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.

a. $N_M = 577$, $df_M = 577 - 4 = 573$, $df_F = 423 - 4 = 419$

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

$\hat{\sigma}_M^2 = \dfrac{SSE_M}{df_M} = \dfrac{97161.9174}{573} \doteq 169.57$ $\quad \hat{\sigma}_F^2 = (12.024)^2 = 144.5766$

$GQ = \dfrac{\hat{\sigma}_M^2}{\hat{\sigma}_F^2} = \dfrac{169.57}{144.5766} = 1.1729$

$F_{(573,419,0.025)} = 0.8377$

$F_{(573,419,0.975)} = 1.1968$

$\Rightarrow \because 0.8377 < 1.1729 < 1.1968$ Do not reject H_0. #

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

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

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.

b. $N_{single} = 400$, $N_{married} = 600$

$df_{sin.} = 400 - 5 = 395$, $df_{Mar.} = 600 - 5 = 595$

$SSE_{sin.} = 56231.0382$, $SSE_{Mar.} = 100,703.0471$

$\begin{cases} H_0 : \sigma_{Mar}^2 = \sigma_{sin}^2 \\ H_1 : \sigma_{Mar}^2 > \sigma_{sin}^2 \end{cases}$

$\hat{\sigma}_{sin.}^2 = \dfrac{56231.0382}{395} = 142.3571$, $\hat{\sigma}_{mar.}^2 = \dfrac{100,703.0471}{595} = 169.2488$

$GQ = \dfrac{\hat{\sigma}_{Mar}^2}{\hat{\sigma}_{sin}^2} = \dfrac{169.2488}{142.3571} = 1.1889$

$F_{(595,395,0.95)} = 1.1647$ $\quad \because 1.1889 > 1.1647 \quad \therefore$ Reject H_0 #

```
> qf(0.95, df1 = 595, df2 = 395)
[1] 1.164705
```

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?

c.

$$\begin{cases} H_0 : \text{homoskedasticity} \\ H_1 : \text{heteroskedasticity} \end{cases}$$

$df = 5-1 = 4$

$\chi^2_{(0.95,4)} = 9.488 < NR^2 = 59.03$

```
> qchisq(0.95,4)
[1] 9.487729
```

∴ Reject the null hypothesis of homoskedasticity in the pooled regression.

d. 原始變數 (4) : EDUC、EXPER、METRO、FEMALE/MALE

連續變數平方項及交乘項 (3) : $EDUC^2$、$EXPER^2$、EDUC × EXPER

連續 × 虛擬變數(4) : EDUC × MARRIED 、EXPER × MARRIED

EDUC × FEMALE 、EXPER × FEMALE

虛擬 × 虛擬 (1) : FEMALE × MARRIED

$df = 4+3+4+1 = 12$

$\chi^2_{(12,0.95)} = 21.0261 < 78.82$

∴ Reject the null hypothesis of homoskedasticity in the pooled regression.

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?

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.

e、

截距項和 EDUC 係數的區間估計變得更寬、

其他變數的區間估計則變得更窄

並不矛盾，∵ robust standard errors 有更能比錯誤的 OLS 標準誤來得更大或更小。

f、

用 MARRIED 檢驗是否對薪資的平均影響顯著，t值為 1 ∴不顯著

但(b)是用 F 檢定 檢驗已婚者和未婚者兩組人在薪資的變動程度上不同

是為異質變異數的來源 。這兩個是不同的概念、並不矛盾。

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)?

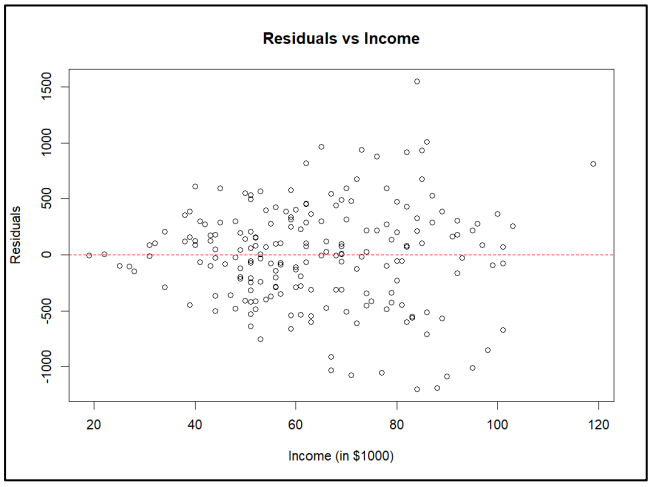
(a)

