

Chapter 7

Heteroscedasticity

Overview

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

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 guide, you should be able to:

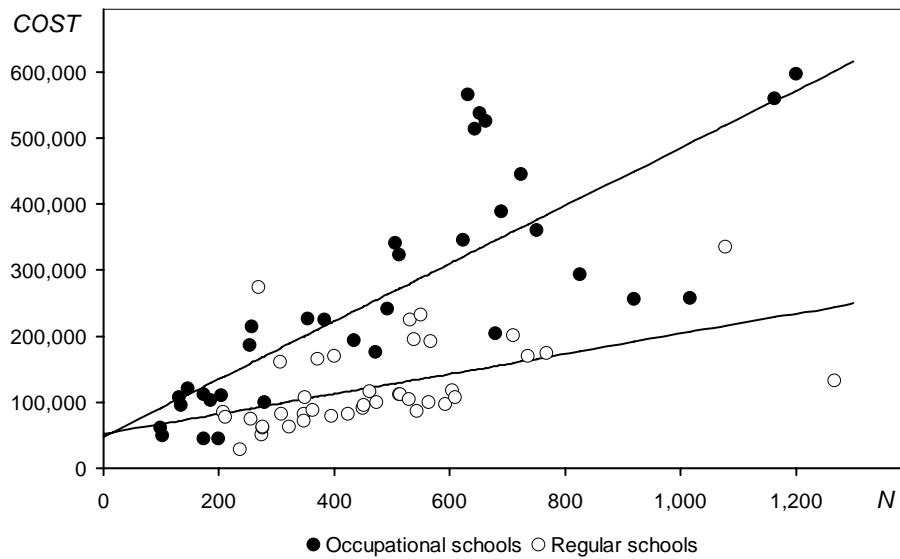
- explain the concepts of homoscedasticity and heteroscedasticity
- describe how the problem of heteroscedasticity may arise
- explain the consequences of heteroscedasticity for OLS estimators, their standard errors, and t and F tests
- perform the Goldfeld–Quandt test for heteroscedasticity
- perform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

Additional exercises

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

Sort the data by *EXPPC*, regress *CATPC* on *EXPPC* and *SIZE*, and perform a Goldfeld–Quandt test to test for heteroscedasticity in the *EXPPC* dimension. Repeat using *LGCATPC* as the dependent variable.

A7.2 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 heteroscedasticity. Download the data set from the heteroscedastic data sets folder on the website and use a Goldfeld–Quandt test to investigate whether this is the case. If the relationship is heteroscedastic, what could be done to alleviate the problem?



A7.3 A researcher hypothesizes 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):

$$\frac{\hat{M}}{G} = 0.37 - 0.000086 G \quad R^2 = 0.12 \quad (1)$$

(0.03) (0.000036)

He plots a scatter diagram, reproduced as Figure 1, and notices that the ratio $\frac{M}{G}$ tends to have relatively high variance when G is small. He also plots a scatter diagram for M and G , reproduced as Figure 2. Defining GSQ as the square of G , he regresses M on G and GSQ :

$$\hat{M} = 7.27 + 0.30 G - 0.000049 GSQ \quad R^2 = 0.86 \quad (2)$$

(10.77) (0.03) (0.000009)

Finally, he plots a scatter diagram for $\log M$ and $\log G$, reproduced as Figure 3, and regresses $\log M$ on $\log G$:

$$\log \hat{M} = -0.14 + 0.80 \log G \quad R^2 = 0.78 \quad (3)$$

(0.37) (0.07)

Having sorted the data by G , he tests for heteroscedasticity 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 summarized in the table.

