQ diff..... 3.661
# med  raw eQQ diff..... 4
# max  raw eQQ diff..... 7

# mean eCDF diff........ 0.075348
# med  eCDF diff........ 0.075538
# max  eCDF diff........ 0.17807

# var ratio (Tr/Co)..... 0.71552
# T-test p-value........ 0.0020402
# KS Bootstrap p-value.. 0.004
# KS Naive p-value...... 0.0087659
# KS Statistic.......... 0.17807
# ***** (V5) srvlng *****
# before matching:
# mean treatment........ 8.5865
# mean control.......... 8.7458
# std mean diff......... -2.1085

# mean raw eQQ diff..... 0.66949
# med  raw eQQ diff..... 0
# max  raw eQQ diff..... 5

# mean eCDF diff........ 0.017181
# med  eCDF diff........ 0.01445
# max  eCDF diff........ 0.051608

# var ratio (Tr/Co)..... 0.77347
# T-test p-value........ 0.85956
# KS Bootstrap p-value.. 0.796
# KS Naive p-value...... 0.97653
# KS Statistic.......... 0.051608


# ***** (V6) demvote *****
# before matching:
# mean treatment........ 0.49929
# mean control.......... 0.50602
# std mean diff......... -5.2747

# mean raw eQQ diff..... 0.011441
# med  raw eQQ diff..... 0.01
# max  raw eQQ diff..... 0.08

# mean eCDF diff........ 0.015928
# med  eCDF diff........ 0.010811
# max  eCDF diff........ 0.048512

# var ratio (Tr/Co)..... 1.1269
# T-test p-value........ 0.61103
# KS Bootstrap p-value.. 0.91
# KS Naive p-value...... 0.98776
# KS Statistic.......... 0.048512


# Before Matching Minimum p.value: 0.0020402
# Variable Name(s): age  Number(s): 4
```

After running a linear regression, the coefficient for the treatment was around -0.4523. More specifically, the treatment effect estimate was around -0.4522678, while the 95% confidence interval entailed number from -4.194060 to 3.289525.

Before matching, some variables had pretty good balance (i.e., KS Bootstrap p-value for srvlng - 0.796, KS Bootstrap p-value for demvote - 0.91), while others didn't (i.e., KS Bootstrap p-value for age - 0.004).

Then, do genetic matching, using the draft code below. Use the same variables as in the regression above.

[ ]:

```
set.seed(2324)

#multivariate matching
X1 = cbind(foo$Dems, foo$Repubs, foo$Christian, foo$age, foo$srvlng, foo$demvote)
#finding weights
genout <- GenMatch(Tr=foo$hasgirls, X=X1, M=1, pop.size=20,
                   max.generations=10, wait.generations=25, nboots=250)

#matching
m1  <- Match(Tr=foo$hasgirls, X=X1, M=1, Weight.matrix = genout)
summary(m1)

# Estimate...  0
# SE.........  0
# T-stat.....  NaN
# p.val......  NA

# Original number of observations..............  430
# Original number of treated obs...............  312
# Matched number of observations...............  312
# Matched number of observations  (unweighted).  312

mb2 <- MatchBalance(hasgirls ~ Dems + Repubs + Christian + age +
                    srvlng + demvote, data = foo, match.out = m1)

# ***** (V1) Dems *****
#                        Before Matching                  After Matching
# mean treatment........    0.45833                      0.45833
# mean control..........    0.50847                      0.46154
# std mean diff.........    -10.047                     -0.64223

# mean raw eQQ diff.....   0.050847                    0.0032051
# med  raw eQQ diff.....          0                            0
# max  raw eQQ diff.....          1                            1

# mean eCDF diff........   0.025071                    0.0016026
# med  eCDF diff........   0.025071                    0.0016026
# max  eCDF diff........   0.050141                    0.0032051
# var ratio (Tr/Co).....    0.98809                      0.99897
# T-test p-value........    0.35571                      0.31731


# ***** (V2) Repubs *****
#                        Before Matching                   After Matching
# mean treatment........    0.53846                      0.53846
# mean control..........    0.49153                      0.53846
# std mean diff.........        9.4                            0

# mean raw eQQ diff.....   0.042373                            0
# med  raw eQQ diff.....          0                            0
# max  raw eQQ diff.....          1                            0

# mean eCDF diff........   0.023468                            0
# med  eCDF diff........   0.023468                            0
# max  eCDF diff........   0.046936                            0

# var ratio (Tr/Co).....    0.98911                            1
# T-test p-value........     0.3873                            1


# ***** (V3) Christian *****
#                        Before Matching                   After Matching
# mean treatment........     0.9391                       0.9391
# mean control..........    0.94915                       0.9391
# std mean diff.........    -4.1958                            0

# mean raw eQQ diff.....   0.016949                            0
# med  raw eQQ diff.....          0                            0
# max  raw eQQ diff.....          1                            0

# mean eCDF diff........   0.005025                            0
# med  eCDF diff........   0.005025                            0
# max  eCDF diff........    0.01005                            0

# var ratio (Tr/Co).....     1.1787                            1
# T-test p-value........    0.68107                            1


# ***** (V4) age *****
#                        Before Matching                   After Matching
# mean treatment........     52.628                       52.628
# mean control..........     49.178                       52.526
# std mean diff.........     38.385                       1.1411
# mean raw eQQ diff.....      3.661                       0.5641
# med  raw eQQ diff.....          4                            1
# max  raw eQQ diff.....          7                            4

# mean eCDF diff........   0.075348                     0.012251
# med  eCDF diff........   0.075538                    0.0096154
# max  eCDF diff........    0.17807                     0.038462

# var ratio (Tr/Co).....    0.71552                      0.99199
# T-test p-value........  0.0020402                       0.5061
# KS Bootstrap p-value..      0.006                        0.892
# KS Naive p-value......  0.0087659                      0.97513
# KS Statistic..........    0.17807                     0.038462


# ***** (V5) srvlng *****
#                        Before Matching                   After Matching
# mean treatment........     8.5865                       8.5865
# mean control..........     8.7458                       8.7019
# std mean diff.........    -2.1085                      -1.5279

# mean raw eQQ diff.....    0.66949                      0.49359
# med  raw eQQ diff.....          0                            0
# max  raw eQQ diff.....          5                            9

# mean eCDF diff........   0.017181                     0.012944
# med  eCDF diff........    0.01445                     0.011218
# max  eCDF diff........   0.051608                     0.051282

# var ratio (Tr/Co).....    0.77347                       0.9505
# T-test p-value........    0.85956                       0.5278
# KS Bootstrap p-value..      0.786                         0.56
# KS Naive p-value......    0.97653                      0.80655
# KS Statistic..........   0.051608                     0.051282


# ***** (V6) demvote *****
#                        Before Matching                   After Matching
# mean treatment........    0.49929                      0.49929
# mean control..........    0.50602                      0.49933
# std mean diff.........    -5.2747                    -0.025149

# mean raw eQQ diff.....   0.011441                    0.0091346
# med  raw eQQ diff.....       0.01                         0.01
# max  raw eQQ diff.....       0.08                         0.08

# mean eCDF diff........   0.015928                     0.013838
# med  eCDF diff........   0.010811                     0.0096154
# max  eCDF diff........   0.048512                      0.044872

# var ratio (Tr/Co).....     1.1269                        1.128
# T-test p-value........    0.61103                      0.98889
# KS Bootstrap p-value..      0.928                        0.778
# KS Naive p-value......    0.98776                      0.91194
# KS Statistic..........   0.048512                      0.044872


