Chapter 7
Heteroskedasticity

7.1 Overview

This chapter begins with a general discussion of homoskedasticity and heteroskedasticity: the meanings of the terms, the reasons why the distribution of a disturbance term may be subject to heteroskedasticity, and the consequences of the problem for OLS estimators. It continues by presenting several tests for heteroskedasticity and methods of alleviating the problem. It shows how apparent heteroskedasticty may be caused by model misspecification. It concludes with a description of the use of heteroskedasticity-consistent standard errors.

7.2 Learning outcomes

After working through the corresponding chapter in the text, studying the corresponding slideshows, and doing the starred exercises in the text and the additional exercises in this subject guide, you should be able to:

- explain the concepts of homoskedasticity and heteroskedasticity
- describe how the problem of heteroskedasticity may arise
- explain the consequences of heteroskedasticity for OLS estimators, their standard errors, and t and F tests
- perform the Goldfeld–Quandt test for heteroskedasticity
- perform the White test for heteroskedasticity
- explain how the problem of heteroskedasticity may be alleviated
- explain why a mathematical misspecification of the regression model may give rise to a problem of apparent heteroskedasticity
- explain the use of heteroskedasticity-consistent standard errors.

7.3 Additional exercises

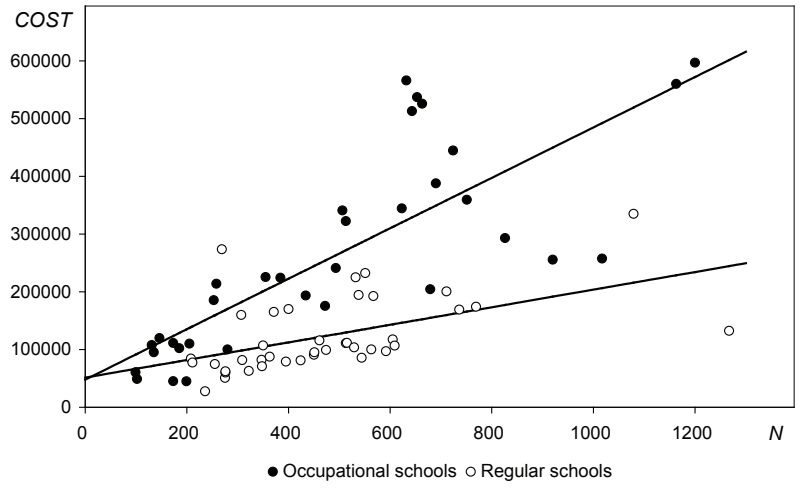
A7.1 Is the disturbance term in your *CES* expenditure function heteroskedastic?

Sort the data by *EXPPC*. Excluding observations for which *EXPPC* is zero, regress *CATPC* on *EXPPC* and *SIZE* (a) for the first three-eighths of the non-zero

observations, and (b) for the last three-eighths. Perform a Goldfeld–Quandt test to test for heteroskedasticity in the *EXPPC* dimension. Repeat using *LGCATPC* as the dependent variable and regressing it on *LGEXPPC* and *LGSIZE*.

A7.2 Repeat Exercise A7.1, using a White test instead of a Goldfeld–Quandt test.

A7.3 The observations for the occupational schools (see Chapter 5 in the text) in the figure suggest that a simple linear regression of cost on number of students, restricted to the subsample of these schools, would be subject to heteroskedasticity. Download the data set from the Online Resource Centre and use a Goldfeld–Quandt test to investigate whether this is the case. If the relationship is heteroskedastic, what could be done to alleviate the problem?



A7.4 A researcher hypothesises that larger economies should be more self-sufficient than smaller ones and that M/G, the ratio of imports, M, to gross domestic product, G, should be negatively related to G:

$$\frac{M}{G} = \beta_1 + \beta_2 G + u$$

with $\beta_2 < 0$. Using data for a sample of 42 countries, with M and G both measured in US$ billion, he fits the regression (standard errors in parentheses):

$$\widehat{\frac{M}{G}} = \underset{\substack{(0.03)}}{0.37} - \underset{\substack{(0.000036)}}{0.000086G} \qquad R^2 = 0.12 \qquad (1)$$

He plots a scatter diagram, reproduced as Figure 7.1, and notices that the ratio M/G tends to have relatively high variance when G is small. He also plots a scatter diagram for M and G, reproduced as Figure 7.2. Defining GSQ as the square of G, he regresses M on G and GSQ:

$$\widehat{M} = \underset{\substack{(10.77)}}{7.27} + \underset{\substack{(0.03)}}{0.30G} - \underset{\substack{(0.000009)}}{0.000049\,GSQ} \qquad R^2 = 0.86 \qquad (2)$$

Finally, he plots a scatter diagram for $\log M$ and $\log G$, reproduced as Figure 7.3, and regresses $\log M$ on $\log G$:

$$\widehat{\log M} = \underset{(0.37)}{-0.14} + \underset{(0.07)}{0.80} \log G \qquad R^2 = 0.78 \qquad (3)$$

Having sorted the data by G, he tests for heteroskedasticity by regressing specifications (1) – (3) first for the 16 countries with smallest G, and then for the 16 countries with the greatest G. RSS_1 and RSS_2, the residual sums of squares for these regressions, are summarised in the following table.

Specification	RSS_1	RSS_2
(1)	0.53	0.21
(2)	3178	71404
(3)	3.45	3.60

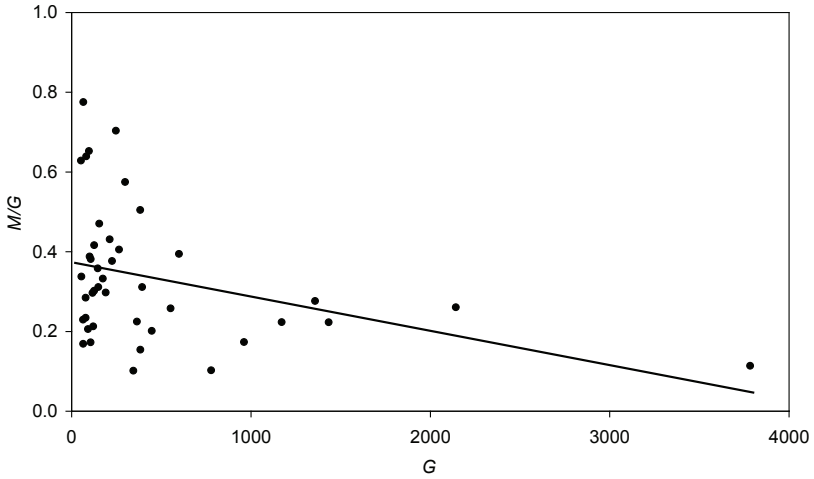


Figure 7.1: Scatter diagram of M/G against G.

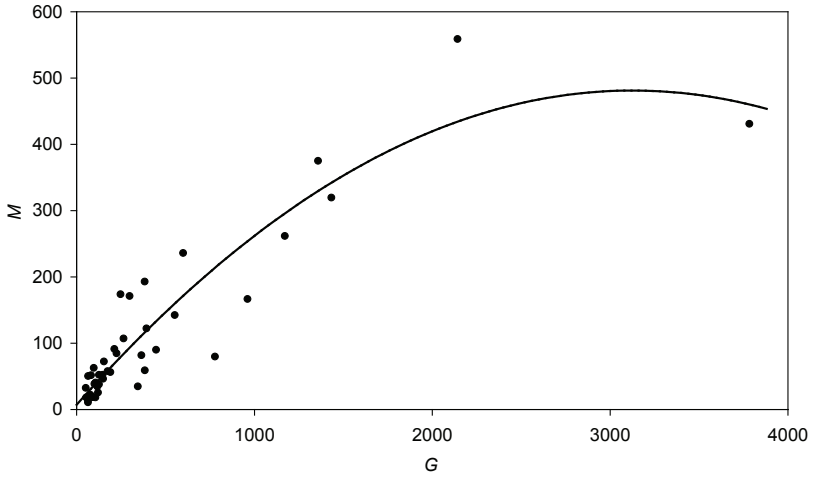


Figure 7.2: Scatter diagram of M against G.

