Homework 3

Stefan Hinkelmann

13 March 2019

R Markdown for Homework 3

This R Markdown document tries to replicate the regression result in equation (19) of Katz and Murphy (1992).

Preliminary setup and data import

```
############### Homework 3 ##################

# clear environment
rm(list = ls())
# changing directory
setwd("C:/Users/s4522444/Dropbox/Master of Advanced Economics/2019/Macro/Homework 3")
# load libraries
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.5.2
```

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## v ggplot2 3.1.0       v purrr   0.3.0
## v tibble  2.0.1       v dplyr   0.8.0.1
## v tidyr   0.8.2       v stringr 1.4.0
## v readr   1.3.1       v forcats 0.4.0
```

```
## Warning: package 'ggplot2' was built under R version 3.5.2
```

```
## Warning: package 'tibble' was built under R version 3.5.2
```

```
## Warning: package 'tidyr' was built under R version 3.5.2
```

```
## Warning: package 'readr' was built under R version 3.5.2
```

```
## Warning: package 'purrr' was built under R version 3.5.2
```

```
## Warning: package 'dplyr' was built under R version 3.5.2
```

```
## Warning: package 'stringr' was built under R version 3.5.2
```

```
## Warning: package 'forcats' was built under R version 3.5.2
```

```
## -- Conflicts ------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

```
## Warning: package 'lubridate' was built under R version 3.5.2
```

```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:base':
##
##     date
```

```r
# import data frame
data_01 <- read.csv("data_01.csv", header = TRUE, sep = ",", quote = "\"",
                    dec = ".", fill = TRUE, comment.char = "")
```

Setting up the data

```r
# Select all males
data_04 <- data_01 %>%
  filter(sex == 1)

# Create education level dummies
data_04 <- data_04 %>%
  mutate(edlvl = ifelse(deduc_1==1,1,ifelse(deduc_2==1,3,
                                            ifelse(deduc_3==1,4,ifelse(deduc_4==1,5,2)))))

# Calculating relative supply
hs_supply <- matrix(0, 2018-1962+1, 2)
coll_supply <- matrix(0, 2018-1962+1, 2)
Y = dim(hs_supply)[1]

for (i in 1:Y){
  summe <- sum(data_04$edlvl == 1 | data_04$edlvl == 2 & data_04$year == i+1961)
  hs_supply[i,1] <- i+1961
  hs_supply[i,2] <- summe
  summe2 <- sum(data_04$edlvl == 3 | data_04$edlvl == 4 |
                  data_04$edlvl == 5 & data_04$year == i+1961)
  coll_supply[i,1] <- i+1961
  coll_supply[i,2] <- summe2
}

relsup <- matrix(0, 2018-1962+1, 2)
relsup[,1] <- hs_supply[,1]
relsup[,2] <- coll_supply[,2]/hs_supply[,2]
colnames(relsup) <- c("year", "relsupply")

# Calculating relative wage (relative log wages)
hs_wage <- matrix(0, 2018-1962+1, 2)
coll_wage <- matrix(0, 2018-1962+1, 2)

for (i in 1:Y){
  lrwagesum_hs <- sum(data_04[which(data_04$edlvl == 1 | data_04$edlvl == 2 &
                                      data_04$year == i+1961), 27])
  summe <- sum(data_04$edlvl == 1 | data_04$edlvl == 2 & data_04$year == i+1961)
  hs_wage[i,1] <- i+1961
  hs_wage[i,2] <- lrwagesum_hs/summe

  lrwagesum_coll <- sum(data_04[which(data_04$edlvl == 3 | data_04$edlvl == 4 |
                                        data_04$edlvl == 5 & data_04$year == i+1961), 27])
  summe2 <- sum(data_04$edlvl == 3 | data_04$edlvl == 4 |
```

```
                  data_04$edlvl == 5 & data_04$year == i+1961)
  coll_wage[i,1] <- i+1961
  coll_wage[i,2] <- lrwagesum_coll/summe2
}

rellrwage <- matrix(0, 2018-1962+1, 2)
rellrwage[,1] <- hs_wage[,1]
rellrwage[,2] <- coll_wage[,2]/hs_wage[,2]
colnames(rellrwage) <- c("year", "rellrwage")

# Calculating relative wage (log relative wages)
hs_wage <- matrix(0, 2018-1962+1, 2)
coll_wage <- matrix(0, 2018-1962+1, 2)

for (i in 1:Y){
  rwagesum_hs <- sum(data_04[which(data_04$edlvl == 1 |
                                     data_04$edlvl == 2 & data_04$year == i+1961), 26])
  summe <- sum(data_04$edlvl == 1 | data_04$edlvl == 2 & data_04$year == i+1961)
  hs_wage[i,1] <- i+1961
  hs_wage[i,2] <- rwagesum_hs/summe

  rwagesum_coll <- sum(data_04[which(data_04$edlvl == 3 | data_04$edlvl == 4 |
                                       data_04$edlvl == 5 & data_04$year == i+1961), 26])
  summe2 <- sum(data_04$edlvl == 3 | data_04$edlvl == 4 |
                  data_04$edlvl == 5 & data_04$year == i+1961)
  coll_wage[i,1] <- i+1961
  coll_wage[i,2] <- rwagesum_coll/summe2
}

relrwage <- matrix(0, 2018-1962+1, 2)
relrwage[,1] <- hs_wage[,1]
relrwage[,2] <- coll_wage[,2]/hs_wage[,2]
colnames(relrwage) <- c("year", "relrwage")

# Combining the information
# the log real wage version
#allthedata <- merge(x = relsup, y = rellrwage, by = c("year"))

# the real wage version
allthedata <- merge(x = relsup, y = relrwage, by = c("year"))
```

