CHAPTER 8

Exercise Solutions

EXERCISE 8.1

When $\sigma_i^2 = \sigma^2$

$$\frac{\sum_{i=1}^{N}\left[(x_i-\bar{x})^2\sigma_i^2\right]}{\left[\sum_{i=1}^{N}(x_i-\bar{x})^2\right]^2} = \frac{\sum_{i=1}^{N}\left[(x_i-\bar{x})^2\sigma^2\right]}{\left[\sum_{i=1}^{N}(x_i-\bar{x})^2\right]^2} = \frac{\sigma^2\sum_{i=1}^{N}(x_i-\bar{x})^2}{\left[\sum_{i=1}^{N}(x_i-\bar{x})^2\right]^2} = \frac{\sigma^2}{\sum_{i=1}^{N}(x_i-\bar{x})^2}$$

EXERCISE 8.2

(a) Multiplying the first normal equation by $\left(\sum \sigma_i^{-1} x_i^*\right)$ and the second one by $\left(\sum \sigma_i^{-2}\right)$ yields

$$\left(\sum \sigma_i^{-1} x_i^*\right)\left(\sum \sigma_i^{-2}\right)\hat{\beta}_1 + \left(\sum \sigma_i^{-1} x_i^*\right)^2 \hat{\beta}_2 = \left(\sum \sigma_i^{-1} x_i^*\right)\sum \sigma_i^{-1} y_i^*$$

$$\left(\sum \sigma_i^{-2}\right)\left(\sum \sigma_i^{-1} x_i^*\right)\hat{\beta}_1 + \left(\sum \sigma_i^{-2}\right)\left(\sum x_i^{*2}\right)\hat{\beta}_2 = \left(\sum \sigma_i^{-2}\right)\sum x_i^* y_i^*$$

Subtracting the first of these two equations from the second yields

$$\left[\left(\sum \sigma_i^{-2}\right)\left(\sum x_i^{*2}\right) - \left(\sum \sigma_i^{-1} x_i^*\right)^2\right]\hat{\beta}_2 = \left(\sum \sigma_i^{-2}\right)\sum x_i^* y_i^* - \left(\sum \sigma_i^{-1} x_i^*\right)\sum \sigma_i^{-1} y_i^*$$

Thus,

$$\hat{\beta}_2 = \frac{\left(\sum \sigma_i^{-2}\right)\sum x_i^* y_i^* - \left(\sum \sigma_i^{-1} x_i^*\right)\left(\sum \sigma_i^{-1} y_i^*\right)}{\left(\sum \sigma_i^{-2}\right)\left(\sum x_i^{*2}\right) - \left(\sum \sigma_i^{-1} x_i^*\right)^2}$$

$$= \frac{\dfrac{\sum \sigma_i^{-2} y_i x_i}{\sum \sigma_i^{-2}} - \left(\dfrac{\sum \sigma_i^{-2} y_i}{\sum \sigma_i^{-2}}\right)\left(\dfrac{\sum \sigma_i^{-2} x_i}{\sum \sigma_i^{-2}}\right)}{\dfrac{\sum \sigma_i^{-2} x_i^2}{\sum \sigma_i^{-2}} - \left(\dfrac{\sum \sigma_i^{-2} x_i}{\sum \sigma_i^{-2}}\right)^2}$$

In this last expression, the second line is obtained from the first by making the substitutions $y_i^* = \sigma_i^{-1} y_i$ and $x_i^* = \sigma_i^{-1} x_i$, and by dividing numerator and denominator by $\left(\sum \sigma_i^{-2}\right)^2$. Solving the first normal equation $\left(\sum \sigma_i^{-2}\right)\hat{\beta}_1 + \left(\sum \sigma_i^{-1} x_i^*\right)\hat{\beta}_2 = \sum \sigma_i^{-1} y_i^*$ for $\hat{\beta}_1$ and making the substitutions $y_i^* = \sigma_i^{-1} y_i$ and $x_i^* = \sigma_i^{-1} x_i$, yields

$$\hat{\beta}_1 = \frac{\sum \sigma_i^{-2} y_i}{\sum \sigma_i^{-2}} - \left(\frac{\sum \sigma_i^{-2} x_i}{\sum \sigma_i^{-2}}\right)\hat{\beta}_2$$

(b) When $\sigma_i^2 = \sigma^2$ for all i, $\sum \sigma_i^{-2} y_i x_i = \sigma^{-2} \sum y_i x_i$, $\sum \sigma_i^{-2} y_i = \sigma^{-2} \sum y_i$, $\sum \sigma_i^{-2} x_i = \sigma^{-2} \sum x_i$, and $\sum \sigma_i^{-2} = N\sigma^{-2}$. Making these substitutions into the expression for $\hat{\beta}_2$ yields

$$\hat{\beta}_2 = \frac{\dfrac{\sigma^{-2}\sum y_i x_i}{N\sigma^{-2}} - \left(\dfrac{\sigma^{-2}\sum y_i}{N\sigma^{-2}}\right)\left(\dfrac{\sigma^{-2}\sum x_i}{N\sigma^{-2}}\right)}{\dfrac{\sigma^{-2}\sum x_i^2}{N\sigma^{-2}} - \left(\dfrac{\sigma^{-2}\sum x_i}{N\sigma^{-2}}\right)^2} = \frac{\dfrac{\sum y_i x_i}{N} - \bar{y}\,\bar{x}}{\dfrac{\sum x_i^2}{N} - \bar{x}^2}$$

and that for $\hat{\beta}_1$ becomes

$$\hat{\beta}_1 = \frac{\sigma^{-2}\sum y_i}{N\sigma^{-2}} - \left(\frac{\sigma^{-2}\sum x_i}{N\sigma^{-2}}\right)\hat{\beta}_2 = \bar{y} - \bar{x}\hat{\beta}_2$$

These formulas are equal to those for the least squares estimators b_1 and b_2. See pages 52 and 83-84 of the text.

Exercise 8.2 (continued)

(c) The least squares estimators b_1 and b_2 are functions of the following averages

$$\bar{x} = \frac{1}{N}\sum x_i \qquad \bar{y} = \frac{1}{N}\sum y_i \qquad \frac{1}{N}\sum x_i y_i \qquad \frac{1}{N}\sum x_i^2$$

For the generalized least squares estimator for $\hat{\beta}_1$ and $\hat{\beta}_2$, these unweighted averages are replaced by the weighted averages

$$\left(\frac{\sum \sigma_i^{-2} x_i}{\sum \sigma_i^{-2}}\right) \qquad \left(\frac{\sum \sigma_i^{-2} y_i}{\sum \sigma_i^{-2}}\right) \qquad \left(\frac{\sum \sigma_i^{-2} y_i x_i}{\sum \sigma_i^{-2}}\right) \qquad \left(\frac{\sum \sigma_i^{-2} x_i^2}{\sum \sigma_i^{-2}}\right)$$

In these weighted averages each observation is weighted by the inverse of the error variance. Reliable observations with small error variances are weighted more heavily than those with higher error variances that make them more unreliable.

EXERCISE 8.3

For the model $y_i = \beta_1 + \beta_2 x_i + e_i$ where $\text{var}(e_i) = \sigma^2 x_i^2$, the transformed model that gives a constant error variance is

$$y_i^* = \beta_1 x_i^* + \beta_2 + e_i^*$$

where $y_i^* = y_i / x_i$, $x_i^* = 1/x_i$, and $e_i^* = e_i / x_i$. This model can be estimated by least squares with the usual simple regression formulas, but with β_1 and β_2 reversed. Thus, the generalized least squares estimators for β_1 and β_2 are

$$\hat{\beta}_1 = \frac{N\sum x_i^* y_i^* - \sum x_i^* \sum y_i^*}{N\sum (x_i^*)^2 - \left(\sum x_i^*\right)^2} \quad \text{and} \quad \hat{\beta}_2 = \bar{y}^* - \hat{\beta}_1 \bar{x}^*$$

Using observations on the transformed variables, we find

$$\sum y_i^* = 7 \qquad \sum x_i^* = 4 \qquad \sum x_i^* y_i^* = 11/2 \qquad \sum (x_i^*)^2 = 7/2$$

With $N = 5$, the generalized least squares estimates are

$$\hat{\beta}_1 = \frac{5(11/2) - 4 \times 7}{5(7/2) - (4)^2} = -0.333$$

and

$$\hat{\beta}_2 = \bar{y}^* - \hat{\beta}_1 \bar{x}^* = (7/5) - (-0.333)\frac{4}{5} = 1.667$$

EXERCISE 8.4

(a) In the plot of the residuals against income the absolute value of the residuals increases as income increases, but the same effect is not apparent in the plot of the residuals against age. In this latter case there is no apparent relationship between the magnitude of the residuals and age. Thus, the graphs suggest that the error variance depends on income, but not age.

(b) Since the residual plot shows that the error variance may increase when income increases, and this is a reasonable outcome since greater income implies greater flexibility in travel, we set up the null and alternative hypotheses as the one tail test $H_0: \sigma_1^2 = \sigma_2^2$ versus $H_1: \sigma_1^2 > \sigma_2^2$, where σ_1^2 and σ_2^2 are artificial variance parameters for high and low income households. The value of the test statistic is

$$F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} = \frac{(2.9471\times 10^7)/(100-4)}{(1.0479\times 10^7)/(100-4)} = 2.8124$$

The 5% critical value for (96, 96) degrees of freedom is $F_{(0.95,96,96)} = 1.401$. Thus, we reject H_0 and conclude that the error variance depends on income.

