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) $\begin{cases} H_0: \sigma_M^2 = \sigma_F^2 \\ H_1: \sigma_M^2 \neq \sigma_F^2 \end{cases}$

$N_M = 577$, $SSE_M = 97161.9174$

$N_F = 423$, $\hat{\sigma}_F = 12.024$

$\alpha = 0.05$

$\hat{\sigma}_M = \sqrt{MSE} = \sqrt{\dfrac{97161.9174}{577-4}} = 13.022$

$F = \dfrac{\sigma_M^2}{\sigma_F^2} \sim F(577-4, 423-4)$ (?)

1.196781   0.837669

$RR = \{ F \geq F_{0.975}(573, 419) \text{ or } F \leq F_{0.025}(573, 419) \}$

$F = \dfrac{13.022^2}{12.024^2} = 1.1729 \notin RR$, do not reject H_0.

$\Rightarrow \sigma_m^2$ isn't significantly differ from σ_F^2 at the 5% level of siginifcance.

(b) $\begin{cases} H_0: \sigma_s^2 = \sigma_m^2 \\ H_1: \sigma_s^2 < \sigma_m^2 \end{cases}$

$N_s = 400$, $SSE_s = 56231.0382$

$N_m = 600$, $SSE_m = 100703.0471$

$\alpha = 0.05$

$F = \dfrac{\sigma_s^2}{\sigma_m^2} \sim F(395, 595)$

$RR = \{ F \leq F_{0.05}(395, 595) = 0.8585867 \}$

$F = \dfrac{\frac{56231.0382}{395}}{\frac{100703.0471}{595}} = 0.8411 \in RR$, reject H_0.

$\Rightarrow \sigma_s^2$ is significantly smaller than σ_m^2 at the 5% level significance.

(c) $\chi^2 = N \times R^2 \sim \chi^2(5-1) = 9.487729$ (右尾檢定)

$NR^2 = 59.03 > \chi^2(4) \in RR$, reject H_0

$\Rightarrow$ we conclude that the heteroskedasticity exist and the error variation is significantly differ from males and females.

(d) White's test ($\sigma_i^2 = \alpha_0$)

$$Y = \beta_1 + \beta_2 X_1 + \beta_3 X_2 + \beta_4 X_3 + \beta_5 X_4$$

$$\begin{cases} H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_n = 0 \text{ (homoscedasticity)} \\ H_1: \text{不全為 } 0 \text{ (heteroscedasticity)} \end{cases}$$

$$Var(u_i) \rightarrow \sigma_i^2 = \alpha_1 + \alpha_2 X_1 + \alpha_3 X_2 + \alpha_4 X_3 + \alpha_5 X_4 + \alpha_6 X_1^2 + \alpha_7 X_2^2 + \alpha_8 X_3^2 + \alpha_9 X_4^2 + \alpha_{10} X_1 X_2 + \alpha_{11} X_1 X_3 + \alpha_{12} X_1 X_4 + \alpha_{13} X_2 X_3 + \alpha_{14} X_2 X_4 + \alpha_{15} X_3 X_4$$

test statistic: $nR^2 \sim \chi^2(k-1)$, k 為 α 個數

$RR = \{ nR^2 \geq \chi^2_{0.95}(14) = 23.68479 \}$

$nR^2 = 78.82 \in RR$, reject H_0.

$\Rightarrow$ we conclude that the heteroskedasticity exist and the error variation is significantly differ from males and females.

(e)

Wider: Intercept, EDUC

Narrower: EXPER, METRO, FEMALE

$\Rightarrow$ Inconsistent

(f)

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)?

