Problem set 3 - Bao Hoang Nguyen

19-03-2019

For this problem set, I do data processing again, since the code for data processing provided for problem set 2 does not exactly match with data description in the paper. The changes are documented as follows:

- Use two data samples: a wage sample and a count sample
- Potential experience is calculated as min(age - years of schooling - 7, age -17)
- Use four education groups (instead of five groups), which are: less than 12, 12, 13-15, and 16 or more years of schooling
- Wage sample include includes full-time wage and salary workers who participated in the labor force at least 39 weeks, worked at least one week, OR did not work past year due to school, retirement, or military service

Since some necessary variables/year are not available in the downloaded data, there are some deviations from data description in the paper as follows:

- All samples start from the year 1964 (since data on 1963 is not available)
- Use weeks worked instead of hours worked to calculate labour supply

The results for each question are presented below.

1. The entire time period (from 1964 to 2017)

The estimation results using the period 1964 - 2017 data are as follows:

```
##
## Call:
## lm(formula = log(relative_wage) ~ log(relative_supply) + t)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.10906 -0.03377  0.00890  0.03761  0.07992
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -0.367430   0.069728  -5.269 2.78e-06 ***
## log(relative_supply)  0.130692   0.046866   2.789  0.00742 **
## t                     0.029176   0.001294  22.539  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04648 on 51 degrees of freedom
## Multiple R-squared:  0.9922, Adjusted R-squared:  0.9919
## F-statistic:  3231 on 2 and 51 DF,  p-value: < 2.2e-16
```

2. The time period from 1964 to 1987

The estimation results using the period 1964 - 1987 data are as follows:

```
##
## Call:
## lm(formula = log(relative_wage) ~ log(relative_supply) + t)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.05105 -0.03452 -0.01291  0.03769  0.07461
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -0.228988   0.305981  -0.748  0.46254
## log(relative_supply)  0.198501   0.177404   1.119  0.27581
## t                     0.025909   0.008868   2.922  0.00815 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04175 on 21 degrees of freedom
## Multiple R-squared:  0.9757, Adjusted R-squared:  0.9734
## F-statistic: 421.9 on 2 and 21 DF,  p-value: < 2.2e-16
```

As can be seen from the results, the coefficient of relative labor supply is insignificant and has opposite sign with the result in Katz and Murphy (1992). The reasons might be as follows:

- Due to the fact that last year working hours are not available in downloaded data, I use working weeks to calculate labour supply
- The paper aggregate male and female, but I only use male sample in this replication.

3. The time period from 1988 to 2017

The estimation results using the period 1988 - 2017 data are as follows:

```
##
## Call:
## lm(formula = log(relative_wage) ~ log(relative_supply) + t)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.102637 -0.030204  0.000708  0.040195  0.071391
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          0.679586   0.249873   2.720 0.011284 *
## log(relative_supply) 0.698577   0.379588   1.840 0.076733 .
## t                    0.022115   0.005672   3.899 0.000578 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.05108 on 27 degrees of freedom
## Multiple R-squared:  0.971,  Adjusted R-squared:  0.9689
## F-statistic: 452.7 on 2 and 27 DF,  p-value: < 2.2e-16
```

As can be seen from the results, the coefficient of relative labor supply is significant at 10% level of significance. Moreover, its value triple the value of relative labor supply's coefficient in the ealier period.

