Assignment 4
Econometrics I

Universidad Carlos III de Madrid

Gabriel Merlo

```
# install packages (if missing)
list_packages <- c("dplyr", "MatchIt", "tidyr")
new_packages <- list_packages[!(list_packages %in% installed.packages()[,"Package"])]
if(length(new_packages)) install.packages(new_packages)

# Load packages
sapply(list_packages, require, character.only = TRUE)
```


```
# install packages (if missing)
list_packages <- c("dplyr", "MatchIt", "tidyr")
new_packages <- list_packages[!(list_packages %in% installed.packages()[,"Package"])]
if(length(new_packages)) install.packages(new_packages)

# Load packages
sapply(list_packages, require, character.only = TRUE)
```


Exercise 1

(a) Read the data and estimate the ATE using the standard difference of sample means and a linear regression using as controls X.

```
# Load data
penn <- as.data.frame(read.table("penn_jae.dat", header = TRUE))

# Keep control group, and treatment group 4
penn4 <- penn %>% filter(tg == 0 | tg == 4)

# Recode treatment variable
penn4$tg <- recode(penn4$tg, `4` = 1L)

# Control variables
x <- c("female", "black", "othrace", "dep", "q2", "q3", "q4", "q5", "q6", "agelt35",
    "agegt54", "durable", "lusd", "husd")

# Log transformation of dependent variable
penn4$l_inuidur1 <- log(penn4$inuidur1)

# ATE from difference of sample means
diff_mean <- penn4 %>% group_by(tg) %>% summarize(mean = mean(l_inuidur1)) %>% spread(tg,
    mean) %>% summarize(diff = `1` - `0`)
diff_mean
```

```
## # A tibble: 1 x 1
```

```
##      diff
##     <dbl>
## 1 -0.0855
```

```
# ATE from linear regression with controls
ate <- lm(as.formula(paste("l_inuidur1 ~ tg +", paste(x, collapse = "+"))), data = penn4)
summary(ate)
```

```
##
## Call:
## lm(formula = as.formula(paste("l_inuidur1 ~ tg +", paste(x, collapse = "+"))),
##     data = penn4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6195 -0.9966  0.3133  1.0400  2.0883
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.178441   0.159001  13.701  < 2e-16 ***
## tg          -0.071659   0.035460  -2.021 0.043349 *
## female       0.125810   0.034780   3.617 0.000301 ***
## black       -0.293971   0.052967  -5.550 3.00e-08 ***
## othrace     -0.470387   0.198281  -2.372 0.017713 *
## dep          0.045993   0.022535   2.041 0.041308 *
## q2           0.073251   0.156807   0.467 0.640420
## q3          -0.039092   0.156454  -0.250 0.802704
## q4          -0.055596   0.156534  -0.355 0.722478
## q5          -0.144996   0.155854  -0.930 0.352243
## q6           0.003035   0.166438   0.018 0.985453
## agelt35     -0.162642   0.036960  -4.401 1.10e-05 ***
## agegt54      0.227801   0.058892   3.868 0.000111 ***
## durable      0.126551   0.048142   2.629 0.008597 **
## lusd        -0.175602   0.040972  -4.286 1.85e-05 ***
## husd        -0.105557   0.044893  -2.351 0.018746 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.199 on 5083 degrees of freedom
## Multiple R-squared:  0.02912,    Adjusted R-squared:  0.02625
## F-statistic: 10.16 on 15 and 5083 DF,  p-value: < 2.2e-16
```

```
# Checking balance of covariates
penn4 %>% group_by(tg) %>% select(x) %>% summarise_all(funs(mean(.)))
```

```
## # A tibble: 2 x 15
##      tg female black othrace   dep    q2    q3    q4    q5     q6 agelt35
##   <int>  <dbl> <dbl>   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl>   <dbl>
## 1     0  0.405 0.121 0.00566 0.438 0.205 0.237 0.220 0.256 0.0698   0.535
## 2     1  0.402 0.123 0.0103  0.442 0.202 0.233 0.237 0.264 0.0499   0.564
## # ... with 4 more variables: agegt54 <dbl>, durable <dbl>, lusd <dbl>,
## #   husd <dbl>
```

The difference in the mean of log of duration of unemployment between treated and control groups is -0.09. This implies that those that receive the treatment spend less time being unemployed than those who don't

get the treatment.

Controlling by observable characteristics of the individuals, the log of duration of unemployment is 0.07 smaller for the individuals that receive the treatment. Once we control by our vector of observables `x`, the effect of the treatment is 0.01 smaller than when comparing using the difference of means (without controls).

Ideally, randomization should balance the distribution of covariates among treated and untreated. One way to check if this is true is by calculating the sample mean difference in covariates between treatment and control groups. The balance is in general quite good but some characteristics are still not very well balanced (we could test if the differences are significant). This can explain the difference in the ATE by the two previous methods.

(b) One way to evaluate if the randomization is sucessful is to test the significance of θ_0 in a Probit specification of the propensity score $p(x) = \Phi(x'\theta_0)$. Run such a test and interpret the results. Discuss the type of test, critical value, etc.

Randomization is used to assure that the participation in the treatment is the only differentiating factor between individuals in both groups. Propensity score matching can be used to evaluate if the randomization process was correctly done by comparing the outcome variable for similar individuals in the treatment group and the ones in the control group.

Matching propensity scores allows to compare similar individuals in and outside the treatment group. By calculating the probability of participation, matching individuals with same probability but in different groups can be used to calculate ATE. However, two assuptions are used in this case:

1- Bias is only on observables. 2- Treatment and control groups have a common support.

