8.6 Consider the wage equation

$$WAGE_i = \beta_1 + \beta_2 EDUC_i + \beta_3 EXPER_i + \beta_4 METRO_i + e_i \quad \text{(XR8.6a)}$$

where wage is measured in dollars per hour, education and experience are in years, and $METRO = 1$ if the person lives in a metropolitan area. We have $N = 1000$ observations from 2013.

a. We are curious whether holding education, experience, and *METRO* constant, there is the same amount of random variation in wages for males and females. Suppose $\text{var}(e_i|\mathbf{x}_i, FEMALE = 0) = \sigma_M^2$ and $\text{var}(e_i|\mathbf{x}_i, FEMALE = 1) = \sigma_F^2$. We specifically wish to test the null hypothesis $\sigma_M^2 = \sigma_F^2$ against $\sigma_M^2 \neq \sigma_F^2$. Using 577 observations on males, we obtain the sum of squared OLS residuals, $SSE_M = 97161.9174$. The regression using data on females yields $\hat{\sigma}_F = 12.024$. Test the null hypothesis at the 5% level of significance. Clearly state the value of the test statistic and the rejection region, along with your conclusion.

b. We hypothesize that married individuals, relying on spousal support, can seek wider employment types and hence holding all else equal should have more variable wages. Suppose $\text{var}(e_i|\mathbf{x}_i, MARRIED = 0) = \sigma^2_{SINGLE}$ and $\text{var}(e_i|\mathbf{x}_i, MARRIED = 1) = \sigma^2_{MARRIED}$. Specify the null hypothesis $\sigma^2_{SINGLE} = \sigma^2_{MARRIED}$ versus the alternative hypothesis $\sigma^2_{MARRIED} > \sigma^2_{SINGLE}$. We add *FEMALE* to the wage equation as an explanatory variable, so that

$$WAGE_i = \beta_1 + \beta_2 EDUC_i + \beta_3 EXPER_i + \beta_4 METRO_i + \beta_5 FEMALE + e_i \quad \text{(XR8.6b)}$$

Using $N = 400$ observations on single individuals, OLS estimation of (XR8.6b) yields a sum of squared residuals is 56231.0382. For the 600 married individuals, the sum of squared errors is 100,703.0471. Test the null hypothesis at the 5% level of significance. Clearly state the value of the test statistic and the rejection region, along with your conclusion.

c. Following the regression in part (b), we carry out the NR^2 test using the right-hand-side variables in (XR8.6b) as candidates related to the heteroskedasticity. The value of this statistic is 59.03. What do we conclude about heteroskedasticity, at the 5% level? Does this provide evidence about the issue discussed in part (b), whether the error variation is different for married and unmarried individuals? Explain.

d. Following the regression in part (b) we carry out the White test for heteroskedasticity. The value of the test statistic is 78.82. What are the degrees of freedom of the test statistic? What is the 5% critical value for the test? What do you conclude?

e. The OLS fitted model from part (b), with usual and robust standard errors, is

$$\widehat{\text{WAGE}} = -17.77 + 2.50EDUC + 0.23EXPER + 3.23METRO - 4.20FEMALE$$

(se)	(2.36)	(0.14)	(0.031)	(1.05)	(0.81)
(robse)	(2.50)	(0.16)	(0.029)	(0.84)	(0.80)

For which coefficients have interval estimates gotten narrower? For which coefficients have interval estimates gotten wider? Is there an inconsistency in the results?

f. If we add *MARRIED* to the model in part (b), we find that its *t*-value using a White heteroskedasticity robust standard error is about 1.0. Does this conflict with, or is it compatible with, the result in (b) concerning heteroskedasticity? Explain.

c. Since we have EDUC, EXPER, METRO and FEMALE four variables, the degree of freedom $= 5-1$.

Thus, the test statistic $NR^2 \sim \chi^2_4$, and with $\alpha = 0.05$, $\chi^2_{0.95,4} = 9.48$

Since $59.03 > 9.48$, we reject the null hypothesis.

Therefore, they related to the heteroskedasticity.

It provide evidence about the issue discussed in part (b), that is, the error variation is not the same for all observation based on the heteroskedasticity.

e. Since the Interval $= \{\hat{b}_i - se(b_i)t_c, \hat{b}_i + se(b_i)t_c\}$, the length is $2se(b_i)t_c$.

Since $2t_c$ is a constant, the standard error is greater, the length is wider.

According to the given se and robust se, For b_1 and b_2, their robust standard error is greater $\Rightarrow$ the interval estimates of b_1 and b_2 are wider.

For b_3, b_4, b_5, their robust standard error is smaller $\Rightarrow$ the interval estimates of b_3, b_4 and b_5 are narrower.

