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.

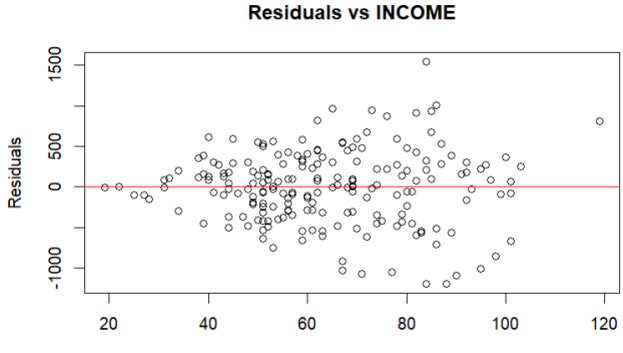
a. Use the data file vacation to estimate the model by OLS. Construct a 95% interval estimate for the effect of one more child on miles traveled, holding the two other variables constant.

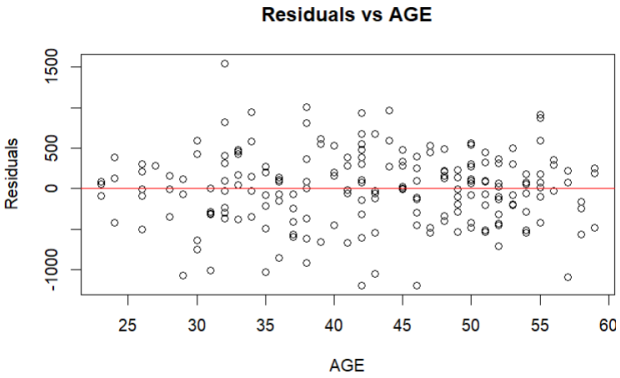
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -391.548    169.775  -2.306   0.0221 *
income        14.201      1.800   7.889 2.10e-13 ***
age           15.741      3.757   4.189 4.23e-05 ***
kids         -81.826     27.130  -3.016   0.0029 **
---
Signif. codes:  0 ‘***’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

```
            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?





Based on the residual plots:
For INCOME, the spread of residuals appears to increase as income increases, suggesting **possible heteroskedasticity**.
For AGE, the residuals seem to be evenly spread, showing **no strong evidence of heteroskedasticity**.
Therefore, we suspect the presence of heteroskedasticity related to INCOME.

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.

```
        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

>
> qf(0.95,86,86)
[1] 1.428617
```

1. H0: σ1=σ2 against H1: σ2>σ1
2. Reject region: t-stat>1.428617
3. The t-stat is 3.1041 in the rr, so we reject h0, conclude that σ2>σ1.

last

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -476.803    548.833  -0.869   0.3874
income        15.556      5.450   2.855   0.0054 *
age           16.388      7.385   2.219   0.0291 *
kids        -116.017     49.861  -2.327   0.0223 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 562 on 86 degrees of freedom
Multiple R-squared:  0.1514,    Adjusted R-squared:  0.1218
F-statistic: 5.116 on 3 and 86 DF,  p-value: 0.002642
```

first

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -392.511    214.166  -1.833  0.07030 .
income        10.960      3.770   2.907  0.00464 **
age           18.869      3.783   4.988 3.14e-06 ***
kids         -70.371     29.138  -2.415  0.01785 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 319 on 86 degrees of freedom
Multiple R-squared:  0.309,     Adjusted R-squared:  0.2849
F-statistic: 12.82 on 3 and 86 DF,  p-value: 5.31e-07
```

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.

```
             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

> ci_robust <- coefci(model_ols, vcov. = robust_se, level = 0.95)
> ci_robust["kids", ]
     2.5 %     97.5 %
-139.32297  -24.32986
```

The interval is wider than (a.)

e. Obtain GLS estimates assuming σi^2 = σ^2**INCOME2**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)?**

```
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, "kids", level = 0.95)
         2.5 %    97.5 %
kids -119.8945 -33.71808

             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, "kids", vcov. = vcovHC(model_gls, type = "HC1"), level = 0.95)
         2.5 %    97.5 %
kids -121.4134 -32.19919
```