4. R code for this problemset

```
library(tidyverse)
library(lubridate)
# read the downloaded
data_00 <- read_fwf(file="data_00.dat",
                    fwf_cols(year      = c(1, 4),
                             serial    = c(5,9),
                             month     = c(10,11),
                             cpsid     = c(12,25),
                             asecflag  = c(26,26),
                             hflag     = c(27,27),
                             asecwth   = c(28,37),
                             pernum    = c(38,39),
                             cpsidp    = c(40,53),
                             asecwt    = c(54,63),
                             age       = c(64,65),
                             sex       = c(66,66),
                             race      = c(67,69),
                             educ      = c(70,72),
                             schlcoll  = c(73,73),
                             indly     = c(74,77),
                             classwly  = c(78,79),
                             wkswork1  = c(80,81),
                             wkswork2  = c(82,82),
                             fullpart  = c(83,83),
                             incwage   = c(84,90)),
                    col_types = cols(year       = "i",
                                     serial     = "n",
                                     month      = "i",
                                     cpsid      = "d",
                                     asecflag   = "i",
                                     hflag      = "i",
                                     asecwth    = "d",
                                     pernum     = "i",
                                     cpsidp     = "d",
                                     asecwt     = "d",
                                     age        = "i",
                                     sex        = "i",
                                     race       = "i",
                                     educ       = "i",
                                     schlcoll   = "i",
                                     indly      = "i",
```

```
                                    classwly   = "i",
                                    wkswork1   = "i",
                                    wkswork2   = "i",
                                    fullpart   = "i",
                                    incwage    = "n"))
data_00$asecwt = data_00$asecwt/10000

# merge cpi data (see Acemoglu and Autor's Data Appendix)
data_cpi <- read_csv(file = "data_cpi.csv",
                     col_names = c("year","cpi"),
                     col_types=cols(year = "D", cpi = "d"),
                     skip = 1)
data_cpi$year <- year(data_cpi$year)
data_cpi <- data_cpi %>%
  mutate(price_1982 = ifelse(year == 1982, cpi, 0)) %>%
  # the base year is 1982
  #(see Acemoglu and Autor's Data Appendix)
  mutate(price_1982 = max(price_1982)) %>%
  mutate(cpi = cpi/price_1982) %>%
  select(year, cpi)
data_00 <- data_00 %>%
  left_join(data_cpi, by = "year")
# replace missing values
data_00 <- data_00 %>%
  mutate(educ = ifelse(educ == 999, NA, educ)) %>%
  mutate(classwly = ifelse(classwly == 99, NA, classwly)) %>%
  mutate(wkswork2 = ifelse(wkswork2 == 999, NA, wkswork2)) %>%
  mutate(incwage = ifelse(incwage == 9999999 | incwage == 9999998, NA,
                          incwage)) %>%
  mutate(race = ifelse(race == 999, NA, race))
# create wrkswork variable: worked weeks are in brackets before 1976
# see Katz and Murphy (1992)
data_00 <- data_00 %>%
  mutate(wkswork = ifelse(year >= 1976, wkswork1, NA)) %>%
