Problem Set 3

Yuqing Liu

20/03/2019

```r
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------------------------- tidyverse 1.2.1
```

```
## 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
```

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

```r
library(lubridate)
```

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

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

```r
data_00 = read.csv("data_01.csv")
```

```r
# creating five experience groups
data_02 = matrix(0,2009-1962+1,2)
Y = dim(data_02)[1]
data_02[,2] = 1962:2009
table1 = data.frame(exp = c(5,15,25,35,45), deduc_1 = c(0), deduc_2 = c(0),
                    deduc_3 = c(0), deduc_4 = c(0), drace_1 = 0, drace_2 = 0)
table2 = data.frame(exp = c(5,15,25,35,45), deduc_1 = c(1), deduc_2 = c(0),
                    deduc_3 = c(0), deduc_4 = c(0), drace_1 = 0, drace_2 = 0)
table3 = data.frame(exp = c(5,15,25,35,45), deduc_1 = c(0), deduc_2 = c(1),
                    deduc_3 = c(0), deduc_4 = c(0), drace_1 = 0, drace_2 = 0)
table4 = data.frame(exp = c(5,15,25,35,45), deduc_1 = c(0), deduc_2 = c(0),
                    deduc_3 = c(1), deduc_4 = c(0), drace_1 = 0, drace_2 = 0)
table5 = data.frame(exp = c(5,15,25,35,45), deduc_1 = c(0), deduc_2 = c(0),
                    deduc_3 = c(0), deduc_4 = c(1), drace_1 = 0, drace_2 = 0)
exp_edu = matrix(0, 10, 5)
mat1 = matrix(0,5,5)
predict_all = array(0, dim = c(10, 5, Y))

# separate female data
data_female <- filter(data_00, sex == 2)

#seperate male data
data_male <- filter(data_00, sex == 1)

# Prediction
# regression lrwage = deduc_1+deduc_2+deduc_3+deduc_4+poly(exp,4,raw=TRUE)+
#                     (deduc_1+deduc_2+deduc_3+deduc_4):poly(exp,4,raw = TRUE)+
#                     drace_1+drace_2+(deduc_1+deduc_2+deduc_3+deduc_4):exp:(drace_1+drace_2)
for (i in 1:Y){
  data_f <- filter(data_female, year == (1963+i))  # predicted values for female (sex == 2)
  coef_f = lm(lrwage ~ deduc_1+deduc_2+deduc_3+deduc_4+
              poly(exp,4,raw=TRUE)+(deduc_1+deduc_2+deduc_3+deduc_4):poly(exp,4,raw = TRUE)+
              drace_1+drace_2+(deduc_1+deduc_2+deduc_3+deduc_4):exp:(drace_1+drace_2), data = data_f)
  pred1_f = predict(coef_f, table1)
  pred2_f = predict(coef_f, table2)
  pred3_f = predict(coef_f, table3)
  pred4_f = predict(coef_f, table4)
  pred5_f = predict(coef_f, table5)
  mat_f = matrix(c(pred1_f,pred2_f,pred3_f,pred4_f,pred5_f),5,5, byrow = T)

  # Predicted values for male (sex == 1)
  data_m <- filter(data_male, year == (1963+i))
  coef_m = lm(lrwage ~ deduc_1+deduc_2+deduc_3+deduc_4+
              poly(exp,4,raw=TRUE)+(deduc_1+deduc_2+deduc_3+deduc_4):poly(exp,4,raw = TRUE)+
              drace_1+drace_2+(deduc_1+deduc_2+deduc_3+deduc_4):exp:(drace_1+drace_2), data = data_m)
  pred1_m = predict(coef_m, table1)
  pred2_m = predict(coef_m, table2)
  pred3_m = predict(coef_m, table3)
  pred4_m = predict(coef_m, table4)
  pred5_m = predict(coef_m, table5)
  mat_m = matrix(c(pred1_m,pred2_m,pred3_m,pred4_m,pred5_m),5,5, byrow = T)
  exp_edu[6:10,] = mat_m
  exp_edu[1:5,] = mat_f
  # predict value for each i (year)
  predict_all[,,i] = exp_edu
}
```

