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.

(a) $H_0: \sigma_M^2 = \sigma_F^2$ $H_1: \sigma_M^2 \neq \sigma_F^2$ $\alpha = 5\%$

Test statistic: $F = \frac{s_M^2}{s_F^2}$

$s_M^2 = \frac{SSE_M}{n_M - K} = \frac{97161.9174}{577 - 4} = 169.567$

$s_F^2 = (12.024)^2 = 144.5766$

$F_0 = \frac{169.567}{144.5766} \approx 1.173 \sim F(573, 1000 - 577 - 4) = F(573, 419)$

$RR = \{|F_0| > F_{(0.975, 573, 419)}\} = \{F > 1.207 \text{ or } F < 0.828\}$

$\because F_0$ & RR $\therefore$ don't reject H_0, there is no evidence to say $\sigma_M^2 \neq \sigma_F^2$

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 $\text{var}(e_i|\mathbf{x}_i, MARRIED = 0) = \sigma_{SINGLE}^2$ and $\text{var}(e_i|\mathbf{x}_i, MARRIED = 1) = \sigma_{MARRIED}^2$. Specify the null hypothesis $\sigma_{SINGLE}^2 = \sigma_{MARRIED}^2$ versus the alternative hypothesis $\sigma_{MARRIED}^2 > \sigma_{SINGLE}^2$. 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.

(b) $H_0: \sigma^2_{SINGLE} = \sigma^2_{MARRIED}$ $\alpha = 5\%$

$H_1: \sigma^2_{MARRIED} > \sigma^2_{SINGLE}$

Test statistic: $F = \frac{s^2_{MARRIED}}{s^2_{SINGLE}}$

$s^2_M = \frac{10703.0471}{600-5} = 169.249$

$s^2_S = \frac{56231.0382}{400-5} = 142.357$

$F_0 = \frac{169.249}{142.357} = 1.1889 \sim F_{(595,395)}$

$RR = \{F_0 > F_{(0.95,595,395)} = 1.1647\}$

$\because F_0 \in RR \therefore$ We reject H_0. there is an evidence to say $\sigma^2_{MARRIED} > \sigma^2_{SINGLE}$

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.

Test statistic: $NR^2 \sim \chi^2_q$ $\alpha = 5\%$

$\chi^2_{0.95}(4) = 9.488$

$\because NR^2 = 59.03 > 9.488 \therefore$ reject H_0

there is an evidence to say existence of

heteroskedasticity.

這個檢定進一步支持了part(b)的結論，誤差變異不是常數，可能與包括婚姻狀態在內的變數有關。

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?

White 的檢定誤差項的變異是否依賴於自變數的任意非線性形式

degree of freedom = 4+2+6 = 12

{ 4個原始變數
2個平方項(∵METRO, FEMALE是dummy variable)
C_2^4 = 6個交乘項

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

∵ $NR^2 = 78.82 > 21.026$ ∴ reject H_0

there is an evidence to say existence of heteroskedasticity.

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?

compare robust SE and usual SE

narrower: EXPER, METRO, FEMALE

wider : intercept . EDUC

not an inconsistency. It suggests that heteroskedasticity affects the estimation precision of different coefficients differently.

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.