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

Figure 1

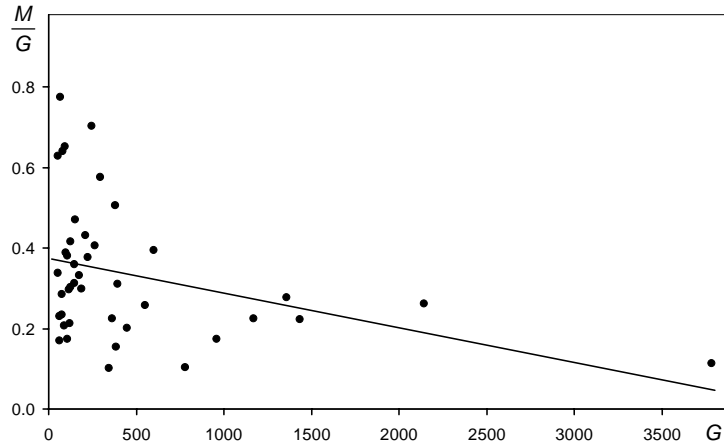


Figure 2

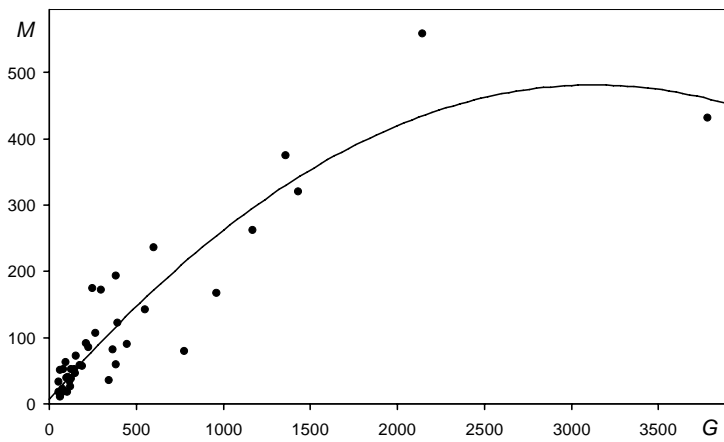
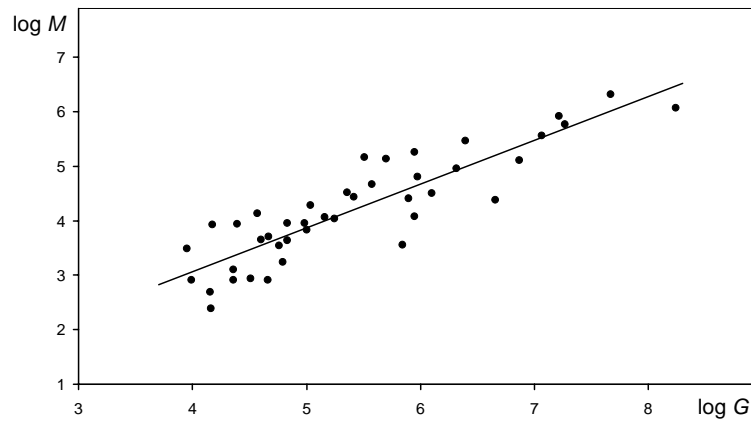


Figure 3



- Discuss whether (1) appears to be an acceptable specification, given the data in the table and Figure 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 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 3.
- What are your conclusions concerning the researcher's hypothesis?

A7.4 A researcher has data on the number of children attending, N , and annual recurrent expenditure, EXP , measured in US dollars, for 50 nursery schools in a US city for 2006 and hypothesizes 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:

$$\hat{EXP} = 17,999 + 1,060 N - 1.29 NSQ \quad R^2=0.74 \quad (1)$$

(12,908) (133) (0.30)

Suspecting that the regression was subject to heteroscedasticity, 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, EXP_N , expenditure per student, as $EXP_N = EXP/N$, and fits the equation

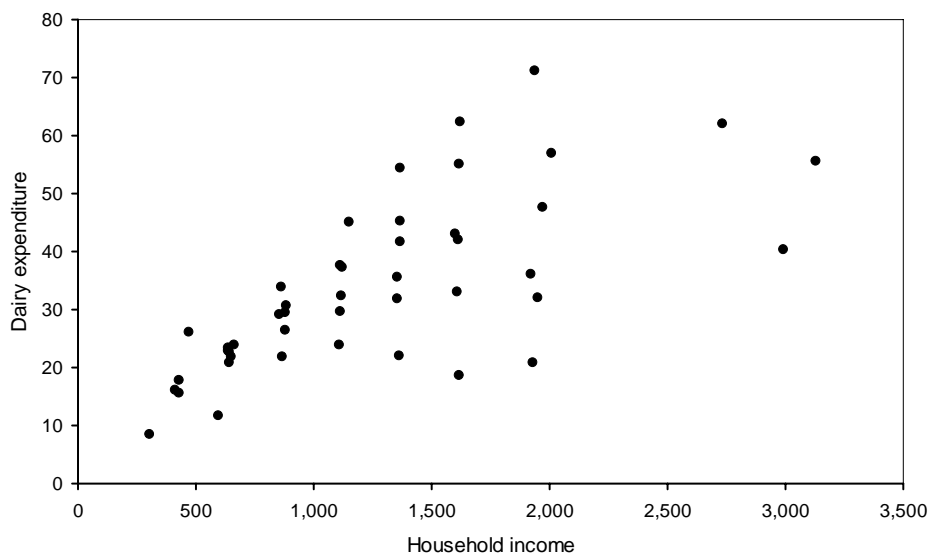
$$\hat{EXP}_N = 1,080 - 1.25 N + 16,114 NREC \quad R^2=0.65 \quad (2)$$

(90) (0.25) (6,000)

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.