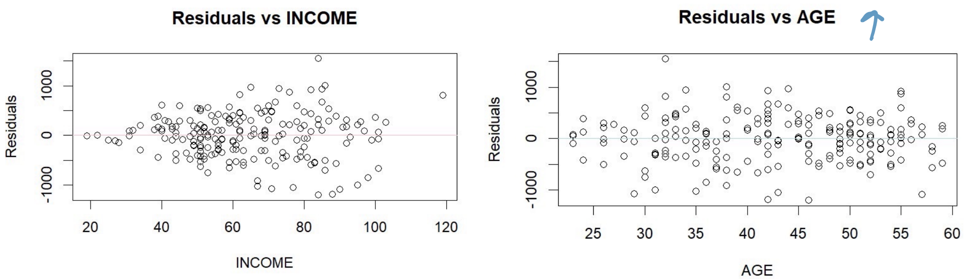
(a)

```
Call:
lm(formula = MILES ~ INCOME + AGE + KIDS, data = vacation)

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

> confint(model_ols, "KIDS", level = 0.95)
         2.5 %    97.5 %
KIDS -135.3298 -28.32302
```

narrow

(b)

homoskedasticity



⇒ In the residual plot of income, as the values of the variables increase, the residuals become more dispersed, indicating the presence of heteroskedasticity.

(c)

```
	Goldfeld-Quandt test

data:  MILES ~ INCOME + AGE + KIDS
GQ = 3.1041, df1 = 86, df2 = 86, p-value = 1.64e-07
alternative hypothesis: variance increases from segment 1 to 2
```

(d)

```
t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)    
(Intercept) -391.5480   142.6548 -2.7447 0.0066190 ** 
INCOME        14.2013     1.9389  7.3246 6.083e-12 ***
AGE           15.7409     3.9657  3.9692 0.0001011 ***
KIDS         -81.8264    29.1544 -2.8067 0.0055112 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> 
> # 自行計算信賴區間 for KIDS
> kid_coef <- coef(model_ols)["KIDS"]
> kid_se_robust <- sqrt(vcovHC(model_ols, type = "HC1")["KIDS",
> ci_lower <- kid_coef - 1.96 * kid_se_robust
> ci_upper <- kid_coef + 1.96 * kid_se_robust
> c(ci_lower, ci_upper)
      KIDS       KIDS 
-138.96900  -24.68383 
```

wide

⇒ The interval estimate of KIDS is wider than OLS model.

(e)

```
Call:
lm(formula = MILES ~ INCOME + AGE + KIDS, data = vacation, weights = 1/(INCOME^2))

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
```

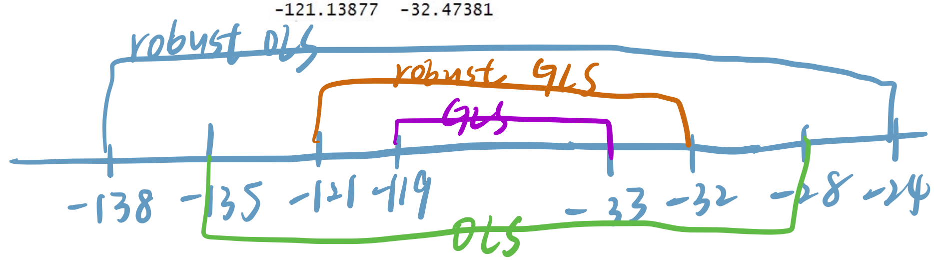
OLS [-135.33, -28.32]

robust OLS [-138.97, -24.68]

GLS [-119.89, -33.71]

robust GLS [-121.14, -32.47]

```
> confint(model_gls, "KIDS", level = 0.95)
         2.5 %    97.5 %
KIDS -119.8945 -33.71808
>
> # 若要使用 robust GLS 標準誤
> robust_gls_se <- coeftest(model_gls, vcov = vcovHC(model_gls, type = "HC1"))
> print(robust_gls_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

> kid_coef <- coef(model_gls)["KIDS"]
> robust_gls_se <- sqrt(vcovHC(model_gls, type = "HC1")["KIDS", "KI
> ci_lower <- kid_coef - 1.96 * robust_gls_se
> ci_upper <- kid_coef + 1.96 * robust_gls_se
> c(ci_lower, ci_upper)
      KIDS       KIDS 
-121.13877  -32.47381 
```



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)

```
        Goldfeld-Quandt test

data:  model
GQ = 0.97487, df1 = 3257, df2 = 3256, p-value = 0.7661
alternative hypothesis: variance increases from segment 1 to 2
```

$RR = \{GQ > F_{0.975}(3257, 3256) = 1.071119$ or $GQ < F_{0.025}(3257, 3256) = 0.9336037\}$

$GQ = 0.97487 \notin RR$, do not reject H_0.