It is compatible, 兩個是互相獨立的檢定 part(b) 是整個變異數檢定，t檢定 MARRIED 變數是否顯著。

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

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

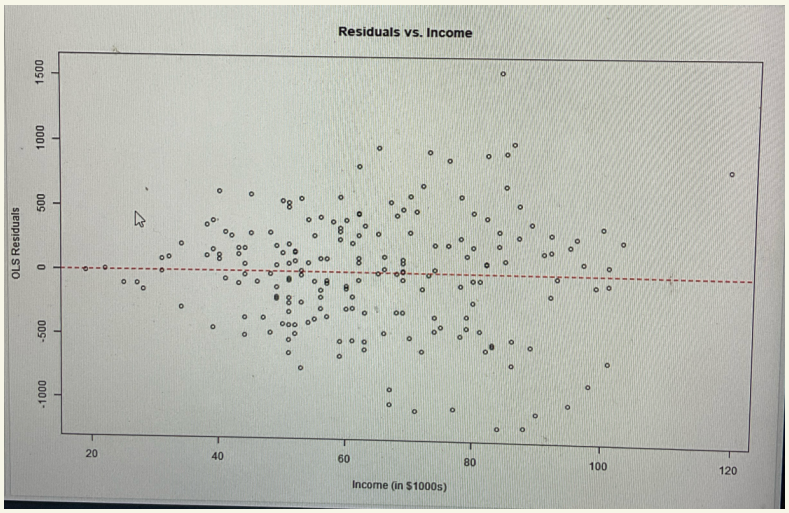
MILES is miles driven per year, *INCOME* is measured in $1000 units, *AGE* is the average age of the adult members of the household, and *KIDS* is the number of children.

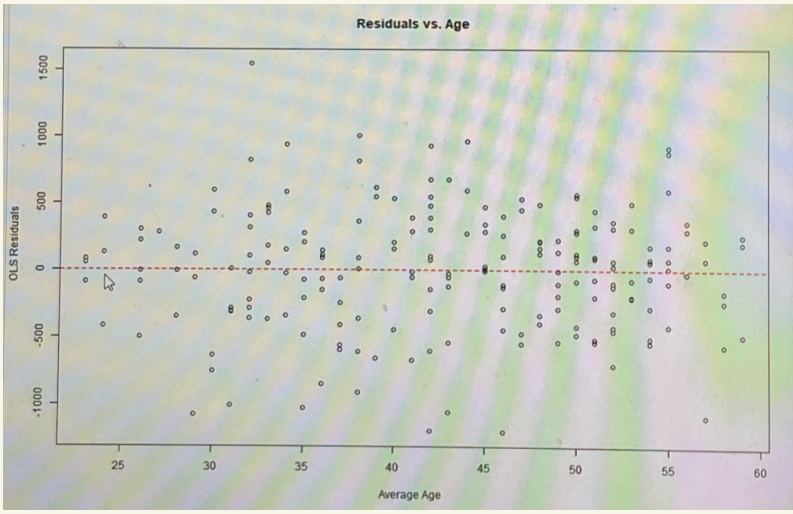
a. Use the data file *vacation* to estimate the model by OLS. Construct a 95% interval estimate for the effect of one more child on miles traveled, holding the two other variables constant.

```
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, "kids", level = 0.95)
          2.5 %    97.5 %
kids -135.3298 -28.32302
>
```

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





the residuals are wider when income become larger

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.

group 1 : first 90 observations (低收入)
group 2 : last 90 observations (高收入)

$H_0: \sigma_1^2 = \sigma_2^2 \quad H_1: \sigma_1^2 < \sigma_2^2$

$$F = \frac{SSE_{group2}/df_2}{SSE_{group1}/df_1} \sim F_{(0.95,86,86)}$$

```
> cat("F statistic:", round(Fstat, 4), "\n")
F statistic: 3.1041
> cat("Critical value (5% level):", round(f_crit, 4), "\n")
Critical value (5% level): 1.4286
> if (Fstat > f_crit) {
+   cat("Reject H0: Evidence of heteroskedasticity.\n")
+ } else {
+   cat("Fail to reject H0: No evidence of heteroskedasticity.\n")
+ }
Reject H0: Evidence of heteroskedasticity.
>
```

∵ Fstat > Fcrit ∴ reject H_0, there is an evidence to say existence of heteroskedasticity.