  mutate(wkswork = ifelse(year < 1976 & wkswork2 == 1, 7, wkswork)) %>%
  mutate(wkswork = ifelse(year < 1976 & wkswork2 == 2, 20, wkswork)) %>%
  mutate(wkswork = ifelse(year < 1976 & wkswork2 == 3, 33, wkswork)) %>%
  mutate(wkswork = ifelse(year < 1976 & wkswork2 == 4, 43.5, wkswork)) %>%
  mutate(wkswork = ifelse(year < 1976 & wkswork2 == 5, 48.5, wkswork)) %>%
  mutate(wkswork = ifelse(year < 1976 & wkswork2 == 6, 51, wkswork))
# handle the top coding issue for income see Katz and Murphy (1992)'s Data section
data_00 <- data_00 %>%
  group_by(year) %>%
  mutate(top_incwage = max(incwage, na.rm = TRUE)) %>%
  mutate(incwage = ifelse(incwage == top_incwage, 1.45*incwage, incwage)) %>%
  ungroup()
# calculate log real wages
data_00 <- data_00 %>%
  mutate(rwage = incwage/cpi/wkswork) %>%
  mutate(lrwage = log(rwage))
# create education duammies
data_00 <- data_00 %>%
  mutate(dfemale = (sex == 2)) # female
```

```r
data_00 <- data_00 %>%
  mutate(deduc_1 = ifelse(educ < 70, 1, 0)) %>%
  # Less than 12 years of schooling
  mutate(deduc_2 = ifelse(educ >= 80 & educ < 110, 1, 0)) %>%
  # 13-15 years of schooling
  mutate(deduc_3 = ifelse(educ >= 110, 1, 0))
  # 16 or more years of schooling
data_00 <- data_00 %>%
  mutate(drace_1 = ifelse(race == 200,1,0)) %>% # black
  mutate(drace_2 = ifelse(race > 200,1,0)) # nonwhite other
# create experience variable: check the IPUMS website for variable definition
# I changed this, since in Katz & Murphy (1992):
# exp = min (age - years of schooling - 7, age -17)
data_00 <- data_00 %>%
  mutate(exp = ifelse(educ == 10, age - 17, NA)) %>%
  mutate(exp = ifelse(educ == 11, age - 17, exp)) %>%
  mutate(exp = ifelse(educ == 12, age - 17, exp)) %>%
  mutate(exp = ifelse(educ == 13, age - 17, exp)) %>%
  mutate(exp = ifelse(educ == 14, age - 17, exp)) %>%
  mutate(exp = ifelse(educ == 20, age - 17, exp)) %>%
  mutate(exp = ifelse(educ == 21, age - 17, exp)) %>%
  mutate(exp = ifelse(educ == 22, age - 17, exp)) %>%
  mutate(exp = ifelse(educ == 30, age - 17, exp)) %>%
  mutate(exp = ifelse(educ == 31, age - 17, exp)) %>%
  mutate(exp = ifelse(educ == 32, age - 17, exp)) %>%
  mutate(exp = ifelse(educ == 40, age - 17, exp)) %>%
  mutate(exp = ifelse(educ == 50, age - 17, exp)) %>%
  mutate(exp = ifelse(educ == 60, age - 17, exp)) %>%
  mutate(exp = ifelse(educ == 70, age - 18, exp)) %>%
  mutate(exp = ifelse(educ == 71, age - 18, exp)) %>%
  mutate(exp = ifelse(educ == 72, age - 18, exp)) %>%
  mutate(exp = ifelse(educ == 73, age - 18, exp)) %>%
  mutate(exp = ifelse(educ == 80, age - 19, exp)) %>%
  mutate(exp = ifelse(educ == 81, age - 19, exp)) %>%
  mutate(exp = ifelse(educ == 90, age - 20, exp)) %>%
  mutate(exp = ifelse(educ == 91, age - 20, exp)) %>%
  mutate(exp = ifelse(educ == 92, age - 20, exp)) %>%
  mutate(exp = ifelse(educ == 100, age - 21, exp)) %>%
  mutate(exp = ifelse(educ == 110, age - 22, exp)) %>%
  mutate(exp = ifelse(educ == 111, age - 22, exp)) %>%
  mutate(exp = ifelse(educ == 120, age - 23.5, exp)) %>%
  mutate(exp = ifelse(educ == 121, age - 23, exp)) %>%
  mutate(exp = ifelse(educ == 122, age - 24, exp)) %>%
  mutate(exp = ifelse(educ == 123, age - 23, exp)) %>%
  mutate(exp = ifelse(educ == 124, age - 23, exp)) %>%
  mutate(exp = ifelse(educ == 125, age - 27, exp))
# sample selection (see Katz and Murphy (1992) and Acemoglu and Autor (2011)'s Data Appendix)
wagedata_all <- data_00 %>%
  filter(rwage >= 67) %>%
  # real wage more than 67 dollars in the 1982 dollar work full-time
  filter(fullpart == 1) %>%