$\Rightarrow$ No inconsistency.

f. Since we state heteroskedasticity in part (b), using robust se makes more reliable. Thus, it does not conflict with part(b).

a. $H_0: \frac{\sigma_M^2}{\sigma_F^2} = 1$

$H_1: \frac{\sigma_M^2}{\sigma_F^2} \neq 1$

, $F_{(0.975,573,419)} = 1.196$, $F_{(0.025,573,419)} = 0.839$

$\Rightarrow$ The rejection region $= \{ t > F_{(0.975,573,419)}$ or $t < F_{(0.025,573,419)} : t \text{ is the test-statistic} \}$

Since $SSE_M = 97161.9174 \Rightarrow \sigma_M^2 = \frac{97161.9174}{575} = 168.97725$

Then the test-statistic $t = \frac{168.97725}{(12.024)^2} = 1.16877$, which is in the rejection region, so we reject H_0.

b. $H_0: \frac{\sigma^2_{married}}{\sigma^2_{single}} = 1$

$H_1: \frac{\sigma^2_{married}}{\sigma^2_{single}} > 1$

, $F_{(0.95,595,395)} = 1.18$

$\Rightarrow$ The rejection region $= \{ t > F_{(595,395)} : t \text{ is the test-statistic} \}$

Since $SSE_{single} = 56231.0382$, $\sigma^2_{single} = \frac{56231.0382}{395} = 142.35705$

Since $SSE_{married} = 100,703.0417$, $\sigma^2_{married} = \frac{100,703.0471}{595} = 169.24882$

Thus, the test-statistic $t = \frac{169.24882}{142.35705} = 1.18894$, which is greater than 1.18 $\Rightarrow$ we reject H_0.

d.

The White test uses: $Z_2 = EDUC$

$Z_3 = EXPER$

$Z_4 = Metro$

$Z_5 = Female$

$Z_6 = EDUC^2$

$Z_7 = EXPER^2$

$Z_8 = EDUC \times EXPER$

$Z_9 = EDUC \times Metro$

$Z_{10} = EDUC \times Female$

$Z_{11} = EXPER \times Metro$

$Z_{12} = EXPER \times Female$

$Z_{13} = Metro \times female$

$\Rightarrow S = 13$, so we have 12 degree of freedom

The critical value $\chi^2_{(0.95,12)} = 21.026$

To sum up, since the test statistic is 78.82, which is greater than 21.026, we reject H_0.

8.16 A sample of 200 Chicago households was taken to investigate how far American households tend to travel when they take a vacation. Consider the model

$$MILES = \beta_1 + \beta_2 INCOME + \beta_3 AGE + \beta_4 KIDS + e$$

MILES is miles driven per year, *INCOME* is measured in $1000 units, *AGE* is the average age of the adult members of the household, and *KIDS* is the number of children.

a. Use the data file *vacation* to estimate the model by OLS. Construct a 95% interval estimate for the effect of one more child on miles traveled, holding the two other variables constant.

b. Plot the OLS residuals versus *INCOME* and *AGE*. Do you observe any patterns suggesting that heteroskedasticity is present?

c. Sort the data according to increasing magnitude of income. Estimate the model using the first 90 observations and again using the last 90 observations. Carry out the Goldfeld–Quandt test for heteroskedastic errors at the 5% level. State the null and alternative hypotheses.

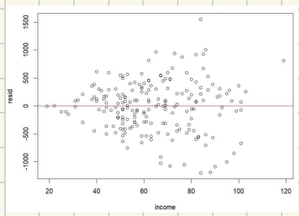
d. Estimate the model by OLS using heteroskedasticity robust standard errors. Construct a 95% interval estimate for the effect of one more child on miles traveled, holding the two other variables constant. How does this interval estimate compare to the one in (a)?

e. Obtain GLS estimates assuming $\sigma_i^2 = \sigma^2 INCOME_i^2$. Using both conventional GLS and robust GLS standard errors, construct a 95% interval estimate for the effect of one more child on miles traveled, holding the two other variables constant. How do these interval estimates compare to the ones in (a) and (d)?

a.

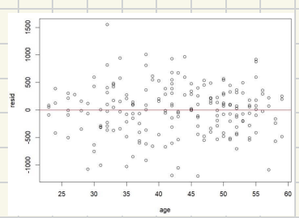
```
> cat(lower, upper)
-135.3298 -28.32302
```

b.



As the income becomes larger, the residual becomes greater.

It suggests that heteroskedasticity is present.



There is no any unusual pattern.

