Code to load dataset and compute regressions

Commenting style follows http://adv-r.had.co.nz/Style.html‘

```r
rm(list = ls())

library(foreign)
library(cobalt)
library(xtable)
library(stargazer)
library(plm)
library(dplyr)
```

List of working directories

Default working directory

```r
cd <- 'C:/Users/sindr/Desktop/Tinbergen/2nd-year/Block 1/Applied Microeconometrics/Assignments/code'
```

Folder containing data and folder for outputs

```r
dt <- 'C:/Users/sindr/Desktop/Tinbergen/2nd-year/Block 1/Applied Microeconometrics/Assignments/data'
tb <- 'C:/Users/sindr/Desktop/Tinbergen/2nd-year/Block 1/Applied Microeconometrics/Assignments/Final'
```

Reading in data

```r
setwd(dt)
df <- as.data.frame(read.dta('searchperiod.dta'))
setwd(tb)
```

Assignment

Naive estimation

***(vii)* Usually higher educated workers have more favorable labor market outcomes. Use education as monotone instrumental variable and compute the bounds.**

Solution

In our data set we have 5 education-related dummy-variables. Let us get a feeling for those variables by considering some simple values for all of them:

```r
# How common is each value?
Nz0 <- sum(df[["educ_primaryorless"]]==1)
Nz0
```

```
## [1] 231
```

```r
Nz1 <- sum(df[["educ_prepvocational"]]==1)
Nz1
```

```
## [1] 349
```

```r
Nz2 <- sum(df[["educ_vocational"]]==1)
Nz2
```

```
## [1] 592
```

```r
Nz3 <- sum(df[["educ_bachelormaster"]]==1)
Nz3
```

```
## [1] 442
```

```r
NzU <- sum(df[["educ_unknown"]]==1)
NzU
```

```
## [1] 51
```

We see that a fairly even amount of applicants have prep-vocational schooling, vocational schooling or a bachelor or master degree. However fewer have only a primary education or less, and the education level is unknown for very few applicants.

We can then consider the mean values for applicants from each of those categories:

```r
# 10 weeks mean for different education levels
E0 <- mean(df[["benefits_week10"]][df[["educ_primaryorless"]]==1])
E0
```

```
## [1] 0.6709957
```

```r
E1 <- mean(df[["benefits_week10"]][df[["educ_prepvocational"]]==1])
E1
```

```
## [1] 0.7191977
```

```r
E2 <- mean(df[["benefits_week10"]][df[["educ_vocational"]]==1])
E2
```

```
## [1] 0.6993243
```

```r
E3 <- mean(df[["benefits_week10"]][df[["educ_bachelormaster"]]==1])
E3
```

```
## [1] 0.6040724
```

```
unk <- mean(df[["benefits_week10"]][df[["educ_unknown"]]==1])
unk
```

```
## [1] 0.2745098
```

```
# 30 weeks mean for different education levels
E03 <- mean(df[["benefits_week30"]][df[["educ_primaryorless"]]==1])
E03
```

```
## [1] 0.5670996
```

```
E13 <- mean(df[["benefits_week30"]][df[["educ_prepvocational"]]==1])
E13
```

```
## [1] 0.5444126
```

```
E23 <- mean(df[["benefits_week30"]][df[["educ_vocational"]]==1])
E23
```

```
## [1] 0.5084459
```

```
E33 <- mean(df[["benefits_week30"]][df[["educ_bachelormaster"]]==1])
E33
```

```
## [1] 0.3891403
```

```
unk3 <- mean(df[["benefits_week30"]][df[["educ_unknown"]]==1])
unk3
```

```
## [1] 0.1960784
```

Again we see that the applicants whose educational level is unkown stick out like a sore thumb, and so we remove those individuals from our sample:

```
df <- df[df$educ_unknown != 1,]
```