The matrix allthedata now contains three columns: year, relative supply of workers (coll/hs) and relative real wages (coll/hs). This is all the information we need to conduct the regression analysis as in Katz and Murphy.

Regression over the entire time period 1962 - 2018:

```
# Regression 1
summary(regression1 <- lm(log(allthedata$relrwage) ~ allthedata$year +
                            log(allthedata$relsup), data = allthedata))
```

```
##
## Call:
## lm(formula = log(allthedata$relrwage) ~ allthedata$year + log(allthedata$relsup),
```

```
##     data = allthedata)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0123893 -0.0050521  0.0008428  0.0043090  0.0186631
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -1.030e+00  1.158e-01  -8.892 3.69e-12 ***
## allthedata$year         6.411e-04  6.324e-05  10.138 4.20e-14 ***
## log(allthedata$relsup) -3.549e-02  4.472e-02  -0.794    0.431
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.006668 on 54 degrees of freedom
## Multiple R-squared:  0.7092, Adjusted R-squared:  0.6984
## F-statistic: 65.85 on 2 and 54 DF,  p-value: 3.291e-15
```

This is the regression result for the entire time period 1962 to 2018. We see that the intercept and the time regressor are highly significant. The time regressor has a positive effect while the effect of relative supply is negative but insignificant. The directions are in line with Katz and Murphy's findings.

Regression over the same time period as Katz and Murphy (1963 - 1987):

```
# Regression 2
summary(regression2 <- lm(log(allthedata$relrwage[2:26]) ~ allthedata$year[2:26] +
                            log(allthedata$relsup[2:26]), data = allthedata))
```

```
##
## Call:
## lm(formula = log(allthedata$relrwage[2:26]) ~ allthedata$year[2:26] +
##     log(allthedata$relsup[2:26]), data = allthedata)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0084200 -0.0046648  0.0001764  0.0039451  0.0163036
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -2.3602877  0.3588057  -6.578 1.29e-06 ***
## allthedata$year[2:26]         0.0012666  0.0001769   7.162 3.52e-07 ***
## log(allthedata$relsup[2:26])  0.1686672  0.0673303   2.505   0.0201 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.006153 on 22 degrees of freedom
## Multiple R-squared:  0.7015, Adjusted R-squared:  0.6744
## F-statistic: 25.86 on 2 and 22 DF,  p-value: 1.674e-06
```

Here the log of the relative supply of workers has a positive and significant effect. This is in contrast to the finding of Katz and Murphy.

Regression over the later period 1988 - 2018:

```
# Regression 3
summary(regression3 <- lm(log(allthedata$relrwage[27:57]) ~ allthedata$year[27:57] +
                            log(allthedata$relsup[27:57]), data = allthedata))
```

```
##
## Call:
## lm(formula = log(allthedata$relrwage[27:57]) ~ allthedata$year[27:57] +
##     log(allthedata$relsup[27:57]), data = allthedata)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -4.753e-04 -1.496e-04  2.440e-06  1.573e-04  5.370e-04
##
## Coefficients:
##                                  Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                     4.799e-01  1.379e-02   34.793  < 2e-16 ***
## allthedata$year[27:57]          5.478e-05  5.974e-06    9.169  6.3e-10 ***
## log(allthedata$relsup[27:57]) -7.041e-01  5.399e-03 -130.418  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0002355 on 28 degrees of freedom
## Multiple R-squared:  0.9991, Adjusted R-squared:  0.999
## F-statistic: 1.481e+04 on 2 and 28 DF,  p-value: < 2.2e-16
```

Here we can see that the coefficient on the relative supply of workers is almost identical to the result in Katz and Murphy: -0.704 vs. -0.709. However, it is not the same time interval.