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

A7.5 The figure displays a plot of monthly household expenditure on dairy produce, D , measured in dollars, on total monthly household expenditure, I , also measured in dollars, for 45 randomly selected households. Explain why an OLS regression of D on I is likely to be subject to heteroscedasticity and perform a Goldfeld–Quandt test for heteroscedasticity. Download the data set from the heteroscedastic data sets folder on the website and investigate whether you can re-specify the model to alleviate the heteroscedasticity. In the data set, N is the number of persons in the household.



A7.6 Explain what is correct, mistaken, confused or in need of further explanation in the following statements relating to heteroscedasticity in a regression model:

- (a) “Heteroscedasticity occurs when the disturbance term in a regression model is correlated with one of the explanatory variables.”
- (b) “In the presence of heteroscedasticity ordinary least squares (OLS) is an inefficient estimation technique and this causes t tests and F tests to be invalid.”
- (c) “OLS remains unbiased but it is inconsistent.”
- (d) “Heteroscedasticity can be detected with a Chow test.”
- (e) “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.”
- (f) “One way of eliminating the problem is to make use of a restriction involving the variable correlated with the disturbance term.”
- (g) “If you can find another variable related to the one responsible for the heteroscedasticity, you can use it as a proxy and this should eliminate the problem.”
- (h) “Sometimes apparent heteroscedasticity 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.”

Answers to the starred exercises in the text

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

$$\hat{COST} = 24,000 + 339N \quad R^2 = 0.39$$

(27,000) (50)

$$\hat{COST} = 51,000 - 4,000OCC + 152N + 284NOCC \quad R^2 = 0.68.$$

(31,000)(41,000) (60) (76)

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 sum 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 heteroscedasticity for the two models and, with reference to Figure 6.5, explain why the problem of heteroscedasticity 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 percent level, and so we reject the null hypothesis of homoscedasticity 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 percent level, and so we (just) do not reject the null hypothesis of homoscedasticity at that significance level.

Why is the problem of heteroscedasticity 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 heteroscedastic, irrespective of whether the disturbance term is truly heteroscedastic or not. The first regression takes no account of this and the Goldfeld–Quandt test therefore indicates significant heteroscedasticity. In the second regression the problem of apparent heteroscedasticity does not arise because the intercept and slope dummy variables allow separate implicit regression lines for the two types of school.

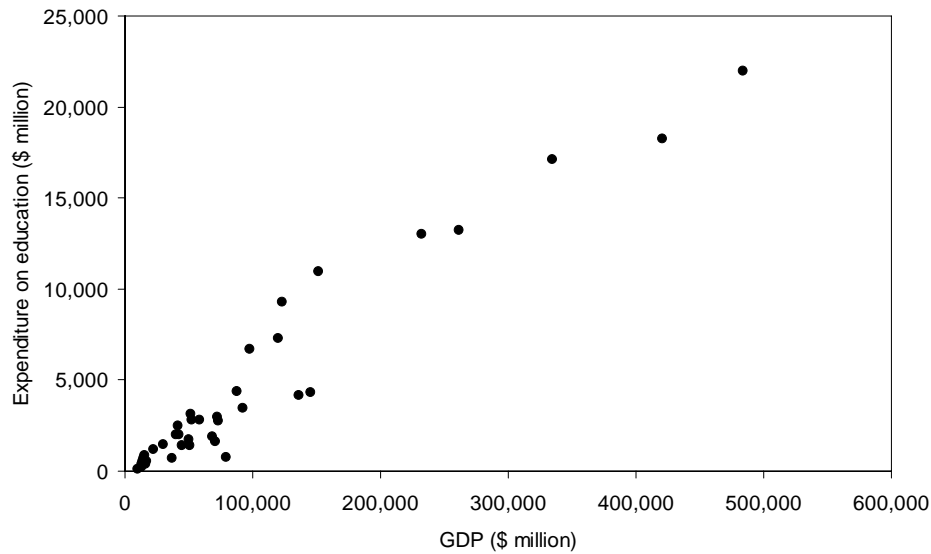
Looking closely at the diagram, the observations for the occupational schools exhibit a classic pattern of true heteroscedasticity, and this would be confirmed by a Goldfeld–Quandt test confined to the subsample of those schools (see Exercise A7.2). However the observations for the regular schools appear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