$\Rightarrow$ σ_M^2 isn't significantly differ from σ_F^2 at the 5% level of siginifcance.

(b)

```
> n <- nrow(cps5)
> R2 <- summary(aux_model)$r.squared
> NR2 <- n * R2
> print(NR2)
[1] 23.55681
> # 1% 顯著水準下的卡方臨界值 (自由度 = 3)
> qchisq(0.99, df = 3)
[1] 11.34487
```

$NR^2 = 23.55681 > \chi^2(3)$, reject H_0

(c)

```
> bptest(model, ~ EDUC + EXPER + I(EXPER^2) + FEMALE +
+ WEST +
+           I(EDUC^2) + I(EXPER^2) + I(FEMALE^2) + I(BLA

        studentized Breusch-Pagan test

data:  model
BP = 133.67, df = 10, p-value < 2.2e-16
```

p-value < 0.05, reject H_0.

(d)

```
> # 95% 信賴區間 (常規 OLS)
> confint(ols_model)
                     2.5 %        97.5 %
(Intercept)   1.1384302204  1.2643338265
EDUC          0.0977830603  0.1046761665
EXPER         0.0270727569  0.0321706349
I(EXPER^2)   -0.0004974407 -0.0003941203
FEMALE       -0.1841810529 -0.1468229075
BLACK        -0.1447358548 -0.0783146449
METRO         0.0948966363  0.1431441846
SOUTH        -0.0723384657 -0.0191724010
MIDWEST      -0.0915893895 -0.0362971859
WEST         -0.0348207138  0.0216425095
```

```
> round(robust_ci, 4)
              Lower Estimate   Upper
(Intercept)  1.1371   1.2014  1.2657  N
EDUC         0.0975   0.1012  0.1050  W
EXPER        0.0270   0.0296  0.0322  W
I(EXPER^2)  -0.0005  -0.0004 -0.0004  N
FEMALE      -0.1841  -0.1655 -0.1469  N
BLACK       -0.1431  -0.1115 -0.0800  N
METRO        0.0963   0.1190  0.1417  N
SOUTH       -0.0730  -0.0458 -0.0185  W
MIDWEST     -0.0908  -0.0639 -0.0370  N
WEST        -0.0351  -0.0066  0.0219  W
```

$\Rightarrow$ inconsistency

(e)

(f)

```
> confint(fgls_model)
                   2.5 %        97.5 %
(Intercept)  1.127694057  1.2515350381
EDUC         0.098351366  0.1052682659
EXPER        0.027590905  0.0326693606
I(EXPER^2)  -0.000509177 -0.0004041652
FEMALE      -0.184317568 -0.1471399412
BLACK       -0.144166923 -0.0776164205
METRO        0.094808099  0.1401225846
SOUTH       -0.071252312 -0.0182311336
MIDWEST     -0.090708494 -0.0358393299
WEST        -0.033747215  0.0226111169
```

```
> round(ci_fgls_robust, 6)
                 Lower  Estimate     Upper
(Intercept)   1.126256  1.189615  1.252973
EDUC          0.098104  0.101810  0.105515
EXPER         0.027574  0.030130  0.032686
I(EXPER^2)   -0.000510 -0.000457 -0.000403
FEMALE       -0.184227 -0.165729 -0.147230
BLACK        -0.141981 -0.110892 -0.079802
METRO         0.094813  0.117465  0.140118
SOUTH        -0.071854 -0.044742 -0.017629
MIDWEST      -0.090142 -0.063274 -0.036406
WEST         -0.033996 -0.005568  0.022860
```

narrower → wider:

EXPER : FGLS → FGLS_robust → OLS-robust → OLS

METRO : ?

(g) robust FGLS