20181016 hoemwork

RyanYip

2018骞10鏈<88>16鏃

### load data

wagexp=read.csv("./wagexp.csv",header = TRUE)

### check NA value and duplicated values

sum(is.na(wagexp)) # there is no NA value

## [1] 0

wagexp=wagexp[!duplicated(wagexp),]

there is 206 observation

predictor=c("AGE","EXP","HSDRoP","HSGRAD","ASSOC","FEMALE")  
all.model=lm(WAGE~AGE+I(AGE^2)+EXP+HSDROP+HSGRAD+ASSOC+FEMALE,data = wagexp)  
model.summary=summary(all.model)$coefficients  
confidents=confint(all.model)  
summary(all.model)

##   
## Call:  
## lm(formula = WAGE ~ AGE + I(AGE^2) + EXP + HSDROP + HSGRAD +   
## ASSOC + FEMALE, data = wagexp)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8.9088 -2.8479 -0.1728 1.9005 15.2639   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.651324 3.630121 -0.179 0.857790   
## AGE 0.684073 0.183864 3.721 0.000259 \*\*\*  
## I(AGE^2) -0.006198 0.001999 -3.101 0.002209 \*\*   
## EXP -0.071574 0.079027 -0.906 0.366205   
## HSDROP -6.583392 1.162073 -5.665 5.12e-08 \*\*\*  
## HSGRAD -4.356037 0.815277 -5.343 2.50e-07 \*\*\*  
## ASSOC -1.959340 0.926987 -2.114 0.035796 \*   
## FEMALE -2.663212 0.645894 -4.123 5.49e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.132 on 198 degrees of freedom  
## Multiple R-squared: 0.3989, Adjusted R-squared: 0.3777   
## F-statistic: 18.77 on 7 and 198 DF, p-value: < 2.2e-16

### question 1

H0:WAGE~a1\**AGE+a2\**(AGE^2)+a4\**HSDROP+a5\**HSGRAD+a6\**ASSOC+a7\**FEMALE+a0

H1:WAGE~a1\**AGE+a2\**(AGE^2)+a3\**EXP+a4\**HSDROP+a5\**HSGRAD+a6\**ASSOC+a7\*FEMALE+a0

H0 成立的前提下: formule: t=beta3\_hat/sqrt(sd(beta3) 服从自由度为n-p的t分布

p{|t.vale|

the correlation coefficient and plot tells us that the variable \*\* AGE \*\* and **Experience** is highly correlated

### question 4

cor(wagexp[,c("AGE","FEMALE","HSDROP","HSGRAD","ASSOC","EXP")])

## AGE FEMALE HSDROP HSGRAD ASSOC  
## AGE 1.00000000 0.14468871 0.1832568 -0.11462502 -0.04763059  
## FEMALE 0.14468871 1.00000000 0.0220865 0.12510689 -0.08417942  
## HSDROP 0.18325684 0.02208650 1.0000000 -0.34381992 -0.16231723  
## HSGRAD -0.11462502 0.12510689 -0.3438199 1.00000000 -0.40404918  
## ASSOC -0.04763059 -0.08417942 -0.1623172 -0.40404918 1.00000000  
## EXP 0.92394171 0.01670644 0.2862939 -0.02402402 -0.09145799  
## EXP  
## AGE 0.92394171  
## FEMALE 0.01670644  
## HSDROP 0.28629389  
## HSGRAD -0.02402402  
## ASSOC -0.09145799  
## EXP 1.00000000

car::vif(all.model)

## AGE I(AGE^2) EXP HSDROP HSGRAD ASSOC FEMALE   
## 57.364423 45.122317 10.620281 1.736977 1.992350 1.394490 1.257599

I consider that this model suffer from multicollinearity: \* 1) the VIF value of AGE,EXP are larger than 10;

* 1. the correlation coefficient between \*\* AGE \*\* and \*\* EXPERIENCE \*\* is larger than 0.9, \*\* HSGRAD\*\* and \*\* ASSOC \*\* also has -0.4
  2. the \*\* EXPERIENCE \*\* estimate coefficient is **-0.071574**,which tell us that EXPERIENCE has negative effect on WAGE,this is incorrect ,it is not consistenct with commond sence.

### question 5

wagexp$ln.wage=log2(wagexp$WAGE)  
new.model=lm(ln.wage ~ log2(AGE)+ED\*FEMALE,data=wagexp)  
summary(new.model)

##   
## Call:  
## lm(formula = ln.wage ~ log2(AGE) + ED \* FEMALE, data = wagexp)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.61265 -0.40513 0.05933 0.43236 1.58232   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.15711 0.54658 -2.117 0.0355 \*   
## log2(AGE) 0.51568 0.09242 5.580 7.72e-08 \*\*\*  
## ED 0.13245 0.02175 6.091 5.63e-09 \*\*\*  
## FEMALE -0.56422 0.44364 -1.272 0.2049   
## ED:FEMALE 0.01495 0.03298 0.453 0.6509   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.606 on 201 degrees of freedom  
## Multiple R-squared: 0.3796, Adjusted R-squared: 0.3673   
## F-statistic: 30.75 on 4 and 201 DF, p-value: < 2.2e-16

new.model.summary=summary(new.model)$Coefficient  
new.model.confident=confint(new.model)

Based on summary of new model,the beta1 coefficiet is 0.51568 , it represent that,ln(AGE)has a positive affect on ln(wage) ,when ln(age) add one unit,and ln(wage) add 0.51568 unit

### question 6

We can know the estimate \*\* beta4 \*\* is 0.01495;When female=1,eudcation add one unit,and ln(wage) add0.1474 unit;when female=0,education add one unit,and ln(wage) add 0.13245 unit;the comparsion tell us that the sexes increse 1.1612535 rate in wage.So sex is exactly correlated with eduaction in wage.