- 7.6 The file *educ.dta* in the heteroscedastic data sets folder on the website 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 heteroscedasticity. Sort the data set by *GDP* and perform a Goldfeld–Quandt test for heteroscedasticity, 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. The observations exhibit heteroscedasticity. Sorting them by *GDP* and regressing *EDUC* on *GDP* for the subsamples of 14 countries with smallest and greatest *GDP*, the residual sum 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 percent level is 7.00, and so we reject the null hypothesis of homoscedasticity.



Expenditure on education and GDP

- 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 heteroscedasticity. Compare the results of regressions using the entire sample and the alternative specifications.

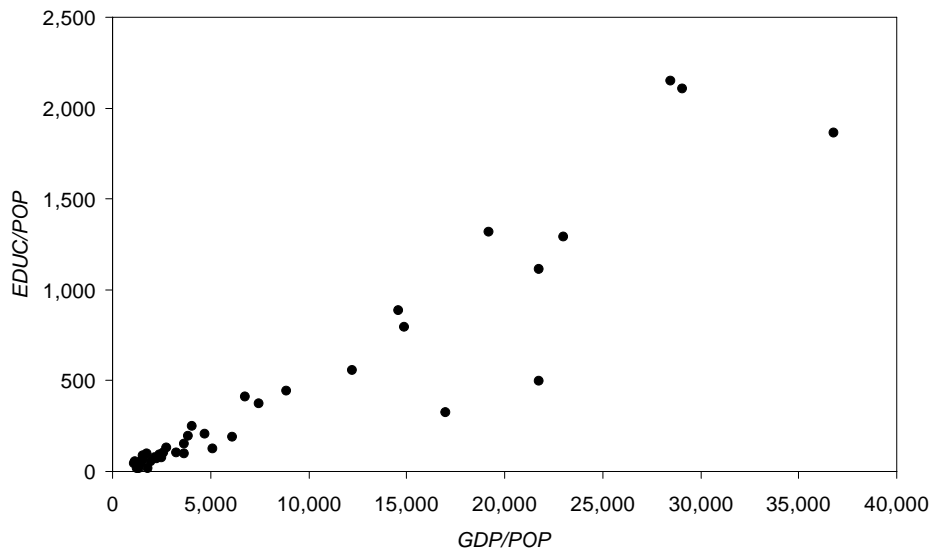
Answer: Dividing through by population, POP , the model becomes

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

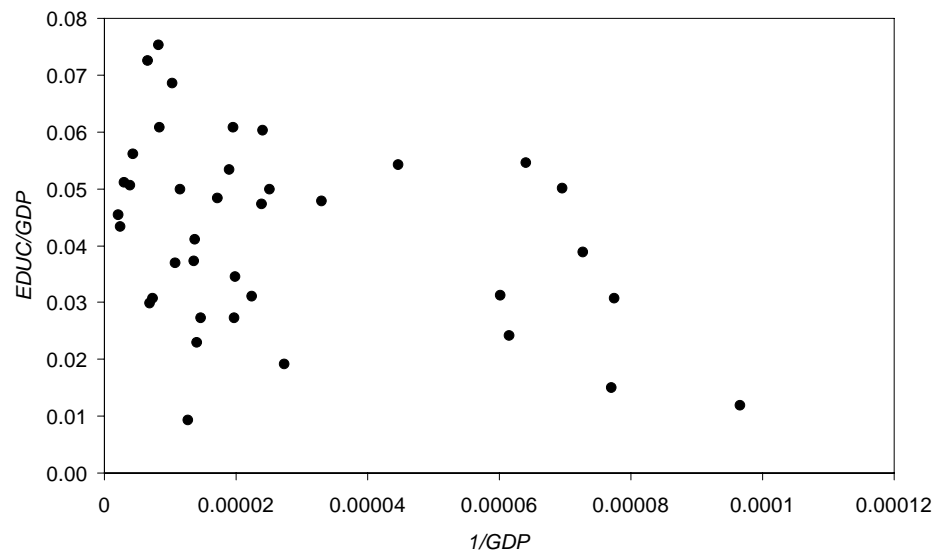
with expenditure on education per capita, denoted $EDUCPOP$, hypothesized to be a function of gross domestic product per capita, $GDPPOP$, and the reciprocal of population, $POPREC$, with no intercept. Sorting the sample by $GDPPOP$ and running the regression for the subsamples of 14 countries with smallest and largest $GDPPOP$, $RSS_1 = 56,541$ and $RSS_2 = 1,415,515$. Now