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)?

```
> print(results)
             Estimate  Std.Error   t.value      p.value Conf.Int.Lower Conf.Int.Upper
(Intercept) -391.54801 141.221027 -2.772590 6.097754e-03    -668.341219     -114.75479
income        14.20133   1.919371  7.398952 3.931796e-12      10.439367       17.96330
age           15.74092   3.925878  4.009530 8.644281e-05       8.046204       23.43564
kids         -81.82642  28.861363 -2.835154 5.060627e-03    -138.394691     -25.25815
> ls()
```

	Kids estimate	95% interval
OLS (a)	-81.826	[-135.330, -28.323]
Robust OLS (d)	-81.826	[-138.395, -25.258]

∵ robust SE ↑ ∴ interval become wider

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)?

GLS model:

```
> confint(gls_model, "kids", level = 0.95)
        2.5 %    97.5 %
kids -119.8945 -33.71808
```

robust GLS model

```
> print(kids_ci)
      2.5 %     97.5 %
-121.13953  -32.47305
```

GLS kids estimate = -76.806 < OLS, ∵ GLS 使用權重 $\frac{1}{INCOME_i}$, 使低 INCOME 類測值有更大權重, 導致估計改變.

GLS estimate have narrower interval than OLS

∵ consider heteroskedasticity problem (more efficient)

8.16 R

```r
data(vacation)
#8.16(a)
model <- lm(miles ~ income + age + kids, data = vacation)
summary(model)
confint(model, "kids", level = 0.95)
#(b)
resid <- residuals(model)

plot(vacation$income, resid,
     xlab = "Income (in $1000s)", ylab = "OLS Residuals",
     main = "Residuals vs. Income")
abline(h = 0, col = "red", lty = 2)


plot(vacation$age, resid,
     xlab = "Average Age", ylab = "OLS Residuals",
     main = "Residuals vs. Age")
abline(h = 0, col = "red", lty = 2)
#(c)
vacation_sorted <- vacation[order(vacation$income), ]
group1 <- vacation_sorted[1:90, ]
group2 <- vacation_sorted[111:200, ]
model1 <- lm(miles ~ income + age + kids, data = group1)
model2 <- lm(miles ~ income + age + kids, data = group2)

SSE1 <- sum(resid(model1)^2)
SSE2 <- sum(resid(model2)^2)
df1 <- 90 - 4
df2 <- 90 - 4
Fstat <- (SSE2 / df2) / (SSE1 / df1)
alpha <- 0.05
f_crit <- qf(1 - alpha, df1, df2)
cat("F statistic:", round(Fstat, 4), "\n")
cat("Critical value (5% level):", round(f_crit, 4), "\n")
if (Fstat > f_crit) {
  cat("Reject H0: Evidence of heteroskedasticity.\n")
} else {
  cat("Fail to reject H0: No evidence of heteroskedasticity.\n")
}

```

```r
#(d)
library(sandwich)
library(lmtest)
model <- lm(miles ~ income + age + kids, data = vacation)
# 計算 White robust 標準誤
residuals <- residuals(model)
X <- model.matrix(model)
n <- nrow(X)
k <- ncol(X)
# 2. 計算 HC0 穩健協方差矩阵
#    HC0 = (X'X)^(-1) (X' diag(e_i^2) X) (X'X)^(-1)
bread <- solve(crossprod(X))  # (X'X)^(-1)
meat <- t(X) %*% diag(residuals^2) %*% X  # X' diag(e_i^2) X
robust_vcov <- bread %*% meat %*% bread  # HC0 估计
# 3. 计算稳健标准误和 t 检验
robust_se <- sqrt(diag(robust_vcov))
t_stats <- coef(model) / robust_se
p_values <- 2 * pt(abs(t_stats), df = n - k, lower.tail = FALSE)
# 4. 输出结果
results <- data.frame(
  Estimate = coef(model),
  Std.Error = robust_se,
  t.value = t_stats,
  p.value = p_values
)
confint_lower <- coef(model) - 1.96 * robust_se
confint_upper <- coef(model) + 1.96 * robust_se
results$Conf.Int.Lower <- confint_lower
results$Conf.Int.Upper <- confint_upper
print(results)

#(e)
# 建立權重變數：income 的平方的倒數
vacation$weight_gls <- 1 / (vacation$income^2)