<u>Remark</u>: An inspection of the file *vacation.dat* after the observations have been ordered according to *INCOME* reveals 7 middle observations with the same value for *INCOME*, namely 62. Thus, when the data are ordered only on the basis of *INCOME*, there is not one unique ordering, and the values for SSE_1 and SSE_2 will depend on the ordering chosen. Those specified in the question were obtained by ordering first by *INCOME* and then by *AGE*.

(c) (i) All three sets of estimates suggest that vacation miles travelled are directly related to household income and average age of all adults members but inversely related to the number of kids in the household.

(ii) The White standard errors are slightly larger but very similar in magnitude to the conventional ones from least squares. Thus, using White's standard errors leads one to conclude estimation is less precise, but it does not have a big impact on assessment of the precision of estimation.

(iii) The generalized least squares standard errors are less than the White standard errors for least squares, suggesting that generalized least squares is a better estimation technique.

EXERCISE 8.5

(a) The table below displays the 95% confidence intervals obtained using the critical *t*-value $t_{(0.975,497)} = 1.965$ and both the least squares standard errors and the White's standard errors. After recognizing heteroskedasticity and using White's standard errors, the confidence intervals for *CRIME*, *AGE* and *TAX* are narrower while the confidence interval for *ROOMS* is wider. However, in terms of the magnitudes of the intervals, there is very little difference, and the inferences that would be drawn from each case are similar. In particular, none of the intervals contain zero and so all of the variables have coefficients that would be judged to be significant no matter what procedure is used.

95% confidence intervals

	Least squares standard errors		White's standard errors	
	Lower	Upper	Lower	Upper
CRIME	– 0.255	– 0.112	– 0.252	– 0.114
ROOMS	5.600	7.143	5.065	7.679
AGE	– 0.076	– 0.020	– 0.070	– 0.026
TAX	– 0.020	– 0.005	– 0.019	– 0.007

(b) Most of the standard errors did not change dramatically when White's procedure was used. Those which changed the most were for the variables *ROOMS*, *TAX*, and *PTRATIO*. Thus, heteroskedasticity does not appear to present major problems, but it could lead to slightly misleading information on the reliability of the estimates for *ROOMS*, *TAX* and *PTRATIO*.

(c) As mentioned in parts (a) and (b), the inferences drawn from use of the two sets of standard errors are likely to be similar. However, keeping in mind that the differences are not great, we can say that, after recognizing heteroskedasticity and using White's standard errors, the standard errors for *CRIME*, *AGE*, *DIST*, *TAX* and *PTRATIO* decrease while the others increase. Therefore, using incorrect standard errors (least squares) understates the reliability of the estimates for *CRIME*, *AGE*, *DIST*, *TAX* and *PTRATIO* and overstates the reliability of the estimates for the other variables.

Remark: Because the estimates and standard errors are reported to 4 decimal places in Exercise 5.5 (Table 5.7), but only 3 in this exercise (Table 8.2), there will be some rounding error differences in the interval estimates in the above table. These differences, when they occur, are no greater than 0.001.

EXERCISE 8.6

(a) *ROOMS* significantly effects the variance of house prices through a relationship that is quadratic in nature. The coefficients for *ROOMS* and $ROOMS^2$ are both significantly different from zero at a 1% level of significance. Because the coefficient of $ROOMS^2$ is positive, the quadratic function has a minimum which occurs at the number of rooms for which

$$\frac{\partial \hat{e}^2}{\partial ROOMS} = \alpha_2 + 2\alpha_3 ROOMS = 0$$

Using the estimated equation, this number of rooms is

$$ROOMS_{\min} = \frac{-\hat{\alpha}_2}{2\hat{\alpha}_3} = \frac{305.311}{2 \times 23.822} = 6.4$$

Thus, for houses of 6 rooms or less the variance of house prices decreases as the number of rooms increases and for houses of 7 rooms or more the variance of house prices increases as the number of rooms increases.

The variance of house prices is also a quadratic function of *CRIME*, but this time the quadratic function has a maximum. The crime rate for which it is a maximum is

$$CRIME_{\max} = \frac{-\hat{\alpha}_4}{2\hat{\alpha}_5} = \frac{2.285}{2 \times 0.039} = 29.3$$

Thus, the variance of house prices increases with the crime rate up to crime rates of around 30 and then declines. There are very few observations for which $CRIME \geq 30$, and so we can say that, generally, the variance increases as the crime rate increases, but at a decreasing rate.

The variance of house prices is negatively related to *DIST*, suggesting that the further the house is from the employment centre, the smaller the variation in house prices.

(b) We can test for heteroskedasticity using the White test. The null and alternative hypotheses are

$$H_0 : \alpha_2 = \alpha_3 = \cdots = \alpha_6 = 0$$

$$H_1 : \text{not all } \alpha_s \text{ in } H_0 \text{ are zero}$$

The test statistic is $\chi^2 = N \times R^2$. We reject H_0 if $\chi^2 > \chi^2_{(0.95,5)}$ where $\chi^2_{(0.95,5)} = 11.07$. The test value is

$$\chi^2 = N \times R^2 = 506 \times 0.08467 = 42.84$$

Since $42.84 > 11.07$, we reject H_0 and conclude that heteroskedasticity exists.

EXERCISE 8.7

(a) Hand calculations yield

$$\sum x_i = 0 \qquad \sum y_i = 31.1 \qquad \sum x_i y_i = 89.35 \qquad \sum x_i^2 = 52.34$$

$$\bar{x} = 0 \qquad \bar{y} = 3.8875$$

The least squares estimates are given by

$$b_2 = \frac{N\sum x_i y_i - \sum x_i \sum y_i}{N\sum x_i^2 - \left(\sum x_i\right)^2} = \frac{8\times 89.35 - 0\times 31.1}{8\times 52.34 - (0)^2} = 1.7071$$

and

$$b_1 = \bar{y} - b_2\bar{x} = 3.8875 - 1.7071\times 0 = 3.8875$$

(b) The least squares residuals $\hat{e}_i = y_i - \hat{y}_i$ and other information useful for part (c) follow

observation	$\hat{e}$	$\ln(\hat{e}^2)$	$z\times\ln(\hat{e}^2)$
1	– 1.933946	1.319125	4.353113
2	0.733822	– 0.618977	– 0.185693
3	9.549756	4.513031	31.591219
4	– 1.714707	1.078484	5.068875
5	– 3.291665	2.382787	4.527295
6	3.887376	2.715469	18.465187
7	– 3.484558	2.496682	5.742369
8	– 3.746079	2.641419	16.905082

(c) To estimate α, we begin by taking logs of both sides of $\sigma_i^2 = \exp(\alpha z_i)$, that yields $\ln(\sigma_i^2) = \alpha z_i$. Then, we replace the unknown σ_i^2 with $\hat{e}_i^2$ to give the estimating equation

$$\ln(\hat{e}_i^2) = \alpha z_i + v_i$$

Using least squares to estimate α from this model is equivalent to a simple linear regression without a constant term. See, for example, Exercise 2.4. The least squares estimate for α is

$$\hat{\alpha} = \frac{\sum_{i=1}^{8}\left(z_i \ln(\hat{e}_i^2)\right)}{\sum_{i=1}^{8} z_i^2} = \frac{86.4674}{178.17} = 0.4853$$

Exercise 8.7 (continued)

(d) Variance estimates are given by the predictions $\hat{\sigma}_i^2 = \exp(\hat{\alpha} z_i) = \exp(0.4853 \times z_i)$. These values and those for the transformed variables

$$y_i^* = \left(\frac{y_i}{\hat{\sigma}_i}\right), \qquad x_i^* = \left(\frac{x_i}{\hat{\sigma}_i}\right)$$

are given in the following table.

observation	$\hat{\sigma}_i^2$	y_i^*	x_i^*
1	4.960560	0.493887	– 0.224494
2	1.156725	– 0.464895	– 2.789371
3	29.879147	3.457624	0.585418
4	9.785981	– 0.287700	– 0.575401
5	2.514531	4.036003	2.144126
6	27.115325	0.345673	– 0.672141
7	3.053260	2.575316	1.373502
8	22.330994	– 0.042323	– 0.042323

(e) From Exercise 8.2, the generalized least squares estimate for β_2 is

$$\hat{\beta}_2 = \frac{\dfrac{\sum y_i^* x_i^*}{\sum \sigma_i^{-2}} - \left(\dfrac{\sum \sigma_i^{-2} y_i}{\sum \sigma_i^{-2}}\right)\left(\dfrac{\sum \sigma_i^{-2} x_i}{\sum \sigma_i^{-2}}\right)}{\dfrac{\sum x_i^{*2}}{\sum \sigma_i^{-2}} - \left(\dfrac{\sum \sigma_i^{-2} x_i}{\sum \sigma_i^{-2}}\right)^2}$$

$$= \frac{\dfrac{15.33594}{2.008623} - 2.193812 \times (-0.383851)}{\dfrac{15.442137}{2.008623} - (-0.383851)^2}$$

$$= \frac{8.477148}{7.540580}$$

$$= 1.1242$$

The generalized least squares estimate for β_1 is

$$\hat{\beta}_1 = \frac{\sum \sigma_i^{-2} y_i}{\sum \sigma_i^{-2}} - \left(\frac{\sum \sigma_i^{-2} x_i}{\sum \sigma_i^{-2}}\right)\hat{\beta}_2 = 2.193812 - (-0.383851) \times 1.1242 = 2.6253$$

EXERCISE 8.8

(a) The regression results with standard errors in parenthesis are

$$\widehat{PRICE} = 5193.15 + 68.3907 SQFT - 217.8433 AGE$$
$$\text{(se)} \quad (3586.64) \quad (2.1687) \qquad (35.0976)$$

These results tell us that an increase in the house size by one square foot leads to an increase in house price of \$63.39. Also, relative to new houses of the same size, each year of age of a house reduces its price by \$217.84.