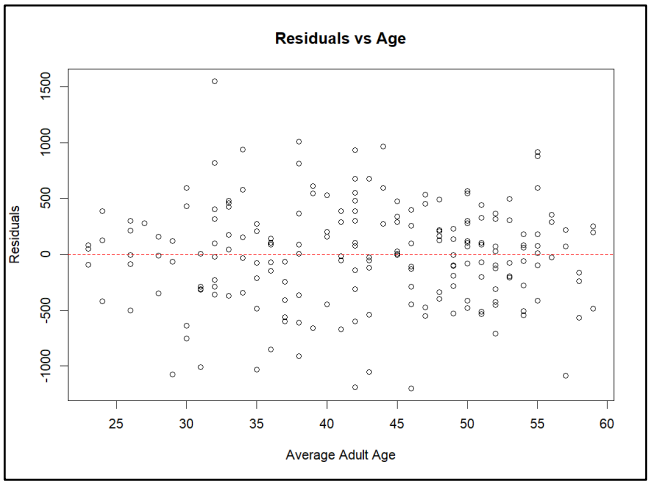
```
> summary(model)

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, 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)

在殘差對收入的圖表中，隨著收入增加，殘差的變異增加。在殘差對年齡的圖表中，這種效應並不明顯。





(c)

我們可以得出結論，誤差變異取決於收入。殘差的變異隨著收入的增加而增加。

```
> cat("F 統計量 =", round(F_stat, 4), "\n")
F 統計量 = 3.1041
> cat("臨界值 =", round(crit_val, 4), "\n")
臨界值 = 1.4286
```

(d)

在(a)中，使用常規 OLS 方法，對於每增加一個孩子的影響估計為-81.82642，標準誤差為 27.1296，95%置信區間估計為[-135.3298，-28.32302]。使用異方差 robust standard error，對於每增加一個孩子的影響估計仍為-81.82642，標準誤差為 29.15438，95%置信區間估計為[-139.323，-24.32986]。 與(a)相比，由於具有較大的標準誤差，其區間更寬。

```
> # 強健標準誤估計
> coeftest(model, vcov. = vcovHC(model, type = "HC1"))

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

>
> # 計算 kids 的 robust CI
> b_kids <- coef(model)["kids"]
> robust_se_kids <- sqrt(vcovHC(model, type = "HC1")["kids", "kids"])
> df <- df.residual(model)
> CI_robust <- b_kids + c(-1, 1) * qt(0.975, df) * robust_se_kids
> CI_robust
[1] -139.32297  -24.32986
```

(e)

```
> # 查看 GLS 結果
> summary(gls_model)