```

```r
# 用 WLS 模型 (即 GLS)
gls_model <- lm(miles ~ income + age + kids, data = vacation, weights = weight_gls)
summary(gls_model)
# 使用 confint() 建立 95% 信賴區間 (常規標準誤)
confint(gls_model, "kids", level = 0.95)

# 获取加权残差和设计矩阵
w <- gls_model$weights  # 提取权重
X <- model.matrix(gls_model)
residuals <- residuals(gls_model) * sqrt(w)  # 加权残差

# 计算 HC0 稳健协方差矩阵
bread <- solve(t(X) %*% diag(w) %*% X)  # (X'WX)^(-1)
meat <- t(X) %*% diag(residuals^2) %*% X  # X' diag(e_i^2) X
robust_vcov <- bread %*% meat %*% bread  # HC0 估计

# 计算稳健标准误和置信区间
robust_se <- sqrt(diag(robust_vcov))
t_stats <- coef(gls_model) / robust_se
df <- nrow(X) - ncol(X)  # 自由度
p_values <- 2 * pt(abs(t_stats), df = df, lower.tail = FALSE)

# 输出结果
results <- data.frame(
  Estimate = coef(gls_model),
  Std.Error = robust_se,
  t.value = t_stats,
  p.value = p_values,
  CI_lower = coef(gls_model) - 1.96 * robust_se,
  CI_upper = coef(gls_model) + 1.96 * robust_se
)
print(results)

# 单独提取 "kids" 的 95% 置信区间
kids_ci <- c(
  coef(gls_model)["kids"] - 1.96 * robust_se["kids"],
  coef(gls_model)["kids"] + 1.96 * robust_se["kids"]
)
names(kids_ci) <- c("2.5 %", "97.5 %")
print(kids_ci)

```

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.

(a)

```
> cat("自由度: ", df_male, "和", df_female, "\n")
自由度:  5415 和 4366
> cat("F統計量值: ", F_stat, "\n")
F統計量值:  1.05076
> cat("5%顯著水準的臨界值: ", F_critical_lower, "和", F_critical_upper, "\n")
5%顯著水準的臨界值:  0.9452566 和 1.058097
> p_value <- 2 * min(pf(F_stat, df_male, df_female, lower.tail = TRUE),
+                    pf(F_stat, df_male, df_female, lower.tail = FALSE))
> cat("p 值: ", p_value, "\n")
p 值:  0.08569168
> if (F_stat < F_critical_lower || F_stat > F_critical_upper) {
+   cat("=> 拒絕虛無假設: 男性與女性的誤差變異數顯著不同. \n")
+ } else {
+   cat("=> 無法拒絕虛無假設: 無足夠證據認為男性與女性的誤差變異數不同. \n")
+ }
=> 無法拒絕虛無假設: 無足夠證據認為男性與女性的誤差變異數不同.
>
```

$F = \frac{SSE_{male}/df_{male}}{SSE_{female}/df_{female}}$

$H_0: \sigma_M^2 = \sigma_F^2$, $H_1: \sigma_M^2 \neq \sigma_F^2$ $\alpha = 5\%$

p-value $= 0.0857 > \alpha = 0.05$ ∴ don't reject H_0

there is no evidence to say $\sigma_M^2 \neq \sigma_F^2$ at 5% significance level

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.

```
> cat("---- 使用 METRO, FEMALE, BLACK 的 NR² 測試 ----\n")
---- 使用 METRO, FEMALE, BLACK 的 NR2 測試 ----
> cat("NR² 統計量:", NR2_1, "\n")
NR2 統計量: 23.55681
> cat("p 值:", p_val1, "\n")
p 值: 3.0909e-05
> if (p_val1 < 0.01) {
+   cat("=> 在 1% 顯著水準下, 拒絕虛無假設, 存在異質變異數. \n")
+ } else {
+   cat("=> 在 1% 顯著水準下, 無法拒絕虛無假設, 無明顯異質變異數. \n")
+ }
=> 在 1% 顯著水準下, 拒絕虛無假設, 存在異質變異數.
>
```