(b) For $SQFT$ = 1600 and AGE = 15

$$PRICE = 5193.15 + 68.3907 \times 1600 - 217.8433 \times 15 = 111{,}351$$

The estimated price for a 1600 square foot house, which is 15 years old, is \$11,351. For $SQFT$ = 2000 and AGE = 15

$$PRICE = 5193.15 + 68.3907 \times 2000 - 217.8433 \times 15 = 138{,}707$$

The estimated price for a 2000 square foot house, which is 15 years old, is \$138,707.

(c) For the White test we estimate the equation

$$\hat{e}_i^2 = \alpha_1 + \alpha_2 SQFT + \alpha_3 AGE + \alpha_4 SQFT^2 + \alpha_5 AGE^2 + \alpha_6 SQFT \times AGE + v_i$$

and test the null hypothesis $H_0: \alpha_2 = \alpha_3 = \cdots = \alpha_6 = 0$. The value of the test statistic is

$$\chi^2 = N \times R^2 = 940 \times 0.0375 = 35.25$$

Since $\chi^2_{(0.95,5)} = 11.07$, the calculated value is larger than the critical value. That is, $\chi^2 > \chi^2_{(0.95,5)}$. Thus, we reject the null hypothesis and conclude that heteroskedasticity exists.

(d) Estimating the regression $\log(\hat{e}_i^2) = \alpha_1 + \alpha_2 SQFT + v_i$ gives the results

$$\hat{\alpha}_1 = 16.3786, \quad \hat{\alpha}_2 = 0.001414$$

With these results we can estimate σ_i^2 as

$$\hat{\sigma}_i^2 = \exp(16.3786 + 0.001414 SQFT)$$

Exercise 8.8 (continued)

(e) Generalized least squares requires us to estimate the equation

$$\left(\frac{PRICE_i}{\sigma_i}\right)=\beta_1\left(\frac{1}{\sigma_i}\right)+\beta_2\left(\frac{SQFT_i}{\sigma_i}\right)+\beta_2\left(\frac{AGE_i}{\sigma_i}\right)+\left(\frac{e_i}{\sigma_i}\right)$$

When estimating this model, we replace the unknown σ_i with the estimated standard deviations $\hat{\sigma}_i$. The regression results, with standard errors in parenthesis, are

$$\begin{array}{llll} \widehat{PRICE} = 8491.14 + & 65.3269 SQFT & - & 187.6587 AGE \\ \text{(se)} \quad (3109.43) & (2.0825) & & (29.2844) \end{array}$$

These results tell us that an increase in the house size by one square foot leads to an increase in house price of \$65.33. Also, relative to new houses of the same size, each year of age of a house reduces its price by \$187.66.

(f) For $SQFT$ = 1600 and AGE = 15

$$PRICE = 8491.14 + 65.3269 \times 1600 - 187.6587 \times 15 = 110{,}199$$

The estimated price for a 1600 square foot house, which is 15 years old, is \$110,199. For $SQFT$ = 2000 and AGE = 15

$$PRICE = 8491.14 + 65.3269 \times 2000 - 187.6587 \times 15 = 136{,}330$$

The estimated price for a 2000 square foot house, which is 15 years old, is \$136,330.

EXERCISE 8.9

(a) (i) Under the assumptions of Exercise 8.8 part (a), the mean and variance of house prices for houses of size $SQFT = 1600$ and $AGE = 15$ are

$$E(PRICE) = \beta_1 + 1600\beta_2 + 15\beta_3 \qquad \text{var}(PRICE) = \sigma^2$$

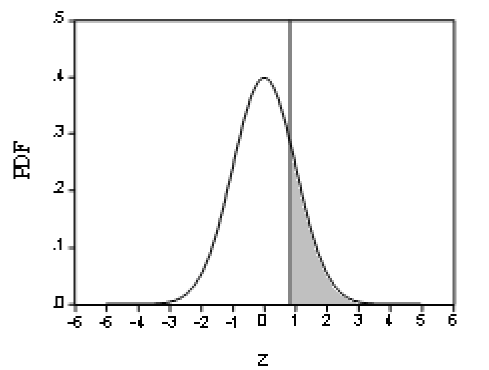
Replacing the parameters with their estimates gives

$$E(PRICE) = 111351 \qquad \text{var}(PRICE) = 22539.63^2$$

Assuming the errors are normally distributed,

$$\begin{aligned} P(PRICE > 115000) &= P\left(Z > \frac{115000 - 111351}{22539.63}\right) \\ &= P(Z > 0.1619) \\ &= 0.436 \end{aligned}$$

where Z is the standard normal random variable $Z \sim N(0,1)$. The probability is depicted as an area under the standard normal density in the following diagram.



The probability that your 1600 square feet house sells for more than \$115,000 is 0.436.

(ii) For houses of size $SQFT = 2000$ and $AGE = 15$, the mean and variance of house prices from Exercise 8.8(a) are

$$E(PRICE) = 138707 \qquad \text{var}(PRICE) = 22539.63^2$$

The required probability is

$$\begin{aligned} P(PRICE < 110000) &= P\left(Z < \frac{110000 - 138707}{22539.63}\right) \\ &= P(Z < -1.2736) \\ &= 0.101 \end{aligned}$$

The probability that your 2000 square feet house sells for less than \$110,000 is 0.101.

Exercise 8.9 (continued)

(b) (i) Using the generalized least squares estimates as the values for β_1, β_2 and β_3, the mean of house prices for houses of size $SQFT = 1600$ and $AGE = 15$ is, from Exercise 8.8(f), $E(PRICE) = 110199$. Using estimates of α_1 and α_2 from Exercise 8.8(d), the variance of these house types is

$$\begin{aligned} \text{var}(PRICE) &= \exp(\alpha_1 + 1.2704 + \alpha_2 \times 1600) \\ &= \exp(16.378549 + 1.2704 + 0.00141417691 \times 1600) \\ &= 4.44131859 \times 10^8 \\ &= 21074.4^2 \end{aligned}$$

Thus,

$$\begin{aligned} P(PRICE > 115000) &= P\left(Z > \frac{115000 - 110199}{21074.4}\right) \\ &= P(Z > 0.2278) \\ &= 0.410 \end{aligned}$$

The probability that your 1600 square feet house sells for more than \$115,000 is 0.410.

(ii) For your larger house where $SQFT = 2000$, we find that $E(PRICE) = 136330$ and

$$\begin{aligned} \text{var}(PRICE) &= \exp(\alpha_1 + 1.2704 + \alpha_2 \times 2000) \\ &= \exp(16.378549 + 1.2704 + 0.00141417691 \times 2000) \\ &= 7.81951143 \times 10^8 \\ &= 27963.4^2 \end{aligned}$$

Thus,

$$\begin{aligned} P(PRICE < 110000) &= P\left(Z < \frac{110000 - 136330}{27963.4}\right) \\ &= P(Z < -0.9416) \\ &= 0.173 \end{aligned}$$

The probability that your 2000 square feet house sells for less than \$110,000 is 0.173.

(c) In part (a) where the heteroskedastic nature of the error term was not recognized, the same standard deviation of prices was used to compute the probabilities for both house types. In part (b) recognition of the heteroskedasticity has led to a standard deviation of prices that is smaller than that in part (a) for the case of the smaller house, and larger than that in part (a) for the case of the larger house. These differences have in turn led to a smaller probability for part (i) where the distribution is less spread out and a larger probability for part (ii) where the distribution has more spread.

EXERCISE 8.10

(a) The transformed model corresponding to the variance assumption $\sigma_i^2 = \sigma^2 x_i$ is

$$\frac{y_i}{\sqrt{x_i}} = \beta_1\left(\frac{1}{\sqrt{x_i}}\right) + \beta_2\sqrt{x_i} + e_i^* \qquad \text{where } e_i^* = \left(\frac{e_i}{\sqrt{x_i}}\right)$$

We obtain the residuals from this model, square them, and regress the squares on x_i to obtain

$$\hat{e}^{*2} = -123.79 + 23.35x \qquad R^2 = 0.13977$$

To test for heteroskedasticity, we compute a value of the χ^2 test statistic as

$$\chi^2 = N \times R^2 = 40 \times 0.13977 = 5.59$$

A null hypothesis of no heteroskedasticity is rejected because 5.59 is greater than the 5% critical value $\chi^2_{(0.95,1)} = 3.84$. Thus, the variance assumption $\sigma_i^2 = \sigma^2 x_i$ was not adequate to eliminate heteroskedasticity.

(b) The transformed model used to obtain the estimates in (8.27) is

$$\frac{y_i}{\hat{\sigma}_i} = \beta_1\left(\frac{1}{\hat{\sigma}_i}\right) + \beta_2\frac{x_i}{\hat{\sigma}_i} + e_i^* \qquad \text{where } e_i^* = \left(\frac{e_i}{\hat{\sigma}_i}\right)$$

and

$$\hat{\sigma}_i = \sqrt{\exp(0.93779596 + 2.32923872 \times \ln(x_i)}$$

We obtain the residuals from this model, square them, and regress the squares on x_i to obtain

$$\hat{e}^{*2} = 1.117 + 0.05896x \qquad R^2 = 0.02724$$

To test for heteroskedasticity, we compute a value of the χ^2 test statistic as

$$\chi^2 = N \times R^2 = 40 \times 0.02724 = 1.09$$

A null hypothesis of no heteroskedasticity is not rejected because 1.09 is less than the 5% critical value $\chi^2_{(0.95,1)} = 3.84$. Thus, the variance assumption $\sigma_i^2 = \sigma^2 x_i^{\gamma}$ is adequate to eliminate heteroskedasticity.