# Before Matching Minimum p.value: 0.0020402
# Variable Name(s): age  Number(s): 4

# After Matching Minimum p.value: 0.31731
# Variable Name(s): Dems  Number(s): 1
```

Summarize (in 5-15 sentences) the genetic matching procedure and results, including what you matched on, what you balanced on, and what your balance results were. Provide a link to your code, and provide output for Match() and MatchBalance() in the body of your submission.

As we could see from the output given in the comments above, since we didn't include Y, we couldn't obtain treatment effect estimates. However, when matching, we don't actually need any information about the outcomes, so it is fine. The genetic matching here was used to find optimal balance and determine the weights of each given covariate. We know that the scale, which we are using for each of the variables, greatly affects which units we are going to match, so genetic matching helps us define the most appropriate weights (scale) for each variable to find the best matches. The results are then fed into the `Match` function, which is used to obtain causal estimates. `GenMatch` basically automates the process of finding suitable matches. Its main advantage is that`GenMatch` directly optimizes covariate balance instead of manually checking covariate balance.

Here, I matched on the 6 variables used for regression above (excluding the treatment variable). I balanced on the same variables, and the results were satisfactory. As we saw before, the worst balanced variable had a p-value of 0.0020402, but after matching, the worst p-value became 0.31731. We want higher values because we want the groups to be as similar as possible (i.e., no difference).

If you obtain high balance, consider rerunning with M = 2 or 3… If/when you are satisfied by balance achieved, then rerun Match() with Y included and obtain the treatment effect estimate, the standard error, and the confidence interval.

Below, I tried matching, including the Y variable, which gave us the treatment effect estimates. I also tried matching two-to-one and three-to-one, but the results of the Match Balance weren't much better. I also included the outcome for M=2 below treatment estimate when M=2 below.

The treatment effect is 1.081731, while the standard error is 2.206799, and subsequently, the 95% confidence interval is 1.081731 ± 2.206799 × 1.96, which is from around -3.244 to 5.407. The confidence interval is a little similar to the one obtained by regression, although the estimate in a linear regression model was negative, while the matching estimate is positive, which suggests that those who have daughters vote more liberally since matching is a more reliable method.

[ ]:

```
m2  <- Match(Y=foo$nowtot, Tr=foo$hasgirls, X=X1, M=1, Weight.matrix = genout)
summary(m2)
# Estimate...  1.0817
# AI SE......  2.2068
# T-stat.....  0.49018
# p.val......  0.62401

# Original number of observations..............  430
# Original number of treated obs...............  312
# Matched number of observations...............  312
# Matched number of observations  (unweighted).  313

m2$est
#           [,1]
# [1,] 1.081731
m2$se
# [1] 2.206799

####### for M=2
genout <- GenMatch(Tr=foo$hasgirls, X=X1, M=2,
                   pop.size=20, max.generations=10, wait.generations=25, nboots=250)
m21  <- Match(Y=foo$nowtot, Tr=foo$hasgirls, X=X1, M=2, Weight.matrix = genout)
summary(m21)
# Estimate...  1.3141
# AI SE......  2.1231
# T-stat.....  0.61896
# p.val......  0.53594

# Original number of observations..............  430
# Original number of treated obs...............  312
# Matched number of observations...............  312
# Matched number of observations  (unweighted).  626

m2$est
#           [,1]
# [1,] 1.314103
m2$se
# [1] 2.123072
```

1.2 PART (B)

Repeat everything you've done for Part (A), including the regression, genetic algorithm, code, output, and 5-15 sentences, EXCEPT this time change the definition of treatment to cover 2

girls, and change the definition of control to cover 2 boys. Exclude all observations that don't meet these requirements. Be sure to explain (in a sentence or two) what you're doing with your new treatment and control definitions. Do your new definitions change anything?

We are testing for a more extreme case, since now we only count people who have 2 daughters as treated and people who have 2 boys as controlled (i.e., they only have 2 children of the same gender and no children of the other gender). We are changing the definitions of the control and treatment, but the overall idea and hypothesis are still the same, and this new assignment mechanism still works. This left us with a total of 59 units, with 31 treated units and 28 control units. Definitely much fewer observations now, but the new definitions just change the extremeness of the treatment effect, but the overall trend is still the same.

[ ]:

```
set.seed(2324)
#extracting data
foo.treat <- foo[which(foo$ngirls == 2 & foo$nboys == 0), ]
foo.ctrl <- foo[which(foo$nboys == 2 & foo$ngirls == 0), ]

#assignment
foo.new <- rbind(foo.ctrl, foo.treat)
foo.new$treat1 <- ifelse(foo.new$ngirls==2, 1, 0)
foo.new$treat1

sum(foo.new$treat1)
# 31 treated units with 2 girls and no boys

lm2 <- lm(nowtot ~ treat1 + Dems + Repubs + Christian + age + srvlng +
          demvote, data=foo.new)
summary(lm2)
# Call:
# lm(formula = nowtot ~ treat1 + Dems + Repubs + Christian + age +
#     srvlng + demvote, data = foo.new)

# Residuals:
#     Min      1Q  Median      3Q     Max
# -54.665  -4.266   0.512   8.109  24.145

# Coefficients: (1 not defined because of singularities)
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept) -25.5073    18.5535  -1.375 0.175089
# treat1       13.2027     3.7620   3.510 0.000936 ***
# Dems         55.9167     4.8958  11.421 8.67e-16 ***
# Repubs            NA         NA      NA       NA
# Christian     1.1113     8.7391   0.127 0.899298
# age          -0.0159     0.2662  -0.060 0.952615
# srvlng        0.1017     0.2846   0.357 0.722309
# demvote      68.0931    18.9100   3.601 0.000708 ***
# ---
# Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

# Residual standard error: 13.42 on 52 degrees of freedom
# Multiple R-squared:  0.8921,        Adjusted R-squared:  0.8797
# F-statistic: 71.68 on 6 and 52 DF,  p-value: < 2.2e-16

confint(lm2)[2,]
#     2.5 %    97.5 %
#  5.653734 20.751634
mean(confint(lm2)[2,])
# [1] 13.20268


mb4 <- MatchBalance(treat1 ~ Dems + Repubs + Christian + age + srvlng +
                    demvote, data=foo.new)

# ***** (V1) Dems *****
# before matching:
# mean treatment........ 0.64516
# mean control.......... 0.42857
# std mean diff......... 44.532

# mean raw eQQ diff..... 0.21429
# med  raw eQQ diff..... 0
# max  raw eQQ diff..... 1

# mean eCDF diff........ 0.10829
# med  eCDF diff........ 0.10829
# max  eCDF diff........ 0.21659

# var ratio (Tr/Co)..... 0.93145
# T-test p-value........ 0.099329


# ***** (V2) Repubs *****
# before matching:
# mean treatment........ 0.35484
# mean control.......... 0.57143
# std mean diff......... -44.532

# mean raw eQQ diff..... 0.21429
# med  raw eQQ diff..... 0
# max  raw eQQ diff..... 1

# mean eCDF diff........ 0.10829
# med  eCDF diff........ 0.10829
# max  eCDF diff........ 0.21659
# var ratio (Tr/Co)..... 0.93145
# T-test p-value........ 0.099329


# ***** (V3) Christian *****
# before matching:
# mean treatment........ 0.90323
# mean control.......... 1
# std mean diff......... -32.2

# mean raw eQQ diff..... 0.10714
# med  raw eQQ diff..... 0
# max  raw eQQ diff..... 1

# mean eCDF diff........ 0.048387
# med  eCDF diff........ 0.048387
# max  eCDF diff........ 0.096774

# var ratio (Tr/Co)..... Inf
# T-test p-value........ 0.083087


# ***** (V4) age *****
# before matching:
# mean treatment........ 48.226
# mean control.......... 49.857
# std mean diff......... -19.026

# mean raw eQQ diff..... 2.4643
# med  raw eQQ diff..... 2.5
# max  raw eQQ diff..... 5

# mean eCDF diff........ 0.061382
# med  eCDF diff........ 0.051843
# max  eCDF diff........ 0.13479

# var ratio (Tr/Co)..... 1.0028
# T-test p-value........ 0.46822
# KS Bootstrap p-value.. 0.808
# KS Naive p-value...... 0.95201
# KS Statistic.......... 0.13479


# ***** (V5) srvlng *****
# before matching:
# mean treatment........ 7.5484
# mean control.......... 9.6071
# std mean diff......... -28.926

# mean raw eQQ diff..... 2.4286
# med  raw eQQ diff..... 1
# max  raw eQQ diff..... 10

# mean eCDF diff........ 0.066172
# med  eCDF diff........ 0.05818
# max  eCDF diff........ 0.17051

# var ratio (Tr/Co)..... 0.60661
# T-test p-value........ 0.34249
# KS Bootstrap p-value.. 0.484
# KS Naive p-value...... 0.7858
# KS Statistic.......... 0.17051


# ***** (V6) demvote *****
# before matching:
# mean treatment........ 0.52677
# mean control.......... 0.50714
# std mean diff......... 15.554

# mean raw eQQ diff..... 0.05
# med  raw eQQ diff..... 0.05
# max  raw eQQ diff..... 0.12

# mean eCDF diff........ 0.10108
# med  eCDF diff........ 0.066244
# max  eCDF diff........ 0.29493

# var ratio (Tr/Co)..... 0.88501
# T-test p-value........ 0.56612
# KS Bootstrap p-value.. 0.09
# KS Naive p-value...... 0.15463
# KS Statistic.......... 0.29493


