COMM8102 Assignment 1 coding part

Yuhao Liu z5097536

```
library(readxl)
dat <- read_excel("cps09mar.xlsx")
```

Excercise 3.24

First we extract the entries from the dataset and construct the variables referring to section 3.22 and section 3.25 from the textbook. We only looking at those entries that are single Asian man with less that 45 years of experience.

```
#3.24
#single Asian man
sam <- (dat[,11]==4)&(dat[,12]==7)&(dat[,2]==0)
datsam <- dat[sam,]
#experience and experience squared
experience<-datsam[,1]-datsam[,4]-6
exp2<-experience^2/100
#less than 45 year experience
sam<-experience<45
datsam <- datsam[sam,]
#logwage
Y<-as.matrix(log(datsam[,5]/(datsam[,6]*datsam[,7])))
#new experience and its squared
experience<-experience[sam]
exp2<-experience^2/100
#X
X<-as.matrix(cbind(datsam[,4],experience,exp2,matrix(1,nrow(datsam),1)))
```

(a)

Now we estimate 3.49.

```
#estimation of 3.49
beta349<-solve(t(X)%*%X,t(X)%*%Y)
rownames(beta349)<-c('education','experience','experience^2/100','intercept')
print(beta349)
```

```
##                     earnings
## education         0.14430729
## experience        0.04263326
## experience^2/100 -0.09505636
## intercept         0.53089068
```

We can see that the coefficients are same as 3.49. The R^2 and **Sum of squared errors** are

```
#R^2 and SSE
Y_hat<-as.matrix(0.144*datsam$education+0.043*experience-0.095/100*experience^2+0.531)
#y bar
Y_bar<-mean(Y)
```

```
#R^2
R_squared<-sum((Y_hat-Y_bar)^2)/sum((Y-Y_bar)^2)
#sum of squared errors
SSE<-sum((Y-Y_hat)^2)
#print result
cat('R^2 = ', R_squared)
```

```
## R^2 =  0.3886973
```

```
cat('SSE = ', SSE)
```

```
## SSE =  82.50921
```

The R^2 is 0.39 and SSE is 82.5.

(b)

Now, we try the residual regression approach. First, we regress **log(wage)** on **experience and its square**.

```
#regress logwage on experience and its sqaure
X1<-as.matrix(cbind(experience,exp2,matrix(1,nrow(datsam),1)))
beta1<-solve(t(X1)%*%X1,t(X1)%*%Y)
```

Then, we we regress **education** on **experience and its square**.

```
#regress education on experience and its sqaure
beta2<-solve(t(X1)%*%X1,t(X1)%*%as.matrix(datsam[,4]))
```

Finally, we regress the residuals on the residuals.

```
#residuals and regression
e_1tilda<-Y-X1%*%beta1
x_2tilda<-as.matrix(datsam[,4]-X1%*%beta2)
xres<-as.matrix(cbind(x_2tilda,matrix(1,nrow(datsam),1)))
xxres<-t(xres)%*%xres
xyres<-t(xres)%*%e_1tilda
beta2_hat<-solve(xxres,xyres)
rownames(beta2_hat)<-c('beta2_hat','intercept')
res_hat<-e_1tilda-xres%*%beta2_hat
print(beta2_hat)
```

```
##                earnings
## beta2_hat  1.443073e-01
## intercept -6.852850e-16
```

The R^2 and **Sum of squared errors** in the residual regression approach is evalued by

```
#sse_NEW
SSE_NEW<-sum(res_hat^2)
#new r sqaured
rsquared_new<-1-SSE_NEW/sum((Y-Y_bar)^2)
#print result
cat('The Re-estimate  slope on education is', beta2_hat[1])
```

```
## The Re-estimate  slope on education is 0.1443073
```

```
cat('The new R^2 is', rsquared_new )
```

```
## The new R^2 is 0.3893207
```

```
cat('The new SSE is', SSE_NEW )
```

```
## The new SSE is 82.505
```

The slope coefficient on education equals to the value in (3.49). When we regress **log(wage)** on **experience and its square**, the residuals are the **log(wage)** change that cannot be explained by **experience and its square**. When we regress **education** on **experience and its square**, the residuals are **education** change that cannot be explained by **experience and its square**. Then we regress the residuals on the residuals, we get the coefficient that **log(wage)** change attribute to only **education**, which is the same as the explaination of the slope coefficient of **education** in OLS.

(c)

The R^2 and **Sum of squared errors** from part (a) and (b) are equal. In both regression, the regressors are the same and they are fully used to explain the total variation, so the residuals should also be the same.

Excercise 3.26

(a)

First, we extract the entries from the dataset.

```
#white male hispanic
wmh<-(dat[,3]==1)&(dat[,11]==1)&(dat[,2]==0)
datwmh<-dat[wmh,]
```

Then construct **log(wage)**, **education** and **experience and its square** as before.

```
#logwage
Y<-as.matrix(log(datwmh[,5]/(datwmh[,6]*datwmh[,7])))
#education
edu<-as.matrix(datwmh[,4])
#experience
expwmh<-as.matrix(datwmh[,1]-datwmh[,4]-6)
#experience sqaure
expwmh2<-as.matrix(expwmh^2/100)
```

Next, we codify the dummy variable for regions. We have three dummy variable for the four different regions. The excluded group is when all dummy variables are 0.

```
#dummy variables for region
#NE
x1<-as.numeric(datwmh[,10]==1)
#S
x2<-as.numeric(datwmh[,10]==3)
#W
x3<-as.numeric(datwmh[,10]==4)
```

We do the similary thing to the marital status with four dummy variables.

```
#dummy variables for marital
#Married
xxx1<-as.numeric(datwmh[,12]==1|datwmh[,12]==2|datwmh[,12]==3)
#Widowed
xxx2<-as.numeric(datwmh[,12]==4)
#Divorced
xxx3<-as.numeric(datwmh[,12]==5)
```

```
#Separated
xxx4<-as.numeric(datwmh[,12]==6)
```

Finally, we do the regression.

```
#regression
X<-cbind(edu,expwmh,expwmh2,x1,x2,x3,xxx1,xxx2,xxx3,xxx4,matrix(1,nrow(datwmh),1))
beta326<-solve(t(X)%*%X,t(X)%*%Y)
rownames(beta326)<-c('education','experience','(experience^2)/100','Northeast',
'South','West','married','widowed','divorced','separated','intercept')
print(beta326)
```

```
##                       earnings
## education           0.08833357
## experience          0.02792293
## (experience^2)/100 -0.03646362
## Northeast           0.06163684
## South              -0.06753707
## West                0.02011173
## married             0.17796669
## widowed             0.24297545
## divorced            0.07870880
## separated           0.01694743
## intercept           1.19175223
```