EXERCISE 8.11

The results are summarized in the following table and discussed below.

	part (a)	part (b)	part (c)
$\hat{\beta}_1$	81.000	76.270	81.009
$\text{se}(\hat{\beta}_1)$	32.822	12.004	33.806
$\hat{\beta}_2$	10.328	10.612	10.323
$\text{se}(\hat{\beta}_2)$	1.706	1.024	1.733
$\chi^2 = N \times R^2$	6.641	2.665	6.955

The transformed models used to obtain the generalized estimates are as follows.

(a) $$\left(\frac{y_i}{x_i^{0.25}}\right) = \beta_1\left(\frac{1}{x_i^{0.25}}\right) + \beta_2\left(\frac{x_i}{x_i^{0.25}}\right) + e_i^* \qquad \text{where } e_i^* = \frac{e_i}{x_i^{0.25}}$$

(b) $$\left(\frac{y_i}{x_i}\right) = \beta_1\left(\frac{1}{x_i}\right) + \beta_2\left(\frac{x_i}{x_i}\right) + e_i^* \qquad \text{where } e_i^* = \frac{e_i}{x_i}$$

(c) $$\left(\frac{y_i}{\sqrt{\ln(x_i)}}\right) = \beta_1\left(\frac{1}{\sqrt{\ln(x_i)}}\right) + \beta_2\left(\frac{x_i}{\sqrt{\ln(x_i)}}\right) + e_i^* \qquad \text{where } e_i^* = \frac{e_i}{\sqrt{\ln(x_i)}}$$

In each case the residuals from the transformed model were squared and regressed on income and income squared to obtain the R^2 values used to compute the χ^2 values. These equations were of the form

$$\hat{e}^{*2} = \alpha_1 + \alpha_2 x + \alpha_3 x^2 + v$$

For the White test we are testing the hypothesis $H_0: \alpha_2 = \alpha_3 = 0$ against the alternative hypothesis $H_1: \alpha_2 \neq 0$ and/or $\alpha_3 \neq 0.$ The critical chi-squared value for the White test at a 5% level of significance is $\chi^2_{(0.95,2)} = 5.991$. After comparing the critical value with our test statistic values, we reject the null hypothesis for parts (a) and (c) because, in these cases, $\chi^2 > \chi^2_{(0.95,2)}$. The assumptions $\text{var}(e_i) = \sigma^2\sqrt{x_i}$ and $\text{var}(e_i) = \sigma^2 \ln(x_i)$ do not eliminate heteroskedasticity in the food expenditure model. On the other hand, we do not reject the null hypothesis in part (b) because $\chi^2 < \chi^2_{(0.95,2)}$. Heteroskedasticity has been eliminated with the assumption that $\text{var}(e_i) = \sigma^2 x_i^2$.

In the two cases where heteroskedasticity has not been eliminated (parts (a) and (c)), the coefficient estimates and their standard errors are almost identical. The two transformations have similar effects. The results are substantially different for part (b), however, particularly the standard errors. Thus, the results can be sensitive to the assumption made about the heteroskedasticity, and, importantly, whether that assumption is adequate to eliminate heteroskedasticity.

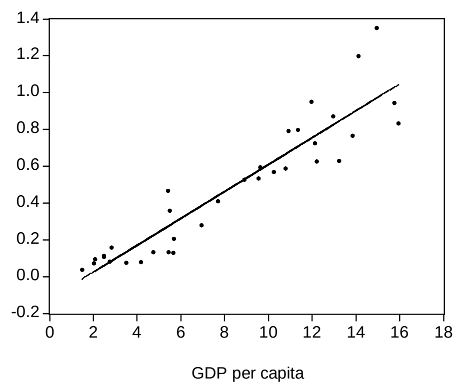
EXERCISE 8.12

(a) This suspicion might be reasonable because richer countries, countries with a higher GDP per capita, have more money to distribute, and thus they have greater flexibility in terms of how much they can spend on education. In comparison, a country with a smaller GDP will have fewer budget options, and therefore the amount they spend on education is likely to vary less.

(b) The regression results, with the standard errors in parentheses are

$$\widehat{\left(\frac{EE_i}{P_i}\right)} = -0.1246 + 0.0732\left(\frac{GDP_i}{P_i}\right)$$
$$\text{(se)} \quad (0.0485) \quad (0.0052)$$

The fitted regression line and data points appear in the following figure. There is evidence of heteroskedasticity. The plotted values are more dispersed about the fitted regression line for larger values of GDP per capita. This suggests that heteroskedasticity exists and that the variance of the error terms is increasing with GDP per capita.



(c) For the White test we estimate the equation

$$\hat{e}_i^2 = \alpha_1 + \alpha_2\left(\frac{GDP_i}{P_i}\right) + \alpha_3\left(\frac{GDP_i}{P_i}\right)^2 + v_i$$

This regression returns an R^2 value of 0.29298. For the White test we are testing the hypothesis $H_0: \alpha_2 = \alpha_3 = 0$ against the alternative hypothesis $H_1: \alpha_2 \neq 0$ and/or $\alpha_3 \neq 0$. The White test statistic is

$$\chi^2 = N \times R^2 = 34 \times 0.29298 = 9.961$$

The critical chi-squared value for the White test at a 5% level of significance is $\chi^2_{(0.95,2)} = 5.991$. Since 9.961 is greater than 5.991, we reject the null hypothesis and conclude that heteroskedasticity exists.

Exercise 8.12 (continued)

(d) Using White's formula:

$$\text{se}(b_1) = 0.040414, \qquad \text{se}(b_2) = 0.006212$$

The 90% confidence interval for β_2 using the conventional least squares standard errors is

$$b_2 \pm t_{(0.95,32)}\text{se}(b_2) = 0.073173 \pm 1.6939 \times 0.005517947 = (0.0644, 0.0819)$$

The 90% confidence interval for β_2 using White's standard errors is

$$b_2 \pm t_{(0.95,32)}\text{se}(b_2) = 0.073173 \pm 1.6939 \times 0.00621162 = (0.0627, 0.0837)$$

In this case, ignoring heteroskedasticity tends to overstate the precision of least squares estimation. The confidence interval from White's standard errors is wider.

(e) Re-estimating the equation under the assumption that $\text{var}(e_i) = \sigma^2 x_i$, we obtain

$$\widehat{\left(\frac{EE_i}{P_i}\right)} = -0.0929 + 0.0693\left(\frac{GDP_i}{P_i}\right)$$
$$\text{(se)} \qquad (0.0289) \quad (0.0044)$$

Using these estimates, the 90% confidence interval for β_2 is

$$b_2 \pm t_{(0.95,32)}\text{se}(b_2) = 0.069321 \pm 1.6939 \times 0.00441171 = (0.0618, 0.0768)$$

The width of this confidence interval is less than both confidence intervals calculated in part (d). Given the assumption $\text{var}(e_i) = \sigma^2 x_i$ is true, we expect the generalized least squares confidence interval to be narrower than that obtained from White's standard errors, reflecting that generalized least squares is more precise than least squares when heteroskedasticity is present. A direct comparison of the generalized least squares interval with that obtained using the conventional least squares standard errors is not meaningful, however, because the least squares standard errors are biased in the presence of heteroskedasticity.

EXERCISE 8.13

(a) For the model $C_{1t} = \beta_1 + \beta_2 Q_{1t} + \beta_3 Q_{1t}^2 + \beta_4 Q_{1t}^3 + e_{1t}$, where $\text{var}(e_{1t}) = \sigma^2 Q_{1t}$, the generalized least squares estimates of β_1, β_2, β_3 and β_4 are:

	estimated coefficient	standard error
β_1	93.595	23.422
β_2	68.592	17.484
β_3	–10.744	3.774
β_4	1.0086	0.2425

(b) The calculated F value for testing the hypothesis that $\beta_1 = \beta_4 = 0$ is 108.4. The 5% critical value from the $F_{(2,24)}$ distribution is 3.40. Since the calculated F is greater than the critical F, we reject the null hypothesis that $\beta_1 = \beta_4 = 0$. The F value can be calculated from

$$F = \frac{(SSE_R - SSE_U)/2}{(SSE_U)/24} = \frac{(61317.65 - 6111.134)/2}{(6111.134)/24} = 108.4$$

(c) The average cost function is given by

$$\frac{C_{1t}}{Q_{1t}} = \beta_1\left(\frac{1}{Q_{1t}}\right) + \beta_2 + \beta_3 Q_{1t} + \beta_4 Q_{1t}^2 + \frac{e_t}{Q_{1t}}$$

Thus, if $\beta_1 = \beta_4 = 0$, average cost is a linear function of output.

(d) The average cost function is an appropriate transformed model for estimation when heteroskedasticity is of the form $\text{var}(e_{1t}) = \sigma^2 Q_{1t}^2$.

EXERCISE 8.14

(a) The least squares estimated equations are

$$\hat{C}_1 = 72.774 + 83.659Q_1 - 13.796Q_1^2 + 1.1911Q_1^3 \qquad \hat{\sigma}_1^2 = 324.85$$
$$\text{(se)} \qquad (23.655) \quad (4.597) \quad (0.2721) \qquad SSE_1 = 7796.49$$

$$\hat{C}_2 = 51.185 + 108.29Q_2 - 20.015Q_2^2 + 1.6131Q_2^3 \qquad \hat{\sigma}_2^2 = 847.66$$
$$\text{(se)} \qquad (28.933) \quad (6.156) \quad (0.3802) \qquad SSE_2 = 20343.83$$

To see whether the estimated coefficients have the expected signs consider the marginal cost function

$$MC = \frac{dC}{dQ} = \beta_2 + 2\beta_3 Q + 3\beta_4 Q^2$$

We expect $MC > 0$ when $Q = 0$; thus, we expect $\beta_2 > 0$. Also, we expect the quadratic MC function to have a minimum, for which we require $\beta_4 > 0$. The slope of the MC function is $d(MC)/dQ = 2\beta_3 + 6\beta_4 Q$. For this slope to be negative for small Q (decreasing MC), and positive for large Q (increasing MC), we require $\beta_3 < 0$. Both our least-squares estimated equations have these expected signs. Furthermore, the standard errors of all the coefficients except the constants are quite small indicating reliable estimates. Comparing the two estimated equations, we see that the estimated coefficients and their standard errors are of similar magnitudes, but the estimated error variances are quite different.

