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

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

(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.

a. $H_0: \sigma_M^2 = \sigma_F^2$

$H_1: \sigma_M^2 \neq \sigma_F^2$

$\alpha = 0.05$

$df_M = n_M - k = 577 - 4 = 573$

$df_F = n_F - k = 1000 - 577 - 4 = 419$

$$\hat{\sigma}_M^2 = \frac{SSE_M}{df_M} = \frac{97161.9174}{573} = 169.57$$

$$F^* = \frac{\hat{\sigma}_M^2}{\hat{\sigma}_F^2} = \frac{169.57}{12.024^2} = 1.17$$

$F(0.975, 573, 419) = 1.96781$

$F(0.025, 573, 419) = 0.837669$

$RR = \{0.837669, 1.967813$

non-reject H_0

b. $H_0: \sigma_S^2 = \sigma_M^2$

$H_1: \sigma_S^2 < \sigma_M^2$

$\alpha = 0.05$

$df_S = 400 - 5 = 395$

$df_M = 600 - 5 = 595$

$$\hat{\sigma}_S^2 = \frac{SSE_S}{df_S} = \frac{56231.0382}{395} = 142.36$$

$$\hat{\sigma}_M^2 = \frac{SSE_M}{df_M} = \frac{100703.0471}{595} = 169.25$$

$$F^* = \frac{\hat{\sigma}_S^2}{\hat{\sigma}_M^2} = \frac{142.36}{169.25} = 0.84$$

$F(0.05, 395, 595) = 0.85\,85867$

$R.R: F^* < 0.85\,85867$

reject H_0

S-1

c. $H_0: \alpha_2=\alpha_3=\alpha_4=\alpha_5=0$ (homoskedasticity)

$q_{chisq}(1-0.05,\ 5-1)$

H_1: not all the α_i in H_0 are zero (heteroskedasticity)

$\alpha=0.05$ $S=5$

$NR^2=59.03 \geq \chi^2_{0.95,5-1}=9.49$

reject H_0 ⇒ 拒绝同质变异的假设

与 b 部分一致

d. EDUC . EXPER . METRO FEMALE

$EDUC^2$ $EXPER^2$ ($METRO^2$, $FEMALE^2$ 为二元变量)

EDUC × EXPER . EDUC × METRO . EDUC × FEMALE

EXPER × METRO EXPER × FEMALE

METRO × FEMALE ⇒ df = 12

$\chi^2_{0.95,12}=21.026$

$NR^2=78.82 > \chi^2_{0.95,12}=21.026$

e. 截距、EDUC : wider

EXPER . EMTRO . FEMALE : narrower

存在不一致

se. robse 差異是因为異质变異的存在

f. $t=0<1.96 \Rightarrow$ 不顯著 (5%)

compatible.

b 檢驗 σ 是否与 MARRIED 相关

f 檢驗 MARRIED 是否影響 WAGE

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)?

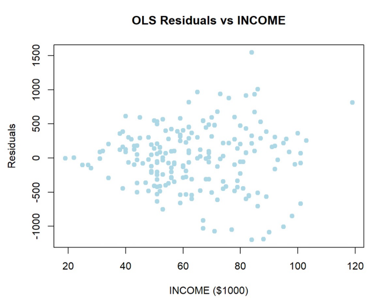
e. Obtain GLS estimates assuming $\sigma_i^2 = \sigma^2 INCOME_i^2$. Using both conventional GLS and robust GLS standard errors, construct a 95% interval estimate for the effect of one more child on miles traveled, holding the two other variables constant. How do these interval estimates compare to the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

a

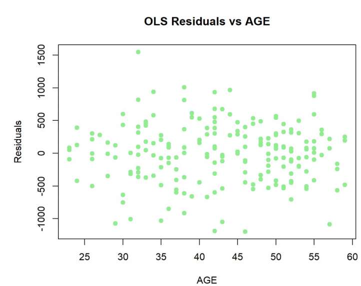
```
Residuals:
     Min       1Q   Median       3Q      Max
-1198.14  -295.31    17.98   287.54  1549.41

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -391.548    169.775  -2.306   0.0221 *
income        14.201      1.800   7.889 2.10e-13 ***
age           15.741      3.757   4.189 4.23e-05 ***
kids         -81.826     27.130  -3.016   0.0029 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '

Residual standard error: 452.3 on 196 degrees of freedom
Multiple R-squared:  0.3406,    Adjusted R-squared:  0.3305
F-statistic: 33.75 on 3 and 196 DF,  p-value: < 2.2e-16
```

```
> confint(model_ols, level = 0.95)
                  2.5 %    97.5 %
(Intercept) -726.36871 -56.72731
income        10.65097  17.75169
age            8.33086  23.15099
kids        -135.32981 -28.32302
```

b.



Yes



c.