# Before Matching Minimum p.value: 0.083087
# Variable Name(s): Christian  Number(s): 3
```

After running a linear regression, the coefficient for the treatment was around 13.2027. More specifically, the treatment effect estimate was around 13.20268, while the 95% confidence interval entailed number from 5.653734 to 20.751634.

Before matching, one variable didn't have a very good balance (i.e., KS Bootstrap p-value for demvote - 0.09), while some did have a satisfactory balance (i.e., KS Bootstrap p-value for age -

0.808, KS Bootstrap p-value for srvlng - 0.484). We could see that the treatment effect estimate in this linear regression with new treatment indicator is positive rather than negative, which suggests that the linear regression is susceptible to bias and outliers, and to estimate the treatment effect, we should use better methods, such as matching.

[ ]:

```
X2 = cbind(foo.new$Dems, foo.new$Repubs, foo.new$Christian, foo.new$age,
           foo.new$srvlng, foo.new$demvote)
genout <- GenMatch(Tr=foo.new$treat1, X=X2, M=1, pop.size=20,
                   max.generations=10, wait.generations=25, nboots=250)

m3  <- Match(Tr=foo.new$treat1, X=X2, M=1, Weight.matrix = genout)
summary(m3)
# Estimate...  0
# SE.........  0
# T-stat.....  NaN
# p.val......  NA

# Original number of observations..............  59
# Original number of treated obs...............  31
# Matched number of observations...............  31
# Matched number of observations  (unweighted).  31

mb5 <- MatchBalance(treat1 ~ Dems + Repubs + Christian + age + srvlng +
                    demvote, data = foo.new, match.out = m3)

# ***** (V1) Dems *****
#                        Before Matching                   After Matching
# mean treatment........    0.64516                          0.64516
# mean control..........    0.42857                           0.6129
# std mean diff.........     44.532                           6.6324

# mean raw eQQ diff.....    0.21429                         0.032258
# med  raw eQQ diff.....          0                                0
# max  raw eQQ diff.....          1                                1

# mean eCDF diff........    0.10829                         0.016129
# med  eCDF diff........    0.10829                         0.016129
# max  eCDF diff........    0.21659                         0.032258

# var ratio (Tr/Co).....    0.93145                          0.96491
# T-test p-value........   0.099329                          0.31751


# ***** (V2) Repubs *****
#                        Before Matching                   After Matching
# mean treatment........    0.35484                          0.35484
# mean control..........    0.57143                           0.3871
# std mean diff.........    -44.532                       -6.6324

# mean raw eQQ diff.....    0.21429                      0.032258
# med  raw eQQ diff.....          0                             0
# max  raw eQQ diff.....          1                             1

# mean eCDF diff........    0.10829                      0.016129
# med  eCDF diff........    0.10829                      0.016129
# max  eCDF diff........    0.21659                      0.032258

# var ratio (Tr/Co).....    0.93145                       0.96491
# T-test p-value........   0.099329                       0.31751


# ***** (V3) Christian *****
#                        Before Matching                    After Matching
# mean treatment........    0.90323                       0.90323
# mean control..........          1                             1
# std mean diff.........      -32.2                         -32.2

# mean raw eQQ diff.....    0.10714                      0.096774
# med  raw eQQ diff.....          0                             0
# max  raw eQQ diff.....          1                             1

# mean eCDF diff........   0.048387                      0.048387
# med  eCDF diff........   0.048387                      0.048387
# max  eCDF diff........   0.096774                      0.096774

# var ratio (Tr/Co).....        Inf                           Inf
# T-test p-value........   0.083087                      0.078364


# ***** (V4) age *****
#                        Before Matching                    After Matching
# mean treatment........     48.226                        48.226
# mean control..........     49.857                            48
# std mean diff.........    -19.026                        2.6336

# mean raw eQQ diff.....     2.4643                       0.80645
# med  raw eQQ diff.....        2.5                             1
# max  raw eQQ diff.....          5                             6

# mean eCDF diff........   0.061382                      0.026882
# med  eCDF diff........   0.051843                      0.032258
# max  eCDF diff........    0.13479                      0.064516

# var ratio (Tr/Co).....     1.0028                        1.1844
# T-test p-value........    0.46822                      0.65424
# KS Bootstrap p-value..      0.788                            1
# KS Naive p-value......    0.95201                            1
# KS Statistic..........    0.13479                     0.064516


# ***** (V5) srvlng *****
#                        Before Matching                   After Matching
# mean treatment........     7.5484                       7.5484
# mean control..........     9.6071                       8.0323
# std mean diff.........    -28.926                      -6.7985

# mean raw eQQ diff.....     2.4286                        1.129
# med  raw eQQ diff.....          1                            0
# max  raw eQQ diff.....         10                            8

# mean eCDF diff........   0.066172                     0.032258
# med  eCDF diff........    0.05818                     0.032258
# max  eCDF diff........    0.17051                     0.096774

# var ratio (Tr/Co).....    0.60661                      0.77575
# T-test p-value........    0.34249                      0.57759
# KS Bootstrap p-value..      0.524                        0.942
# KS Naive p-value......     0.7858                      0.99866
# KS Statistic..........    0.17051                     0.096774


# ***** (V6) demvote *****
#                        Before Matching                   After Matching
# mean treatment........    0.52677                      0.52677
# mean control..........    0.50714                         0.52
# std mean diff.........     15.554                       5.3674

# mean raw eQQ diff.....       0.05                     0.044839
# med  raw eQQ diff.....       0.05                         0.05
# max  raw eQQ diff.....       0.12                         0.09

# mean eCDF diff........    0.10108                     0.087702
# med  eCDF diff........   0.066244                     0.080645
# max  eCDF diff........    0.29493                      0.22581

# var ratio (Tr/Co).....    0.88501                       1.0549
# T-test p-value........    0.56612                      0.56739
# KS Bootstrap p-value..      0.096                        0.332
# KS Naive p-value......    0.15463                       0.4081
# KS Statistic..........    0.29493                      0.22581
# Before Matching Minimum p.value: 0.083087
# Variable Name(s): Christian  Number(s): 3

# After Matching Minimum p.value: 0.078364
# Variable Name(s): Christian  Number(s): 3
```

Here, I matched on the 6 variables used for regression above (excluding the treatment variable). I balanced on the same variables, and the results were satisfactory. As we saw before, the worst balanced variable had a p-value of 0.083087, but after matching, the worst p-value became even worse 0.078364. This might have happened because of the lack of observations as we only had 59 observations, and it was hard to match appropriately. However, the KS Bootstrapped p-value for srvlng, age and demvote became better (i.e., 0.788 to 1 for age, 0.524 to 0.942 for srvlng and 0.096 to 0.332 for demvote), which could suggest that matching helped to balance out other variables at the expense of one.

Below, I tried matching, including the Y variable, which gave us the treatment effect estimates. The treatment effect is 15.16129, while the standard error is 4.847144, and subsequently, the 95% confidence interval is 15.16129 ± 4.847144 × 1.96, which is from around 5.6609 to 24.6617. This is similar to the one obtained by regression, and the estimate in both regression and matching was large and positive, which suggests that those who have 2 daughters vote more liberally than those who have 2 boys. The treatment effect here was larger than in the previous treatment assignment. This might further support the hypothesis that those who have daughters vote more liberally than those who have boys, which is shown to the extreme through this treatment assignment.

[ ]:

```
m4  <- Match(Y=foo.new$nowtot, Tr=foo.new$treat1, X=X2, M=1, Weight.matrix = genout)
summary(m4)
# Estimate...  15.161
# AI SE......  4.8471
# T-stat.....  3.1279
# p.val......  0.0017607

# Original number of observations..............  59
# Original number of treated obs...............  31
# Matched number of observations...............  31
# Matched number of observations  (unweighted).  31

m4$est
#          [,1]
# [1,] 15.16129
m4$se
# [1] 4.847144
```

Do NOT, under any circumstances, match or try to balance on “hasgirls”. I don’t think it’s wise to match or balance on “totchi” for a related reason. (What is the reason?)

The whole point of doing matching is to approximating an RCT and matching two similar units but with different assignments. Since `hasgirls` is the treatment in our case, and it is highly related to the second type of assignment too, matching on the treatment would make no sense, since we would end up matching units with the same treatment (or its absence) and it wouldn't give any reliable ground for causal inference. `totchi` in this case is again correlated with the treatment, and it wouldn't help us in matching, but instead could bias the matching and the results.

2 QUESTION 2: "Business Lending in Indonesia"

Your task is to change the code, specifically the matching algorithm. You won't be changing all the code.

Your new matching algorithm should differ from the one in the paper in 2 ways:

1. instead of matching on postal code (postal code is not important), match on district code (exact matches on first 2 digits of district code):
2. match on date_of_birth within a 1-year (365-day) caliper

2.1 Part (A)

In the body of your submitted assignment (not in a link), submit the code you added/modified AND provide your Match() and MatchBalance() output (which must now also additionally show balance on the district code variable).

[1]:

```
#copy-pasted all the needed code from Prof. Diamond's memo
#below is the part I changed

#made a new column with the first two digits from the district code, since we␣
 ↪need exact matching on these
foo$district_code_new <- substring(as.character(foo$district_code), 1, 2)

#finding the sd of the date of births
caliper1 <- 365/sd(foo$R_date_of_birth)
# 0.000305872


X <- cbind(foo$R_date_of_birth, foo$gender, foo$marital_status,
           foo$education, foo$occupation, foo$district_code_new,
           foo$worker, foo$capital, foo$credit_proposal,
           foo$worker_transformed, foo$capital_transformed,␣
 ↪foo$credit_proposal_transformed,
           foo$missing_date_of_birth,
           foo$NA_capital,
           foo$NA_credit_proposal)
```

```
#this is needed because for some reason all the entries are characters, so I␣
 ↪needed to
#turn them into numeric so that genetic matching can be performed
X <- mapply(X, FUN=as.numeric)
X <- matrix(data=X, ncol=15, nrow=76770)

Tr <- foo$treat
BalanceMat <- X

#generic matching, exact matches on all the same variables and district code␣
 ↪(excluding postal code)
#matching using calipers for birth date

genout <- GenMatch(Tr=Tr, X=X, BalanceMatrix=BalanceMat, estimand="ATT", M=1,
                   pop.size=16, max.generations=5, wait.generations=3,
                   caliper = c(caliper1, 1e20, 1e20,
                                    1e20, 1e20, 1e20, 1e20, 1e20,
                                    1e20, 1e20, 1e20, 1e20, 1e20, 1e20, 1e20),
                   exact = c(FALSE, TRUE, TRUE,
                             TRUE, TRUE, TRUE,
                             FALSE, FALSE, FALSE,
                             TRUE, TRUE, TRUE,
                             TRUE, TRUE, TRUE))

#matching using the weight matrix from the genetic matching function
mout <- Match(Tr=Tr, X=X, estimand="ATT", M=1,
              exact = c(FALSE, TRUE, TRUE,
                        TRUE, TRUE, TRUE,
                        FALSE, FALSE, FALSE,
                        TRUE, TRUE, TRUE,
                        TRUE, TRUE, TRUE),
             caliper = c(caliper1, 1e20, 1e20,
                                    1e20, 1e20, 1e20, 1e20, 1e20,
                                    1e20, 1e20, 1e20, 1e20, 1e20, 1e20, 1e20),
              Weight.matrix = genout)



summary(mout)
# Estimate...  0
# SE.........  0
# T-stat.....  NaN
# p.val......  NA

# Original number of observations..............  76770
# Original number of treated obs...............  15872
# Matched number of observations...............  11429
# Matched number of observations  (unweighted).  149920

# Number of obs dropped by 'exact' or 'caliper'  4443

#testing the balance
mb <- MatchBalance(foo$treat~
                     foo$R_date_of_birth + foo$gender + foo$marital_status +
                     foo$education + foo$occupation + foo$district_code_new +
                     foo$worker + foo$capital + foo$credit_proposal +
                     foo$worker_transformed + foo$capital_transformed +
                     foo$credit_proposal_transformed + foo$missing_date_of_birth␣
 ↪+
                     foo$NA_capital + foo$NA_credit_proposal,
                     match.out=mout, nboots=500)



# ***** (V1) foo$R_date_of_birth *****
#                        Before Matching                       After Matching
# mean treatment........     -139148                         -173571
# mean control..........     -139530                         -173574
# std mean diff.........    0.031967                      0.00027479

# mean raw eQQ diff.....      1726.3                          14.401
# med  raw eQQ diff.....         969                               9
# max  raw eQQ diff.....     9346761                             310

# mean eCDF diff........    0.054966                      0.00085783
# med  eCDF diff........    0.053922                      0.00055363
# max  eCDF diff........     0.11401                       0.0050227

# var ratio (Tr/Co).....      1.0053                               1
# T-test p-value........     0.97137                         0.03946
# KS Bootstrap p-value.. < 2.22e-16                            0.042
# KS Naive p-value...... < 2.22e-16                          0.04555
# KS Statistic..........     0.11401                       0.0050227


# ***** (V2) foo$genderLAKI-LAKI *****
#                        Before Matching                       After Matching
# mean treatment........     0.64579                         0.68055
# mean control..........     0.62621                         0.68055
# std mean diff.........      4.0938                               0

# mean raw eQQ diff.....    0.019594                               0
# med  raw eQQ diff.....           0                               0
# max  raw eQQ diff.....           1                               0
# mean eCDF diff........  0.0097901                            0
# med  eCDF diff........  0.0097901                            0
# max  eCDF diff........    0.01958                            0

# var ratio (Tr/Co).....    0.97729                            1
# T-test p-value........ 4.6148e-06                            1


# ***** (V3) foo$genderPEREMPUAN *****
#                        Before Matching                  After Matching
# mean treatment........    0.35364                      0.31945
# mean control..........    0.37318                      0.31945
# std mean diff.........    -4.0868                            0

# mean raw eQQ diff.....   0.019531                            0
# med  raw eQQ diff.....          0                            0
# max  raw eQQ diff.....          1                            0

# mean eCDF diff........  0.0097699                            0
# med  eCDF diff........  0.0097699                            0
# max  eCDF diff........    0.01954                            0

# var ratio (Tr/Co).....    0.97723                            1
# T-test p-value........ 4.7914e-06                            1


# ***** (V4) foo$marital_statusBELUM KAWIN *****
#                        Before Matching                  After Matching
# mean treatment........   0.031943                      0.01085
# mean control..........    0.02647                      0.01085
# std mean diff.........      3.112                            0

# mean raw eQQ diff.....  0.0054814                            0
# med  raw eQQ diff.....          0                            0
# max  raw eQQ diff.....          1                            0

# mean eCDF diff........  0.0027363                            0
# med  eCDF diff........  0.0027363                            0
# max  eCDF diff........  0.0054726                            0

# var ratio (Tr/Co).....        1.2                            1
# T-test p-value........ 0.00038068                            1


# ***** (V5) foo$marital_statusKAWIN *****
#                        Before Matching                  After Matching
# mean treatment........    0.93473                      0.96264
# mean control..........    0.92854                      0.96264
# std mean diff.........     2.5066                            0

# mean raw eQQ diff.....  0.0062374                            0
# med  raw eQQ diff.....          0                            0
# max  raw eQQ diff.....          1                            0

# mean eCDF diff........  0.0030958                            0
# med  eCDF diff........  0.0030958                            0
# max  eCDF diff........  0.0061916                            0

# var ratio (Tr/Co).....    0.91949                            1
# T-test p-value........  0.0053164                            1


# ***** (V6) foo$educationDIPLOMA *****
#                        Before Matching                  After Matching
# mean treatment........   0.019153                    0.0069122
# mean control..........   0.013137                    0.0069122
# std mean diff.........     4.3894                            0

# mean raw eQQ diff.....  0.0059854                            0
# med  raw eQQ diff.....          0                            0
# max  raw eQQ diff.....          1                            0

# mean eCDF diff........  0.0030083                            0
# med  eCDF diff........  0.0030083                            0
# max  eCDF diff........  0.0060165                            0

# var ratio (Tr/Co).....     1.4492                            1
# T-test p-value........ 3.5894e-07                            1


# ***** (V7) foo$educationLAINNYA *****
#                        Before Matching                  After Matching
# mean treatment........   0.038495                     0.034386
# mean control..........   0.034155                     0.034386
# std mean diff.........     2.2558                            0

# mean raw eQQ diff.....  0.0043473                            0
# med  raw eQQ diff.....          0                            0
# max  raw eQQ diff.....          1                            0

# mean eCDF diff........    0.00217                            0
# med  eCDF diff........    0.00217                            0
# max  eCDF diff........    0.00434                            0
# var ratio (Tr/Co).....     1.1221                            1
# T-test p-value........   0.010471                            1