(b) Testing $H_0: \sigma_1^2 = \sigma_2^2$ against $H_1: \sigma_1^2 \neq \sigma_2^2$ is a two-tail test. The critical values for performing a two-tail test at the 10% significance level are $F_{(0.05,24,24)} = 0.0504$ and $F_{(0.95,24,24)} = 1.984$. The value of the F statistic is

$$F = \frac{\hat{\sigma}_2^2}{\hat{\sigma}_1^2} = \frac{847.66}{324.85} = 2.61$$

Since $F > F_{(0.95,24,24)}$, we reject H_0 and conclude that the data do not support the proposition that $\sigma_1^2 = \sigma_2^2$.

(c) Since the test outcome in (b) suggests $\sigma_1^2 \neq \sigma_2^2$, but we are assuming both firms have the same coefficients, we apply generalized least squares to the combined set of data, with the observations transformed using $\hat{\sigma}_1$ and $\hat{\sigma}_2$. The estimated equation is

$$\hat{C} = 67.270 + 89.920Q - 15.408Q^2 + 1.3026Q^3$$
$$\text{(se)} \qquad (16.973) \quad (3.415) \quad (0.2065)$$

<u>Remark</u>: Some automatic software commands will produce slightly different results if the transformed error variance is restricted to be unity or if the variables are transformed using variance estimates from a pooled regression instead of those from part (a).

Exercise 8.14 (continued)

(d) Although we have established that $\sigma_1^2 \neq \sigma_2^2$, it is instructive to first carry out the test for

$$H_0 : \beta_1 = \delta_1, \quad \beta_2 = \delta_2, \quad \beta_3 = \delta_3, \quad \beta_4 = \delta_4$$

under the assumption that $\sigma_1^2 = \sigma_2^2$, and then under the assumption that $\sigma_1^2 \neq \sigma_2^2$.

Assuming that $\sigma_1^2 = \sigma_2^2$, the test is equivalent to the Chow test discussed on pages 268-270 of the text. The test statistic is

$$F = \frac{\left(SSE_R - SSE_U\right)/J}{SSE_U/\left(N-K\right)}$$

where SSE_U is the sum of squared errors from the full dummy variable model. The dummy variable model does not have to be estimated, however. We can also calculate SSE_U as the sum of the SSE from separate least squares estimation of each equation. In this case

$$SSE_U = SSE_1 + SSE_2 = 7796.49 + 20343.83 = 28140.32$$

The restricted model has not yet been estimated under the assumption that $\sigma_1^2 = \sigma_2^2$. Doing so by combining all 56 observations yields $SSE_R = 28874.34$. The *F*-value is given by

$$F = \frac{\left(SSE_R - SSE_U\right)/J}{SSE_U/\left(N-K\right)} = \frac{(28874.34 - 28140.32)/4}{28140.32/(56-8)} = 0.313$$

The corresponding χ^2-value is $\chi^2 = 4 \times F = 1.252$. These values are both much less than their respective 5% critical values $F_{(0.95,4,48)} = 2.565$ and $\chi^2_{(0.95,4)} = 9.488$. There is no evidence to suggest that the firms have different coefficients. In the formula for *F*, note that the number of observations *N* is the total number from both firms, and *K* is the number of coefficients from both firms.

The above test is not valid in the presence of heteroskedasticity. It could give misleading results. To perform the test under the assumption that $\sigma_1^2 \neq \sigma_2^2$, we follow the same steps, but we use values for *SSE* computed from transformed residuals. For restricted estimation from part (c) the result is $SSE_R^* = 49.2412$. For unrestricted estimation, we have the interesting result

$$SSE_U^* = \frac{SSE_1}{\hat{\sigma}_1^2} + \frac{SSE_2}{\hat{\sigma}_2^2} = \frac{(N_1 - K_1)\times\hat{\sigma}_1^2}{\hat{\sigma}_1^2} + \frac{(N_2 - K_2)\times\hat{\sigma}_2^2}{\hat{\sigma}_2^2} = N_1 - K_1 + N_2 - K_2 = 48$$

Thus,

$$F = \frac{(49.2412 - 48)/4}{48/48} = 0.3103 \qquad \text{and} \qquad \chi^2 = 1.241$$

The same conclusion is reached. There is no evidence to suggest that the firms have different coefficients.

EXERCISE 8.15

(a) To estimate the two variances using the variance model specified, we first estimate the equation

$$WAGE_i = \beta_1 + \beta_2 EDUC_i + \beta_3 EXPER_i + \beta_4 METRO_i + e_i$$

From this equation we use the squared residuals to estimate the equation

$$\ln(\hat{e}_i^2) = \alpha_1 + \alpha_2 METRO_i + v_i$$

The estimated parameters from this regression are $\hat{\alpha}_1 = 1.508448$ and $\hat{\alpha}_2 = 0.338041$. Using these estimates, we have

$$METRO = 0 \quad \Rightarrow \quad \hat{\sigma}_R^2 = \exp(1.508448 + 0.338041 \times 0) = 4.519711$$

$$METRO = 1, \quad \Rightarrow \quad \hat{\sigma}_M^2 = \exp(1.508448 + 0.338041 \times 1) = 6.337529$$

These error variance estimates are much smaller than those obtained from separate sub-samples ($\hat{\sigma}_M^2 = 31.824$ and $\hat{\sigma}_R^2 = 15.243$). One reason is the bias factor from the exponential function – see page 317 of the text. Multiplying $\hat{\sigma}_M^2 = 6.3375$ and $\hat{\sigma}_R^2 = 4.5197$ by the bias factor $\exp(1.2704)$ yields $\hat{\sigma}_M^2 = 22.576$ and $\hat{\sigma}_R^2 = 16.100$. These values are closer, but still different from those obtained using separate sub-samples. The differences occur because the residuals from the combined model are different from those from the separate sub-samples.

(b) To use generalized least squares, we use the estimated variances above to transform the model in the same way as in (8.35). After doing so the regression results are, with standard errors in parentheses

$$\begin{array}{lllll} \widehat{WAGE}_i = & -9.7052 + & 1.2185 EDUC_i + & 0.1328 EDUC_i + & 1.5301 METRO_i \\ \text{(se)} & (1.0485) & (0.0694) & (0.0150) & (0.3858) \end{array}$$

The magnitudes of these estimates and their standard errors are almost identical to those in equation (8.36). Thus, although the variance estimates can be sensitive to the estimation technique, the resulting generalized least squares estimates of the mean function are much less sensitive.

(c) The regression output using White standard errors is

$$\begin{array}{lllll} \widehat{WAGE}_i = & -9.9140 + & 1.2340 EDUC_i + & 0.1332 EDUC_i + & 1.5241 METRO_i \\ \text{(se)} & (1.2124) & (0.0835) & (0.0158) & (0.3445) \end{array}$$

With the exception of that for *METRO*, these standard errors are larger than those in part (b), reflecting the lower precision of least squares estimation.

EXERCISE 8.16

(a) Separate least squares estimation gives the error variance estimates $\hat{\sigma}_G^2 = 2.899215 \times 10^{-4}$ and $\hat{\sigma}_A^2 = 15.36132 \times 10^{-4}$.

(b) The critical values for testing the hypothesis $H_0 : \sigma_G^2 = \sigma_A^2$ against the alternative $H_1 : \sigma_G^2 \neq \sigma_A^2$ at a 1% level of significance are $F_{(0.005,15,15)} = 0.246$ and $F_{(0.995,15,15)} = 4.070$. The value of the F-statistic is

$$F = \frac{\hat{\sigma}_A^2}{\hat{\sigma}_G^2} = \frac{15.36132 \times 10^{-4}}{2.899215 \times 10^{-4}} = 5.298$$

Since 5.298 > 4.070, we reject the null hypothesis and conclude that the error variances of the two countries, Austria and Germany, are not the same.

(c) The estimates of the coefficients using generalized least squares are

	estimated coefficient	standard error
β_1 [const]	2.0268	0.4005
β_2 [ln(*INC*)]	–0.4466	0.1838
β_3 [ln(*PRICE*)]	–0.2954	0.1262
β_4 [ln(*CARS*)]	0.1039	0.1138

(d) Testing the null hypothesis that demand is price inelastic, i.e., $H_0 : \beta_3 \geq -1$ against the alternative $H_1 : \beta_3 < -1$, is a one-tail t test. The value of our test statistic is

$$t = \frac{-0.2954 - (-1)}{0.1262} = 5.58$$

The critical t value for a one-tail test and 34 degrees of freedom is $t_{(0.01,34)} = -2.441$. Since $5.58 > -2.441$, we do not reject the null hypothesis and conclude that there is not enough evidence to suggest that demand is elastic.