  filter(wkswork >= 40|(wkswork==0 & (schlcoll == 5|
        ((year >= 1992 & year <= 2002) & (indly >= 940 & indly <= 960))|
```

```
        (year >= 2003 & indly == 9890))) %>%
  # particiapted in labour force at least 39 weeks, worked 1 week
  # or not work due to school, retirement, military service

  filter(classwly != 10 | classwly != 13 | classwly != 14) %>%
  # not self-employed
  filter(exp >= 1 & exp <= 40) %>%
  # from 1 to 40 years of experience
  filter(year >= 1964)

countdata_all <- data_00 %>%
  filter(wkswork >= 1) %>%
  filter(exp >=1 & exp <= 40) %>%
  filter(year >= 1964)

##############################################################################################
##############################Sample: 1964 - 2017#############################################

# Select sample

countdata <- countdata_all %>%
  filter(year >= 1964 & year <= 2017) %>%
  filter (dfemale == FALSE)
wagedata <- wagedata_all %>%
  filter(year >= 1964 & year <= 2017) %>%
  filter (dfemale == FALSE)

# Count data (Since hours worked lats year are not available,
# I use the weeks worked instead)

work_week_dataframe <- countdata %>%
  group_by(year,deduc_3,deduc_2,deduc_1,exp) %>%
  summarise(total = sum(wkswork*asecwt))

work_week_matrix <- as.matrix(work_week_dataframe[,6])
dim(work_week_matrix) <- c(160,54)

work_week_year <- t(as.matrix(.colSums(work_week_matrix,160,54)))

work_week_year_matrix <- matrix(work_week_year, nrow = 160, ncol = 54, byrow = TRUE)

work_week_matrix_deflated <- work_week_matrix/work_week_year_matrix

fixed_employment_share <- as.matrix(.rowMeans(work_week_matrix_deflated, 160, 54))

# Wage data

rwage_week_dataframe <- wagedata %>%
  group_by(year,deduc_3,deduc_2,deduc_1,exp) %>%
  summarise(total = sum(rwage))

rwage_week_matrix <- as.matrix(rwage_week_dataframe[,6])
dim(rwage_week_matrix) <- c(160,54)
```

```r
average_wage <- as.matrix(.rowMeans(rwage_week_matrix,160,54))

# Relative wage

high_school_wage <- rwage_week_matrix[1:40,]
high_school_fixed_weight <- fixed_employment_share[1:40]/
                            sum(fixed_employment_share[1:40])

graduate_wage <- rwage_week_matrix[121:160,]
graduate_fixed_weight <- fixed_employment_share[121:160]/
                         sum(fixed_employment_share[121:160])

aggregate_high_school_wage <- high_school_fixed_weight %*% high_school_wage
aggregate_graduate_wage <- graduate_fixed_weight %*% graduate_wage

relative_wage <- as.vector(aggregate_graduate_wage/aggregate_high_school_wage)

# College and highschool equivalents

high_school_supply <- work_week_matrix[1:40,]
high_school_supply_weight <- average_wage[1:40]/sum(average_wage[1:40])
aggregate_high_school_supply <- high_school_supply_weight %*% high_school_supply

high_school_dropout_supply <- work_week_matrix[41:80,]
high_school_dropout_supply_weight <- average_wage[41:80]/sum(average_wage[41:80])
aggregate_high_school_dropout_supply <- high_school_dropout_supply_weight %*%
                                        high_school_dropout_supply

some_college_supply <- work_week_matrix[81:120,]
some_college_supply_weight <- average_wage[81:120]/sum(average_wage[81:120])
aggregate_some_college_supply <- some_college_supply_weight %*% some_college_supply

graduate_supply <- work_week_matrix[121:160,]
graduate_supply_weight <- average_wage[121:160]/sum(average_wage[121:160])
aggregate_graduate_supply <- graduate_supply_weight %*% graduate_supply

high_school_equivalent <- aggregate_high_school_supply + 0.69 * aggregate_some_college_supply +