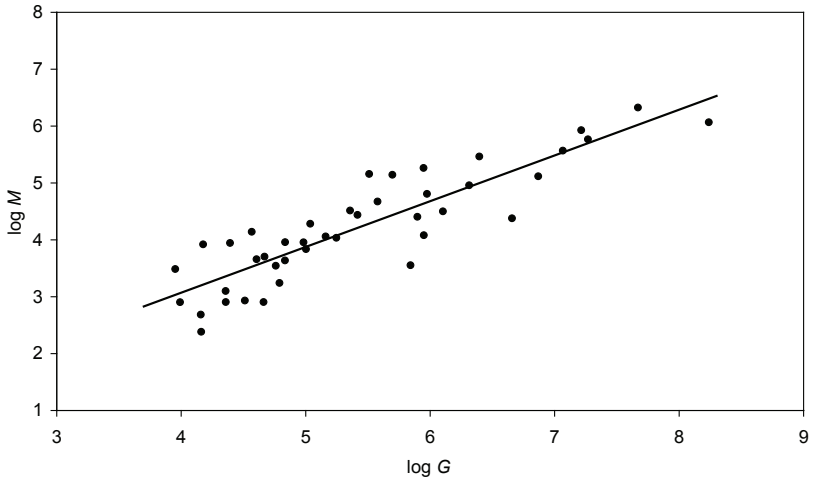


Figure 7.3: Scatter diagram of $\log M$ against $\log G$.

- Discuss whether (1) appears to be an acceptable specification, given the data in the table and Figure 7.1.
- Explain what the researcher hoped to achieve by running regression (2).
- Discuss whether (2) appears to be an acceptable specification, given the data in the table and Figure 7.2.
- Explain what the researcher hoped to achieve by running regression (3).
- Discuss whether (3) appears to be an acceptable specification, given the data in the table and Figure 7.3.
- What are your conclusions concerning the researcher's hypothesis?

A7.5 A researcher has data on the number of children attending, N, and annual recurrent expenditure, EXP, measured in US$, for 50 nursery schools in a US city for 2006 and hypothesises that the cost function is of the quadratic form:

$$EXP = \beta_1 + \beta_2 N + \beta_3 NSQ + u$$

where NSQ is the square of N, anticipating that economies of scale will cause β_3 to be negative. He fits the following equation:

$$\begin{array}{rcll} \widehat{EXP} & = & 17999 + 1060N - 1.29NSQ & R^2 = 0.74 \qquad (1) \\ & & (12908)\ \ (133) \qquad (0.30) & \end{array}$$

Suspecting that the regression was subject to heteroskedasticity, the researcher runs the regression twice more, first with the 19 schools with lowest enrolments, then with the 19 schools with the highest enrolments. The residual sums of squares in the two regressions are 8.0 million and 64.0 million, respectively.

The researcher defines a new variable, $EXPN$, expenditure per student, as $EXPN = EXP/N$, and fits the equation:

$$\begin{array}{rcll} \widehat{EXPN} & = & 1080 - 1.25N + 16114NREC & R^2 = 0.65 \qquad (2) \\ & & (90) \quad (0.25) \quad (6000) & \end{array}$$

where $NREC = 1/N$. He again runs regressions with the 19 smallest schools and the 19 largest schools and the residual sums of squares are 900,000 and 600,000.

- Perform a Goldfeld–Quandt test for heteroskedasticity on both of the regression specifications.
- Explain why the researcher ran the second regression.
- R^2 is lower in regression (2) than in regression (1). Does this mean that regression (1) is preferable?

A7.6 This is a continuation of Exercise A6.5.

- When the researcher presents her results at a seminar, one of the participants says that, since I and G have been divided by Y, (2) is less likely to be subject to heteroskedasticity than (1). Evaluate this suggestion.

A7.7 A researcher has data on annual household expenditure on food, F, and total annual household expenditure, E, both measured in dollars, for 400 households in the United States for 2010. The scatter plot for the data is shown as Figure 7.4. The basic model of the researcher is:

$$F = \beta_1 + \beta_2 E + u \qquad (1)$$

where u is a disturbance term. The researcher suspects heteroskedasticity and performs a Goldfeld–Quandt test and a White test. For the Goldfeld–Quandt test, she sorts the data by size of E and fits the model for the subsample with the 150 smallest values of E and for the subsample with the 150 largest values. The residual sums of squares (RSS) for these regressions are shown in column (1) of the table. She also fits the regression for the entire sample, saves the residuals, and then fits an auxiliary regression of the squared residuals on E and its square. R^2 for this regression is also shown in column (1) in the table. She performs parallel tests of heteroskedasticity for two alternative models:

$$\frac{F}{A} = \beta_1 \frac{1}{A} + \beta_2 \frac{E}{A} + v \qquad (2)$$

$$\log F = \beta_1 + \beta_2 \log E + w. \qquad (3)$$

A is household size in terms of equivalent adults, giving each adult a weight of 1 and each child a weight of 0.7. The scatter plot for F/A and E/A is shown as Figure 7.5, and that for $\log F$ and $\log E$ as Figure 7.6. The data for the heteroskedasticity tests for models (2) and (3) are shown in columns (2) and (3) of the table.

Specification	(1)	(2)	(3)
Goldfeld–Quandt test			
RSS smallest 150	200 million	40 million	20.0
RSS largest 150	820 million	240 million	21.0
White test			
R^2 from auxiliary regression	0.160	0.140	0.001

- Perform the Goldfeld–Quandt test for each model and state your conclusions.

- Explain why the researcher thought that model (2) might be an improvement on model (1).
- Explain why the researcher thought that model (3) might be an improvement on model (1).
- When models (2) and (3) are tested for heteroskedasticity using the White test, auxiliary regressions must be fitted. State the specification of this auxiliary regression for model (2).
- Perform the White test for the three models.
- Explain whether the results of the tests seem reasonable, given the scatter plots of the data.

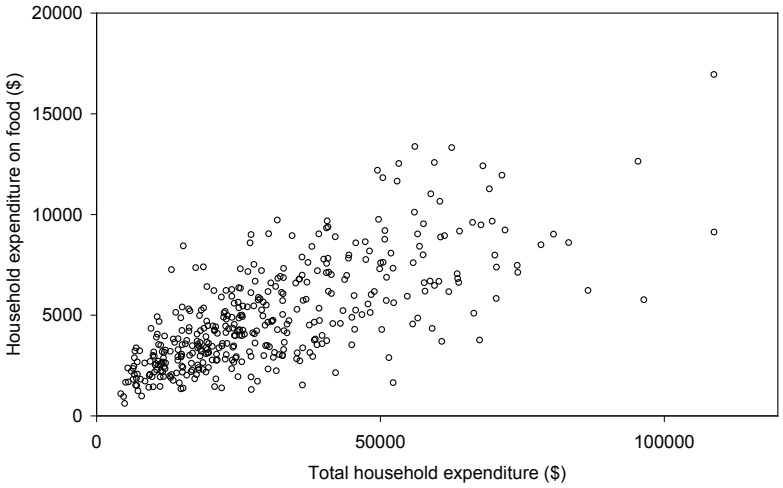


Figure 7.4: Scatter diagram of household expenditure on food against total household expenditure.