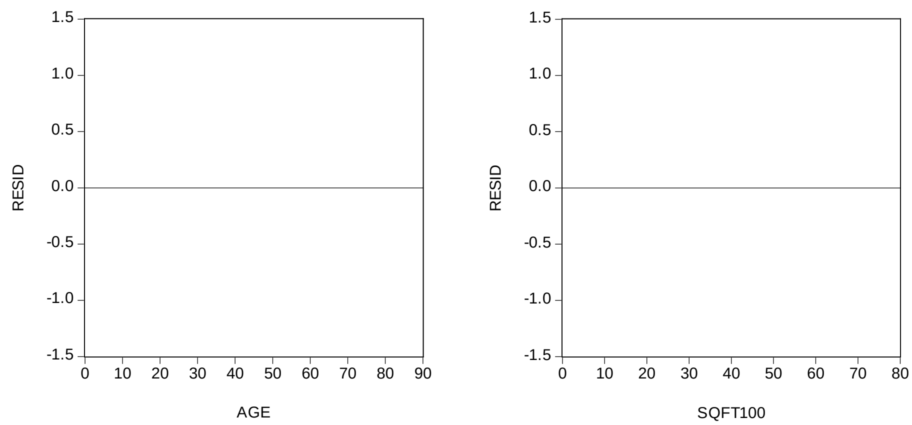
EXERCISE 8.17

(a) The estimated regression is

$$\widehat{\ln(PRICE)} = 11.1196 + 0.03876 SQFT100 - 0.01756 AGE + 0.0001734 AGE^2$$
$$\text{(se)} \qquad (0.274) \quad (0.00087) \qquad\qquad (0.00136) \qquad\quad (0.0000227)$$

(b) The residual plots are given in the figures below. The absolute magnitude of the residuals increases as *AGE* increases, suggesting heteroskedasticity, with the variance dependent on the age of the house. Conversely, the absolute magnitude of the residuals appears to decrease as *SQFT100* increases, although this pattern is less pronounced. The variance might decrease as the house size increases, but we cannot be certain.

Figure xr8.17(b)



Plot of residuals against *AGE* **Plot of residuals against *SQFT100***

(c) We set up the model $\text{var}(e) = h(\alpha_1 + \alpha_2 AGE + \alpha_3 SQFT100)$ and test the hypotheses:

$$H_0: \alpha_2 = 0,\ \alpha_3 = 0 \qquad\qquad H_1: \alpha_2 = 0 \text{ and/or } \alpha_3 = 0$$

The test statistic value is

$$\chi^2 = N \times R^2 = 1080 \times 0.1082 = 116.876$$

The critical chi-squared value at a 1% level of significance is $\chi^2_{(0.99,2)} = 9.210$. Since 116.88 is greater than 9.210, we reject the null hypothesis and conclude that heteroskedasticity exists.

(d) The estimated variance function is given as

$$\hat{\sigma}_i^2 = \exp(-4.7139 + 0.02177 AGE_i + 0.006377 SQFT100_i)$$

The robust standard errors for *AGE* and *SQFT100* are 0.00404 and 0.006945, respectively. Corresponding *p*-values are 0.0000 and 0.3589. We can conclude that *AGE* has a significant effect on variance while *SQFT100* is not significant. This conclusion agrees with our speculation from inspecting the figures in part (b), although in part (b) we did suggest the sign of *SQFT100* might be negative.

Exercise 8.17 (continued)

(e) The estimated generalized least squares model is

$$\widehat{\ln(PRICE)} = 11.105 + 0.03881SQFT100 - 0.01540AGE + 0.0001297AGE^2$$
$$\text{(se)} \quad (0.024) \quad (0.00082) \qquad (0.00136) \qquad (0.0000272)$$

(f)

	b_1	b_2	b_3	b_4
(i) Least Squares	11.120 (0.027)	0.03876 (0.00087)	–0.01756 (0.00136)	0.0001734 (0.0000227)
(ii) with HC standard errors	11.120 (0.033)	0.03876 (0.00123)	–0.01756 (0.00175)	0.0001734 (0.0000372)
(iii) GLS	11.105 (0.024)	0.03881 (0.00082)	–0.01540 (0.00136)	0.0001297 (0.0000272)
(iv) with HC standard errors	11.105 (0.028)	0.03881 (0.00105)	–0.01540 (0.00144)	0.0001297 (0.0000314)

The coefficient estimates from least squares and GLS are similar, with the greatest differences being those for *AGE* and AGE^2. The heteroskedasticity-consistent (HC) standard errors are higher than the conventional standard errors for both least squares and GLS, and for all coefficients. The conventional GLS standard errors are smaller than the least squares HC standard errors, suggesting that GLS has improved the efficiency of estimation. The GLS HC standard errors are slightly larger than the conventional GLS ones; this could be indicative of some remaining heteroskedasticity.

(g) The Breusch-Pagan test statistic obtained by regressing the squares of the transformed residuals on *AGE* and *SQFT100* is

$$\chi^2 = N \times R^2 = 1080 \times 0.018169 = 19.62$$

The 5% critical value is $\chi^2_{(0.95,2)} = 5.99$ and the *p*-value of the test is 0.0001. Thus we reject a null hypothesis of homoskedastic errors. The variance function that we used does not appear to have been adequate to eliminate the heteroskedasticity.

EXERCISE 8.18

(a) $COKE_{ij}$ is a binary variable which assigns 1 if the shopper buys coke and zero otherwise. Therefore, the total number of shoppers who buy coke in store *i* is given by $\sum_{j=1}^{N_i} COKE_{ij}$ and the proportion will be given by $\frac{1}{N_i}\sum_{j=1}^{N_i} COKE_{ij}$, which is $\overline{COKE}_i$.

(b)

$$E\left(\overline{COKE}_i\right) = \frac{1}{N_i} E\left(\sum_{j=1}^{N_i} COKE_{ij}\right)$$

$$= \frac{1}{N_i}\sum_{j=1}^{N_i} E\left(COKE_{ij}\right)$$

$$= \frac{1}{N_i} N_i p_i = p_i$$

$$\text{var}\left(\overline{COKE}_i\right) = \frac{1}{N_i^2}\text{var}\left(\sum_{j=1}^{N_i} COKE_{ij}\right)$$

$$= \frac{1}{N_i^2}\sum_{j=1}^{N_i}\text{var}\left(COKE_{ij}\right) + \text{zero covariance terms}$$

$$= \frac{1}{N_i^2}\sum_{j=1}^{N_i} p_i(1-p_i)$$

$$= \frac{N_i}{N_i^2} p_i(1-p_i) = \frac{p_i(1-p_i)}{N_i}$$

(c) p_i is the population proportion of customers in store *i* who purchase Coke. We can think of it as the proportion evaluated for a large number of customers in store *i*, or the probability that a customer in store *i* will purchase Coke. We can write

$$p_i = \beta_1 + \beta_2 PRATIO_i + \beta_3 DISP\_COKE_i + \beta_4 DISP\_PEPSI_i$$

(d) The estimated regression is:

$$\widehat{\overline{COKE}}_i = 0.5196 - 0.06594 PRATIO_i + 0.08571 DISP\_COKE_i - 0.1097 DISP\_PEPSI_i$$

(se) (0.3207) (0.31199) (0.04671) (0.0469)

The results suggest that *PRATIO* and *DISP_PEPSI* have negative impacts on the probability of purchasing coke, although the coefficient of the price ratio is not significantly different from zero at a 5% significance level; *DISP_COKE* has a positive impact on the probability of purchasing coke. Both *DISP_PEPSI* and *DISP_COKE* have significant coefficients if one-tail tests and a 5% significance level are used.

Exercise 8.18 (continued)

(e) The null and alternative hypotheses are

$$H_0: \text{errors are homoskedastic}$$

$$H_1: \text{errors are heteroskedastic}$$

The test statistic is

$$\chi^2 = N \times R^2 = 50 \times 0.15774 = 7.887$$

The critical chi-squared value for the White test at a 5% level of significance is $\chi^2_{(0.95,7)} = 14.067$. Since 7.887 < 14.067, we do not reject the null hypothesis. There is insufficient evidence to conclude that the errors are heteroskedastic. The *p*-value of the test is 0.343.

The variance of the error term is

$$\begin{aligned}\operatorname{var}\left(\overline{COKE_i}\right) &= \frac{p_i(1-p_i)}{N_i} \\ &= \left(\beta_1 + \beta_2 PRATIO_i + \beta_3 DISP\_COKE_i + \beta_4 DISP\_PEPSI_i\right) \\ &\quad \times \left(\beta_1 + \beta_2 PRATIO_i + \beta_3 DISP\_COKE_i + \beta_4 DISP\_PEPSI_i\right) / N_i\end{aligned}$$

The product in the above equation means that the variance will depend on each of the variables and their cross products. Thus, it makes sense to include the cross-product terms when carrying out the White test. It is surprising that the White test did not pick up any heteroskedasticity. Perhaps the variation in p_i is not sufficient, or the sample size is too small, for the test to be conclusive. Or the omission of N_i could be masking the effect of the variables.

(f) The estimated results are reported in the table below:

	Mean	Standard Deviation	Maximum	Minimum
$\hat{p}$	0.4485	0.04135	0.5459	0.3385

(g) The estimated GLS regression is:

$$\widehat{COKE_i} = 0.5503 - 0.09673 PRATIO + 0.07831 DISP\_COKE - 0.1009 DISP\_PEPSI$$

$$\text{(se)} \quad (0.3099)\ (0.30205) \qquad (0.04568) \qquad (0.0449)$$

The results are very similar to those obtained in part (d), both in terms of coefficient magnitudes and significance. The coefficient of *PRATIO* is a mild exception; it is larger in absolute value than its least squares counterpart, but remains insignificant. Given the relative importance of *PRATIO*, this insignificance is puzzling. It could be attributable to the small variation in *PRATIO.*

EXERCISE 8.19

(a) The estimated least square regression with heteroskedasticity-robust standard errors is

$$\widehat{\ln(WAGE)} = 0.5297 + 0.1272EDUC + 0.06298EXPER - 0.0007139EXPER^2$$
$$\text{(se)} \quad (0.2528) \ (0.0170) \quad (0.01138) \quad (0.0000920)$$
$$-0.001322EXPER \times EDUC$$
$$(0.000637)$$

(b) Adding marriage to the equation yields

$$\widehat{\ln(WAGE)} = 0.5411 + 0.1261EDUC + 0.06137EXPER - 0.0006933EXPER^2$$
$$\text{(se)} \quad (0.2542) \ (0.0171) \quad (0.01159) \quad (0.0000956)$$
$$-0.001309EXPER \times EDUC + 0.0403MARRIED$$
$$(0.000638) \qquad (0.03392)$$

The null and alternative hypotheses for testing whether married workers get higher wages are given by

$$H_0: \beta_6 \le 0 \qquad H_1: \beta_6 > 0$$

The test value is:

$$t = \frac{b_6}{\text{se}(b_6)} = \frac{0.04029}{0.00339} = 1.188$$

The corresponding *p*-value is 0.1176. Also, the critical value at the 1% level of significance is 2.330. Since the test value is less than the critical value (or because the *p*-value is less than 0.01), we do not reject the null hypothesis at the 1% level. We conclude that there is insufficient evidence to show that wages of married workers are greater than those of unmarried workers.