$$F(12,12) = \frac{RSS_2}{RSS_1} = \frac{1415515}{56541} = 25.04.$$

Thus the model is still subject to heteroscedasticity at the 0.1 percent level. This is evident in the figure.



Expenditure on education per capita and GDP per capita



Expenditure on education as a proportion of GDP and the reciprocal of GDP

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 $EDUCGDP$, hypothesized 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 heteroscedasticity under the hypothesis that the standard deviation of the disturbance term is inversely related to $GDPREC$:

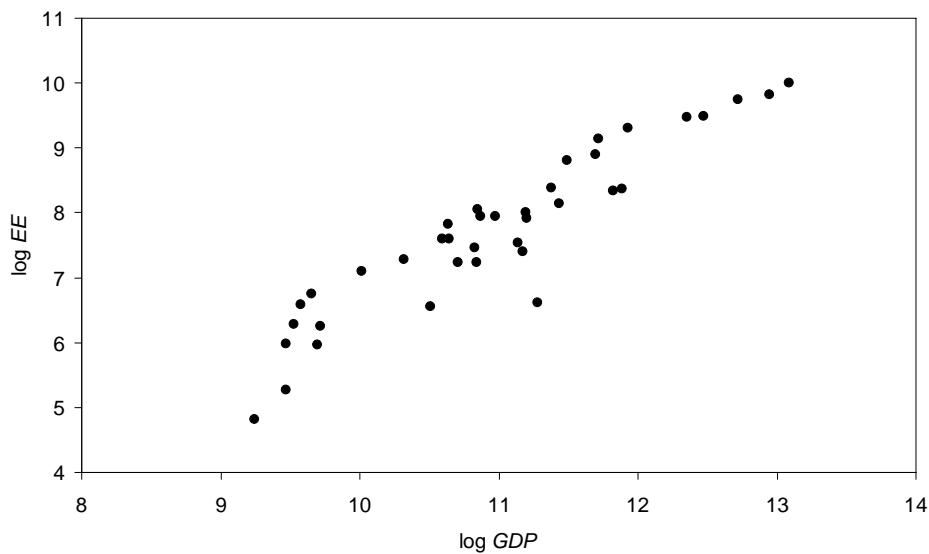
$$F(12,12) = \frac{RSS_1}{RSS_2} = \frac{0.00413}{0.00238} = 1.74$$

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

Finally, we will consider a logarithmic specification. If the true relationship is logarithmic, and homoscedastic, it would not be surprising that the linear model appeared heteroscedastic. 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 homoscedasticity.



Expenditure on education and GDP, logarithmic

The third and fourth models both appear to be free from heteroscedasticity. 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			
Model	.001348142	1	.001348142	Number of obs =	38	
Residual	.008643037	36	.000240084	F(1, 36) =	5.62	
Total	.009991179	37	.000270032	Prob > F =	0.0233	
				R-squared =	0.1349	
				Adj R-squared =	0.1109	
				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

Dougherty: Introduction to Econometrics 3e
Study Guide

```
. reg LGEE LGGDP
```

Source	SS	df	MS			
Model	51.9905508	1	51.9905508	Number of obs =	38	
Residual	7.6023197	36	.211175547	F(1, 36) =	246.20	
				Prob > F =	0.0000	
				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

$$\frac{\hat{EDUC}}{GDP} = 0.048 - 234.1 \frac{1}{GDP} \quad R^2 = 0.13$$

(0.004) (98.8)

Multiplying through by GDP , it may be rewritten

$$\hat{EDUC} = -234.1 + 0.048GDP$$

It implies that expenditure on education accounts for 4.8 percent 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 heteroscedasticity:

```
. reg EDUC GDP
```