$\sim \chi^2_{0.99}(3)$

H_0: 誤差項的變異數與METRO, FEMALE, BLACK無關

H_1: 誤差項的變異數與METRO, FEMALE, BLACK有關

Test statistic $NR^2 = nXR^2 \sim \chi^2_{0.99}(3)$ $\alpha = 0.01$

∵ p-value $= 3.099e{-}05 < \alpha = 0.01$ ∴ reject H_0

there is an evidence to say there exist heteroskedasticity

```
---- 使用所有解釋變數的 NR2 測試 ----
> cat("NR² 統計量:", NR2_2, "\n")
NR2 統計量: 109.4243
> cat("p 值:", p_val2, "\n")
p 值: 0
> if (p_val2 < 0.01) {
+   cat("=> 在 1% 顯著水準下，拒絕虛無假設，存在異質變異數。\n")
+ } else {
+   cat("=> 在 1% 顯著水準下，無法拒絕虛無假設，無明顯異質變異數。\n")
+ }
=> 在 1% 顯著水準下，拒絕虛無假設，存在異質變異數。
>
```

H_0: 誤差項的變異數與任何解釋變數無關

H_1: 誤差項的變異數與至少一個解釋變數有關

Test statistic $NR^2 = nXR^2 \sim \chi^2_{0.99}(9)$ $\alpha = 0.01$

∵ p-value $= 0 < \alpha = 0.01$ ∴ reject H_0

there is an evidence to say there exist heteroskedasticity

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

```
        studentized Breusch-Pagan test

data:  model_ols
BP = 194.44, df = 44, p-value < 2.2e-16

> # 臨界值（5%顯著水準）
> df_white <- white_test$parameter  # 使用 bptest 中的自由度
> critical_value_white <- qchisq(1-0.05, df_white)
> cat("White 檢定 NR^2:", white_test$statistic, "\n")
White 檢定 NR^2: 194.4447
> cat("自由度:", white_test$parameter, "\n")
自由度: 44
> cat("5% 顯著水準的臨界值:", critical_value_white, "\n")
5% 顯著水準的臨界值: 60.48089
>
```

H_0: No heteroskedasticity

H_1: heteroskedasticity $\alpha = 0.05$

$\because NR^2 = 194.4447 > 60.481 = \chi^2_{0.95}(44)$

$\therefore$ reject H_0, there exist heteroskedasticity

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?

```
> print(round(comparison, 4))
            Estimate SE_OLS SE_White CI_low_OLS CI_high_OLS CI_low_White CI_high_White
(Intercept)   1.2014 0.0321   0.0328     1.1384      1.2643       1.1371        1.2656
educ          0.1012 0.0018   0.0019     0.0978      0.1047       0.0975        0.1050
exper         0.0296 0.0013   0.0013     0.0271      0.0322       0.0270        0.0322
I(exper^2)   -0.0004 0.0000   0.0000    -0.0005     -0.0004      -0.0005       -0.0004
female       -0.1655 0.0095   0.0095    -0.1842     -0.1468      -0.1841       -0.1469
black        -0.1115 0.0169   0.0161    -0.1447     -0.0783      -0.1431       -0.0800
metro         0.1190 0.0123   0.0116     0.0949      0.1431       0.0963        0.1417
south        -0.0458 0.0136   0.0139    -0.0723     -0.0192      -0.0730       -0.0185
midwest      -0.0639 0.0141   0.0137    -0.0916     -0.0363      -0.0908       -0.0371
west         -0.0066 0.0144   0.0145    -0.0348      0.0216      -0.0351        0.0219
```

variable: black, metro, midwest se↓, narrower 估計更有信心

variable: educ, south, west se↑, wider 考慮更多變異數後，估計不確定性提高

In terms of coefficient estimates, there is no inconsistency.