What we then note, is that the difference between the outcomes for applicants in the categories “primaryorless”, “prepvocational”, and “vocational”, is far smaller and less pronounced than the difference between the outcomes for those applications and those from “bachelormaster”. Simple t-tests reveal that only “bachelormaster” is always significantly different from all other categories with regards to outcomes.

```
# 10 weeks: prepvocational - vocational
t.test(df[["benefits_week10"]][df[["educ_prepvocational"]]==1]-df[["benefits_week10"]][df[["educ_vocati
```

```
## Warning in df[["benefits_week10"]][df[["educ_prepvocational"]] == 1] -
## df[["benefits_week10"]][df[["educ_vocational"]] == : longer object length is not
## a multiple of shorter object length
```

```
##
##  One Sample t-test
##
## data:  df[["benefits_week10"]][df[["educ_prepvocational"]] == 1] - df[["benefits_week10"]][df[["educ_
## t = 1.0907, df = 591, p-value = 0.2758
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.02299089  0.08042333
## sample estimates:
##  mean of x
## 0.02871622
```

```
# 30 weeks: primaryorless - prepvocational
t.test(df[["benefits_week30"]][df[["educ_primaryorless"]]==1]-df[["benefits_week30"]][df[["educ_prepvoca
```

```
## Warning in df[["benefits_week30"]][df[["educ_primaryorless"]] == 1] -
## df[["benefits_week30"]][df[["educ_prepvocational"]] == : longer object length is
## not a multiple of shorter object length
```

```
##
##  One Sample t-test
##
## data:  df[["benefits_week30"]][df[["educ_primaryorless"]] == 1] - df[["benefits_week30"]][df[["educ_p
## t = -0.076363, df = 348, p-value = 0.9392
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.07666479  0.07093414
## sample estimates:
##   mean of x
## -0.00286533
```

It thus seems logical to use the variable "educ_bachelormaster" as our monotone instrumental variable against all other forms of education. In this way, we also avoid splitting our fairly limited dataset of only 1614 observations into too many cells. So, we set $Z = 1$ if "educ_bachelormaster"$= 1$.

The Monotone Instrumental Variable assumption we make is that $E[Y_d^*|Z = 0] \geq E[Y_d^*|Z = 1], \quad d = 0, 1$. Keep in mind that a lower value of Y is preferred by the case worker.

To calculate the bounds, we take the following as a starting point:

$$E[Y_1^*] = Pr(Z = 0)E[Y_1^*|Z = 0] + Pr(Z = 1)E[Y_1^*|Z = 1]$$

The probabilities are easily calculated:

```
Pz1 <- sum(df[["educ_bachelormaster"]]==1)/nrow(df)
Pz1
```

```
## [1] 0.2738538
```

```
Pz0 <- 1-Pz1
Pz0
```

```
## [1] 0.7261462
```

The best lower bound for $E[Y_1^*|Z=0]$ is the maximum value of the bounds $LB(d=1,Z=0)$ and $LB(d=1,Z=1)$. We calculate these lower bounds in the following way:

$$E[Y_d^*|Z=0] \geq LB(d,Z=0) = E[Y_d|z=0]Pr(Z=0) + y_{min}Pr(Z=1), \quad d \in 0,1$$

For $Z=0$, and:

$$E[Y_d^*|Z=1] \geq LB(d,Z=1) = E[Y_d|Z=1]Pr(Z=1) + y_{min}Pr(Z=0), \quad d \in 0,1$$

Since we have that $y_min = 0$, the above equations simplify to just $E[Y_1^*|Z=i] = E[Y_1|z=i], \quad i \in 1,2$. We will call the better/higher lower bound $LB_{max}(1,z)$, and the worse lower bound $LB_{min}(1,z)$ for $z \in 0,1$. Let us calculate those values:

```
# 10 weeks
LB10d1z0 <- mean(df[df$searchperiod == 1 & df$educ_bachelormaster == 0,]$benefits_week10)*Pz0
LB10d1z1 <- mean(df[df$searchperiod == 1 & df$educ_bachelormaster == 1,]$benefits_week10)*Pz1

LB10d0z0 <- mean(df[df$searchperiod == 0 & df$educ_bachelormaster == 0,]$benefits_week10)*Pz0
LB10d0z1 <- mean(df[df$searchperiod == 0 & df$educ_bachelormaster == 1,]$benefits_week10)*Pz1

# 30 weeks
LB30d1z0 <- mean(df[df$searchperiod == 1 & df$educ_bachelormaster == 0,]$benefits_week30)*Pz0
LB30d1z1 <- mean(df[df$searchperiod == 1 & df$educ_bachelormaster == 1,]$benefits_week30)*Pz1

LB30d0z0 <- mean(df[df$searchperiod == 0 & df$educ_bachelormaster == 0,]$benefits_week30)*Pz0
LB30d0z1 <- mean(df[df$searchperiod == 0 & df$educ_bachelormaster == 1,]$benefits_week30)*Pz1
```