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

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

>
> # 常規 GLS 的 95% 信賴區間
> confint(gls_model, "kids", level = 0.95)
         2.5 %    97.5 %
kids -119.8945 -33.71808
```

	OLS	robust standard errors	GLS
$\beta 4$	-81.826	-81.826	-76.806
Se($\beta 4$)	27.13	29.1544	21.848
95% interval	[-135.3298,-28.323]	[-139.3223,-24.3299]	[-119.8945,-33.7181]

使用 GLS robust standard error，對於每增加一個孩子的影響估計為-76.80629，標準誤為 21.84844，95%區間估計為[-119.8945，-33.71808]。和 OLS、robust standard errors 相比，GLS 的標準誤更小、區間相對較窄。

8.28 In this exercise you will create some simulated data and try out estimation and testing methods. Use your software to create a new data set, or "workfile," with $N = 100$ observations. All modern software has functions, called random number generators, to create uniformly distributed and normally distributed random values. Follow these steps.

1. Create $X2 = 1 + 5 \times U1$, where $U1$ is a random number between zero and one.
2. Create $X3 = 1 + 5 \times U2$, where $U2$ is another random number between zero and one.
3. Create $E = \sqrt{\exp(2 + 0.6X2)} \times Z$, where $Z \sim N(0, 1)$.
4. Create $Y = 5 + 4X2 + E$

You should now have 100 values for Y, $X2$, and $X3$. Note: Your results should be different from your classmates, and your results might change from one experiment to the next. To prevent this from happening, you can set the random number's "seed." See your software documentation for instructions.

a. Regress Y on $X2$ and $X3$ and obtain conventional OLS standard errors. Compare the estimated coefficients to the true values of the regression parameters, $\beta_1 = 5$, $\beta_2 = 4$, $\beta_3 = 0$. Do the t-values suggest that the coefficients are significantly different from 0 at the 5% level?

b. Calculate the least squares residuals $\hat{e}$ from the OLS estimation in (a) and regress $\hat{e}^2$ on $X2$ and $X3$. What evidence, if any, do you find for the presence of heteroskedasticity?

c. Regress Y on $X2$ and $X3$ and obtain robust standard errors. Compare these to the conventional standard errors in (a).

d. Assume the heteroskedasticity pattern is $\sigma^2 X2^2$. Obtain GLS estimates with conventional and robust standard errors. Are the GLS parameter estimates closer to the true parameter values or not? Which set of standard errors should be used?

e. Assume the multiplicative heteroskedasticity model $\exp(\alpha_1 + \alpha_2 X2 + \alpha_3 X3)$. Obtain FGLS estimates with conventional and robust standard errors. Are the FGLS estimates closer to the true parameter values than the GLS or OLS estimates? Which set of standard errors should be used?

(a)

```
Call:
lm(formula = Y ~ X2 + X3)

Residuals:
    Min      1Q  Median      3Q     Max
-22.317  -4.606  -0.723   3.872  23.531

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.0863     3.2068   1.586    0.116
X2            3.9329     0.5720   6.876    6e-10 ***
X3           -0.1839     0.6180  -0.298    0.767
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.079 on 97 degrees of freedom
Multiple R-squared:  0.3315,    Adjusted R-squared:  0.3177
F-statistic: 24.05 on 2 and 97 DF,  p-value: 3.301e-09
```

```
T value1: 0.02690609
 T value2: -0.1173035
 T value3: -0.2975644
```

The result shows that beta_1 is not significantly different from 5 at the 5% significance level, beta_2 is not significantly different from 4 at the 5% significance level, and beta_3 is not significantly different from 0 at the 5% significance level.

(b)

$\widehat{e^2}$ = -29.3859 + 26.996X2 - 0.4187X3

The regression R^2=0.138, so that the BP test statistic NR^2 = 13.8

Set alpha = 0.05, H0:Homoskedasticity vs H1:Hetroskedasticity

Since the p-value in the BP test is 0.000203 less than alpha=0.05, and $\chi^2_{(2,0.95)} = 5.99146$, we reject H0 and conclude that heteroskedasticity may exist.

```
> # 計算 OLS 殘差與平方
> resid <- residuals(mod1)
> resid_squared <- resid^2
>
> # 進行殘差平方回歸
> resid_mod1 <- lm(resid_squared ~ X2 + X3)
> summary(resid_mod1)

Call:
lm(formula = resid_squared ~ X2 + X3)

Residuals:
    Min      1Q  Median      3Q     Max
-118.68  -58.02  -16.90   16.03  427.98

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -29.3859    38.5668  -0.762 0.447939
X2           26.9660     6.8788   3.920 0.000165 ***
X3           -0.4187     7.4322  -0.056 0.955189
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 97.16 on 97 degrees of freedom
Multiple R-squared:  0.138,     Adjusted R-squared:  0.1202
F-statistic: 7.764 on 2 and 97 DF,  p-value: 0.0007453
```

```
> # Breusch-Pagan 檢定的計算邏輯
> N <- nobs(resid_mod1)
> gresid_mod1 <- glance(resid_mod1)  # 使用 broom 的 glance() 取得摘要
> S <- gresid_mod1$df.residual + length(coef(resid_mod1))  # 模型自由度總數
> Rsqres <- gresid_mod1$r.squared
> chisq <- N * Rsqres
> pval <- 1 - pchisq(chisq, S - 1)
>
> # 顯示結果
> cat('Breusch-Pagan 檢定結果：\n')
Breusch-Pagan 檢定結果：
> cat('Chi-square statistic:', round(chisq, 4), '\n')
Chi-square statistic: 13.799
> cat('p-value:', round(pval, 5), '\n')
p-value: 1
```

(c)

The robust standard errors are larger for constant term and X2 but not for the X3.