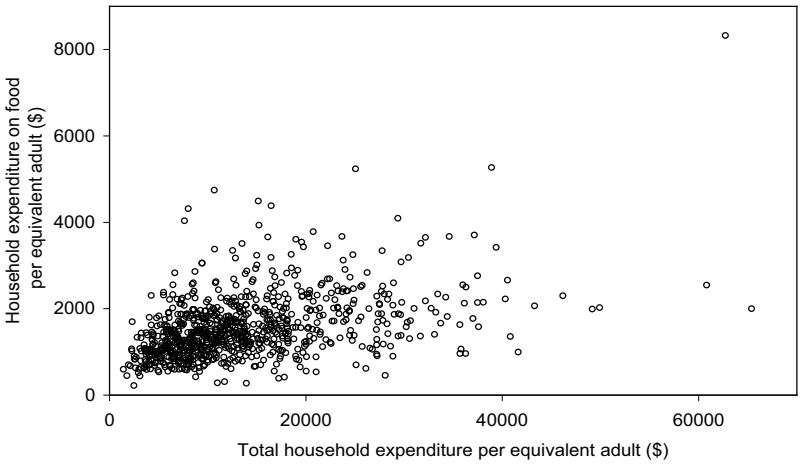


Figure 7.5: Scatter diagram of household expenditure on food per equivalent adult against total household expenditure per equivalent adult.

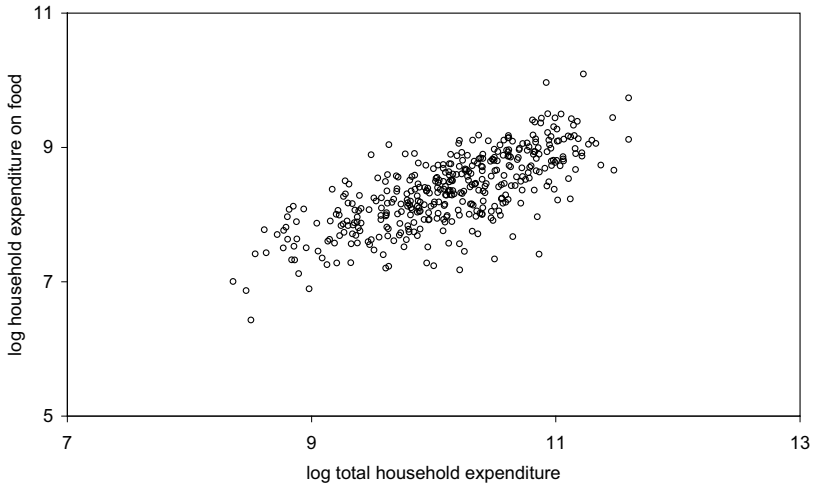


Figure 7.6: Scatter diagram of log household expenditure on food against log total household expenditure.

A7.8 Explain what is correct, mistaken, confused or in need of further explanation in the following statements relating to heteroskedasticity in a regression model:

- 'Heteroskedasticity occurs when the disturbance term in a regression model is correlated with one of the explanatory variables.'
- 'In the presence of heteroskedasticity ordinary least squares (OLS) is an inefficient estimation technique and this causes t tests and F tests to be invalid.'
- 'OLS remains unbiased but it is inconsistent.
- 'Heteroskedasticity can be detected with a Chow test.'
- 'Alternatively one can compare the residuals from a regression using half of the observations with those from a regression using the other half and see if there is a significant difference. The test statistic is the same as for the Chow test.'
- 'One way of eliminating the problem is to make use of a restriction involving the variable correlated with the disturbance term.'
- 'If you can find another variable related to the one responsible for the heteroskedasticity, you can use it as a proxy and this should eliminate the problem.'
- 'Sometimes apparent heteroskedasticity can be caused by a mathematical misspecification of the regression model. This can happen, for example, if the dependent variable ought to be logarithmic, but a linear regression is run.'

7.4 Answers to the starred exercises in the textbook

7.5 The following regressions were fitted using the Shanghai school cost data introduced in Section 6.1 (standard errors in parentheses):

$$\begin{aligned} \widehat{COST} &= 24000 + 339N \qquad R^2 = 0.39 \\ &\quad (27000)\ \ (50) \\ \widehat{COST} &= 51000 - 4000OCC + 152N + 284NOCC \qquad R^2 = 0.68. \\ &\quad (31000)\ (41000)\ \ (60)\ \ (76) \end{aligned}$$

where $COST$ is the annual cost of running a school, N is the number of students, OCC is a dummy variable defined to be 0 for regular schools and 1 for occupational schools, and $NOCC$ is a slope dummy variable defined as the product of N and OCC. There are 74 schools in the sample. With the data sorted by N, the regressions are fitted again for the 26 smallest and 26 largest schools, the residual sums of squares being as shown in the table.

	26 smallest	26 largest
First regression	7.8×10^{10}	54.4×10^{10}
Second regression	6.7×10^{10}	13.8×10^{10}

Perform a Goldfeld–Quandt test for heteroskedasticity for the two models and, with reference to Figure 6.5, explain why the problem of heteroskedasticity is less severe in the second model.

Answer:

For both regressions RSS will be denoted RSS_1 for the 26 smallest schools and RSS_2 for the 26 largest schools. In the first regression, $RSS_2/RSS_1 = (54.4 \times 10^{10})/(7.8 \times 10^{10}) = 6.97$. There are 24 degrees of freedom in each subsample (26 observations, 2 parameters estimated). The critical value of $F(24, 24)$ is approximately 3.7 at the 0.1 per cent level, and so we reject the null hypothesis of homoskedasticity at that level. In the second regression, $RSS_2/RSS_1 = (13.8 \times 10^{10})/(6.7 \times 10^{10}) = 2.06$. There are 22 degrees of freedom in each subsample (26 observations, 4 parameters estimated). The critical value of $F(22, 22)$ is 2.05 at the 5 per cent level, and so we (just) do not reject the null hypothesis of homoskedasticity at that significance level.

Why is the problem of heteroskedasticity less severe in the second regression? The figure in Exercise A7.2 reveals that the cost function is much steeper for the occupational schools than for the regular schools, reflecting their higher marginal cost. As a consequence the two sets of observations diverge as the number of students increases and the scatter is bound to appear heteroskedastic, irrespective of whether the disturbance term is truly heteroskedastic or not. The first regression takes no account of this and the Goldfeld–Quandt test therefore indicates significant heteroskedasticity. In the second regression the problem of apparent heteroskedasticity does not arise because the intercept and slope dummy variables allow separate implicit regression lines for the two types of school.

Looking closely at the diagram, the observations for the occupational schools exhibit a classic pattern of true heteroskedasticity, and this would be confirmed by a Goldfeld–Quandt test confined to the subsample of those schools (see Exercise A7.2). However the observations for the regular schools appear to be homoskedastic and this accounts for the fact that we did not (quite) reject the null hypothesis of homoskedasticity for the combined sample.

7.6 The file educ.dta on the website contains contains international cross-sectional data on aggregate expenditure on education, *EDUC*, gross domestic product, *GDP*, and population, *POP*, for a sample of 38 countries in 1997. *EDUC* and *GDP* are measured in US$ million and *POP* is measured in thousands. Download the data set, plot a scatter diagram of *EDUC* on *GDP*, and comment on whether the data set appears to be subject to heteroskedasticity. Sort the data set by *GDP* and perform a Goldfeld–Quandt test for heteroskedasticity, running regressions using the subsamples of 14 countries with the smallest and greatest *GDP*.