In the same way we acquire the upper bounds for $E[Y_1^*|Z=0]$ is the minimum value of the bounds $UB(d=1,Z=0)$ and $UB(d=1,Z=1)$. Which we calculate in the following way:

$$E[Y_1^*|Z=0] \leq UB(d=1,Z=0) = E[Y_1|z=0]Pr(Z=0) + y_{max}Pr(Z=1)$$

For $Z=0$, and:

$$E[Y_1^*|Z=1] \leq UB(d=1,Z=1) = E[Y_1|Z=1]Pr(Z=1) + y_{max}Pr(Z=0).$$

We can calculate the upper bounds:

```
# 10 weeks
UB10d1z0 <- (LB10d1z0+Pz1)
UB10d1z1 <- (LB10d1z1+Pz0)

UB10d0z0 <- (LB10d0z0+Pz1)
UB10d0z1 <- (LB10d0z1+Pz0)

# 30 weeks
UB30d1z0 <- (LB30d1z0+Pz1)
UB30d1z1 <- (LB30d1z1+Pz0)

UB30d0z0 <- (LB30d0z0+Pz1)
UB30d0z1 <- (LB30d0z1+Pz0)
```

Using this to make some substitutions, we get:

$$E[Y_1^*] \geq Pr(Z=0)LB_{max}(1,0) + Pr(Z=1)LB_{min}(1,1)$$

Which gives us the complete lower bound:

```
# 10 weeks
LB10_d1 <- Pz0*max(LB10d1z1, LB10d1z0)+Pz1*min(LB10d1z1, LB10d1z0)
LB10_d0 <- Pz0*max(LB10d0z1, LB10d0z0)+Pz1*min(LB10d0z1, LB10d0z0)

# 30 weeks
LB30_d1 <- Pz0*max(LB30d1z1, LB30d1z0)+Pz1*min(LB30d1z1, LB30d1z0)
LB30_d0 <- Pz0*max(LB30d0z1, LB30d0z0)+Pz1*min(LB30d0z1, LB30d0z0)
```

Similarly for the upper bound we get:

$$E[Y_1^*] \leq Pr(Z=0)UB_{max}(1,0) + Pr(Z=1)UB_{min}(1,1)$$

Which gives us the complete upper bound:

```
# 10 weeks
UB10_d1 <- Pz0*max(UB10d1z1, UB10d1z0)+Pz1*min(UB10d1z1, UB10d1z0)
UB10_d0 <- Pz0*max(UB10d0z1, UB10d0z0)+Pz1*min(UB10d0z1, UB10d0z0)

# 30 weeks
UB30_d1 <- Pz0*max(UB30d1z1, UB30d1z0)+Pz1*min(UB30d1z1, UB30d1z0)
UB30_d0 <- Pz0*max(UB30d0z1, UB30d0z0)+Pz1*min(UB30d0z1, UB30d0z0)
```

Putting it all together, we see that the bounds for $E[Y_0^*]$ are:

$$E[Y_d^*] \geq Pr(Z=0)LB_{max}(0,0)+Pr(Z=1)LB_{min}(0,1)E[Y_d^*] \leq Pr(Z=0)UB_{max}(0,0)+Pr(Z=1)UB_{min}(0,1)$$

This gives us the final, complete MIV bounds as:

```
LB10_MIV <- LB10_d1 - UB10_d0
UB10_MIV <- UB10_d1 - LB10_d0
```

So we see that using education as a monotone instrumental variable gives us the bounds: -0.524 $\leq E[Y^*_{10weeks}] \leq 0.374$.

```
LB30_MIV <- LB30_d1 - UB30_d0
LB30_MIV
```

```
## [1] -0.5350021
```

```
UB30_MIV <- UB30_d1 - LB30_d0
UB30_MIV
```

```
## [1] 0.4187329
```

So we see that using education as a monotone instrumental variable gives us the bounds: -0.535 $\leq E[Y^*_{30weeks}] \leq 0.419$.

As we can see the bounds are slightly tighter than for the no-assumption bounds, but the difference is very small. This leads us to believe that using education as a monotone variable as constructed by us is not very helpful. The results might differ if we use an education variable where we use four categories.