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)?

```
> print(round(comparison_ci, 4))
            CI_low_White CI_high_White CI_low_FGLS CI_high_FGLS
(Intercept)       1.1371        1.2656      1.1303       1.2541
educ              0.0975        0.1050      0.0982       0.1051
exper             0.0270        0.0322      0.0275       0.0326
I(exper^2)       -0.0005       -0.0004     -0.0005      -0.0004
female           -0.1841       -0.1469     -0.1848      -0.1476
black            -0.1431       -0.0800     -0.1442      -0.0775
metro             0.0963        0.1417      0.0953       0.1402
south            -0.0730       -0.0185     -0.0713      -0.0183
midwest          -0.0908       -0.0371     -0.0906      -0.0358
west             -0.0351        0.0219     -0.0337       0.0227
>
```

variable: educ, exper, south CI narrower
variable: black, midwest CI wider
FGLS provides more precise inference under the assumption that the variance model is well specified (明顯要管變量數)

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)?

```
t test of coefficients:

              Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  1.1922e+00  3.2360e-02  36.8422 < 2.2e-16 ***
educ         1.0166e-01  1.8928e-03  53.7107 < 2.2e-16 ***
exper        3.0090e-02  1.3046e-03  23.0643 < 2.2e-16 ***
I(exper^2)  -4.5614e-04  2.7408e-05 -16.6423 < 2.2e-16 ***
female      -1.6621e-01  9.4381e-03 -17.6109 < 2.2e-16 ***
black       -1.1085e-01  1.5869e-02  -6.9857 3.020e-12 ***
metro        1.1777e-01  1.1563e-02  10.1851 < 2.2e-16 ***
south       -4.4843e-02  1.3834e-02  -3.2414  0.001193 **
midwest     -6.3192e-02  1.3713e-02  -4.6083 4.111e-06 ***
west        -5.4938e-03  1.4509e-02  -0.3787  0.704951
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> print(round(fgls_robust_ci, 4))
             2.5 %  97.5 %
(Intercept) 1.1288  1.2556
educ        0.0980  0.1054
exper       0.0275  0.0326
I(exper^2) -0.0005 -0.0004
female     -0.1847 -0.1477
black      -0.1420 -0.0797
metro       0.0951  0.1404
south      -0.0720 -0.0177
midwest    -0.0901 -0.0363
west       -0.0339  0.0229
>
```

(d) (e) (f)

變數	OLS Robust CI	FGLS CI	FGLS + Robust CI
(Intercept)	1.1371 ~ 1.2656	1.1303 ~ 1.2541	1.1288 ~ 1.2556
educ	0.0975 ~ 0.1050	0.0982 ~ 0.1051	0.0980 ~ 0.1054
exper	0.0270 ~ 0.0322	0.0275 ~ 0.0326	0.0275 ~ 0.0326
I(exper²)	-0.0005 ~ -0.0004	-0.0005 ~ -0.0004	-0.0005 ~ -0.0004
female	-0.1841 ~ -0.1469	-0.1848 ~ -0.1476	-0.1847 ~ -0.1477
black	-0.1431 ~ -0.0800	-0.1442 ~ -0.0775	-0.1420 ~ -0.0797
metro	0.0963 ~ 0.1417	0.0953 ~ 0.1402	0.0951 ~ 0.1404
south	-0.0730 ~ -0.0185	-0.0713 ~ -0.0183	-0.0720 ~ -0.0177
midwest	-0.0908 ~ -0.0371	-0.0906 ~ -0.0358	-0.0901 ~ -0.0363
west	-0.0351 ~ 0.0219	-0.0337 ~ 0.0227	-0.0339 ~ 0.0229