                          0.93 *  aggregate_high_school_dropout_supply
graduate_equivalent <- aggregate_graduate_supply + 0.29 * aggregate_some_college_supply -
                       0.05 *  aggregate_high_school_dropout_supply

relative_supply <- as.vector(graduate_equivalent/high_school_equivalent)

# Time trend

t <- as.vector(c(1:54))

# Regression

model19_1 <- lm(log(relative_wage) ~ log(relative_supply) + t)
summary(model19_1)

##############################Sample: 1964 - 1987##############################################
```

```r
# Select sample

countdata <- countdata_all %>%
  filter(year >= 1964 & year <= 1987) %>%
  filter (dfemale == FALSE)
wagedata <- wagedata_all %>%
  filter(year >= 1964 & year <= 1987) %>%
  filter (dfemale == FALSE)

# Count data (Since hours worked lats year are not available,
# I use the weeks worked instead)

work_week_dataframe <- countdata %>%
  group_by(year,deduc_3,deduc_2,deduc_1,exp) %>%
  summarise(total = sum(wkswork*asecwt))

work_week_matrix <- as.matrix(work_week_dataframe[,6])
dim(work_week_matrix) <- c(160,24)

work_week_year <- t(as.matrix(.colSums(work_week_matrix,160,24)))

work_week_year_matrix <- matrix(work_week_year, nrow = 160, ncol = 24, byrow = TRUE)

work_week_matrix_deflated <- work_week_matrix/work_week_year_matrix

fixed_employment_share <- as.matrix(.rowMeans(work_week_matrix_deflated, 160, 24))

# Wage data

rwage_week_dataframe <- wagedata %>%
  group_by(year,deduc_3,deduc_2,deduc_1,exp) %>%
  summarise(total = sum(rwage))

rwage_week_matrix <- as.matrix(rwage_week_dataframe[,6])
dim(rwage_week_matrix) <- c(160,24)

average_wage <- as.matrix(.rowMeans(rwage_week_matrix,160,24))

# Relative wage

high_school_wage <- rwage_week_matrix[1:40,]
high_school_fixed_weight <- fixed_employment_share[1:40]/
                            sum(fixed_employment_share[1:40])

graduate_wage <- rwage_week_matrix[121:160,]
graduate_fixed_weight <- fixed_employment_share[121:160]/
                         sum(fixed_employment_share[121:160])

aggregate_high_school_wage <- high_school_fixed_weight %*% high_school_wage
aggregate_graduate_wage <- graduate_fixed_weight %*% graduate_wage

relative_wage <- as.vector(aggregate_graduate_wage/aggregate_high_school_wage)
```

```r
# College and highschool equivalents

high_school_supply <- work_week_matrix[1:40,]
high_school_supply_weight <- average_wage[1:40]/sum(average_wage[1:40])
aggregate_high_school_supply <- high_school_supply_weight %*% high_school_supply

high_school_dropout_supply <- work_week_matrix[41:80,]
high_school_dropout_supply_weight <- average_wage[41:80]/sum(average_wage[41:80])
aggregate_high_school_dropout_supply <- high_school_dropout_supply_weight %*%
                                        high_school_dropout_supply

some_college_supply <- work_week_matrix[81:120,]
some_college_supply_weight <- average_wage[81:120]/sum(average_wage[81:120])
aggregate_some_college_supply <- some_college_supply_weight %*%
                                 some_college_supply

graduate_supply <- work_week_matrix[121:160,]
graduate_supply_weight <- average_wage[121:160]/sum(average_wage[121:160])
aggregate_graduate_supply <- graduate_supply_weight %*% graduate_supply

high_school_equivalent <- aggregate_high_school_supply + 0.69 * aggregate_some_college_supply +
                          0.93 *  aggregate_high_school_dropout_supply
graduate_equivalent <- aggregate_graduate_supply + 0.29 * aggregate_some_college_supply -
                       0.05 *  aggregate_high_school_dropout_supply