Source	SS	df	MS			
Model	1.0571e+09	1	1.0571e+09	Number of obs =	38	
Residual	74645819.2	36	2073494.98	F(1, 36) =	509.80	
				Prob > F =	0.0000	
				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, 0.48, is identical to two decimal places. This is not entirely a surprise, since heteroscedasticity 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 heteroscedasticity. 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

$$\log \hat{EDUC} = -5.03 + 1.17 \log GDP \quad R^2 = 0.87$$

(0.82) (0.07)

implying that the elasticity of expenditure on education with regard to gross domestic product is 1.17. 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 heteroscedasticity-consistent estimate of the standard error of the coefficient of *GDP* in equation (7.13) was 0.18. Explain why the corresponding standard error in equation (7.15) ought to be lower and comment on the fact that it is not.

Answer: (7.15), unlike (7.13) appears to be free from heteroscedasticity 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 heteroscedasticity-consistent estimator to be a good guide.

Answers to the additional exercises

- A7.1 The first step is to sort the data set by *EXPPC*. Then, if there were no zero-expenditure observations, the subsample regressions should use approximately the first and last 326 observations, 326 being three-eighths of 869. This procedure has been adopted anyway, on the assumption that the zero observations are distributed randomly and that the first and last 326 observations capture about three-eighths of the available ones. 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. However this procedure does not work well for those categories with many zero observations because there is a tendency for the number of zero observations to be relatively great for low *EXPPC* (*LOCT* being an understandable exception). It would have been better to have saved the data set under a new name for this exercise, with the zero observations dropped, and to have identified the smallest and largest three-eighths properly. However it is doubtful that the outcome would have been much different.

```
. sort EXPPC
```

```
. reg FDHOPC EXPPC SIZE in 1/326 if FDHO>0
```

Source	SS	df	MS			
Model	20263081.5	2	10131540.8	Number of obs =	326	
Residual	65763256.7	323	203601.414	F(2, 323) =	49.76	
Total	86026338.3	325	264696.425	Prob > F =	0.0000	
				R-squared =	0.2355	
				Adj R-squared =	0.2308	
				Root MSE =	451.22	

FDHOPC	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
EXPPC	.0634414	.0142184	4.46	0.000	.035469	.0914138
SIZE	-103.9873	14.90612	-6.98	0.000	-133.3126	-74.66197
_cons	1104.826	114.3772	9.66	0.000	879.8079	1329.845

Dougherty: Introduction to Econometrics 3e
Study Guide

. reg FDHOPC EXPPC SIZE in 544/869 if FDHO>0

Source	SS	df	MS			
Model	27679828.3	2	13839914.1	Number of obs =	325	
Residual	221573281	322	688115.778	F(2, 322) =	20.11	
Total	249253109	324	769299.719	Prob > F =	0.0000	
				R-squared =	0.1111	
				Adj R-squared =	0.1055	
				Root MSE =	829.53	

FDHOPC	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
EXPPC	.0263627	.0057528	4.58	0.000	.0150448	.0376806
SIZE	-147.5727	46.11894	-3.20	0.002	-238.3052	-56.84019
_cons	1578.133	178.0577	8.86	0.000	1227.829	1928.436

The F statistic for the linear specification is

$$F(322,323) = \frac{221.57/322}{65.76/323} = 3.38$$

The corresponding F statistic for the logarithmic specification is 1.54. The critical value of $F(300,200)$ at the 0.1 percent level is 1.48. The critical value for $F(322,323)$ must be lower. Thus in both cases the null hypothesis of homoscedasticity is rejected, but the problem appears to much less severe for the logarithmic specification.