Answer:

The figure plots expenditure on education, *EDUC*, and gross domestic product, *GDP*, for the 38 countries in the sample, measured in $ billion rather than $ million. The observations exhibit heteroskedasticity. Sorting them by *GDP* and regressing *EDUC* on *GDP* for the subsamples of 14 countries with smallest and greatest *GDP*, the residual sums of squares for the first and second subsamples, denoted RSS_1 and RSS_2, respectively, are 1,660,000 and 63,113,000, respectively. Hence:

$$F(12,12) = \frac{RSS_2}{RSS_1} = \frac{63113000}{1660000} = 38.02.$$

The critical value of $F(12,12)$ at the 0.1 per cent level is 7.00, and so we reject the null hypothesis of homoskedasticity.

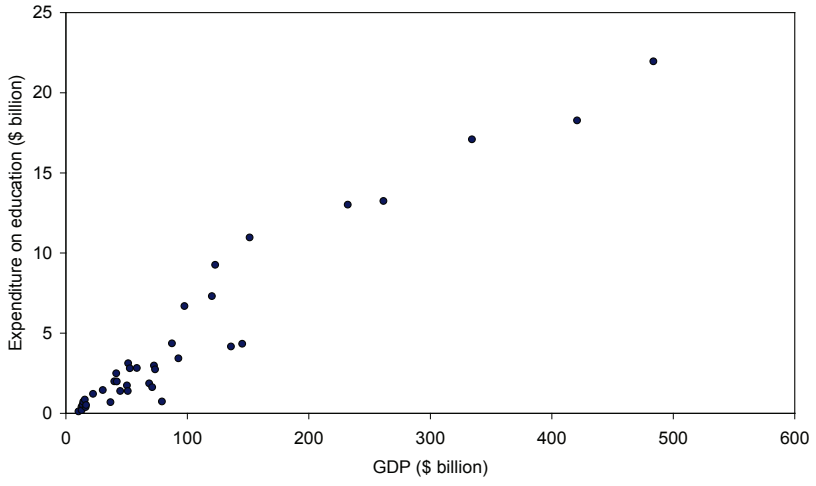


Figure 7.7: Expenditure on education and GDP ($ billion).

7.9 Repeat Exercise 7.6, using the Goldfeld–Quandt test to investigate whether scaling by population or by *GDP*, or whether running the regression in logarithmic form,

would eliminate the heteroskedasticity. Compare the results of regressions using the entire sample and the alternative specifications.

Answer:

Dividing through by population, POP, the model becomes:

$$\frac{EDUC}{POP} = \beta_1 \frac{1}{POP} + \beta_2 \frac{GDP}{POP} + \frac{u}{POP}$$

with expenditure on education per capita, denoted $EDUCPOP$, hypothesised to be a function of gross domestic product per capita, $GDPPOP$, and the reciprocal of population, $POPREC$, with no intercept. Sorting the sample by $GDPPOP$ and running the regression for the subsamples of 14 countries with smallest and largest $GDPPOP$, $RSS_1 = 0.006788$ and $RSS_2 = 1.415516$. Now:

$$F(12,12) = \frac{RSS_2}{RSS_1} = \frac{1.415516}{0.006788} = 208.5.$$

Thus the model is still subject to heteroskedasticity at the 0.1 per cent level. This is evident in Figure 7.8.

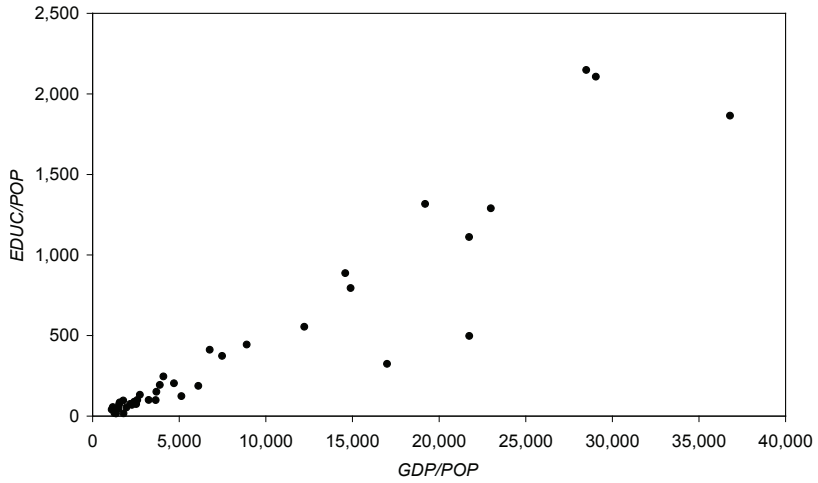


Figure 7.8: Expenditure on education per capita and GDP per capita ($ per capita).

Dividing through instead by GDP, the model becomes:

$$\frac{EDUC}{GDP} = \beta_1 \frac{1}{GDP} + \beta_2 + \frac{u}{GDP}$$

with expenditure on education as a share of gross domestic product, denoted $EDUCGDP$, hypothesised to be a simple function of the reciprocal of gross domestic product, $GDPREC$, with no intercept. Sorting the sample by $GDPREC$ and running the regression for the subsamples of 14 countries with smallest and largest $GDPREC$, $RSS_1 = 0.00413$ and $RSS_2 = 0.00238$. Since RSS_2 is less than RSS_1, we test for heteroskedasticity under the hypothesis that the standard deviation of the disturbance term is inversely related to $GDPREC$:

$$F(12,12) = \frac{RSS_1}{RSS_2} = \frac{0.00413}{0.00238} = 1.74.$$

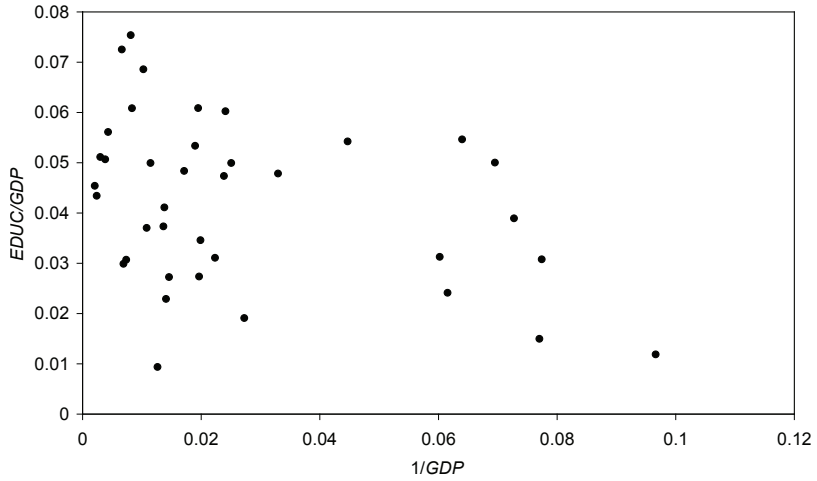


Figure 7.9: Expenditure on education as a proportion of GDP and the reciprocal of GDP (measured in $ billion).

The critical value of $F(12, 12)$ at the 5 per cent level is 2.69, so we do not reject the null hypothesis of homoskedasticity. Could one tell this from Figure 7.9? It is a little difficult to say.

Finally, we will consider a logarithmic specification. If the true relationship is logarithmic, and homoskedastic, it would not be surprising that the linear model appeared heteroskedastic. Sorting the sample by *GDP*, RSS_1 and RSS_2 are 2.733 and 3.438 for the subsamples of 14 countries with smallest and greatest *GDP*. The F statistic is:

$$F(12, 12) = \frac{RSS_1}{RSS_2} = \frac{3.438}{2.733} = 1.26.$$

Thus again we would not reject the null hypothesis of homoskedasticity.