The second assumption can be solved by restricting the sample to a common support. We will use the `MatchIt` package to do the propensity score matching. The algorithm method selected for matching is the nearest neighbor method (it matches individuals with the closest propensity score from the other group).

```
# Match observations
match <- matchit(
  as.formula(paste("tg ~ ", paste(x, collapse = "+"))),
  method = "nearest",
  link = "probit",
  replace = TRUE,
  data = penn4)
summary(match)
```

```
##
## Call:
## matchit(formula = as.formula(paste("tg ~ ", paste(x, collapse = "+"))),
##     data = penn4, method = "nearest", link = "probit", replace = TRUE)
##
## Summary of Balance for All Data:
##          Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean
## distance        0.3451        0.3407          0.1399     1.0154    0.0308
## female          0.4017        0.4052         -0.0071          .    0.0035
## black           0.1232        0.1213          0.0057          .    0.0019
## othrace         0.0103        0.0057          0.0460          .    0.0047
## dep             0.4424        0.4383          0.0054     1.0110    0.0014
## q2              0.2017        0.2048         -0.0078          .    0.0031
## q3              0.2332        0.2367         -0.0083          .    0.0035
```

```
## q4              0.2372        0.2200          0.0405          .    0.0172
## q5              0.2642        0.2564          0.0176          .    0.0078
## q6              0.0499        0.0698         -0.0915          .    0.0199
## agelt35         0.5639        0.5352          0.0579          .    0.0287
## agegt54         0.1100        0.1091          0.0029          .    0.0009
## durable         0.1427        0.1509         -0.0234          .    0.0082
## lusd            0.2762        0.2531          0.0516          .    0.0231
## husd            0.2126        0.2218         -0.0225          .    0.0092
##          eCDF Max
## distance   0.0627
## female     0.0035
## black      0.0019
## othrace    0.0047
## dep        0.0028
## q2         0.0031
## q3         0.0035
## q4         0.0172
## q5         0.0078
## q6         0.0199
## agelt35    0.0287
## agegt54    0.0009
## durable    0.0082
## lusd       0.0231
## husd       0.0092
##
##
## Summary of Balance for Matched Data:
##          Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean
## distance        0.3451        0.3452         -0.0003     0.9887    0.0001
## female          0.4017        0.3994          0.0047          .    0.0023
## black           0.1232        0.1181          0.0157          .    0.0052
## othrace         0.0103        0.0097          0.0057          .    0.0006
## dep             0.4424        0.4447         -0.0030     1.0005    0.0023
## q2              0.2017        0.1971          0.0114          .    0.0046
## q3              0.2332        0.2367         -0.0081          .    0.0034
## q4              0.2372        0.2332          0.0094          .    0.0040
## q5              0.2642        0.2711         -0.0156          .    0.0069
## q6              0.0499        0.0487          0.0053          .    0.0011
## agelt35         0.5639        0.5610          0.0058          .    0.0029
## agegt54         0.1100        0.1072          0.0092          .    0.0029
## durable         0.1427        0.1284          0.0410          .    0.0143
## lusd            0.2762        0.2756          0.0013          .    0.0006
## husd            0.2126        0.2115          0.0028          .    0.0011
##          eCDF Max Std. Pair Dist.
## distance   0.0017          0.0010
## female     0.0023          0.0655
## black      0.0052          0.0715
## othrace    0.0006          0.0057
## dep        0.0046          0.0693
## q2         0.0046          0.0257
## q3         0.0034          0.0352
## q4         0.0040          0.0498
## q5         0.0069          0.0416
## q6         0.0011          0.0211
```

```
## agelt35    0.0029          0.0497
## agegt54    0.0029          0.0678
## durable    0.0143          0.0737
## lusd       0.0006          0.0577
## husd       0.0011          0.0336
##
## Percent Balance Improvement:
##          Std. Mean Diff. Var. Ratio eCDF Mean eCDF Max
## distance            99.8       25.6      99.7     97.3
## female              33.9          .      33.9     33.9
## black             -177.1          .    -177.1   -177.1
## othrace             87.7          .      87.7     87.7
## dep                 44.4       95.6     -66.7    -62.3
## q2                 -47.4          .     -47.4    -47.4
## q3                   1.6          .       1.6      1.6
## q4                  76.7          .      76.7     76.7
## q5                  11.5          .      11.5     11.5
## q6                  94.2          .      94.2     94.2
## agelt35             90.0          .      90.0     90.0
## agegt54           -216.5          .    -216.5   -216.5
## durable            -75.3          .     -75.3    -75.3
## lusd                97.5          .      97.5     97.5
## husd                87.6          .      87.6     87.6
##
## Sample Sizes:
##               Control Treated
## All           3354.      1745
## Matched (ESS)  132.07    1745
## Matched        375.      1745
## Unmatched     2979.         0
## Discarded        0.         0
```

```r
# Dataframe with only matched observations
penn4m <- match.data(match)
dim(penn4m)
```

```
## [1] 2120   26
```

```r
# ATE from difference of sample means from matched data
diff_mean_matched <- penn4m %>%
  group_by(tg) %>%
  summarize(mean = mean(l_inuidur1)) %>%
  spread(tg, mean) %>%
  summarize(diff = `1` - `0`)
diff_mean_matched
```

```
## # A tibble: 1 x 1
##     diff
##    <dbl>
## 1 -0.140
```