# ***** (V8) foo$educationSARJANA *****
#                        Before Matching                  After Matching
# mean treatment........    0.02936                    0.0077872
# mean control..........   0.027144                    0.0077872
# std mean diff.........     1.3127                            0

# mean raw eQQ diff.....  0.0022051                            0
# med  raw eQQ diff.....          0                            0
# max  raw eQQ diff.....          1                            0

# mean eCDF diff........  0.0011081                            0
# med  eCDF diff........  0.0011081                            0
# max  eCDF diff........  0.0022161                            0

# var ratio (Tr/Co).....     1.0792                            1
# T-test p-value........    0.13775                            1


# ***** (V9) foo$educationSD *****
#                        Before Matching                  After Matching
# mean treatment........    0.20319                        0.205
# mean control..........    0.24462                        0.205
# std mean diff.........    -10.297                            0

# mean raw eQQ diff.....   0.041457                            0
# med  raw eQQ diff.....          0                            0
# max  raw eQQ diff.....          1                            0

# mean eCDF diff........   0.020717                            0
# med  eCDF diff........   0.020717                            0
# max  eCDF diff........   0.041434                            0

# var ratio (Tr/Co).....    0.87622                            1
# T-test p-value........ < 2.22e-16                            1


# ***** (V10) foo$educationSMP *****
#                        Before Matching                  After Matching
# mean treatment........    0.18214                      0.17613
# mean control..........    0.18634                      0.17613
# std mean diff.........    -1.0881                            0
# mean raw eQQ diff.....  0.0042213                            0
# med  raw eQQ diff.....          0                            0
# max  raw eQQ diff.....          1                            0

# mean eCDF diff........  0.0020999                            0
# med  eCDF diff........  0.0020999                            0
# max  eCDF diff........  0.0041997                            0

# var ratio (Tr/Co).....    0.98255                            1
# T-test p-value........    0.22298                            1


# ***** (V11) foo$educationSMU *****
#                         Before Matching                   After Matching
# mean treatment........    0.52413                      0.56943
# mean control..........    0.49299                      0.56943
# std mean diff.........     6.2355                            0

# mean raw eQQ diff.....   0.031124                            0
# med  raw eQQ diff.....          0                            0
# max  raw eQQ diff.....          1                            0

# mean eCDF diff........   0.015571                            0
# med  eCDF diff........   0.015571                            0
# max  eCDF diff........   0.031142                            0

# var ratio (Tr/Co).....    0.99791                            1
# T-test p-value........  2.716e-12                            1


# ***** (V12) foo$occupationLAIN-LAIN/BADAN USAHA *****
#                         Before Matching                   After Matching
# mean treatment........    0.07718                      0.07551
# mean control..........   0.063927                      0.07551
# std mean diff.........     4.9659                            0

# mean raw eQQ diff.....   0.013231                            0
# med  raw eQQ diff.....          0                            0
# max  raw eQQ diff.....          1                            0

# mean eCDF diff........  0.0066267                            0
# med  eCDF diff........  0.0066267                            0
# max  eCDF diff........   0.013253                            0

# var ratio (Tr/Co).....     1.1903                            1
# T-test p-value........ 1.4739e-08                            1
# ***** (V13) foo$occupationNELAYAN *****
#                        Before Matching                   After Matching
# mean treatment........   0.016381                     0.013212
# mean control..........   0.013662                     0.013212
# std mean diff.........     2.1418                            0

# mean raw eQQ diff..... 0.0027092                            0
# med  raw eQQ diff.....         0                            0
# max  raw eQQ diff.....         1                            0

# mean eCDF diff........ 0.0013594                            0
# med  eCDF diff........ 0.0013594                            0
# max  eCDF diff........ 0.0027189                            0

# var ratio (Tr/Co).....     1.1958                            1
# T-test p-value........   0.014491                            1


# ***** (V14) foo$occupationPEDAGANG *****
#                        Before Matching                   After Matching
# mean treatment........    0.20073                      0.19276
# mean control..........    0.20733                      0.19276
# std mean diff.........    -1.6476                            0

# mean raw eQQ diff..... 0.0066154                            0
# med  raw eQQ diff.....         0                            0
# max  raw eQQ diff.....         1                            0

# mean eCDF diff........ 0.0032997                            0
# med  eCDF diff........ 0.0032997                            0
# max  eCDF diff........ 0.0065994                            0

# var ratio (Tr/Co).....    0.97628                            1
# T-test p-value........   0.065187                            1


# ***** (V15) foo$occupationPENSIUNAN/PURNAWIRAWAN *****
#                        Before Matching                   After Matching
# mean treatment........ 0.00031502                            0
# mean control.......... 0.00045979                            0
# std mean diff.........   -0.81574                            0

# mean raw eQQ diff..... 0.00018901                            0
# med  raw eQQ diff.....         0                            0
# max  raw eQQ diff.....         1                            0
# mean eCDF diff........ 7.2383e-05                            0
# med  eCDF diff........ 7.2383e-05                            0
# max  eCDF diff........ 0.00014477                            0

# var ratio (Tr/Co).....    0.68528                          NaN
# T-test p-value........    0.38173                            1


# ***** (V16) foo$occupationPETANI *****
#                        Before Matching                   After Matching
# mean treatment........    0.16942                      0.19083
# mean control..........     0.1929                      0.19083
# std mean diff.........    -6.2587                            0

# mean raw eQQ diff.....   0.023501                            0
# med  raw eQQ diff.....          0                            0
# max  raw eQQ diff.....          1                            0

# mean eCDF diff........   0.011739                            0
# med  eCDF diff........   0.011739                            0
# max  eCDF diff........   0.023478                            0

# var ratio (Tr/Co).....    0.90388                            1
# T-test p-value........ 3.8256e-12                            1


# ***** (V17) foo$occupationPNS *****
#                        Before Matching                   After Matching
# mean treatment........  0.0011971                            0
# mean control..........  0.0019212                            0
# std mean diff.........    -2.0942                            0

# mean raw eQQ diff..... 0.00075605                            0
# med  raw eQQ diff.....          0                            0
# max  raw eQQ diff.....          1                            0

# mean eCDF diff........ 0.00036208                            0
# med  eCDF diff........ 0.00036208                            0
# max  eCDF diff........ 0.00072417                            0

# var ratio (Tr/Co).....    0.62355                          NaN
# T-test p-value........   0.026721                            1


# ***** (V18) foo$occupationPROFESIONAL *****
#                        Before Matching                   After Matching
# mean treatment........   0.012916                     0.015049
# mean control..........   0.015583                     0.015049
# std mean diff.........    -2.3625                            0

# mean raw eQQ diff.....  0.0027092                            0
# med  raw eQQ diff.....          0                            0
# max  raw eQQ diff.....          1                            0

# mean eCDF diff........  0.0013338                            0
# med  eCDF diff........  0.0013338                            0
# max  eCDF diff........  0.0026676                            0

# var ratio (Tr/Co).....     0.8311                            1
# T-test p-value........  0.0094122                            1


# ***** (V19) foo$occupationTNI/POLRI *****
#                       Before Matching                   After Matching
# mean treatment........ 6.3004e-05                            0
# mean control.......... 8.2105e-05                            0
# std mean diff.........   -0.24064                            0

# mean raw eQQ diff..... 6.3004e-05                            0
# med  raw eQQ diff.....          0                            0
# max  raw eQQ diff.....          1                            0

# mean eCDF diff........ 9.5502e-06                            0
# med  eCDF diff........ 9.5502e-06                            0
# max  eCDF diff........   1.91e-05                            0

# var ratio (Tr/Co).....    0.76741                          NaN
# T-test p-value........    0.79338                            1


# ***** (V20) foo$occupationWIRASWASTA *****
#                       Before Matching                   After Matching
# mean treatment........    0.51184                      0.51011
# mean control..........    0.49261                      0.51011
# std mean diff.........     3.8478                            0

# mean raw eQQ diff.....   0.019216                            0
# med  raw eQQ diff.....          0                            0
# max  raw eQQ diff.....          1                            0

# mean eCDF diff........  0.0096171                            0
# med  eCDF diff........  0.0096171                            0
# max  eCDF diff........   0.019234                            0
# var ratio (Tr/Co).....     0.9997                            1
# T-test p-value........ 1.5851e-05                            1


# ***** (V21) foo$district_code_new21 *****
#                        Before Matching                   After Matching
# mean treatment........   0.028982                     0.010762
# mean control..........   0.012086                     0.010762
# std mean diff.........     10.072                            0