relative_supply <- as.vector(graduate_equivalent/high_school_equivalent)

# Time trend

t <- as.vector(c(1:24))

# Regression

model19_2 <- lm(log(relative_wage) ~ log(relative_supply) + t)

summary(model19_2)

##############################Sample: 1988 - 2017##############################################

# Select sample

countdata <- countdata_all %>%
  filter(year >= 1988 & year <= 2017) %>%
  filter (dfemale == FALSE)
wagedata <- wagedata_all %>%
  filter(year >= 1988 & year <= 2017) %>%
  filter (dfemale == FALSE)

# Count data (Since hours worked lats year are not available,
# I use the weeks worked instead)

work_week_dataframe <- countdata %>%
  group_by(year,deduc_3,deduc_2,deduc_1,exp) %>%
```

```
  summarise(total = sum(wkswork*asecwt))

work_week_matrix <- as.matrix(work_week_dataframe[,6])
dim(work_week_matrix) <- c(160,30)

work_week_year <- t(as.matrix(.colSums(work_week_matrix,160,30)))

work_week_year_matrix <- matrix(work_week_year, nrow = 160, ncol = 30, byrow = TRUE)

work_week_matrix_deflated <- work_week_matrix/work_week_year_matrix

fixed_employment_share <- as.matrix(.rowMeans(work_week_matrix_deflated, 160, 30))

# Wage data

rwage_week_dataframe <- wagedata %>%
  group_by(year,deduc_3,deduc_2,deduc_1,exp) %>%
  summarise(total = sum(rwage))

rwage_week_matrix <- as.matrix(rwage_week_dataframe[,6])
dim(rwage_week_matrix) <- c(160,30)

average_wage <- as.matrix(.rowMeans(rwage_week_matrix,160,30))

# Relative wage

high_school_wage <- rwage_week_matrix[1:40,]
high_school_fixed_weight <- fixed_employment_share[1:40]/
                            sum(fixed_employment_share[1:40])

graduate_wage <- rwage_week_matrix[121:160,]
graduate_fixed_weight <- fixed_employment_share[121:160]/
                         sum(fixed_employment_share[121:160])

aggregate_high_school_wage <- high_school_fixed_weight %*% high_school_wage
aggregate_graduate_wage <- graduate_fixed_weight %*% graduate_wage

relative_wage <- as.vector(aggregate_graduate_wage/aggregate_high_school_wage)

# College and highschool equivalents

high_school_supply <- work_week_matrix[1:40,]
high_school_supply_weight <- average_wage[1:40]/sum(average_wage[1:40])
aggregate_high_school_supply <- high_school_supply_weight %*% high_school_supply

high_school_dropout_supply <- work_week_matrix[41:80,]
high_school_dropout_supply_weight <- average_wage[41:80]/sum(average_wage[41:80])
aggregate_high_school_dropout_supply <- high_school_dropout_supply_weight %*%
                                        high_school_dropout_supply

some_college_supply <- work_week_matrix[81:120,]
some_college_supply_weight <- average_wage[81:120]/sum(average_wage[81:120])
aggregate_some_college_supply <- some_college_supply_weight %*%
```

```
                                some_college_supply

graduate_supply <- work_week_matrix[121:160,]
graduate_supply_weight <- average_wage[121:160]/sum(average_wage[121:160])
aggregate_graduate_supply <- graduate_supply_weight %*% graduate_supply

high_school_equivalent <- aggregate_high_school_supply + 0.69 * aggregate_some_college_supply +
                          0.93 *  aggregate_high_school_dropout_supply
graduate_equivalent <- aggregate_graduate_supply + 0.29 * aggregate_some_college_supply -
                          0.05 *  aggregate_high_school_dropout_supply

relative_supply <- as.vector(graduate_equivalent/high_school_equivalent)

# Time trend

t <- as.vector(c(1:30))

# Regression

model19_3 <- lm(log(relative_wage) ~ log(relative_supply) + t)
summary(model19_3)
```