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

Goldfeld–Quandt tests								
	<i>linear</i>					<i>logarithmic</i>		
	n_1	n_2	$RSS_1 \times 10^{-6}$	$RSS_2 \times 10^{-6}$	F	RSS_1	RSS_2	F
<i>FDHO</i>	326	325	65.76	221.57	3.38	40.07	61.95	1.54
<i>FDAW</i>	292	324	5.76	280.94	43.96	240.91	219.69	1.01
<i>HOUS</i>	324	326	192.79	2097.6	10.81	260.60	146.59	1.77*
<i>TELE</i>	320	324	6.05	75.29	12.29	134.51	112.27	1.18*
<i>DOM</i>	136	189	11.74	491.32	30.11	357.60	536.39	2.08
<i>TEXT</i>	151	206	0.13	15.86	89.43	163.28	284.78	2.38
<i>FURN</i>	86	155	7.64	69.07	5.02	175.07	301.58	3.10
<i>MAPP</i>	70	97	0.93	16.60	12.88	79.55	104.63	1.82
<i>SAPP</i>	141	203	0.30	1.09	2.52	172.05	190.50	1.59
<i>CLOT</i>	308	325	12.11	179.26	14.03	299.14	223.20	1.27*
<i>FOOT</i>	246	273	0.28	2.40	7.72	235.30	210.13	1.01*
<i>GASO</i>	283	311	12.20	59.98	4.47	163.61	110.68	1.35*
<i>TRIP</i>	59	173	0.90	122.07	46.26	125.87	250.34	5.83
<i>LOCT</i>	82	52	2.09	2.39	1.80	199.72	126.57	2.49*
<i>HEAL</i>	293	318	68.92	375.78	5.02	536.75	428.11	1.16*
<i>ENT</i>	298	323	15.48	861.87	51.37	298.52	251.60	1.09*
<i>FEES</i>	216	289	1.00	296.56	221.65	310.61	502.18	2.16
<i>TOYS</i>	206	237	3.49	20.25	5.04	298.88	303.10	1.17
<i>READ</i>	255	313	0.37	4.15	9.14	292.09	340.67	1.43
<i>EDUC</i>	107	106	2.98	300.44	101.77	233.77	337.45	1.43
<i>TOB</i>	146	125	4.38	9.09	2.42	148.74	122.19	1.42*

* indicates $RSS_2 < RSS_1$

A7.2 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 percent level must be a little below 8.75, the critical value for $F(10,10)$, and so the null hypothesis of homoscedasticity is rejected at that significance level.

One possible way of alleviating the heteroscedasticity 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 percent 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			
Model	27010.3792	1	27010.3792	Number of obs =	34	
Residual	1164624.44	32	36394.5138	F(1, 32) =	0.74	
Total	1191634.82	33	36110.1461	Prob > F =	0.3954	
				R-squared =	0.0227	
				Adj R-squared =	-0.0079	
				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

$$\frac{\hat{COST}}{N} = 524.8 + 10976 \frac{1}{N} \quad R^2 = 0.03$$

(53.9) (12741)

Multiplying through by N , it may be rewritten

$$\hat{EDUC} = 10976 + 524.8N.$$

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 heteroscedasticity is to hypothesize that the true relationship is logarithmic, in which case the use of an inappropriate linear specification would give rise to apparent heteroscedasticity. 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 percent level. The regression output for this specification using the full sample is shown.

. reg LGCOST LGN

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

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

- A7.3 (a) Using the Goldfeld–Quandt test to test specification (1) for heteroscedasticity 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 percent level is 2.48, so we just reject the null hypothesis of homoscedasticity at that level. Figure 1 does strongly suggest heteroscedasticity. Thus (1) does not appear to be an acceptable specification.
- (b) If it is true that the standard deviation of u is inversely proportional to G , the heteroscedasticity 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.
- (c) $F(13,13) = \frac{71404}{3178} = 22.47$. The critical value of $F(13,13)$ at the 0.1 percent level is about 6.4, so the null hypothesis of homoscedasticity is rejected. Figure 2 confirms the heteroscedasticity.
- (d) Heteroscedasticity 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 heteroscedasticity in the second specification.
- (e) Obviously there is no problem with the Goldfeld–Quandt test, since $F(14,14) = \frac{3.60}{3.45} = 1.04$. Figure 3 looks free from heteroscedasticity.
- (f) 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 1 and 2 also strongly suggest that on balance larger economies have lower import ratios than smaller ones.

A7.4 (a) The F statistics for the G-Q test for the two specifications are $F(16,16) = \frac{64/16}{8/16} = 8.0$ and

$F(16,16) = \frac{900/16}{600/16} = 1.5$. The critical value of $F(16,16)$ is 2.33 at the 5 percent level and 5.20 at the 0.1 percent level. Hence one would reject the null hypothesis of homoscedasticity at the 0.1 percent level for regression 1 and one would not reject it even at the 5 percent level for regression 2.

(b) He hypothesized 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 homoscedastic, since

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

and is therefore the same for all i .

(c) 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 heteroscedasticity and therefore ought to tend to yield more precise estimates of the coefficients with valid standard errors.