Problem Set 3

highshool dropout: deduc_1

some college: deduc_2

4 years college: deduc_3

more than college: deuc_4

1963——2017

```r
# Create high school and college equivalent
data_00 <- data_00 %>%
  mutate(edu = ifelse(deduc_1==1,1,ifelse(deduc_2==1,3,ifelse(deduc_3==1,4,ifelse(deduc_4==1,5,2)))))
attach(data_00)

data_002 <- filter(data_00, year >= 1963 & year <= 2017, edu==1, sex == 2)
data_hd<-data_002 %>%
  group_by(year, edu) %>%
  summarise(meanwage_hd = mean(rwage))

data_003 <- filter(data_00, year >= 1963 & year <= 2017, edu==2, sex == 2)
data_hs<-data_003 %>%
  group_by(year, edu) %>%
  summarise(meanwage_hs = mean(rwage))

data_004 <- filter(data_00, year >= 1963 & year <= 2017, edu==3, sex == 2)
data_sc<-data_004 %>%
  group_by(year, edu) %>%
  summarise(meanwage_sc = mean(rwage))

data_005 <- filter(data_00, year >= 1963 & year <= 2017, edu==4, sex == 2)
data_cg<-data_005 %>%
  group_by(year, edu) %>%
  summarise(meanwage_cg = mean(rwage))

m1<- merge(x=data_sc,y=data_cg,by=c("year"))
m2<- merge(x=data_hd,y=data_hs,by=c("year"))
m<-merge(x=m1, y=m2, by=c("year"))


lm(meanwage_hd~meanwage_hs+ meanwage_cg, data = m)
```

```
##
## Call:
## lm(formula = meanwage_hd ~ meanwage_hs + meanwage_cg, data = m)
##
## Coefficients:
## (Intercept)  meanwage_hs  meanwage_cg
##     18.4900       0.9382      -0.1192
```

```
lm(meanwage_sc~ meanwage_hs + meanwage_cg, data = m)
```

```
##
## Call:
## lm(formula = meanwage_sc ~ meanwage_hs + meanwage_cg, data = m)
##
## Coefficients:
## (Intercept)  meanwage_hs  meanwage_cg
##      0.3143       0.7381       0.2596
```

```
df_hd=data.frame(aggregate(edu~year,FUN = length, data=data_002))
names(df_hd)[names(df_hd)=="edu"] <- "num_hd"
df_hs=data.frame(aggregate(edu~year,FUN = length, data=data_003))
names(df_hs)[names(df_hs)=="edu"] <- "num_hs"
df_sc=data.frame(aggregate(edu~year,FUN = length, data=data_004))
names(df_sc)[names(df_sc)=="edu"] <- "num_sc"
df_cg=data.frame(aggregate(edu~year,FUN = length, data=data_005))
names(df_cg)[names(df_cg)=="edu"] <- "num_cg"

df_h<-merge(x=df_hd,y=df_hs,by=c("year"))
df_c<-merge(x=df_sc,y=df_cg,by=c("year"))
df<-merge(x=df_h,y=df_c,by=c("year"))
```

```
df_1<-df%>%
  mutate(hs_equivalent = num_hs + 0.9382*num_hd+0.7381*num_sc)%>%
  mutate(co_equivalent = num_cg - 0.1192*num_hd+0.2596*num_sc)%>%
  mutate(supply=log(co_equivalent)-log(hs_equivalent))

dataa <- merge(x=df_1, y=to_present, by=c("year"))
dataa <- dataa %>%
  mutate(time=year-1963)
c1 <- lm(wage_gap~supply+time, data = dataa)
summary(c1)
```

```
##
## Call:
## lm(formula = wage_gap ~ supply + time, data = dataa)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.7170  -0.2347  -0.0684   0.5468   3.2087
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.722791   3.071281   0.235    0.815
## supply       0.291653   1.555346   0.188    0.852
## time        -0.004657   0.043879  -0.106    0.916
##
## Residual standard error: 2.089 on 43 degrees of freedom
## Multiple R-squared:  0.001051,   Adjusted R-squared:  -0.04541
## F-statistic: 0.02261 on 2 and 43 DF,  p-value: 0.9777
```

Results for the year 1963-2017

Coefficients:

(Intercept) supply time

0.722791 0.291653 -0.004657

Regressing the average wage series for highschool dropout and some college group on the wage series for high school graduates and for college graduates over the 1963 - 2017 period.

The regression results suggest that one person with some college is equivalent to a total of 0.73 of a high school graduate and 0.25 of a college graduate, while a high school dropout is equivalent to 0.9382of a high school graduate and -0.12 of a college graduate.

Running the regression of highschool - college wage gap on the supply ratio and time series gives the coefficients of 0.291 on the supply and -0.0046 on the time series, the constnt is 0.722. The standard error for the supply is 3.07 and 0.043 for the time series variable. P-value is 0.9777.