```
> # 使用 robust 標準誤估計（HC1 是 White 常用的）
> coeftest(model, vcov. = vcovHC(model, type = "HC1"))

t test of coefficients:

            Estimate Std. Error t value  Pr(>|t|)
(Intercept)  5.08628    3.21726  1.5809    0.1171
X2           3.93291    0.66486  5.9154 4.967e-08 ***
X3          -0.18389    0.57317 -0.3208    0.7490
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(NEXT PAGE)

(d)

```
> # 顯示常規標準誤下的 GLS 結果
> summary(gls_model)

Call:
lm(formula = Y ~ X2 + X3, weights = weights)

Weighted Residuals:
    Min      1Q  Median      3Q     Max
-6.4724 -1.6926 -0.2707  1.3218  7.9946

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.5334     2.6965   1.310    0.193
X2            3.8748     0.5587   6.935 4.53e-10 ***
X3            0.2937     0.5154   0.570    0.570
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.486 on 97 degrees of freedom
Multiple R-squared:  0.3403,    Adjusted R-squared:  0.3267
F-statistic: 25.02 on 2 and 97 DF,  p-value: 1.726e-09
```

```
> # 強健標準誤的 GLS 結果
> coeftest(gls_model, vcov. = vcovHC(gls_model, type = "HC1"))

t test of coefficients:

            Estimate Std. Error t value  Pr(>|t|)
(Intercept)  3.53341    3.11628  1.1339    0.2596
X2           3.87482    0.68961  5.6189 1.839e-07 ***
X3           0.29366    0.53670  0.5471    0.5855
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

	True value	OLS	Robust OLS	GLS	Robust GLS
C	5	5.0863 (3.2068)	5.08628 (3.21726)	3.5334 (2.6965)	3.53341 (3.11628)
X2	4	3.9329 (0.5720)	3.93291 (0.66486)	3.8748 (0.5587)	3.87482 (0.68961)
X3	0	-0.1839 (0.6180)	-0.18389 (0.57317)	0.2937 (0.5154)	0.29366 (0.53670)

Compared to OLS estimates, GLS estimates are not closer to the true parameter values in all estimates.

(e)

$\widehat{\ln(e^2)}$ = 1.35562 + 0.46679X2 - 0.04322X3

```
> fgls_model <- lm(Y ~ X2 + X3, weights = weights_fgls)
> summary(fgls_model)

Call:
lm(formula = Y ~ X2 + X3, weights = weights_fgls)

Weighted Residuals:
    Min      1Q  Median      3Q     Max
-3.4490 -1.3154 -0.1989  1.1252  5.7961

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.53612    2.74603   1.652    0.102
X2           3.88073    0.54587   7.109 1.99e-10 ***
X3           0.01142    0.52536   0.022    0.983
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.799 on 97 degrees of freedom
Multiple R-squared:  0.3516,    Adjusted R-squared:  0.3382
F-statistic:  26.3 on 2 and 97 DF,  p-value: 7.507e-10
```

```
> coeftest(fgls_model, vcov. = vcovHC(fgls_model, type = "HC1"))

t test of coefficients:

            Estimate Std. Error t value  Pr(>|t|)
(Intercept) 4.536118   2.903965  1.5620    0.1215
X2          3.880729   0.559686  6.9338 4.568e-10 ***
X3          0.011417   0.527819  0.0216    0.9828
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

	True value	FGL	Robust FGL
C	5	4.53612 (2.74603)	4.536118 (2.903965)
X2	4	3.88073 (0.54587)	3.880729 (0.559686)
X3	0	0.01142 (0.52536)	0.011417 (0.527819)

FGLS estimates are in fact based on a skedastic function family that includes the true one. Compared to GLS, all estimates in FGLS are closer to the true parameter.