(c) The residual plot

Figure xr8.19(c) Plot of least squares residuals against marriage

The residual plot suggests the variance of wages for married workers is greater than that for unmarried workers. Thus, there is the evidence of heteroskedasticity.

Exercise 8.19 (continued)

(d) The estimated regression when $MARRIED = 1$ is

$$\widehat{\ln(WAGE)} = 0.9197 + 0.1008 EDUC + 0.05069 EXPER - 0.0007088 EXPER^2$$
$$\text{(se)} \quad (0.3558) \quad (0.0222) \qquad (0.01493) \qquad (0.0001379)$$
$$-0.0004620 EXPER \times EDUC$$
$$(0.0007478)$$

The estimated regression when $MARRIED = 0$ is

$$\widehat{\ln(WAGE)} = 0.1975 + 0.1513 EDUC + 0.07284 EXPER - 0.0007014 EXPER^2$$
$$\text{(se)} \quad (0.2945) \quad (0.0194) \qquad (0.01271) \qquad (0.0001193)$$
$$-0.002145 EXPER \times EDUC$$
$$(0.000654)$$

The Goldfeld-Quandt test

The null and alternative hypotheses are:

$$H_0: \sigma_M^2 = \sigma_U^2 \text{ against } H_1: \sigma_M^2 \neq \sigma_U^2$$

The value of the F statistic is

$$F = \frac{\hat{\sigma}_U^2}{\hat{\sigma}_M^2} = \frac{0.21285}{0.28658} = 0.743$$

The critical values are $F_{Lc} = F_{(0.005,414,576)} = 0.789$ and $F_{Uc} = F_{(0.995,414,576)} = 1.263$. Because $0.743 = F < F_{Lc} = 0.789$, we reject H_0 and conclude that the error variances for married and unmarried women are different.

(e) The generalized least squares estimated regression is

$$\widehat{\ln(WAGE)} = 0.4780 + 0.1309 EDUC + 0.06452 EXPER - 0.0007128 EXPER^2$$
$$\text{(se)} \quad (0.2212) \quad (0.0144) \qquad (0.00932) \qquad (0.0000862)$$
$$-0.001443 EXPER \times EDUC$$
$$(0.000484)$$

There are no major changes in the values of the coefficient estimates. However, the standard errors in the GLS-estimated equation are all less than their counterparts in the least squares-estimated equation, reflecting the increased efficiency of least squares estimation.

Exercise 8.19 (continued)

(f) The marginal effect for a worker with 25 years of experience is given by

$$\frac{\partial E\big(\ln(WAGE)\big)}{\partial EDUC} = \beta_2 + \beta_5 EXPER = \beta_2 + 25\beta_5$$

The estimate for the marginal effect calculated using the regression in part (a) is

$$\frac{\partial E\big(\ln(WAGE)\big)}{\partial EDUC} = 0.127195 - 0.0013224 \times 25 = 0.09414$$

Its standard error is $\mathrm{se}(b_2 + 25b_5) = 0.006471$.

The estimate for the marginal effect calculated using the regression in part (e) is

$$\frac{\partial E\big(\ln(WAGE)\big)}{\partial EDUC} = 0.130853 - 0.0014426 \times 25 = 0.09479$$

Its standard error is $\mathrm{se}\big(\hat{\beta}_2 + 25\hat{\beta}_5\big) = 0.006033$.

The *t*-value for computing the interval estimates is $t_c = t_{(0.975,995)} = 1.962$.

Thus, the two interval estimates are as follows.

From the least squares-estimated equation in part (a):

$$\mathrm{me} \pm t_c \mathrm{se}(b_2 + 25b_5) = 0.09414 \pm 1.962 \times 0.006471 = (0.0814{,}0.1068)$$

From the GLS-estimated equation in part (e):

$$\mathrm{me} \pm t_c \mathrm{se}\big(\hat{\beta}_2 + 25\hat{\beta}_5\big) = 0.09479 \pm 1.962 \times 0.006033 = (0.0830{,}0.1066)$$

The interval estimate from the GLS equation is slightly narrower than its least squares counterpart, but overall, there is very little difference.

EXERCISE 8.20

(a) The residual plots against *EDUC* and *EXPER* are as follows

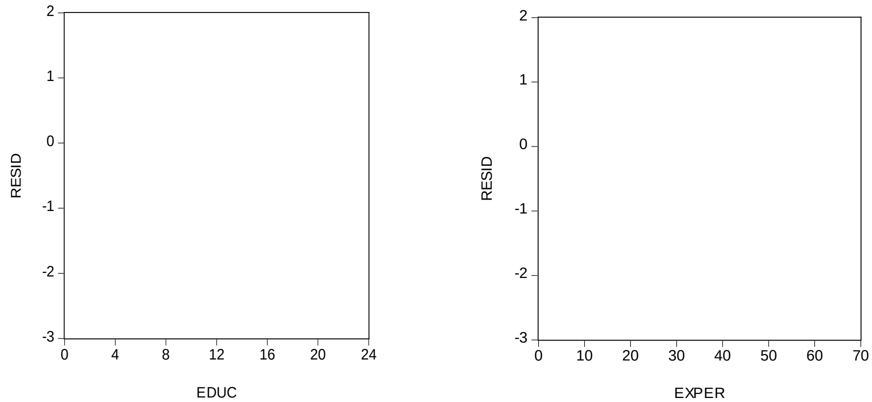


Figure xr8.20 Residual polots against *EDUC* and *EXPER*

Both residual plots exhibit a pattern in which the absolute magnitudes of the residuals tend to increase as the values of *EDUC* and *EXPER* increase, although for *EXPER* the increase is not very pronounced. Thus, the plots suggest there is heteroskedasticity with the variance dependent on *EDUC* and possibly *EXPER*.

(b) The null and alternative hypotheses are

$$H_0: \text{errors are homoskedastic}$$

$$H_1: \text{errors are heteroskedastic}$$

with H_1 implying the error variance depends on one or more of *EXPER*, *EDUC* or *MARRIED*. The value of the test statistic is

$$\chi^2 = N \times R^2 = 1000 \times 0.01465 = 14.65$$

The critical chi-squared value at a 5% level of significance is $\chi^2_{(0.95,3)} = 7.815$. Since 14.65 is greater than 7.815, we reject the null hypothesis and conclude that heteroskedasticity exists. The *p*-value of the test is 0.0021.

(c) The estimated variance function is

$$\hat{\sigma}_i^2 = \exp(-3.0255 + 0.01391 EDUC_i + 0.00516 EXPER_i + 0.04547 MARRIED_i)$$

The standard deviations for each observation are calculated by getting the square roots of the forecast values from the above equation. The first ten estimates are presented in the following table.

Exercise 8.20(c) (continued)

Observation	Standard deviation
1.	0.27856
2.	0.24957
3.	0.26049
4.	0.24982
5.	0.27944
6.	0.26470
7.	0.27217
8.	0.26745
9.	0.27287
10.	0.26123

(d) The generalized least squares estimated regression is

$$\widehat{\ln(WAGE)} = \underset{(0.2203)}{0.5265} + \underset{(0.0144)}{0.1274}EDUC + \underset{(0.00944)}{0.06365}EXPER - \underset{(0.0000887)}{0.0007151}EXPER^2 \quad \text{(se)}$$

$$- \underset{(0.000492)}{0.001369}EXPER \times EDUC$$

The least squares estimated equation with heteroskedasticity-robust standard errors is

$$\widehat{\ln(WAGE)} = \underset{(0.2528)}{0.5297} + \underset{(0.0170)}{0.1272}EDUC + \underset{(0.01138)}{0.06298}EXPER - \underset{(0.0000920)}{0.0007139}EXPER^2 \quad \text{(se)}$$

$$- \underset{(0.000637)}{0.001322}EXPER \times EDUC$$

The coefficient estimates in both equations are very similar. However, the standard errors in the GLS-estimated equation are all less than their counterparts in the least squares-estimated equation, reflecting the increased efficiency of least squares estimation.

Exercise 8.20 (continued)

(e) The marginal effect for a worker with 16 years of education and 20 years of experience is given by

$$\frac{\partial E\left(\ln(WAGE)\right)}{\partial EXPER} = \beta_3 + 2\beta_4 EXPER + \beta_5 EDUC = \beta_3 + 40\beta_4 + 16\beta_5$$

The least squares estimate for the marginal effect is

$$\frac{\widehat{\partial E\left(\ln(WAGE)\right)}}{\partial EDUC} = 0.062981 - 40 \times 0.0007139386 - 16 \times 0.001322388$$

$$= 0.013265$$

Its standard error is $\text{se}\left(b_3 + 40b_4 + 16b_5\right) = 0.002020$.