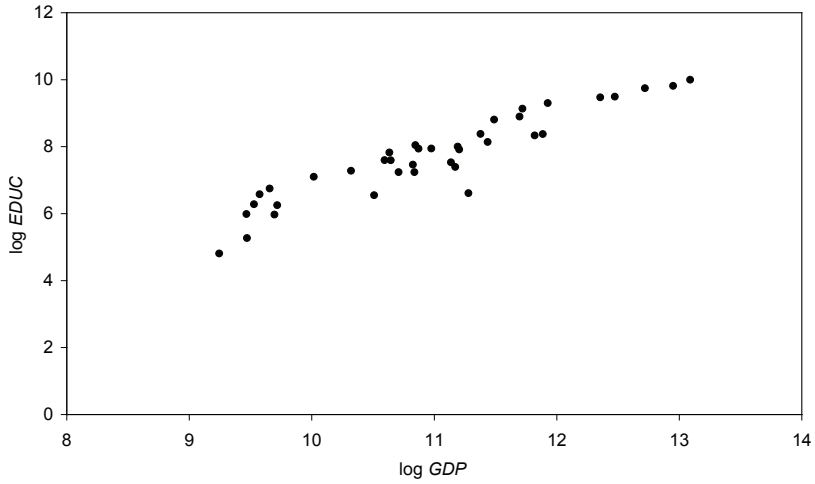


Figure 7.10: Expenditure on education and GDP, logarithmic.

The third and fourth models both appear to be free from heteroskedasticity. How do we choose between them? We will examine the regression results, shown for the two models with the full sample:

```
. reg EDUCGDP GDPREC

  Source |       SS       df       MS                 Number of obs =      38
---------+------------------------------               F(  1,  36)   =    5.62
   Model |  .001348142     1  .001348142               Prob > F      =  0.0233
Residual |  .008643037    36  .000240084               R-squared     =  0.1349
---------+------------------------------               Adj R-squared =  0.1109
   Total |  .009991179    37  .000270032               Root MSE      =  .01549

------------------------------------------------------------------------------
 EDUCGDP |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
  GDPREC |  -234.0823   98.78309     -2.370   0.023      -434.4236   -33.74086
   _cons |   .0484593   .0036696     13.205   0.000       .0410169    .0559016
------------------------------------------------------------------------------
```

```
. reg LGEE LGGDP

  Source |       SS       df       MS                 Number of obs =      38
---------+------------------------------               F(  1,    36) =  246.20
   Model |  51.9905508     1  51.9905508               Prob > F      =  0.0000
Residual |   7.6023197    36  .211175547               R-squared     =  0.8724
---------+------------------------------               Adj R-squared =  0.8689
   Total |  59.5928705    37  1.61061812               Root MSE      =  .45954

------------------------------------------------------------------------------
    LGEE |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
   LGGDP |   1.160594   .0739673     15.691   0.000       1.010582    1.310607
   _cons |  -5.025204   .8152239     -6.164   0.000      -6.678554   -3.371853
------------------------------------------------------------------------------
```

In equation form, the first regression is:

$$\frac{\widehat{EDUC}}{GDP} = \underset{(0.004)}{0.048} - \underset{(98.8)}{234.1}\frac{1}{GDP} \qquad R^2 = 0.13$$

Multiplying through by GDP, it may be rewritten:

$$\widehat{EDUC} = -234.1 + 0.048\,GDP.$$

It implies that expenditure on education accounts for 4.8 per cent of gross domestic product at the margin. The constant does not have any sensible interpretation. We will compare this with the output from an OLS regression that makes no attempt to eliminate heteroskedasticity:

```
. reg EDUC GDP

  Source |       SS       df       MS                  Number of obs =      38
---------+------------------------------                F(  1,    36) =  509.80
   Model |  1.0571e+09     1  1.0571e+09                Prob > F      =  0.0000
Residual |  74645819.2    36  2073494.98                R-squared     =  0.9340
---------+------------------------------                Adj R-squared =  0.9322
   Total |  1.1317e+09    37  30586911.0                Root MSE      =  1440.0

------------------------------------------------------------------------------
    EDUC |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     GDP |   .0480656   .0021288     22.579   0.000       .0437482     .052383
   _cons |  -160.4669    311.699     -0.515   0.610      -792.6219     471.688
------------------------------------------------------------------------------
```

The slope coefficient, 0.048, is identical to three decimal places. This is not entirely a surprise, since heteroskedasticity does not give rise to bias and so there should be no systematic difference between the estimate from an OLS regression and that from a specification that eliminates heteroskedasticity. Of course, it is a surprise that the estimates are so close. Generally there would be some random difference, and of course the OLS estimate would tend to be less accurate. In this case, the main difference is in the estimated standard error. That for the OLS regression is actually smaller than that for the regression of *EDUCGDP* on *GDPREC*, but it is misleading. It is incorrectly calculated and we know that, since OLS is inefficient, the true standard error for the OLS estimate is actually larger.

The logarithmic regression in equation form is:

$$\begin{aligned} \widehat{\log EDUC} &= -5.03 + 1.16 \log GDP \qquad R^2 = 0.87 \\ &\quad (0.82)\ \ (0.07) \end{aligned}$$

implying that the elasticity of expenditure on education with regard to gross domestic product is 1.16. In substance the interpretations of the models are similar, since both imply that the proportion of GDP allocated to education increases slowly with GDP, but the elasticity specification seems a little more informative and probably serves as a better starting point for further exploration. For example, it would be natural to add the logarithm of population to see if population had an independent effect.

7.10 It was reported above that the heteroskedasticity-consistent estimate of the standard error of the coefficient of *GDP* in equation (7.18) was 0.18. Explain why the corresponding standard error in equation (7.20) ought to be lower and comment on the fact that it is not.

Answer:

(7.20), unlike (7.18) appears to be free from heteroskedasticity and therefore should provide more efficient estimates of the coefficients, reflected in lower standard errors when computed correctly. However the sample may be too small for the heteroskedasticity-consistent estimator to be a good guide.

7.11 A health economist plans to evaluate whether screening patients on arrival or spending extra money on cleaning is more effective in reducing the incidence of

infections by the MRSA bacterium in hospitals. She hypothesises the following model:

$$MRSA_i = \beta_1 + \beta_2 S_i + \beta_3 C_i + u_i$$

where, in hospital i, $MRSA$ is the number of infections per thousand patients, S is expenditure per patient on screening, and C is expenditure per patient on cleaning. u_i is a disturbance term that satisfies the usual regression model assumptions. In particular, u_i is drawn from a distribution with mean zero and constant variance σ^2. The researcher would like to fit the relationship using a sample of hospitals. Unfortunately, data for individual hospitals are not available. Instead she has to use regional data to fit:

$$\overline{MRSA_j} = \beta_1 + \beta_2 \overline{S_j} + \beta_3 \overline{C_j} + \overline{u_j}$$

where $\overline{MRSA_j}$, $\overline{S_j}$, $\overline{C_j}$, and $\overline{u_j}$ are the averages of $MRSA$, S, C, and u for the hospitals in region j. There were different numbers of hospitals in the regions, there being n_j hospitals in region j.

Show that the variance of $\overline{u_j}$ is equal to σ^2/n_j and that an OLS regression using the grouped regional data to fit the relationship will be subject to heteroskedasticity.

Assuming that the researcher knows the value of n_j for each region, explain how she could re-specify the regression model to make it homoskedastic. State the revised specification and demonstrate mathematically that it is homoskedastic. Give an intuitive explanation of why the revised specification should tend to produce improved estimates of the parameters.

