Test Exercise 4

zouxia

Endogenety

1. Use OLS to estimate the parameters of the mode

logw=β1+ β2 educ+β3 exper+ β4 exper^2+ β5 smsa+ β6 south

Give an interpretation to the estimated β2 coefficient.

```
setwd("D:/R programming/econometric")
datatest<-read.csv("TestExer4_Wage-round1.csv",header = T,stringsAsFact
ors = F)
names(datatest)
```

```
## [1] "logw"  "educ"  "age"   "exper" "smsa"  "south" "nearc" "daded"
"momed"
```

```
datatest$squareexper<-(datatest$exper)^2
est1<-summary(lm(logw~educ+exper+squareexper+smsa+south,data=datatest))
est1$coefficients
```

```
##                 Estimate   Std. Error    t value      Pr(>|t|)
## (Intercept)  4.611014446 0.0678950310  67.913872  0.000000e+00
## educ         0.081579706 0.0034990436  23.314859 1.191902e-110
## exper        0.083835685 0.0067735171  12.376980  2.360425e-34
## squareexper -0.002202115 0.0003238327  -6.800161  1.255782e-11
## smsa         0.150800573 0.0158359911   9.522648  3.351249e-21
## south       -0.175176080 0.0146486420 -11.958520  3.122054e-32
```

And the β2 equals 0.081579706,which means in average increase one year of education will increase the wage 0.081579%.

2. OLS may be inconsistent in this case as educ and exper may be endogenous. Give a reason why this may be the case. Also indicate whether the estimate in part (a) is still useful. The educ may be endogenous because some variable like smsa which indicating a person live in metropolitan area or not will both influence the year of education and the wages. It's obivious that if a person live in metropolitan area he is more like to receive more education and the wage in metropolitan is more high. The exper may be endogenouds because the variable educ will both influence the educ and logw. The reason is that person who is educated will have less work experience in year and the wages tend to increase with more education as the exercise mentioned. So the estimate in Part(a) is not useful anymore.Because the coefficient of educ and exper is biased.

3. Give a motivation why age and age^2 can be used as instruments for exper and exper^2. As we know, the older you are the longer you work. So the age and age^2 are definetly correlate with the exper and exper^2 respectively. However how much the company will give you wouldn't depend on the age.

4. Run the first-stage regression for educ for the two-stage least squares estimation of the parameters in the model above when age, age^2, nearc, dadeduc, and momeduc are used as additional instruments. What do you conclude about the suitability of these instruments for schooling?

```
datatest$squareage<-(datatest$age)^2
lsts1<-summary(lm(educ~age+squareage+nearc+daded+momed,data = datatest))
lsts1$coefficients
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -5.92327296 4.01050220 -1.476940 1.397964e-01
## age          0.99255040 0.28105960  3.531459 4.194907e-04
## squareage   -0.01707535 0.00487832 -3.500252 4.715804e-04
## nearc        0.52875137 0.09269788  5.704029 1.283849e-08
## daded        0.20204775 0.01566472 12.898267 4.361368e-37
## momed        0.24837851 0.01703554 14.580016 1.414420e-46
```

As we can see in the result, the t-statistic for age squareage nearc,daded,and momed are quite large,which indicating these three instruments are significantly correlated with educ. However the estimate coefficient for variable squareage is quite small. So i think all the variables are suitable.

5.Estimate the parameters of the model for log wage using two-stage least squares. Compare your result to the estimate in part (a).

```
#using the tsls function in sem package to do the two-stage least squar
es estimation.
library(sem)
tsls<-summary(tsls(logw~educ+exper+squareexper+smsa+south,~smsa+south+n
earc+daded+momed+age+squareage,data = datatest))
tsls$coefficients
```

```
##                 Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept)  4.416903900 0.1154207774  38.267840 0.000000e+00
## educ         0.099842919 0.0065738329  15.187931 0.000000e+00
## exper        0.072866858 0.0167133759   4.359793 1.345584e-05
## squareexper -0.001639293 0.0008381474  -1.955853 5.057503e-02
## smsa         0.134937031 0.0167695244   8.046563 1.332268e-15
## south       -0.158986861 0.0156854387 -10.135953 0.000000e+00
```

```
est1$coefficients
```

```
##                 Estimate   Std. Error    t value      Pr(>|t|)
## (Intercept)  4.611014446 0.0678950310  67.913872  0.000000e+00
## educ         0.081579706 0.0034990436  23.314859 1.191902e-110
## exper        0.083835685 0.0067735171  12.376980  2.360425e-34
```

```
## squareexper -0.002202115 0.0003238327  -6.800161  1.255782e-11
## smsa         0.150800573 0.0158359911   9.522648  3.351249e-21
## south       -0.175176080 0.0146486420 -11.958520  3.122054e-32
```

As we can see in the result, the β2 is large in the tsls method. And the β3 for exper is a litter bit small. The squareexper became much more close to 0 and the t value became smaller. On the other hand, the beta2 increase in the two satge estimation which indicate we have under esitimate it in the ols. This means there have factor that have positive relation with logw while have negative effect on educ. Using this way, we can say we have overestimate the β3 and β4.

6.Perform the Sargan test for validity of the instruments. What is your conclusion?

```
#calculate the predicted value of logw
logwhat<-4.416903900+0.099842919*(datatest$educ)+0.072866858*(datatest
$exper)-0.001639293*(datatest$squareexper)+0.134937031*(datatest$smsa)-
0.158986861*(datatest$south)
#calculate the residual
resid<-datatest$logw-logwhat
#regress residual on the instruments
sargantest<-summary(lm(resid~smsa+south+nearc+daded+momed+age+squareage,
data = datatest))
n<-3010
m<-7
k<-5
stat<-sargantest$r.squared*n
chisq<-qchisq(0.95,m-k)
stat
```

```
## [1] 3.702389
```

```
chisq
```

```
## [1] 5.991465
```

As we can see, the nR^2 stat is smaller than the chisq in 95% confidence interval. So i include that the instrument is unrelate with the residual.