The generalized least squares estimate for the marginal effect is

$$\frac{\widehat{\partial E\left(\ln(WAGE)\right)}}{\partial EDUC} = 0.063646 - 40 \times 0.0007151398 - 16 \times 0.00136903$$

$$= 0.013136$$

Its standard error is $\text{se}\left(b_3 + 40b_4 + 16b_5\right) = 0.001898$.

The t-value for computing the interval estimates is $t_c = t_{(0.975,995)} = 1.962$.

Thus, the two interval estimates are as follows.

From the least squares-estimated equation:

$$\widehat{me} \pm t_c \text{se}\left(b_3 + 40b_4 + 16b_5\right) = 0.013265 \pm 1.962 \times 0.002020 = (0.00930, 0.01723)$$

From the GLS-estimated equation in part (d):

$$\widehat{me} \pm t_c \text{se}\left(b_3 + 40b_4 + 16b_5\right) = 0.013136 \pm 1.962 \times 0.001898 = (0.00941, 0.01686)$$

The interval estimate from the GLS equation is slightly narrower than its least squares counterpart, but overall, there is very little difference.

EXERCISE 8.21

(a) Using the natural predictor, the forecast wage for a married worker with 18 years of education and 16 years of experience is

$$\begin{aligned}\widehat{WAGE}_n &= \exp(0.526482 + 0.127412\times 18 + 0.0636458\times 16 \\ &\qquad - 0.00071513983\times 16^2 - 0.00136903402\times 16\times 18) \\ &= 26.072\end{aligned}$$

To compute the forecast using the corrected predictor, we first need to estimate the variance for a married worker with 18 years of education and 16 years of experience. This estimate is given by

$$\begin{aligned}\hat{\sigma}^2 &= \exp(-3.025504 + 0.01391\times 18 + 0.0051605\times 16 + 0.0454734) \\ &= 0.0708577\end{aligned}$$

Then the forecast from the corrected predictor is

$$\begin{aligned}\widehat{WAGE}_c &= \widehat{WAGE}_n \exp\left(\hat{\sigma}^2/2\right) \\ &= 26.072\times \exp\left(0.0708577/2\right) \\ &= 27.012\end{aligned}$$

(b) The 95% forecast interval is given by

$$\begin{aligned}\exp\left(\ln\left(\widehat{WAGE}_n\right) \pm t_c \mathrm{se}(f)\right) &= \exp\left(3.260868 \pm 1.962\times \sqrt{0.0708577}\right) \\ &= (15.464, 43.958)\end{aligned}$$

EXERCISE 8.22

(a) The estimated linear probability model is

$$\widehat{DELINQUENT} = \underset{(0.2112)}{0.6885} + \underset{(0.000785)}{0.001624}LVR - \underset{(0.02383)}{0.05932}REF - \underset{(0.0236)}{0.4816}INSUR + \underset{(0.00860)}{0.03438}RATE$$

(se)

$$+\underset{(0.01267)}{0.02377}AMOUNT - \underset{(0.0002018)}{0.0004419}CREDIT - \underset{(0.00354)}{0.01262}TERM + \underset{(0.0319)}{0.1283}ARM$$

The White test

The null and alternative hypotheses are

H_0 : errors are homoskedastic

H_1 : errors are heteroskedastic

Under H_1 we are assuming that the error variance depends on one or more of the explanatory variables, their squares and their cross products. The cross product terms are included because in the linear probability model

$$\text{var}(DELINQUENT) = E(DELINQUENT) \times \left(1 - E(DELINQUENT)\right)$$

where $E(DELINQUENT)$ is a linear function of all the explanatory variables, as expressed in the estimated equation.

The value of the test statistic is

$$\chi^2 = N \times R^2 = 1000 \times 0.21997 = 219.974$$

The critical chi-squared value for the White test at a 5% level of significance is $\chi^2_{(0.95,40)} = 55.758$. Since 219.974 is greater than 55.758, we reject the null hypothesis and conclude that heteroskedasticity exists.

(b) The error variances are estimated using

$$\widehat{\text{var}(D}ELINQUENT) = \widehat{DELINQUENT} \times \left(1 - \widehat{DELINQUENT}\right)$$

The number of observations where $\widehat{\text{var}(D}ELINQUENT) \geq 1$ is zero.

The number of observations where $\widehat{\text{var}(D}ELINQUENT) \leq 0$ is 135.

The number of observations where $\widehat{\text{var}(D}ELINQUENT) < 0.01$ is 158.

Exercise 8.22 (continued)

(c)

	LVR	*REF*	*INSUR*	*RATE*	*AMOUNT*	*CREDIT*	*TERM*	*ARM*
(i)	0.00162	–0.0593	–0.4816	0.0344	0.0238	–0.000442	–0.0126	0.1283
LS	(0.00078)	(0.0238)	(0.0236)	(0.0086)	(0.0127)	(0.000202)	(0.0035)	(0.0319)
(ii)	0.00162	–0.0593	–0.4816	0.0344	0.0238	–0.000442	–0.0126	0.1283
LS-HC	(0.00068)	(0.0240)	(0.0304)	(0.0098)	(0.0145)	(0.000207)	(0.0036)	(0.0277)
(iii)	0.00159	–0.0571	–0.5016	0.0413	0.0258	–0.000382	–0.0190	0.2089
<0.01	(0.00081)	(0.0211)	(0.0292)	(0.0082)	(0.0121)	(0.000184)	(0.0041)	(0.0407)
(iv)	0.00086	–0.0327	–0.4770	0.0204	0.0187	–0.000162	–0.0065	0.0419
=0.01	(0.00038)	(0.0146)	(0.0297)	(0.0057)	(0.0099)	(0.000118)	(0.0021)	(0.0140)
(v)	0.00054	–0.0267	–0.5127	0.0002	-0.0045	–0.000024	–0.0018	0.0188
=0.00001	(0.00024)	(0.0105)	(0.4086)	(0.0048)	(0.0089)	(0.000085)	(0.0018)	(0.0109)

For most of the coefficients the least squares and generalized least squares estimates are similar, providing the GLS estimates are obtained by discarding observations with variances less than 0.01. Moreover, the standard errors from the first three sets of estimates are sufficiently similar for the same conclusions to be reached about the significance of estimated coefficients; an exception is *AMOUNT* whose coefficient is not significantly different from zero in the least squares estimations.

The magnitudes of the coefficients change considerably when variances less than 0.01, or less than 0.00001, are set equal to one of these threshold values; and the estimates are very sensitive to the threshold which is chosen. In the extreme case where variances less than 0.00001 are set equal to 0.00001, only two of the estimated coefficients are significantly different from zero. In the other cases almost all of the 8 coefficients were significant. Setting small and negative variances equal to a small number seems to be a practice fraught with danger. It places very heavy weights on a relatively few number of observations.

(d) *LVR*: The estimated coefficient is 0.00086. This suggests that, holding other variables constant, a one unit increase in the ratio of the loan amount to the value of property increases the probability of delinquency by 0.00086. The positive sign is reasonable as a higher ratio of the amount of loan to the value of the property will lead to a higher probability of delinquency. The coefficient of *LVR* is significantly different from zero at the 5% level.

REF: The estimated coefficient is –0.0327. This suggests that, holding other variables constant, if the loan was for refinancing, the probability of delinquency decreases by 0.0327. The negative sign is reasonable as refinancing the loan is usually done to make repayments easier to manage, which has a negative impact upon the loan delinquency. The coefficient of *REF* is significantly different from zero at the 5% level.

Exercise 8.22(d) (continued)

INSUR: The estimated coefficient is –0.4770. This suggests that, holding other variables constant, if a mortgage carries mortgage insurance, the probability of delinquency decreases by 0.4770. The negative sign is reasonable; taking insurance is an indication that a borrower is more reliable, reducing the probability of delinquency. The coefficient of *INSUR* is significantly different from zero at the 5% level.

RATE: The estimated coefficient is 0.0204. This suggests that, holding other variables constant, a one unit increase in the initial interest rate of the mortgage increases the probability of delinquency by 0.0204. The positive sign is reasonable as a higher interest rate will result in a higher probability of delinquency. The coefficient of *RATE* is significantly different from zero at the 5% level.

AMOUNT: The estimated coefficient is 0.0187. This suggests that, holding other variables constant, a one unit increase in the amount of the mortgage increases the probability of delinquency by 0.0187. The positive sign is reasonable because, as the amount of the mortgage gets larger, the borrower is more likely to face delinquency. The coefficient of *AMOUNT* is not significantly different from zero at the 5% level.

CREDIT: The estimated coefficient is –0.000162. This suggests that, holding other variables constant, a one unit increase in the credit score decreases the probability of delinquency by 0.000162. The negative sign is reasonable as a borrower with a higher credit rating will have a lower probability of delinquency. The coefficient of *CREDIT* is not significantly different from zero at the 5% level.

TERM: The estimated coefficient is –0.0065. This suggests that, holding other variables constant, a one year-increase in the term between disbursement of the loan, and the date it is expected to be fully repaid, decreases the probability of delinquency by 0.0065. The negative sign is reasonable because, given *AMOUNT* is constant, the longer the term of the loan, the less likely it is that the borrower will face delinquency. The coefficient of *TERM* is significantly different from zero at the 5% level.

ARM: The estimated coefficient is 0.0419. This suggests that, holding other variables constant, if the mortgage interest rate is adjustable, the probability of delinquency increases by 0.0419. The positive sign is reasonable because, with the adjustable rate, the interest rate may rise above what the borrower is able to repay, which leads to a higher probability of delinquency. The coefficient of *ARM* is significantly different from zero at the 5% level.