Answer:

$$\text{var}(\overline{u_j}) = \text{var}\left(\frac{1}{n}\sum_{k=1}^{n_j} u_{jk}\right) = \left(\frac{1}{n_j}\right)^2 \text{var}\left(\sum_{k=1}^{n_j} u_{jk}\right) = \left(\frac{1}{n_j}\right)^2 \sum_{k=1}^{n_j} \text{var}(u_{jk})$$

since the covariance terms are all 0. Hence:

$$\text{var}(\overline{u_j}) = \left(\frac{1}{n_j}\right)^2 n_j \sigma^2 = \frac{\sigma^2}{n_j}.$$

To eliminate the heteroskedasticity, multiply observation j by $\sqrt{n_j}$. The regression becomes:

$$\sqrt{n_j}\overline{MRSA_j} = \beta_1\sqrt{n_j} + \beta_2\sqrt{n_j}\overline{S_j} + \beta_3\sqrt{n_j}\overline{C_j} + \sqrt{n_j}\,\bar{u}_j.$$

The variance of the disturbance term is now:

$$\text{var}\left(\sqrt{n_j}\overline{u_j}\right) = \left(\sqrt{n_j}\right)^2 \text{var}(\overline{u_j}) = n_j \frac{\sigma^2}{n_j} = \sigma^2$$

and is thus the same for all observations.

From the expression for $\text{var}(\overline{u_j})$, we see that, the larger the group, the more reliable should be its observation (the closer its observation should tend to be to the population relationship). The scaling gives greater weight to the more reliable observations and the resulting estimators should be more efficient.

7.5 Answers to the additional exercises

A7.1 The first step is to drop the zero-observations from the data set and sort it by *EXPPC*. The F statistic is then computed as:

$$F(n_2 - k, n_1 - k) = \frac{RSS_2/(n_2 - k)}{RSS_1/(n_1 - k)}$$

where n_1 and n_2 are the number of available observations and k is the number of parameters in the regression specification.

```
.  drop if FDHO == 0
(0 observations deleted)
. gen EXPPC = EXP/SIZE
. sort EXPPC
. gen LGEXPPC = ln(EXPPC)
. gen LGSIZE = ln(SIZE)
. gen FDHOPC = FDHO/SIZE
. gen LGFDHOPC = ln(FDHOPC)

. reg FDHOPC EXPPC SIZE in 1/2375
------------------------------------------------------------------------------
      Source |       SS       df       MS              Number of obs =    2375
-------------+------------------------------           F(  2,  2372) =  278.36
       Model |  7382348.18     2  3691174.09           Prob> F       =  0.0000
    Residual |  31453534.1  2372  13260.3432           R-squared     =  0.1901
-------------+------------------------------           Adj R-squared =  0.1894
       Total |  38835882.2  2374  16358.8383           Root MSE      =  115.15
------------------------------------------------------------------------------
      FDHOPC |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       EXPPC |   .1107869   .0051862    21.36   0.000     .1006169    .1209569
        SIZE |  -4.462209   1.438899    -3.10   0.002    -7.283838   -1.640579
       _cons |   85.38055   9.590628     8.90   0.000     66.57366    104.1874
------------------------------------------------------------------------------

. reg FDHOPC EXPPC SIZE in 3960/6334
------------------------------------------------------------------------------
      Source |       SS       df       MS              Number of obs =    2375
-------------+------------------------------           F(  2,  2372) =  170.94
       Model |  40643447.8     2  20321723.9           Prob> F       =  0.0000
    Residual |   281980931  2372  118878.976           R-squared     =  0.1260
-------------+------------------------------           Adj R-squared =  0.1252
       Total |   322624379  2374  135899.064           Root MSE      =  344.79
------------------------------------------------------------------------------
      FDHOPC |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       EXPPC |   .0286606   .0019716    14.54   0.000     .0247944    .0325268
        SIZE |  -54.33452   7.047302    -7.71   0.000    -68.15403   -40.51501
       _cons |   508.6148   22.37631    22.73   0.000     464.7356    552.4939
------------------------------------------------------------------------------
```

```
. reg LGFDHOPC LGEXPPC LGSIZE in 1/2375
------------------------------------------------------------------------------
    Source |       SS       df       MS              Number of obs =    2375
-----------+------------------------------           F(  2,  2372) =  369.49
     Model |  207.241064     2  103.620532           Prob> F       =  0.0000
  Residual |  665.204785  2372  .280440466           R-squared     =  0.2375
-----------+------------------------------           Adj R-squared =  0.2369
     Total |  872.445849  2374  .367500357           Root MSE      =  .52957
------------------------------------------------------------------------------
  LGFDHOPC |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------+------------------------------------------------------------------
   LGEXPPC |   .6510802   .0265608    24.51   0.000     .5989953     .703165
    LGSIZE |  -.0567001   .0198997    -2.85   0.004    -.0957227   -.0176775
     _cons |   .6450249   .1965331     3.28   0.001     .2596305    1.030419
------------------------------------------------------------------------------
```

```
. reg LGFDHOPC LGEXPPC LGSIZE in 3960/6334
------------------------------------------------------------------------------
    Source |       SS       df       MS              Number of obs =    2375
-----------+------------------------------           F(  2,  2372) =  138.91
     Model |  94.0495475     2  47.0247737           Prob> F       =  0.0000
  Residual |  802.969196  2372  .338519897           R-squared     =  0.1048
-----------+------------------------------           Adj R-squared =  0.1041
     Total |  897.018744  2374  .377851198           Root MSE      =  .58182
------------------------------------------------------------------------------
  LGFDHOPC |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------+------------------------------------------------------------------
   LGEXPPC |   .4072631   .0297285    13.70   0.000     .3489666    .4655596
    LGSIZE |  -.1426229   .0247966    -5.75   0.000    -.1912482   -.0939976
     _cons |   2.742439   .2635057    10.41   0.000     2.225714    3.259165
------------------------------------------------------------------------------
```

The F statistic for the linear specification is:

$$F(2372,\, 2372) = \frac{281980931/2372}{31453534/2372} = 8.97.$$

This is significant at the 0.1 per cent level. The corresponding F statistic for the logarithmic specification is 1.21. The critical value of $F(200, 200)$ at the 5 per cent level is 1.26. The critical value for $F(2372,\, 2372)$ must be lower, so the null hypothesis of homoskedasticity is probably rejected at that level. However, the problem has evidently been largely eliminated.

The logarithmic specification in general appears to be much less heteroskedastic than the linear one and for some categories the null hypothesis of homoskedasticity would not be rejected. Note that for a few of these $RSS_2 < RSS_1$ for the logarithmic specification.

Goldfeld–Quandt tests								
			Linear			Logarithmic		
	n_1	n_2	$RSS_1 \times 10^{-6}$	$RSS_2 \times 10^{-6}$	F	RSS_1	RSS_2	F
ADM	1,056	1,056	1.95	62.93	32.30	1,324.96	1,593.31	1.20
CLOT	1,688	1,688	7.17	316.80	44.17	2,107.28	2,196.79	1.04
DOM	623	623	7.23	238.90	33.05	1,571.19	1,505.92	1.04*
EDUC	210	210	11.70	376.01	32.15	495.12	507.27	1.02
ELEC	2,186	2,186	7.55	33.34	4.41	1,034.70	923.18	1.12*
FDAW	1,913	1,913	9.00	278.13	30.89	1,136.09	1,361.12	1.20
FDHO	2,375	2,375	31.45	281.98	8.97	665.20	802.97	1.21
FOOT	685	685	0.55	5.74	10.37	513.08	514.24	1.00
FURN	183	183	7.17	258.26	36.00	322.50	368.42	1.14
GASO	2,141	2,141	11.06	159.54	14.43	921.26	1,245.55	1.35
HEAL	1,801	1,801	32.91	876.72	26.64	2,233.73	2,192.92	1.02*
HOUS	2,334	2,334	105.48	3,031.19	28.74	2,129.27	1,475.02	1.44*
LIFE	470	470	2.85	48.37	16.95	503.19	667.14	1.33
LOCT	260	260	0.58	5.32	9.13	366.16	409.90	1.12
MAPP	150	150	2.85	37.01	12.96	211.71	243.18	1.15
PERS	1,431	1,431	0.47	9.01	19.34	1,045.70	1,204.31	1.15
READ	858	858	0.36	4.95	13.69	1,076.35	1,085.38	1.01
SAPP	389	389	0.56	10.68	19.04	396.41	433.37	1.09
TELE	2,171	2,171	3.27	26.80	8.19	1,133.43	1,123.46	1.01*
TEXT	372	372	0.57	2.05	3.61	410.29	393.80	1.04*
TOB	433	433	1.56	27.81	17.84	312.71	338.28	1.08
TOYS	939	939	6.83	87.65	12.83	1,079.76	1,064.92	1.01*
TRIP	194	194	9.62	77.65	8.07	300.70	335.75	1.12