1963 - 1987

```
data_00 <- data_00 %>%
  mutate(edu = ifelse(deduc_1==1,1,ifelse(deduc_2==1,3,ifelse(deduc_3==1,4,ifelse(deduc_4==1,5,2)))))
```

```r
attach(data_00)

data_002 <- filter(data_00, year >= 1963 & year <= 1987, edu==1, sex == 2)
data_hd<-data_002 %>%
  group_by(year, edu) %>%
  summarise(meanwage_hd = mean(rwage))

data_003 <- filter(data_00, year >= 1963 & year <= 1987, edu==2, sex == 2)
data_hs<-data_003 %>%
  group_by(year, edu) %>%
  summarise(meanwage_hs = mean(rwage))

data_004 <- filter(data_00, year >= 1963 & year <= 1987, edu==3, sex == 2)
data_sc<-data_004 %>%
  group_by(year, edu) %>%
  summarise(meanwage_sc = mean(rwage))

data_005 <- filter(data_00, year >= 1963 & year <= 1987, edu==4, sex == 2)
data_cg<-data_005 %>%
  group_by(year, edu) %>%
  summarise(meanwage_cg = mean(rwage))

m1<- merge(x=data_sc,y=data_cg,by=c("year"))
m2<- merge(x=data_hd,y=data_hs,by=c("year"))
m<-merge(x=m1, y=m2, by=c("year"))


lm(meanwage_hd~meanwage_hs+ meanwage_cg, data = m)
```

```
##
## Call:
## lm(formula = meanwage_hd ~ meanwage_hs + meanwage_cg, data = m)
##
## Coefficients:
## (Intercept)  meanwage_hs  meanwage_cg
##     3.85792      0.91191     -0.06771
```

```r
lm(meanwage_sc~ meanwage_hs + meanwage_cg, data = m)
```

```
##
## Call:
## lm(formula = meanwage_sc ~ meanwage_hs + meanwage_cg, data = m)
##
## Coefficients:
## (Intercept)  meanwage_hs  meanwage_cg
##      6.7643       0.6512       0.3021
```

```r
df_hd=data.frame(aggregate(edu~year,FUN = length, data=data_002))
names(df_hd)[names(df_hd)=="edu"] <- "num_hd"
df_hs=data.frame(aggregate(edu~year,FUN = length, data=data_003))
names(df_hs)[names(df_hs)=="edu"] <- "num_hs"
df_sc=data.frame(aggregate(edu~year,FUN = length, data=data_004))
names(df_sc)[names(df_sc)=="edu"] <- "num_sc"
df_cg=data.frame(aggregate(edu~year,FUN = length, data=data_005))
names(df_cg)[names(df_cg)=="edu"] <- "num_cg"
```

```r
df_h<-merge(x=df_hd,y=df_hs,by=c("year"))
df_c<-merge(x=df_sc,y=df_cg,by=c("year"))
df<-merge(x=df_h,y=df_c,by=c("year"))

df_1<-df%>%
  mutate(hs_equivalent = num_hs + 0.91191*num_hd+0.6512*num_sc)%>%
  mutate(co_equivalent = num_cg - 0.06771*num_hd+0.3021*num_sc)%>%
  mutate(supply=log(co_equivalent)-log(hs_equivalent))

dataa <- merge(x=df_1, y=to_eighty_seven, by=c("year"))
dataa <- dataa %>%
  mutate(time=year-1963)
c2<-lm(wage_gap~supply+time, data = dataa)
summary(c2)
```

```
##
## Call:
## lm(formula = wage_gap ~ supply + time, data = dataa)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.92319 -0.04189 -0.01677  0.03781  1.11131
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.84624    0.85616  -3.324  0.00322 **
## supply      -1.37304    0.42727  -3.213  0.00417 **
## time         0.07675    0.02068   3.711  0.00129 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3189 on 21 degrees of freedom
## Multiple R-squared:  0.3967, Adjusted R-squared:  0.3393
## F-statistic: 6.906 on 2 and 21 DF,  p-value: 0.004957
```

Results for the year 1963-1987

Regressing the average wage series for highschool dropout and some college group on the wage series for high school graduates and for college graduates over the 1963 - 1987 period.

The regression results suggest that one person with some college is equivalent to a total of 0.65 of a high school graduate and 0.30 of a college graduate, while a high school dropout is equivalent to 0.91191 of a high school graduate and -0.06 of a college graduate.

Running the regression of highschool - college wage gap on the supply ratio and time series gives the coefficients of -1.373 on the supply and 0.076 on the time series, the constant is -2.84. The standard error for the supply is 0.427 and 0.020 for the time series variable.

All te coefficients are aignificant at the 0.001 level. Therefore, the equation is,

```
log(w1/w2) = -1.373 log (x1/x2) + 0.076 time + comstant,
            (0.427)              (0.020)
```

The results is different from the results in the paper. One reason is the difference in classifying different educational level. Another reason is the difference in calculating the wage gap. In this regression, I use the wage gap results from problem set 2, however, the wage gap series is not exactly the same as the wage series in the paper. Therefore, the coefficients might also be different.