8.18 Consider the wage equation,

$$\ln(WAGE_i) = \beta_1 + \beta_2 EDUC_i + \beta_3 EXPER_i + \beta_4 EXPER_i^2 + \beta_5 FEMALE_i + \beta_6 BLACK + \beta_7 METRO_i + \beta_8 SOUTH_i + \beta_9 MIDWEST_i + \beta_{10} WEST + e_i$$

where *WAGE* is measured in dollars per hour, education and experience are in years, and *METRO* = 1 if the person lives in a metropolitan area. Use the data file *cps5* for the exercise.

a. We are curious whether holding education, experience, and *METRO* equal, there is the same amount of random variation in wages for males and females. Suppose $\text{var}(e_i|\mathbf{x}_i, FEMALE = 0) = \sigma_M^2$ and $\text{var}(e_i|\mathbf{x}_i, FEMALE = 1) = \sigma_F^2$. We specifically wish to test the null hypothesis $\sigma_M^2 = \sigma_F^2$ against $\sigma_M^2 \neq \sigma_F^2$. Carry out a Goldfeld–Quandt test of the null hypothesis at the 5% level of significance. Clearly state the value of the test statistic and the rejection region, along with your conclusion.

b. Estimate the model by OLS. Carry out the NR^2 test using the right-hand-side variables *METRO*, *FEMALE*, *BLACK* as candidates related to the heteroskedasticity. What do we conclude about heteroskedasticity, at the 1% level? Do these results support your conclusions in (a)? Repeat the test using all model explanatory variables as candidates related to the heteroskedasticity.

c. Carry out the White test for heteroskedasticity. What is the 5% critical value for the test? What do you conclude?

d. Estimate the model by OLS with White heteroskedasticity robust standard errors. Compared to OLS with conventional standard errors, for which coefficients have interval estimates gotten narrower? For which coefficients have interval estimates gotten wider? Is there an inconsistency in the results?

e. Obtain FGLS estimates using candidate variables *METRO* and *EXPER*. How do the interval estimates compare to OLS with robust standard errors, from part (d)?

f. Obtain FGLS estimates with robust standard errors using candidate variables *METRO* and *EXPER*. How do the interval estimates compare to those in part (e) and OLS with robust standard errors, from part (d)?

g. If reporting the results of this model in a research paper which one set of estimates would you present? Explain your choice.

a.
```
> cat(flc, fuc, fstat)
0.9451213 1.057883 0.9489479
```
The rejection region = {fstat < flc or fstat > fuc}. According to the picture,

Since flc = 0.9451 < 0.9489 = fstat < 1.057 = fuc, we fail to reject

null hypothesis.

b.
```
> cat(nRsq, qchisq(0.99, 3))
20.5628 11.34487
```
```
> cat(nRsq_all, qchisq(0.99, 3))
100.5157 11.34487
```
Since $NR^2 = 20.5628 > 11.34487$ = critical value, also $NR^2 = 100.5157 > 11.34487$

we reject Ho, and conclude that heteroskedasticity exists, which leads

a contradiction to conclusion in (a).

c.
```
> cat(nrow(theData)*R_sq, qchisq(0.95, 46))
151.1188 62.82962
```
The 5% level critical is 62.82 and the $NR^2 = 151.1188 > 62.82$,

we conclude that heteroskedasticity exists.

d.
```
> se
 (Intercept)          educ         exper    I(exper^2)        female         black         metro         south       midwest          west
3.211489e-02 1.758260e-03 1.300342e-03 2.635448e-05 9.529136e-03 1.694240e-02 1.230675e-02 1.356134e-02 1.410367e-02 1.440237e-02
> se_robust
 (Intercept)          educ         exper    I(exper^2)        female         black         metro         south       midwest          west
3.277743e-02 1.904848e-03 1.314237e-03 2.758278e-05 9.483417e-03 1.608548e-02 1.157624e-02 1.389454e-02 1.371725e-02 1.454941e-02
```
For $b_1, b_2, b_3, b_4, b_8, b_{10}$, their robust standard error are greater than their conventional se,

thus, their interval estimates get wider. For b_5, b_6, b_7, b_9, their robust standard

error are smaller than their conventional se, thus, their interval estimates get narrower. ∴ No inconsistency.

e.
```
> cat(lowerG, upperG)
-118.7511 -30.19027
> cat(lower_rseG, upper_rseG)
-122.0045 -26.93683
```
Comparing with (a) and (d),

The interval with conventional GLS standard error

has biggest lower bound and smallest upper bound,

which makes the length of interval becomes narrowest.

c.
```
> se_FGLS
 (Intercept)          educF         experF    I(experF^2)        femaleF         blackF         metroF         southF       midwestF          westF
3.675397e-02 4.247869e-04 3.392941e-04 1.725129e-06 2.285369e-03 4.096109e-03 3.141467e-03 3.259356e-03 3.370808e-03 3.465278e-03
> se_robust
 (Intercept)          educ         exper    I(exper^2)        female         black         metro         south       midwest          west
3.277743e-02 1.904848e-03 1.314237e-03 2.758278e-05 9.483417e-03 1.608548e-02 1.157624e-02 1.389454e-02 1.371725e-02 1.454941e-02
```
For b_1, its standard error of FGLS is greater than robust se with OLS.