FGLS + Robust SE 的 CI 多數介於 OLS Robust 和 FGLS 之間

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

FGLS + Robust SE，考慮可能的異質變異數問題，這種方法最能反映出模型的真實關係，並提升估計與檢定的可靠性。

8.18R

```r
data <- cps5
str(data)
# 8.18(a)
model_ols <- lm(log(wage) ~ educ + exper + I(exper^2) + female + black +
                  metro + south + midwest + west, data = data)

# 分割資料為男性和女性子樣本
data_male <- subset(data, female == 0)
data_female <- subset(data, female == 1)

model_male <- lm(log(wage) ~ educ + exper + I(exper^2) + black +
                   metro + south + midwest + west, data = data_male)
model_female <- lm(log(wage) ~ educ + exper + I(exper^2) + black +
                     metro + south + midwest + west, data = data_female)

# 獲取每個模型的殘差平方和(SSE)和自由度(df)
sse_male <- sum(model_male$residuals^2)
sse_female <- sum(model_female$residuals^2)
df_male <- model_male$df.residual
df_female <- model_female$df.residual

# 計算F統計量 (方差比)
# 以男性為分子，女性為分母計算F統計量
F_stat <- (sse_male/df_male)/(sse_female/df_female)

# 計算 F 分佈臨界值 (5%顯著水準)
F_critical_upper <- qf(1 - 0.05/2, df_male, df_female)
F_critical_lower <- qf(0.05/2, df_male, df_female)

cat("自由度: ", df_male, "和", df_female, "\n")
cat("F統計量值: ", F_stat,"\n")
cat("5%顯著水準的臨界值: ", F_critical_lower,"和", F_critical_upper, "\n")
p_value <- 2 * min(pf(F_stat, df_male, df_female, lower.tail = TRUE),
                   pf(F_stat, df_male, df_female, lower.tail = FALSE))

cat("p 值: ", p_value, "\n")

if (F_stat < F_critical_lower || F_stat > F_critical_upper) {
  cat("=> 拒絕虛無假設：男性與女性的誤差變異數顯著不同。\n")
} else {
  cat("=> 無法拒絕虛無假設：無足夠證據認為男性與女性的誤差變異數不同。\n")
}
```

```r
(b)
# 取得殘差平方
data$e2 <- resid(model_ols)^2

# 第二步：以 e2 作為依變數，對 METRO, FEMALE, BLACK 做迴歸
aux_model1 <- lm(e2 ~ metro + female + black, data = data)

# NR² 計算
n <- nrow(data)
R2_aux1 <- summary(aux_model1)$r.squared
NR2_1 <- n * R2_aux1

# 卡方臨界值與 p 值 (自由度 = 3)
p_val1 <- 1 - pchisq(NR2_1, df = 3)

cat("---- 使用 METRO, FEMALE, BLACK 的 NR² 測試 ----\n")
cat("NR² 統計量:", NR2_1, "\n")
cat("p 值:", p_val1, "\n")

if (p_val1 < 0.01) {
  cat("=> 在 1% 顯著水準下，拒絕虛無假設，存在異質變異數。\n")
} else {
  cat("=> 在 1% 顯著水準下，無法拒絕虛無假設，無明顯異質變異數。\n")
}
# 第二個檢定：用所有解釋變數
aux_model2 <- lm(e2 ~ educ + exper + I(exper^2) + female + black +
                   metro + south + midwest + west, data = data)

R2_aux2 <- summary(aux_model2)$r.squared
NR2_2 <- n * R2_aux2

# 自由度 = 9 (所有右側變數數量)
p_val2 <- 1 - pchisq(NR2_2, df = 9)

cat("\n---- 使用所有解釋變數的 NR² 測試 ----\n")
cat("NR² 統計量:", NR2_2, "\n")
cat("p 值:", p_val2, "\n")

if (p_val2 < 0.01) {
  cat("=> 在 1% 顯著水準下，拒絕虛無假設，存在異質變異數。\n")
} else {
  cat("=> 在 1% 顯著水準下，無法拒絕虛無假設，無明顯異質變異數。\n")
}
```

```r
#(c)
# 進行 White 檢定，輔助回歸包含原始變數、平方項和所有交互項
white_test <- bptest(model_ols,
                     ~ educ + exper + I(exper^2) + female + black + metro + south + midwest + west +