# mean raw eQQ diff.....   0.016885                            0
# med  raw eQQ diff.....          0                            0
# max  raw eQQ diff.....          1                            0

# mean eCDF diff........   0.008448                            0
# med  eCDF diff........   0.008448                            0
# max  eCDF diff........   0.016896                            0

# var ratio (Tr/Co).....     2.3571                            1
# T-test p-value........ < 2.22e-16                            1


# ***** (V22) foo$district_code_new32 *****
#                        Before Matching                   After Matching
# mean treatment........    0.28314                      0.28979
# mean control..........    0.29027                      0.28979
# std mean diff.........     -1.583                            0

# mean raw eQQ diff.....  0.0071195                            0
# med  raw eQQ diff.....          0                            0
# max  raw eQQ diff.....          1                            0

# mean eCDF diff........  0.0035661                            0
# med  eCDF diff........  0.0035661                            0
# max  eCDF diff........  0.0071321                            0

# var ratio (Tr/Co).....    0.98528                            1
# T-test p-value........   0.076152                            1


# ***** (V23) foo$district_code_new33 *****
#                        Before Matching                   After Matching
# mean treatment........    0.21478                      0.24167
# mean control..........    0.21311                      0.24167
# std mean diff.........    0.40671                            0

# mean raw eQQ diff.....  0.0016381                            0
# med  raw eQQ diff.....          0                             0
# max  raw eQQ diff.....          1                             0

# mean eCDF diff........ 0.00083515                             0
# med  eCDF diff........ 0.00083515                             0
# max  eCDF diff........  0.0016703                             0

# var ratio (Tr/Co).....     1.0057                             1
# T-test p-value........    0.64794                             1


# ***** (V24) foo$district_code_new34 *****
#                        Before Matching                    After Matching
# mean treatment........    0.04026                      0.030536
# mean control..........   0.036717                      0.030536
# std mean diff.........     1.8021                             0

# mean raw eQQ diff.....  0.0035282                             0
# med  raw eQQ diff.....          0                             0
# max  raw eQQ diff.....          1                             0

# mean eCDF diff........  0.0017712                             0
# med  eCDF diff........  0.0017712                             0
# max  eCDF diff........  0.0035424                             0

# var ratio (Tr/Co).....     1.0925                             1
# T-test p-value........   0.041359                             1


# ***** (V25) foo$district_code_new51 *****
#                        Before Matching                    After Matching
# mean treatment........   0.094884                      0.074197
# mean control..........   0.017932                      0.074197
# std mean diff.........     26.258                             0

# mean raw eQQ diff.....   0.076928                             0
# med  raw eQQ diff.....          0                             0
# max  raw eQQ diff.....          1                             0

# mean eCDF diff........   0.038476                             0
# med  eCDF diff........   0.038476                             0
# max  eCDF diff........   0.076952                             0

# var ratio (Tr/Co).....      4.877                             1
# T-test p-value........ < 2.22e-16                             1
# ***** (V26) foo$district_code_new52 *****
#                        Before Matching                   After Matching
# mean treatment........  0.0016381                  0.00043748
# mean control.......... 0.00088673                  0.00043748
# std mean diff.........     1.8579                           0

# mean raw eQQ diff..... 0.00075605                           0
# med  raw eQQ diff.....          0                           0
# max  raw eQQ diff.....          1                           0

# mean eCDF diff........ 0.00037569                           0
# med  eCDF diff........ 0.00037569                           0
# max  eCDF diff........ 0.00075138                           0

# var ratio (Tr/Co).....     1.8461                           1
# T-test p-value........   0.028454                           1


# ***** (V27) foo$district_code_new63 *****
#                        Before Matching                   After Matching
# mean treatment........    0.13388                     0.14944
# mean control..........    0.23324                     0.14944
# std mean diff.........    -29.177                           0

# mean raw eQQ diff.....   0.099357                           0
# med  raw eQQ diff.....          0                           0
# max  raw eQQ diff.....          1                           0

# mean eCDF diff........   0.049679                           0
# med  eCDF diff........   0.049679                           0
# max  eCDF diff........   0.099359                           0

# var ratio (Tr/Co).....    0.64842                           1
# T-test p-value........ < 2.22e-16                           1


# ***** (V28) foo$district_code_new73 *****
#                        Before Matching                   After Matching
# mean treatment........    0.20079                      0.2029
# mean control..........    0.19388                      0.2029
# std mean diff.........     1.7255                           0

# mean raw eQQ diff.....  0.0069304                           0
# med  raw eQQ diff.....          0                           0
# max  raw eQQ diff.....          1                           0

# mean eCDF diff........  0.0034561                           0
# med  eCDF diff........  0.0034561                            0
# max  eCDF diff........  0.0069123                            0

# var ratio (Tr/Co).....     1.0268                            1
# T-test p-value........   0.052229                            1


# ***** (V29) foo$worker *****
#                         Before Matching                   After Matching
# mean treatment........      2.821                       3.0091
# mean control..........     2.2934                        3.055
# std mean diff.........     2.7533                     -0.22906

# mean raw eQQ diff.....    0.52186                    0.0090115
# med  raw eQQ diff.....          0                            0
# max  raw eQQ diff.....        123                           46

# mean eCDF diff........   0.001357                   3.0761e-05
# med  eCDF diff........ 0.00082379                    1.334e-05
# max  eCDF diff........   0.067892                    0.0014408

# var ratio (Tr/Co).....     1.3945                      0.94567
# T-test p-value........   0.001455                      0.25935
# KS Bootstrap p-value.. < 2.22e-16                        0.582
# KS Naive p-value...... < 2.22e-16                      0.99771
# KS Statistic..........   0.067892                    0.0014408


# ***** (V30) foo$capital *****
#                         Before Matching                   After Matching
# mean treatment........   24968672                     22186190
# mean control..........   21207819                     18791999
# std mean diff.........     4.0378                       3.9982

# mean raw eQQ diff.....   17180626                       363239
# med  raw eQQ diff.....      2e+06                            0
# max  raw eQQ diff.....   1.69e+11                        4e+09

# mean eCDF diff........   0.042912                   0.00077792
# med  eCDF diff........   0.046603                   0.00070037
# max  eCDF diff........   0.085289                    0.0019077

# var ratio (Tr/Co).....   0.017031                        3.278
# T-test p-value........    0.20773                   1.9847e-05
# KS Bootstrap p-value.. < 2.22e-16                        0.578
# KS Naive p-value...... < 2.22e-16                      0.94787
# KS Statistic..........   0.085289                    0.0019077
# ***** (V31) foo$credit_proposal *****
#                        Before Matching                   After Matching
# mean treatment........   15612098                     14419894
# mean control..........   12550889                     12825612
# std mean diff.........     5.8458                       5.6267

# mean raw eQQ diff.....    2986037                       223025
# med  raw eQQ diff.....          0                            0
# max  raw eQQ diff.....    2.5e+09                        4e+08

# mean eCDF diff........   0.029315                   0.00054517
# med  eCDF diff........   0.029082                   0.00028015
# max  eCDF diff........   0.065578                    0.0021411

# var ratio (Tr/Co).....      4.739                       2.6487
# T-test p-value........ 7.7804e-13                   2.6645e-15
# KS Bootstrap p-value.. < 2.22e-16                         0.43
# KS Naive p-value...... < 2.22e-16                      0.88199
# KS Statistic..........   0.065578                    0.0021411


# ***** (V32) foo$worker_transformed *****
#                        Before Matching                   After Matching
# mean treatment........     2.4635                       2.3516
# mean control..........     2.3237                       2.3516
# std mean diff.........     10.159                            0

# mean raw eQQ diff.....    0.13974                            0
# med  raw eQQ diff.....          0                            0
# max  raw eQQ diff.....          1                            0

# mean eCDF diff........   0.015535                            0
# med  eCDF diff........   0.010613                            0
# max  eCDF diff........   0.067892                            0

# var ratio (Tr/Co).....     1.0915                            1
# T-test p-value........ < 2.22e-16                            1
# KS Bootstrap p-value.. < 2.22e-16                            1
# KS Naive p-value...... < 2.22e-16                            1
# KS Statistic..........   0.067892                   5.0822e-21


# ***** (V33) foo$capital_transformed *****
#                        Before Matching                   After Matching
# mean treatment........     5.9757                       5.8634
# mean control..........     5.7172                       5.8634
# std mean diff.........     9.9813                            0

# mean raw eQQ diff.....    0.29801                            0
# med  raw eQQ diff.....          0                            0
# max  raw eQQ diff.....          8                            0

