Assignment 4
Econometrics I

Universidad Carlos III de Madrid

Gabriel Merlo

```
# install packages (if missing)
list_packages <- c("dplyr", "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 <- "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
#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
```

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

```
##
## Call:
## lm(formula = as.formula(paste("l_inuidur1 ~ tg+", x)), 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
```

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).

(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