* indicates $RSS_2 < RSS_1$

A7.2 The table shows the construction of the White test statistics for the linear and logarithmic specifications for each category of expenditure. The regressors in the auxiliary regression were expenditure per capita and its square, size and its square, and the product of expenditure per capita and size. Hence there were five degrees of freedom for the chi-squared test. The critical values are 11.1 and 15.1 at the 5 per cent and 1 per cent levels. Thus there is strong evidence of heteroskedasticity for all of the categories in the linear specification. There is also evidence for some categories in the logarithmic specification. It is possible that the White test, being more general, is finding evidence of heteroskedasticity not detected by the Goldfeld–Quandt test.

White tests					
		Linear		Logarithmic	
	n	R^2	nR^2	R^2	nR^2
ADM	2,815	0.1710	481.4	0.0097	27.3
CLOT	4,500	0.0180	81.0	0.0074	33.3
DOM	1,661	0.0191	31.7	0.0062	10.3
EDUC	561	0.1432	80.3	0.0078	4.4
ELEC	5,828	0.0487	283.8	0.0090	52.5
FDAW	5,102	0.1072	546.9	0.0067	34.2
FDHO	6,334	0.1143	724.0	0.0129	81.7
FOOT	1,827	0.0191	34.9	0.0023	4.2
FURN	487	0.3287	160.1	0.0197	9.6
GASO	5,710	0.0575	328.3	0.0152	86.8
HEAL	4,802	0.0608	292.0	0.0021	10.1
HOUS	6,223	0.2002	1,245.8	0.0120	74.7
LIFE	1,253	0.0535	67.0	0.0132	16.5
LOCT	692	0.0388	26.8	0.0192	13.3
MAPP	399	0.0882	35.2	0.0168	6.7
PERS	3,817	0.0607	231.7	0.0086	32.8
READ	2,287	0.0158	36.1	0.0072	16.5
SAPP	1,037	0.0221	22.9	0.0032	3.3
TELE	5,788	0.0724	419.1	0.0021	12.2
TEXT	992	0.0183	18.2	0.0049	4.9
TOB	1,155	0.0235	27.1	0.0061	7.0
TOYS	2,504	0.0347	86.9	0.0026	6.5
TRIP	516	0.0571	29.5	0.0047	2.4

A7.3 Having sorted by N, the number of students, RSS_1 and RSS_2 are 2.02×10^{10} and 22.59×10^{10}, respectively, for the subsamples of the 13 smallest and largest schools. The F statistic is 11.18. The critical value of $F(11, 11)$ at the 0.1 per cent level must be a little below 8.75, the critical value for $F(10, 10)$, and so the null hypothesis of homoskedasticity is rejected at that significance level.

One possible way of alleviating the heteroskedasticity is by scaling through by the number of students. The dependent variable now becomes the unit cost per student year, and this is likely to be more uniform than total recurrent cost. Scaling through by N, and regressing *UNITCOST*, defined as *COST* divided by N, on *NREC*, the reciprocal of N, having first sorted by *NREC*, RSS_1 and RSS_2 are now 349,000 and 504,000. The F statistic is therefore 1.44, and this is not significant even at the 5 per cent level since the critical value must be a little above 2.69, the critical value for $F(12, 12)$. The regression output for this specification using the full sample is shown.

```
. reg UNITCOST NREC

  Source |       SS       df       MS              Number of obs =       34
---------+------------------------------           F(  1,    32) =     0.74
   Model |  27010.3792     1  27010.3792           Prob > F      =   0.3954
Residual |  1164624.44    32  36394.5138           R-squared     =   0.0227
---------+------------------------------           Adj R-squared =  -0.0079
   Total |  1191634.82    33  36110.1461           Root MSE      =   190.77
```

```
------------------------------------------------------------------------------
UNITCOST |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
    NREC |   10975.91    12740.7      0.861   0.395      -14976.04    36927.87
   _cons |    524.813   53.88367      9.740   0.000       415.0556    634.5705
------------------------------------------------------------------------------
```

In equation form, the regression is:

$$\frac{\widehat{COST}}{N} = \underset{(53.9)}{524.8} + \underset{(12741)}{10976}\frac{1}{N} \qquad R^2 = 0.03$$

Multiplying through by N, it may be rewritten:

$$\widehat{COST} = 10976 + 524.8N.$$

The estimate of the marginal cost is somewhat higher than the estimate of 436 obtained using OLS in Section 5.3 of the text.

A second possible way of alleviating the heteroskedasticity is to hypothesise that the true relationship is logarithmic, in which case the use of an inappropriate linear specification would give rise to apparent heteroskedasticity. Scaling through by N, and regressing $LGCOST$, the (natural) logarithm of $COST$, on LGN, the logarithm of N, RSS_1 and RSS_2 are 2.16 and 1.58. The F statistic is therefore 1.37, and again this is not significant even at the 5 per cent level. The regression output for this specification using the full sample is shown.

```
. reg LGCOST LGN

  Source |       SS       df       MS              Number of obs =      34
---------+------------------------------           F(  1,    32) =  100.98
   Model |  14.7086057     1  14.7086057            Prob > F      =  0.0000
Residual |  4.66084501    32  .145651406            R-squared     =  0.7594
---------+------------------------------           Adj R-squared =  0.7519
   Total |  19.3694507    33   .58695305            Root MSE      =  .38164

------------------------------------------------------------------------------
  LGCOST |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     LGN |    .909126   .0904681     10.049   0.000       .7248485    1.093404
   _cons |   6.808312   .5435035     12.527   0.000       5.701232    7.915393
------------------------------------------------------------------------------
```

The estimate of the elasticity of cost with respect to number of students, 0.91, is less than 1 and thus suggests that the schools are subject to economies of scale. However, we are not able to reject the null hypothesis that the elasticity is equal to 1 and thus that costs are proportional to numbers, the t statistic for the null hypothesis being too low:

$$t = \frac{0.909 - 1.000}{0.091} = -1.00.$$

A7.4 *Discuss whether (1) appears to be an acceptable specification, given the data in the table and Figure 7.1.*

Using the Goldfeld–Quandt test to test specification (1) for heteroskedasticity assuming that the standard deviation of u is inversely proportional to G, we have:

$$F(14,14) = \frac{0.53}{0.21} = 2.52.$$

The critical value of $F(14,14)$ at the 5 per cent level is 2.48, so we just reject the null hypothesis of homoskedasticity at that level. Figure 7.1 does strongly suggest heteroskedasticity. Thus (1) does not appear to be an acceptable specification.

Explain what the researcher hoped to achieve by running regression (2).

If it is true that the standard deviation of u is inversely proportional to G, the heteroskedasticity could be eliminated by multiplying through by G. This is the motivation for the second specification. An intercept that in principle does not exist has been added, thereby changing the model specification slightly.