1988 - 2017

```
data_00 <- data_00 %>%
  mutate(edu = ifelse(deduc_1==1,1,ifelse(deduc_2==1,3,ifelse(deduc_3==1,4,ifelse(deduc_4==1,5,2)))))
attach(data_00)

data_002 <- filter(data_00, year >= 1988 & year <= 2017, edu==1, sex == 2)
data_hd<-data_002 %>%
  group_by(year, edu) %>%
  summarise(meanwage_hd = mean(rwage))

data_003 <- filter(data_00, year >= 1988 & year <= 2017, edu==2, sex == 2)
data_hs<-data_003 %>%
  group_by(year, edu) %>%
  summarise(meanwage_hs = mean(rwage))

data_004 <- filter(data_00, year >= 1988 & year <= 2017, edu==3, sex == 2)
data_sc<-data_004 %>%
  group_by(year, edu) %>%
  summarise(meanwage_sc = mean(rwage))

data_005 <- filter(data_00, year >= 1988 & year <= 2017, edu==4, sex == 2)
data_cg<-data_005 %>%
  group_by(year, edu) %>%
  summarise(meanwage_cg = mean(rwage))

m1<- merge(x=data_sc,y=data_cg,by=c("year"))
m2<- merge(x=data_hd,y=data_hs,by=c("year"))
m<-merge(x=m1, y=m2, by=c("year"))


lm(meanwage_hd~meanwage_hs+ meanwage_cg, data = m)
```

```
##
## Call:
## lm(formula = meanwage_hd ~ meanwage_hs + meanwage_cg, data = m)
##
## Coefficients:
## (Intercept)  meanwage_hs  meanwage_cg
##     38.9794       0.9342      -0.2286
```

```
lm(meanwage_sc~ meanwage_hs + meanwage_cg, data = m)
```

```
##
## Call:
## lm(formula = meanwage_sc ~ meanwage_hs + meanwage_cg, data = m)
##
## Coefficients:
## (Intercept)  meanwage_hs  meanwage_cg
##      6.6354       0.6908       0.2541
```

```
df_hd=data.frame(aggregate(edu~year,FUN = length, data=data_002))
names(df_hd)[names(df_hd)=="edu"] <- "num_hd"
df_hs=data.frame(aggregate(edu~year,FUN = length, data=data_003))
names(df_hs)[names(df_hs)=="edu"] <- "num_hs"
df_sc=data.frame(aggregate(edu~year,FUN = length, data=data_004))
names(df_sc)[names(df_sc)=="edu"] <- "num_sc"
df_cg=data.frame(aggregate(edu~year,FUN = length, data=data_005))
names(df_cg)[names(df_cg)=="edu"] <- "num_cg"

df_h<-merge(x=df_hd,y=df_hs,by=c("year"))
df_c<-merge(x=df_sc,y=df_cg,by=c("year"))
df<-merge(x=df_h,y=df_c,by=c("year"))

df_1<-df%>%
  mutate(hs_equivalent = num_hs + 0.9342*num_hd+0.6908*num_sc)%>%
  mutate(co_equivalent = num_cg - 0.2286*num_hd+0.2541*num_sc)%>%
  mutate(supply=log(co_equivalent)-log(hs_equivalent))

dataa <- merge(x=df_1, y=eight_to_present, by=c("year"))
dataa <- dataa %>%
  mutate(time=year-1963)
c3<-lm(wage_gap~supply+time, data = dataa)
summary(c3)
```

```
##
## Call:
## lm(formula = wage_gap ~ supply + time, data = dataa)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.4891   0.0116   0.7048   0.9757   1.5744
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.0389    14.0304   0.502    0.622
## supply        2.7947     6.1177   0.457    0.653
## time         -0.1019     0.2065  -0.493    0.627
##
## Residual standard error: 3.001 on 19 degrees of freedom
## Multiple R-squared:  0.0128, Adjusted R-squared:  -0.09112
## F-statistic: 0.1232 on 2 and 19 DF,  p-value: 0.8848
```

Results for the year 1988-2017

Regressing the average wage series for highschool dropout and some college group on the wage series for high school graduates and for college graduates over the 1988 - 2017 period.

The regression results suggest that one person with some college is equivalent to a total of 0.69 of a high school graduate and 0.254 of a college graduate, while a high school dropout is equivalent to 0.9342 of a high school graduate and -0.2286 of a college graduate.

Running the regression of highschool - college wage gap on the supply ratio and time series gives the coefficients of -2.79 on the supply and -0.10 on the time series, the constant is 7.03. The coefficients are not very significant.

In fact, the log wage gap series we derived from problem set 2 for the year from 1988 to 2017 are not very accurate. There are several abnormal fluctuations on the curve. It might be the problem of the code or the data set. That could be a reason for the insignificance of the coefficients.