# mean eCDF diff........   0.035086                            0
# med  eCDF diff........   0.044051                            0
# max  eCDF diff........   0.050318                            0

# var ratio (Tr/Co).....    0.85043                            1
# T-test p-value........ < 2.22e-16                            1
# KS Bootstrap p-value.. < 2.22e-16                            1
# KS Naive p-value...... < 2.22e-16                            1
# KS Statistic..........   0.050318                   2.2023e-20


# ***** (V34) foo$credit_proposal_transformed *****
#                        Before Matching                   After Matching
# mean treatment........     4.6789                       4.4668
# mean control..........     4.1435                       4.4668
# std mean diff.........     18.603                            0

# mean raw eQQ diff.....    0.53541                            0
# med  raw eQQ diff.....          0                            0
# max  raw eQQ diff.....          4                            0

# mean eCDF diff........    0.05593                            0
# med  eCDF diff........   0.060235                            0
# max  eCDF diff........   0.099566                            0

# var ratio (Tr/Co).....     1.1027                            1
# T-test p-value........ < 2.22e-16                            1
# KS Bootstrap p-value.. < 2.22e-16                            1
# KS Naive p-value...... < 2.22e-16                            1
# KS Statistic..........   0.099566                   2.2023e-20


# ***** (V35) foo$missing_date_of_birth *****
#                        Before Matching                   After Matching
# mean treatment........   0.014491                     0.017937
# mean control..........   0.014418                     0.017937
# std mean diff.........     0.0614                            0

# mean raw eQQ diff..... 6.3004e-05                            0
# med  raw eQQ diff.....          0                            0
# max  raw eQQ diff.....           1                            0

# mean eCDF diff........  3.6688e-05                            0
# med  eCDF diff........  3.6688e-05                            0
# max  eCDF diff........  7.3377e-05                            0

# var ratio (Tr/Co).....      1.0051                            1
# T-test p-value........     0.94505                            1


# ***** (V36) foo$NA_capital *****
#                        Before Matching                   After Matching
# mean treatment........     0.02224                     0.018024
# mean control..........    0.024713                     0.018024
# std mean diff.........      -1.677                            0

# mean raw eQQ diff.....   0.0024572                            0
# med  raw eQQ diff.....           0                            0
# max  raw eQQ diff.....           1                            0

# mean eCDF diff........   0.0012365                            0
# med  eCDF diff........   0.0012365                            0
# max  eCDF diff........    0.002473                            0

# var ratio (Tr/Co).....     0.90226                            1
# T-test p-value........    0.062759                            1


# ***** (V37) foo$NA_credit_proposal *****
#                        Before Matching                   After Matching
# mean treatment........     0.24357                       0.2155
# mean control..........     0.17695                       0.2155
# std mean diff.........       15.52                            0

# mean raw eQQ diff.....    0.066595                            0
# med  raw eQQ diff.....           0                            0
# max  raw eQQ diff.....           1                            0

# mean eCDF diff........    0.033311                            0
# med  eCDF diff........    0.033311                            0
# max  eCDF diff........    0.066622                            0

# var ratio (Tr/Co).....      1.2651                            1
# T-test p-value........  < 2.22e-16                            1


# Before Matching Minimum p.value: < 2.22e-16
# Variable Name(s): foo$R_date_of_birth foo$educationSD foo$district_code_new21␣
↪foo$district_code_new51 foo$district_code_new63 foo$worker foo$capital␣
↪foo$credit_proposal foo$worker_transformed foo$capital_transformed␣
↪foo$credit_proposal_transformed foo$NA_credit_proposal  Number(s): 1 9 21 25␣
↪27 29 30 31 32 33 34 37

# After Matching Minimum p.value: 2.6645e-15
# Variable Name(s): foo$credit_proposal  Number(s): 31
```

I created an additional column with the first two digits of the district code so that we can match exactly on those. For the birth date, I used a caliper, which is basically a distance that is acceptable for any match given in standard deviations for that variable. In our case, we calculate the standard deviation for the existing values and then find how much of a standard deviation 365 days make up - the fraction or a ratio of the year to the standard deviation will be our caliper.

3 Part (B)

Provide a few sentences talking about the number of treated units dropped, and a few more sentences talking about the balance obtained. Be sure to assess the balance obtained. Be sure to also explain what it means to match exactly on the first 2 digits of the district code.

The original method discarded 1075 treated units, while the new one discarded 4443 units, which is roughly 4 times more. We probably dropped more units because we matched on the date of birth within a caliper, which makes it harder to find matches since before the changes, there wasn't any caliper put on the date of birth matching neither it was exactly matched on. Also, here we matched exactly on the district code's first two digits. The balance in the original method was very good since most of the KS bootstrapped p-values became either 1 or much closer to 1 than before matching. The balance in the tweaked matching was still good. In essence, most variables had a KS Bootstrapped p-value of 1 or very close to 1, which means that there was no difference at all between these specific covariates for treatment and control units when matching. The worst balanced variable has a p-value of 2.6645e-15 (i.e., credit proposal), which is not really good but better compared to the balance before matching. This might have happened because other variables were well matched at the expense of some other variables. Also, we saw that the stricter the matching is (i.e., the acceptable difference, caliper, or exact matching), the more treated units we drop if our pool of control units is not diverse enough and doesn't match well.

The four-digit district codes represent the city and regency region codes. The first two digits represent the provincial region code, while the last two digits represent the district/city code. Since it is really hard to find matched that also live in the same city, it would be problematic to match exactly on the district code, since we would have rejected many pairs as we would have to match on the exact city. However, it is still a vital covariate, so it is easier to match on a region exactly, and we will get somewhat accurate matches. We also cannot really do non-exact matching on district code, since although it is given in numbers, it is not a numeric variable (i.e., the numbers don't have numerical meaning(, it is a categorical variable where each possible value out of limited possibilities is linked to some nominal category (i.e., region or district). This property makes it meaningless to do non-exact matches with some kind of small differences.

4 Question 3

Install the "agridat" package in R, and obtain the "diggle.cow" data set. Run a regression with only a single independent variable: the treatment that was randomized in the original study. The dependent variable is the outcome in the original study. Now, identify the critical value of Gamma, as we discussed in Lesson 10.2.

[ ]:

```
#loading the libraries
library(rbounds)
library(sensitivitymv)
library(agridat)
#loading the data
data("diggle.cow")

#deleting the NA weight observation
diggle.cow <- diggle.cow[-81,]

#setting the seed
set.seed(2324)

#converting Iron to be the treatment, while the no iron is control; same is for␣
 ↪the infected
diggle.cow$iron_new <- ifelse(diggle.cow$iron=="NoIron", 0, 1)
diggle.cow$infect_new <- ifelse(diggle.cow$infect=="NonInfected", 0, 1)

#logistic regression for the weight
lm3 <- lm(weight ~ iron_new + infect_new, data=diggle.cow)
summary(lm3)
# Call:
# lm(formula = weight ~ iron_new + infect_new, data = diggle.cow)

# Residuals:
#     Min      1Q  Median      3Q     Max
# -174.68  -60.67   -0.67   59.47  185.32

# Coefficients:
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept)  284.682      6.705  42.458  < 2e-16 ***
# iron_new     -20.572      6.338  -3.246  0.00124 **
# infect_new   -34.013      7.165  -4.747 2.59e-06 ***
# ---
# Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

# Residual standard error: 77.11 on 593 degrees of freedom
# Multiple R-squared:  0.05697,        Adjusted R-squared:  0.05379
# F-statistic: 17.91 on 2 and 593 DF,  p-value: 2.796e-08
```

```
#regression only for the iron
lm4 <- lm(weight ~ iron_new, data=diggle.cow)
confint(lm4)[2,]
#     2.5 %    97.5 %
# -35.65881 -10.40107
mean(confint(lm4)[2,])
# [1] -23.02994


#regression only for the infect
lm5 <- lm(weight ~ infect_new, data=diggle.cow)
confint(lm5)[2,]
#     2.5 %    97.5 %
# -50.04985 -21.77538
mean(confint(lm5)[2,])
# [1] -35.91261
```

The diggle cow is a 2-by-2 factorial experiment, which means that there are two treatments - the iron/no iron and the infected/non-infected microorganisms. The assignment was randomized, so I did a linear regression on three cases: only iron, only infected, and both of them on the right-hand side, while the weight is the outcome variable. As seen from the data pre-processing and the regression models, both treatment coefficients are negative, which might suggest that given the 596 observations, those cows who took iron and had infected microorganisms weighed less than those who didn't take iron and who had non-infected organisms, on average.

[ ]:

```
#finding the weights of the treated and control units for iron treatment
trt <- diggle.cow[which(diggle.cow$iron_new==1),]$weight
trt <- sample(trt, 298, replace=FALSE)

ctrl <- diggle.cow[which(diggle.cow$iron_new==0),]$weight

hlsens(trt, ctrl, Gamma=1.8, GammaInc=0.05)
#  Gamma Lower bound Upper bound
#   1.00       -22.5 -2.2500e+01
#   1.50       -42.5   -2.400000
#   1.55       -45.0    0.099982

#finding the weights of the treated and control units for infected treatment
trt1 <- diggle.cow[which(diggle.cow$infect_new==1),]$weight
trt1 <- sample(trt, 160, replace=FALSE)
ctrl1 <- diggle.cow[which(diggle.cow$infect_new==0),]$weight

hlsens(trt1, ctrl1, Gamma=2.3, GammaInc=0.1)
#  Gamma Lower bound Upper bound
#   1.00       -40.0 -4.0000e+01
#   2.20       -75.1  3.6446e-05
#   2.25       -75.1  3.6446e-05
```

Then, write a paragraph explaining what you found. Your paragraph should include numbers obtained from your analysis and explain what those numbers mean as simply as you can. (No need to write about the details of the sensitivitymv package, b/c that gets very complicated quite quickly.) As always, link to your code.

For the `iron` treatment, there were 299 treated units, which is why I had to randomly sample 298 observations since there was one missing weight data point in the dataset, and an overall number of observations was 597.

The function hlsens provides Rosenbaum's bounds for the additive effect due to the treatment, which in this case could be interpreted roughly (i.e., they are not the same thing) as the difference in medians for treatment and control groups. When there is no hidden bias, meaning that gamma is 1 (the ratio of the probability of being assigned treatment and control), the median difference in the weights of the cows is -22.5, meaning that cows who were not treated iron weighed 22.5 pounds more than those who were treated, considering the medians. We see that for gamma values around 1.55, the difference might be as low as -45 or as high as 0.099982, which is a small positive number. This is a considerably small gamma value for bracketing zero, so while the general conclusion is that the iron had a negative treatment effect, the finding is sensitive to possible hidden bias due to some unobserved confounders.

As for the `infected organisms` treatment, there were only 160 controlled units, while others were treated. Once again, I randomly sampled 160 observations for the weight in the treatment group to make the function work.

When there is no hidden bias, meaning that gamma is 1, the median difference in the weights of the cows is -40, meaning that cows who were not treated by infected microorganisms weighed 40 pounds more than those who were treated, considering the medians. We see that for gamma values around 2.2, the difference might be as low as -75.1 or as high as 3.6446e-05, which is a small positive number. This is a medium critical gamma value for bracketing zero, so while the general conclusion is that the infected microorganisms had a negative treatment effect, the finding is relatively sensitive to possible hidden bias due to some unobserved confounders. However, it is less sensitive to hidden bias than the iron treatment, which might suggest that there is a stronger relationship between the infected microorganisms and weight rather than the iron intake and weight.

[ ]:

```r
#since the assignment was random, the propensity score should be equal to 0.5
X = c()
for (i in 1:597) {
  X[[i]] = 0.5
}
Y <- diggle.cow$weight

#psens for iron treatment
Tr <- diggle.cow$iron_new

genout <- GenMatch(Tr=Tr, X=X, estimand="ATT", M=1,
                    pop.size=100, max.generations=10, wait.generations=1, print.
 ↪level=0)
mout <- Match(Y=Y, Tr=Tr, X=X, M=1, estimand="ATT", Weight.matrix=genout)
```

```
summary(mout)
# Estimate...  -23.194
# AI SE......  6.1285
# T-stat.....  -3.7847
# p.val......  0.0001539

# Original number of observations..............  597
# Original number of treated obs...............  299
# Matched number of observations...............  299
# Matched number of observations  (unweighted).  89102

psens(mout, Gamma=1.6, GammaInc=.002)$bounds
#     Gamma Lower bound Upper bound
# 1   1.000           0      0.0000
# 283 1.564           0      0.0396
# 284 1.566           0      0.0552

#psens for infect treatment
Tr <- diggle.cow$infect_new

genout <- GenMatch(Tr=Tr, X=X, estimand="ATT", M=1,
                   pop.size=100, max.generations=10, wait.generations=1, print.
 ↪level=0)
mout <- Match(Y=Y, Tr=Tr, X=X, M=1, estimand="ATT", Weight.matrix=genout)
summary(mout)

# Estimate...  -35.913
# AI SE......  7.2957
# T-stat.....  -4.9225
# p.val......  8.5461e-07

psens(mout, Gamma=2.6, GammaInc=.005)$bounds
#      Gamma Lower bound Upper bound
# 1    1.000           0      0.0000
# 296 2.0325           0      0.0365
# 297 2.0360           0      0.0538
```

Since in this study, the treatment assignment was random, we know that the propensity score is 0.5 for all units.

`Iron Treatment.`

The function psens provides Rosenbaum's bounds for the p-values from Wilcoxon's signed-rank test, and when the gamma is 1, and there is no hidden bias, the p-value is close to that estimated in the matching analysis (i.e., very small). As we see, between the gamma values of 1.564 and 1.566, the p-value exceeds the given standard threshold of 0.05. This means that even if the odds of a cow being assigned iron intake is 1.566 times higher because of an unobserved covariate, the inference changes and the results are no longer statistically significant. This is a relatively small gamma value, which could suggest that the results are sensitive to bias.

`Infection Treatment.`

When the gamma is 1, and there is no hidden bias, the p-value is close to that estimated in the matching analysis (i.e., super small). As we see, between the gamma values of 2.0325 and 2.0360, the p-value exceeds the given common threshold of 0.05. This means that if the odds of a cow being assigned infected microorganisms are roughly 2 times higher because of an unobserved covariate, the inference changes, and the results are no longer statistically significant. This is a moderate gamma value, which could suggest that the results are moderately sensitive to bias. This gamma, similarly to the analysis above, is larger than for the iron treatment, which may suggest that there is a stronger relationship between the infected microorganisms and weight rather than the iron intake and the weight.

Generally, the smaller the original p-value is with no hidden bias, the better.

[ ]:

```r
difference <- trt-ctrl
difference1 <- trt1-ctrl1

#the difference is negative since we are considering that the outcomes for the treated units is smaller than the outcome for the control units
senmv(-difference, gamma=1)
# $pval
# [1] 0.0002282084

# $deviate
# [1] 3.505111

# $statistic
# [1] 22.94966

# $expectation
# [1] 0

# $variance
# [1] 42.86957

senmv(-difference1, gamma=1)
# $pval
# [1] 1.934986e-05

# $deviate
# [1] 4.115109

# $statistic
# [1] 37.11875

# $expectation
# [1] 0
# $variance
# [1] 81.36246

for (gamma in seq(1, 1.5, 0.01))
  cat(sprintf("Gamma: %.3f\tp-val: %.5f\n",
              gamma, senmv(-difference, gamma=gamma, method='t')$pval))

# Gamma: 1.000        p-val: 0.00023
# Gamma: 1.300        p-val: 0.04926
# Gamma: 1.310        p-val: 0.05497

for (gamma in seq(1, 2.1, 0.01))
  cat(sprintf("Gamma: %.3f\tp-val: %.5f\n",
              gamma, senmv(-difference1, gamma=gamma, method='t')$pval))

# Gamma: 1.000        p-val: 0.00002
# Gamma: 1.670        p-val: 0.04816
# Gamma: 1.680        p-val: 0.05098
```

I also used a different library senmv to test the critical gamma and see whether the general conclusions match. I took the negative difference as the argument, since here out treatment weight is smaller than the control weight. As expected, the initial p-value for the iron treatment (i.e., 0.0002) is slightly smaller than the infection treatment (i.e., 1.934986e-05), which explains why the critical gamma for the infection is slightly bigger than for iron treatment. The critical gammas are slightly smaller than with a different package - for iron treatment, the critical gamma is around 1.31, and for infection treatment, it is around 1.68. However, this still confirms that the study results are quite sensitive to the hidden bias and unobserved covariates. As mentioned by Professor, the critical gamma should be above 2 to ensure some kind of robustness and small sensitivity to hidden bias, but since both of our critical gammas are not even close to 2, the study results are indeed sensitive to hidden bias and results might not be as reliable.

5 Code

You can find all code here.