Discuss whether (2) appears to be an acceptable specification, given the data in the table and Figure 7.2.

$$F(13,13) = \frac{71404}{3178} = 22.47.$$

The critical value of $F(13,13)$ at the 0.1 per cent level is about 6.4, so the null hypothesis of homoskedasticity is rejected. Figure 7.2 confirms the heteroskedasticity.

Explain what the researcher hoped to achieve by running regression (3).

Heteroskedasticity can appear to be present in a regression in natural units if the true relationship is logarithmic. The disturbance term in a logarithmic regression is effectively increasing or decreasing the value of the dependent variable by random proportions. Its effect in absolute terms will therefore tend to be greater, the larger the value of G. The researcher is checking to see if this is the reason for the heteroskedasticity in the second specification.

Discuss whether (3) appears to be an acceptable specification, given the data in the table and Figure 7.3.

Obviously there is no problem with the Goldfeld–Quandt test, since:

$$F(14,14) = \frac{3.60}{3.45} = 1.04.$$

Figure 7.3 looks free from heteroskadasticity.

What are your conclusions concerning the researcher's hypothesis?

Evidence in support of the hypothesis is provided by (3) where, with:

$$t = \frac{0.80 - 1}{0.07} = -2.86$$

the elasticity is significantly lower than 1. Figures 7.1 and 7.2 also strongly suggest that on balance larger economies have lower import ratios than smaller ones.

A7.5 *Perform a Goldfeld–Quandt test for heteroskedasticity on both of the regression specifications.*

The F statistics for the G–Q test for the two specifications are:

$$F(16,16) = \frac{64/16}{8/16} = 8.0 \quad \text{and} \quad F(16,16) = \frac{900/16}{600/16} = 1.5.$$

The critical value of $F(16,16)$ is 2.33 at the 5 per cent level and 5.20 at the 0.1 per cent level. Hence one would reject the null hypothesis of homoskedasticity at the 0.1 per cent level for regression 1 and one would not reject it even at the 5 per cent level for regression 2.

Explain why the researcher ran the second regression.

He hypothesised that the standard deviation of the disturbance term in observation i was proportional to N_i : $\sigma_i = \lambda N_i$ for some λ. If this is the case, dividing through by N_i makes the specification homoskedastic, since:

$$\text{var}\left(\frac{u_i}{N_i}\right) = \frac{1}{N^2}\text{var}(u_i) = \frac{1}{N_i^2}(\lambda N_i)^2 = \lambda^2$$

and is therefore the same for all i.

R^2 is lower in regression (2) than in regression (1). Does this mean that regression (1) is preferable?

R^2 is not comparable because the dependent variable is different in the two regressions. Regression (2) is to be preferred since it is free from heteroskedasticity and therefore ought to tend to yield more precise estimates of the coefficients with valid standard errors.

A7.6 *When the researcher presents her results at a seminar, one of the participants says that, since I and G have been divided by Y, (2) is less likely to be subject to heteroskedasticity than (1). Evaluate this suggestion.*

If the restriction is valid, imposing it will have no implications for the disturbance term and so it could not lead to any mitigation of a potential problem of heteroskedasticity. [If there were heteroskedasticity, and if the specification were linear, scaling through by a variable proportional in observation i to the standard deviation of u_i in observation i would lead to the elimination of heteroskedasticity. The present specification is logarithmic and dividing I and G by Y does not affect the disturbance term.]

A7.7 *Perform the Goldfeld–Quandt test for each model and state your conclusions.*

The ratios are 4.1, 6.0, and 1.05. In each case we should look for the critical value of $F(148,148)$. The critical values of $F(150,150)$ at the 5 per cent, 1 per cent, and 0.1 per cent levels are 1.31, 1.46, and 1.66, respectively. Hence we reject the null hypothesis of homoskedasticity at the 0.1 per cent level (1 per cent is OK) for models (1) and (2). We do not reject it even at the 5 per cent level for model (3).

Explain why the researcher thought that model (2) might be an improvement on model (1).

If the assumption that the standard deviation of the disturbance term is proportional to household size, scaling through by A should eliminate the heteroskedasticity, since:

$$E(v^2) = E\left(\left[\frac{u}{A}\right]^2\right) = \frac{1}{A^2}E(u^2) = \lambda^2$$

if the standard deviation of $u = \lambda A$.

Explain why the researcher thought that model (3) might be an improvement on model (1).

It is possible that the (apparent) heteroskedasticity is attributable to mathematical misspecification. If the true model is logarithmic, a homoskedastic disturbance term would appear to have a heteroskedastic effect if the regression is performed in the original units.

When models (2) and (3) are tested for heteroskedasticity using the White test, auxiliary regressions must be fitted. State the specification of this auxiliary regression for model (2).

The dependent variable is the squared residuals from the model regression. The explanatory variables are the reciprocal of A and its square, E/A and its square, and the product of the reciprocal of A and E/A. (No constant.)

Perform the White test for the three models.

nR^2 is 64.0, 56.0, and 0.4 for the three models. Under the null hypothesis of homoskedasticity, this statistic has a chi-squared distribution with degrees of freedom equal to the number of terms on the right side of the regression, minus one. This is two for models (1) and (3). The critical value of chi-squared with two degrees of freedom is 5.99, 9.21, and 13.82 at the 5, 1, and 0.1 per cent levels. Hence H0 is rejected at the 0.1 per cent level for model (1), and not rejected even at the 5 per cent level for model (3). In the case of model (2), there are five terms on the right side of the regression. The critical value of chisquared with four degrees of freedom is 18.47 at the 0.1 per cent level. Hence H_0 is rejected at that level.

Explain whether the results of the tests seem reasonable, given the scatter plots of the data.

Absolutely. In Figures 7.1 and 7.2, the variances of the dispersions of the dependent variable clearly increase with the size of the explanatory variable. In Figure 7.3, the dispersion is much more even.

A7.8 *'Heteroskedasticity occurs when the disturbance term in a regression model is correlated with one of the explanatory variables.'*

This is false. Heteroskedasticity occurs when the variance of the disturbance term is not the same for all observations.

'In the presence of heteroskedasticity ordinary least squares (OLS) is an inefficient estimation technique and this causes t tests and F tests to be invalid.'

It is true that OLS is inefficient and that the t and F tests are invalid, but 'and this causes' is wrong.

'OLS remains unbiased but it is inconsistent.'

It is true that OLS is unbiased, but false that it is inconsistent.

'Heteroskedasticity can be detected with a Chow test.'

This is false.

'Alternatively one can compare the residuals from a regression using half of the observations with those from a regression using the other half and see if there is a significant difference. The test statistic is the same as for the Chow test.'

The first sentence is basically correct with the following changes and clarifications: one is assuming that the standard deviation of the disturbance term is proportional to one of the explanatory variables; the sample should first be sorted according to the size of the explanatory variable; rather than split the sample in half, it would be better to compare the first three-eighths (or one third) of the observations with the last three-eighths (or one third); 'comparing the residuals' is too vague: the F statistic is $F(n' - k, n' - k) = RSS_2/RSS_1$ assuming n' observations and k parameters in each subsample regression, and placing the larger RSS over the smaller.

The second sentence is false.

'One way of eliminating the problem is to make use of a restriction involving the variable correlated with the disturbance term.'

This is nonsense.

'If you can find another variable related to the one responsible for the heteroskedasticity, you can use it as a proxy and this should eliminate the problem.'

This is more nonsense.

'Sometimes apparent heteroskedasticity can be caused by a mathematical misspecification of the regression model. This can happen, for example, if the dependent variable ought to be logarithmic, but a linear regression is run.'

True. A homoskedastic disturbance term in a logarithmic regression, which is responsible for proportional changes in the dependent variable, may appear to be heteroskedastic in a linear regression because the absolute changes in the dependent variable will be proportional to its size.