8.6 Consider the wage equation

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

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.

$H_0: \sigma_m^2 = \sigma_F^2$, $H_1: \sigma_m^2 \neq \sigma_F^2$ $\quad \sigma_m^2 = \dfrac{SSE_1}{N_1 - k_1} = \dfrac{97161.9174}{577-4} = 169.567$

$GQ = \dfrac{\sigma_m^2}{\sigma_F^2} = \dfrac{169.567}{12.024^2} = 1.1729$ * $\sim F_{(573,414)}$

⇒ 5% region region = $GQ \geq F_{(0.975, 573, 414)} = 1.1968$

or

$GQ \leq F_{(0.025, 573, 414)} = 0.8377$ *

⇒ since the test statistic does not fall in the reject region

⇒ not reject H_0

⇒ no significant differnce in wage variance between males and female. *

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

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

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.

$H_0: \sigma^2_{\text{married}} = \sigma^2_{\text{single}} \qquad H_1: \sigma^2_{\text{Married}} > \sigma^2_{\text{Single}}$

$$\hat{\sigma}^2_{\text{married}} = \frac{100703.0471}{600-5} = 169.2488$$

$$\hat{\sigma}^2_{\text{single}} = \frac{56231.0382}{400-5} = 142.3571$$

$$\Rightarrow F = \frac{\hat{\sigma}^2_{\text{married}}}{\hat{\sigma}^2_{\text{single}}} = 1.1889 > F_{(0.95,595,395)} = 1.1647$$

reject region: $F \geq F_{(0.95,595,395)} = 1.1647$

$\Rightarrow$ reject H_0 *

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.

$H_0: \alpha_2 = \alpha_3 = \cdots = \alpha_5 = 0 \Rightarrow$ homoskedasticity

$H_1: \alpha_2 \alpha_3 \cdots \alpha_5 \neq 0 \Rightarrow$ heteroskedasticity

$BP = NR^2 = 59.03 > \chi^2_{0.95,4} = 9.488 \Rightarrow$ reject $H_0 \Rightarrow$ heteroskedasticity *

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

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

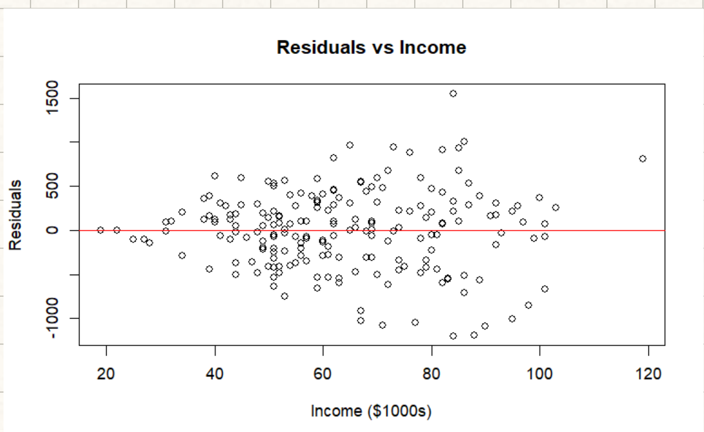
MILES is miles driven per year, *INCOME* is measured in \$1000 units, *AGE* is the average age of the adult members of the household, and *KIDS* is the number of children.

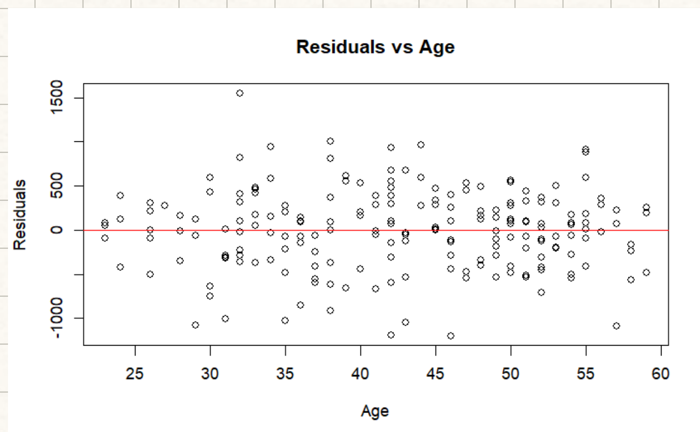
a. Use the data file *vacation* to estimate the model by OLS. Construct a 95% interval estimate for the effect of one more child on miles traveled, holding the two other variables constant.

```
         2.5 %    97.5 %
kids -135.3298 -28.32302
```

⇒ 95% interval = $[-135.3298, -28.32302]$ *

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





⇒ In both plot, there appears to be no strong evidence of serious heteroskedasticity. *

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.

$H_0 : \sigma_{low}^2 = \sigma_{high}^2$ $\quad H_1 : \sigma_{low}^2 \neq \sigma_{high}^2$

```
F statistic: 3.104061
p-value: 1.64001e-07
```

⇒ p-value $\ll 0.05$

⇒ reject H_0, there is a strong evidence 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)?

```
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

      2.5 %      97.5 %
-139.32297   -24.32986
```

⇒ $\widehat{MILES} = -391.5480 + 14.2013 \cdot INCOME + 15.7409 \cdot AGE - 81.8264 \cdot KIDS$

⇒ 95% interval = $[-139.32297, -24.32986]$

⇒ slightly wider than in part (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)?

```
                 2.5 %     97.5 %
I(kids * w) -118.7497 -30.19167
```

⇒ 95% interval of conventional GLS $= [-118.7497, -30.19167]$

```
      2.5 %      97.5 %
-122.00303   -26.93833
```

⇒ 95% interval of robust GLS $= [-122.00303, -26.93833]$

The GLS intervals are slightly narrower than the OLS-based ones, indicating a potential gain in efficiency. However, the robust intervals are more conservative, especially under heteroskedasticity, ensuring inference validity. *

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.

$H_0: \sigma_M^2/\sigma_F^2 = 1$, $H_1: \sigma_M^2/\sigma_F^2 \neq 1$

```
[1] 1.05076
[1] 0.9450933 1.0580966
```

$\because$ $GQ = 1.05076 > 0.9450933$ and

$GQ = 1.05076 < 1.0580966$

$\Rightarrow$ not reject H_0

$\Rightarrow$ There is no statistically significant difference in the wage variance between males and females.

*

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.

```
[1] 23.55681
[1] 3.0909e-05
[1] 109.4243
[1] 1.925849e-19
```

$\Rightarrow$ p-value < 0.01

$\Rightarrow$ p-value < 0.01

$\Rightarrow$ Strong evidence of heteroskedasticity *

This contrasts with part (a), where the GQ test found no significant difference in wage variance between males and females. Therefore, heteroskedasticity likely arises from other variables, not just gender.

*

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
BP = 165.46, df = 41, p-value < 2.2e-16

[1] 56.94239
```

$\Rightarrow \because 165.46 \gg 56.94239 \Rightarrow$ reject H_0

$\Rightarrow$ There is strong evidence of heteroskedasticity in the model