$H_0: \sigma_{high}^2 = \sigma_{low}^2$

$H_1: \sigma_{high}^2 > \sigma_{low}^2$

$\alpha = 0.05$

$df_{high} = n_{high} - k = 90 - 4 = 86$

$df_{low} = n_{low} - k = 90 - 4 = 86$

GQ $F^* = 3.1041$

$F(0.95, 86, 86) = 1.4286$

$F^* > F$ reject H_0

d.

```
Residuals:
     Min       1Q   Median       3Q      Max
-1198.14  -295.31    17.98   287.54  1549.41

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -391.548    169.775  -2.306   0.0221 *
income        14.201      1.800   7.889 2.10e-13 ***
age           15.741      3.757   4.189 4.23e-05 ***
kids         -81.826     27.130  -3.016   0.0029 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 452.3 on 196 degrees of freedom
Multiple R-squared:  0.3406,	Adjusted R-squared:  0.3305
F-statistic: 33.75 on 3 and 196 DF,  p-value: < 2.2e-16
```

估計值一致

標準誤變大

C.I. 變寬

```
> coefci(model_ols, vcov. = robust_se, level = 0.95)
                  2.5 %     97.5 %
(Intercept) -672.883378 -110.21263
income        10.377633   18.02503
age            7.919934   23.56191
kids        -139.322973  -24.32986
```

e.

```
Call:
lm(formula = miles ~ income + age + kids, data = vacation, weights = weights)

Weighted Residuals:
     Min       1Q   Median       3Q      Max 
-15.1907  -4.9555   0.2488   4.3832  18.5462 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -424.996    121.444  -3.500 0.000577 ***
income        13.947      1.481   9.420  < 2e-16 ***
age           16.717      3.025   5.527 1.03e-07 ***
kids         -76.806     21.848  -3.515 0.000545 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.765 on 196 degrees of freedom
Multiple R-squared:  0.4573,	Adjusted R-squared:  0.449 
F-statistic: 55.06 on 3 and 196 DF,  p-value: < 2.2e-16

> confint(model_gls, level = 0.95)  # GLS 信賴區間
                 2.5 %     97.5 %
(Intercept) -664.50116 -185.49119
income        11.02744   16.86718
age           10.75260   22.68240
kids        -119.89450  -33.71808
```

→ 最窄

```
> 
> # Robust SE for GLS model
> gls_robust_se <- vcovHC(model_gls, type = "HC1")
> coeftest(model_gls, vcov. = gls_robust_se)

t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)    
(Intercept) -424.9962    95.8035 -4.4361 1.526e-05 ***
income        13.9473     1.3470 10.3545 < 2.2e-16 ***
age           16.7175     2.7974  5.9761 1.061e-08 ***
kids         -76.8063    22.6186 -3.3957 0.0008286 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> coefci(model_gls, vcov. = gls_robust_se, level = 0.95)    # GLS 信賴區間（以 robust SE）
                 2.5 %     97.5 %
(Intercept) -613.93428 -236.05807
income        11.29086   16.60376
age           11.20062   22.23438
kids        -121.41339  -32.19919
> 
```

→ 較 OLS 寬

(a) OLS

(d) robust OLS

GLS

robust GLS

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. $F^* = 0.9489479$

RR : $\{ F^* < 0.9457157$ or $F^* > 1.057889 \}$

non-reject H_0

b.

$NR^2 = 23.55681$

$\chi^2_{(0.99,3)} = 11.34487$

$NR^2 > \chi^2$ reject H_0

$\Rightarrow$ heteroskedasticity

c. p-value $< 2.2 \times 10^{-16} < 0.05$

$\Rightarrow$ reject H_0 $\Rightarrow$ heteroskedasticity

d.

變數	OLS.標準誤	Robust.標準誤	改變幅度....
(Intercept)	0.0321	0.0328	+2.12%
educ	0.0018	0.0019	+8.39%
exper	0.0013	0.0013	+1.12%
I(exper^2)	0.0000	0.0000	+4.71%
female	0.0095	0.0095	-0.43%
black	0.0169	0.0161	-5.01%
metro	0.0123	0.0116	-5.89%
south	0.0136	0.0139	+2.51%
midwest	0.0141	0.0137	-2.69%
west	0.0144	0.0146	+1.07%

no

e.

```
> print(fgls_table, row.names = FALSE)
```

變數	估計值	標準誤	X95..信賴區間
(Intercept)	1.1922	0.0316	[1.1303, 1.2541]
educ	0.1017	0.0018	[0.0982, 0.1051]
exper	0.0301	0.0013	[0.0275, 0.0326]
I(exper^2)	-0.0005	0.0000	[-5e-04, -4e-04]
female	-0.1662	0.0095	[-0.1848, -0.1476]
black	-0.1109	0.0170	[-0.1442, -0.0775]
metro	0.1178	0.0115	[0.0953, 0.1402]
south	-0.0448	0.0135	[-0.0713, -0.0183]
midwest	-0.0632	0.0140	[-0.0906, -0.0358]
west	-0.0055	0.0144	[-0.0337, 0.0227]

修正異質變異的問題

f.

變數	FGLS.係數	FGLS.SE	Robust.SE	SE.變化....
(Intercept)	1.1922	0.0316	0.0324	+2.43%
educ	0.1017	0.0018	0.0019	+7.26%
exper	0.0301	0.0013	0.0013	+0.55%
I(exper^2)	-0.0005	0.0000	0.0000	+2.31%
female	-0.1662	0.0095	0.0094	-0.45%
black	-0.1109	0.0170	0.0159	-6.61%
metro	0.1178	0.0115	0.0116	+0.9%
south	-0.0448	0.0135	0.0138	+2.31%
midwest	-0.0632	0.0140	0.0137	-1.94%
west	-0.0055	0.0144	0.0145	+0.92%

接近 ⇒ 有效控制

g.

FGLS + Robust SE

存在異質變異數、加強穩健性