                       I(educ^2) + I(exper^4) +
                       educ:exper + educ:female + educ:black + educ:metro + educ:south + educ:midwest + educ:west + educ:I(exper^2) +
                       exper:female + exper:black + exper:metro + exper:south + exper:midwest + exper:west + exper:I(exper^2) +
                       female:black + female:metro + female:south + female:midwest + female:west + female:I(exper^2) +
                       black:metro + black:south + black:midwest + black:west + black:I(exper^2) +
                       metro:south + metro:midwest + metro:west + metro:I(exper^2) +
                       south:I(exper^2) + midwest:I(exper^2) + west:I(exper^2),
                     data = data)

print(white_test)

# 臨界值 (5%顯著水準)
df_white <- white_test$parameter  # 使用 bptest 中的自由度
critical_value_white <- qchisq(1-0.05, df_white)

cat("White 檢定 NR^2:", white_test$statistic, "\n")
cat("自由度:", white_test$parameter, "\n")
cat("5% 顯著水準的臨界值:", critical_value_white, "\n")

#(d)
library(sandwich)
summary_ols <- summary(model_ols)
coeftest_white <- coeftest(model_ols, vcov = vcovHC(model_ols, type = "HC0"))
# 傳統 OLS
coef_ols <- coef(summary_ols)
ci_ols <- cbind(
  coef_ols[, 1] - 1.96 * coef_ols[, 2],
  coef_ols[, 1] + 1.96 * coef_ols[, 2]
)
colnames(ci_ols) <- c("CI_low_OLS", "CI_high_OLS")

# White robust
coef_white <- coeftest_white
ci_white <- cbind(
  coef_white[, 1] - 1.96 * coef_white[, 2],
  coef_white[, 1] + 1.96 * coef_white[, 2]
)
```

```r
colnames(ci_white) <- c("CI_low_White", "CI_high_White")

# 合併
comparison <- cbind(
  Estimate = coef_ols[, 1],
  SE_OLS = coef_ols[, 2],
  SE_White = coef_white[, 2],
  ci_ols,
  ci_white
)

print(round(comparison, 4))

#(e)
# 殘差平方的 log
data$resid2 <- residuals(model_ols)^2
data$log_resid2 <- log(data$resid2)
# 用來建構誤差變異數模型的輔助回歸
aux_model <- lm(log_resid2 ~ metro + exper, data = data)

# 預測出 log(sigma^2_hat)
log_sigma2_hat <- predict(aux_model)

# 轉換成權重 (w = 1 / sigma_hat)
weights_fgls <- 1 / exp(log_sigma2_hat)
# 使用權重回歸 (FGLS)
model_fgls <- lm(log(wage) ~ educ + exper + I(exper^2) + female + black +
                   metro + south + midwest + west,
                 data = data,
                 weights = weights_fgls)

summary_fgls <- summary(model_fgls)
# 95% 信賴區間
ci_fgls <- confint(model_fgls)

# 合併比較 FGLS 與 White robust SE 的信賴區間
comparison_ci <- cbind(
  CI_low_White = ci_white[,1],
  CI_high_White = ci_white[,2],
  CI_low_FGLS = ci_fgls[,1],
  CI_high_FGLS = ci_fgls[,2]
)

print(round(comparison_ci, 4))

#(f)
se_fgls_robust <- coeftest(model_fgls, vcov = vcovHC(model_fgls, type = "HC1"))

# 提取 robust CI (使用 robust SE)
fgls_robust_ci <- coefci(model_fgls, vcov. = vcovHC(model_fgls, type = "HC1"))

# 查看 robust 結果
print(se_fgls_robust)
print(round(fgls_robust_ci, 4))
```