### question 7

Run the following regression model:

* step 1) ln(wage )~ b\_0+ b\_1\**ln(AGE) + b\_2\**FEMALE + b\_3\**ED+ be\_4\*FEMALE*ED+sigma
* step 2) sigma\_hat~phi\_0+phi\_1\*ln(wage)\_hat+phi\_2\*ln(wage)^2
* step 3) Retain the R-squared value from the regression(step 2):R-squre\_phi
* step 4)Calculate the F-statistic or the chi-squared statistic: chi=n\*R-squre\_phi ~chi(3)

If either of these test statistics is significant, then you have evidence of heteroskedasticity. If not, you fail to reject the null hypothesis of homoskedasticity

lnwage\_hat=fitted(new.model) # y\_hat  
df=data.frame(res=residuals(new.model),lnwage\_hat=lnwage\_hat)  
  
  
new2.model=lm(res~lnwage\_hat+I(lnwage\_hat^2),data=df)  
summary(new2.model)

##   
## Call:  
## lm(formula = res ~ lnwage\_hat + I(lnwage\_hat^2), data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.63820 -0.38806 0.05901 0.43509 1.62120   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 1.6086 1.2686 1.268 0.206  
## lnwage\_hat -1.0809 0.8367 -1.292 0.198  
## I(lnwage\_hat^2) 0.1773 0.1365 1.299 0.195  
##   
## Residual standard error: 0.6005 on 203 degrees of freedom  
## Multiple R-squared: 0.008248, Adjusted R-squared: -0.001523   
## F-statistic: 0.8441 on 2 and 203 DF, p-value: 0.4314

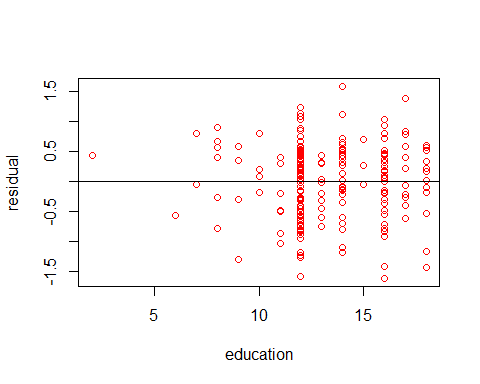
R.squre=summary(new2.model)$r.square  
  
chi=dim(df)[1]\*R.squre  
pchisq(chi,df=3)

## [1] 0.3628579

the p-value of White’s test for heteroskedasticity is 0.5723848,which is larger than 0.05.It is not sinificant.So we have not enough reason to believe that the model has heteroskedasticity

### question 8

Res=residuals(new.model)  
plot(wagexp[,"ED"],Res,col="red",xlab="education",ylab="residual")  
abline(h = 0)

 The plot show us that the distribution of the scattered points on both sides of the residual error is uniform.So I think the education variable do not look heteroskedastic.

### question 9

We can run a regression with White’s corrected standard errors(a robust linear regression),which cabn remove the effect of heteroskedasticity.

require(MASS)

## Loading required package: MASS

robust.model=rlm(ln.wage ~ log2(AGE)+ED\*FEMALE,data=wagexp)  
summary(robust.model)

##   
## Call: rlm(formula = ln.wage ~ log2(AGE) + ED \* FEMALE, data = wagexp)  
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.63101 -0.40197 0.06531 0.40302 1.59067   
##   
## Coefficients:  
## Value Std. Error t value  
## (Intercept) -1.4996 0.5535 -2.7093  
## log2(AGE) 0.5760 0.0936 6.1545  
## ED 0.1366 0.0220 6.2015  
## FEMALE -0.4904 0.4493 -1.0915  
## ED:FEMALE 0.0083 0.0334 0.2472  
##   
## Residual standard error: 0.6017 on 201 degrees of freedom

From the summary result of robust model ,wo know that beta1 is 0.5760,and in the origin model is 0.5157;and Std. Error,t- value between them are also similar.Compared to question 5, we can conclude that there is almost no difference.

### question 10

I do not consider this model suffers from heteroskedasticity.

* 1. According to the question 7,we know that White’s test for heteroskedasticity on this model is not significant.So there is not enough reason to believe that this model has heteroskedasticity.And we consider it has not.
  2. According to the question 8,the plot of residual against eduacation tells us,this model does not have heteroskedasticity.
  3. The coefficients between this model and robust model is are so similar.It is an another reason .