For other estimators, their standard error of FGLS is smaller than robust se with OLS.

f.
```
> se_FGLS_robust
 (Intercept)          educF         experF    I(experF^2)        femaleF         blackF         metroF         southF       midwestF          westF
3.728837e-02 4.549622e-04 3.376888e-04 1.766250e-06 2.263801e-03 3.856254e-03 2.973540e-03 3.323117e-03 3.271000e-03 3.483108e-03
```
Comparing with se_FGLS in (e),

For $b_1, b_2, b_4, b_8, b_{10}$, their standard error of FGLS is smaller than robust

se with FGLS

For b_3, b_5, b_6, b_7, b_9, their standard error of FGLS is greater than robust

se with FGLS.

Comparing with se_robust in (d),

For b_1, its robust se with OLS is smaller than robust se with GLS.

For other estimators, their robust se with OLS is greater than robust se with GLS.

g. I preser FGLS estimation. Since it has narrower estimate interval

and according to (b)·(c), the heteroskedasticity exists, the FGLS

model could fit the data better.

d.
```
> cat(lower_rse, upper_rse)
-139.323 -24.32986
```
Comparing with (a): `> cat(lower, upper)` `-135.3298 -28.32302`

Since -139.323 < -135.3298 and -24.32986 > -28.32302,

the interval with robust standard error seems wider.

No16

```r
miles<-c(vacation$miles)
income<-c(vacation$income)
age<-c(vacation$age)
kids<-c(vacation$kids)

theData<-data.frame(miles, income, age, kids)
modOLS<-lm(miles~income+age+kids, data=theData)
b1<-modOLS$coefficients[[1]]
b2<-modOLS$coefficients[[2]]
b3<-modOLS$coefficients[[3]]
b4<-modOLS$coefficients[[4]]

#a
t<-qt(0.975, length(miles)-4)
lower<-b4-t*summary(modOLS)$coef[4, 2]
upper<-b4+t*summary(modOLS)$coef[4, 2]
cat(lower, upper)

#b
y_head<-b1+b2*income+b3*age+b4*kids
resid<-miles-y_head
#income
plot(income, resid)
mod_resid<-lm(resid~income, data=data.frame(resid, income))
beta1<-mod_resid$coef[[1]]
beta2<-mod_resid$coef[[2]]
abline(beta1, beta2, col="red")
#age
plot(age, resid)
mod_resid<-lm(resid~age, data=data.frame(resid, age))
beta1<-mod_resid$coef[[1]]
beta2<-mod_resid$coef[[2]]
abline(beta1, beta2, col="red")

#c
theSortingData<-theData[order(theData$income), ]
theSDF<-theSortingData[1:90, ]
theSDL<-theSortingData[-90:-1, ]
modfirst90<-lm(miles~income+age+kids, data=theSDF)
modlast90<-lm(miles~income+age+kids, data=theSDL)
sig1<-glance(modfirst90)$sigma^2
sig2<-glance(modlast90)$sigma^2
fstat<-sig1/sig2
flc<-qf(0.05/2, 86, 86)
fuc<-qf(1-0.05/2, 86, 86)
cat(flc, fuc, fstat)

#d
cov1<-hccm(modOLS, type="hc1")
vcv<-coeftest(modOLS, vcov.=cov1)
lower_rse<-b4-t*vcv[4, 2]
upper_rse<-b4+t*vcv[4, 2]
cat(lower_rse, upper_rse)

#e
w<-1/(income)#################
milesG<-miles*w
incomeG<-income*w
ageG<-age*w
kidsG<-kids*w
theDataG<-data.frame(milesG, incomeG, ageG, kidsG)
modGLS<-lm(milesG~incomeG+ageG+kidsG, data=theDataG)
b1G<-summary(modGLS)$coef[[1]]
b3G<-summary(modGLS)$coef[[2]]
b4G<-summary(modGLS)$coef[[3]]
seb4G<-summary(modGLS)$coef[3, 2]
lowerG<-b4G-t*seb4G
upperG<-b4G+t*seb4G

cov2<-hccm(modGLS, type="hc1")
vcv2<-coeftest(modGLS, vcov.=cov2)
lower_rseG<-b4G-t*vcv2[3, 2]
upper_rseG<-b4G+t*vcv2[3, 2]
cat(lowerG, upperG)
cat(lower_rseG, upper_rseG)
```

N018

```r
wage<-c(cps5$wage)
educ<-c(cps5$educ)
exper<-c(cps5$exper)
metro<-c(cps5$metro)
female<-c(cps5$female)

black<-c(cps5$black)
south<-c(cps5$south)
midwest<-c(cps5$midwest)
west<-c(cps5$west)
theData<-data.frame(wage, educ, exper, metro, female, black, south, midwest, west)

#a broom
theDataA<-data.frame(wage, educ, exper, metro, female)
dataF<-theDataA[which(theDataA$female==1), ]
dataM<-theDataA[which(theDataA$female==0), ]
modF<-lm(log(wage)~educ+exper+I(exper^2)+female+metro, data=dataF)
modM<-lm(log(wage)~educ+exper+I(exper^2)+female+metro, data=dataM)
sig1<-glance(modF)$sigma^2
sig2<-glance(modM)$sigma^2
fstat<-sig1/sig2
flc<-qf(0.025, nrow(dataF)-4, nrow(dataM)-4)
fuc<-qf(1-0.05/2, nrow(dataF)-4, nrow(dataM)-4)
cat(flc, fuc, fstat)

#b
modOLS<-lm(log(wage)~educ+exper+I(exper^2)+female+black+metro+south+midwest+west, data=theData)
theB<-summary(modOLS)$coef[, 1]
wage_head<-exp(theB[1]+theB[2]*educ+theB[3]*exper+theB[4]*(exper^2)+theB[5]*female
  +theB[6]*black+theB[7]*metro+theB[8]*south+theB[9]*midwest+theB[10]*west)
se<-summary(modOLS)$coef[, 2]

resid<-summary(modOLS)$resid
residSq<-resid^2
modB<-lm(residSq~metro+female+black, data=theData)
nRsq<-nrow(theData)*(summary(modB)$adj.r.squared)
cat(nRsq, qchisq(0.99, 3))
modOLSB<-lm(residSq~educ+exper+I(exper^2)+female+black+metro+south+midwest+west, data=theData)
nRsq_all<-nrow(theData)*(summary(modOLSB)$adj.r.squared)
cat(nRsq_all, qchisq(0.99, 3))

#c
modC<-lm(residSq~educ +exper +I(exper^2) +female +black +metro +south +midwest +west
        +I(educ^2) +I(exper^4)
        +educ*exper +educ*I(exper^2) +educ*female +educ*black +educ*metro +educ*south +educ*midwest +educ*west
        +I(exper^3) +exper*female +exper*black +exper*metro +exper*south +exper*midwest +exper*west
        +I(exper^2)*female +I(exper^2)*black +I(exper^2)*metro +I(exper^2)*south +I(exper^2)*midwest +I(exper^2)*west
        +female*black +female*metro +female*south +female*midwest +female*west
        +black*metro +black*south +black*midwest +black*west
        +metro*south +metro*midwest +metro*west +south*midwest +south*west +midwest*west
        , data=theData)
R_sq<-summary(modC)$adj.r.squared
cat(nrow(theData)*R_sq, qchisq(0.95, 46))

#d cat and lmtest
cov1<-hccm(modOLS, type="hc0")
vcv<-coeftest(modOLS, vcov.=cov1)
vcv
se_robust<-vcv[ ,2]
se
se_robust

#e
modE<-lm(log(residSq)~metro+exper, data=theData)
summary(modE)
theA<-summary(modE)$coef[, 1]
h_head<-theA[1]+theA[2]*metro+theA[3]*exper
w<-1/(sqrt(exp(h_head)))
wageF<-wage*w
educF<-educ*w
experF<-exper*w
metroF<-metro*w
femaleF<-female*w
blackF<-black*w
southF<-south*w
midwestF<-midwest*w
westF<-west*w
theDataE<-data.frame(wageF, educF, experF, metroF, femaleF, blackF, southF, midwestF, westF)
modFGLS<-lm(log(wageF)~educF+experF+I(experF^2)+femaleF+blackF+metroF+southF+midwestF+westF, data=theDataE)
se_FGLS<-summary(modFGLS)$coef[, 2]
se_FGLS
se_robust

#f
covF<-hccm(modFGLS, type="hc1")
vcvF<-coeftest(modFGLS, vcov.=covF)
se_FGLS_robust<-vcvF[, 2]
se_FGLS_robust
```