A7.5 Sorting the observations by size of household income and regressing expenditure on dairy products on income, RSS_1 and RSS_2 , the residual sums of squares for regressions using the 17 households with smallest and largest income, are 267.5 and 3316.2, respectively. The ratio is 12.40. The critical value of $F(14,15)$ at the 0.1 percent level is 5.62. The critical value for $F(15,15)$ must be lower, and hence we reject the null hypothesis of homoscedasticity at that significance level.

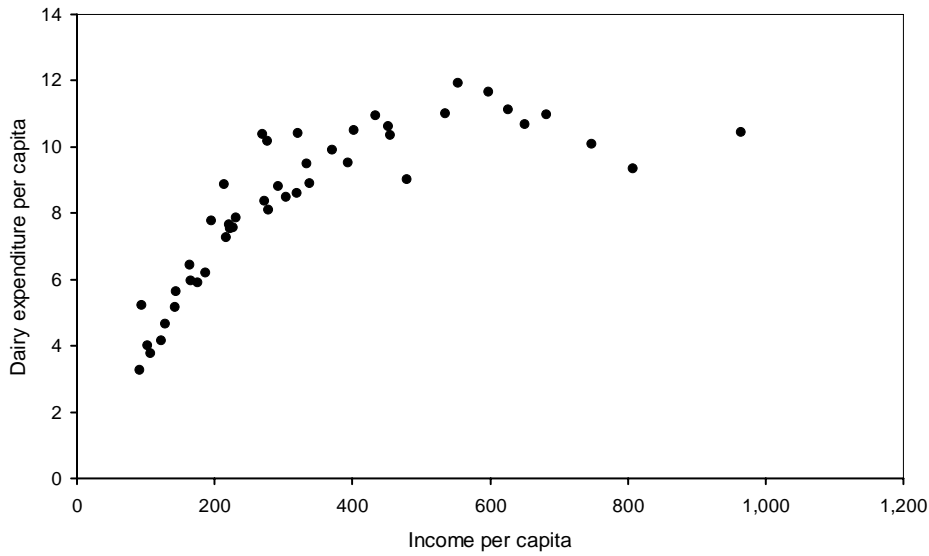
One possible way to alleviate the problem of heteroscedasticity might be to scale by the number of persons in the household. If the original model was

$$D = \beta_1 + \beta_2 I + u,$$

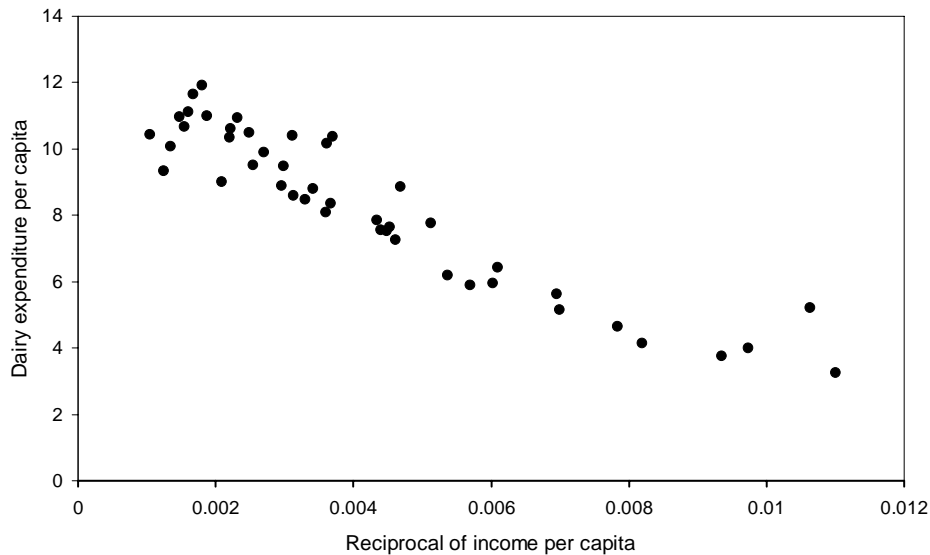
the scaled model is

$$\frac{D}{N} = \beta_1 \frac{1}{N} + \beta_2 \frac{I}{N} + \frac{u}{N}.$$

Sorting the observations by $\frac{I}{N}$, RSS_1 and RSS_2 are 5.92 and 21.34, respectively. The ratio is 3.60. The critical value of $F(14,15)$ at the 1 percent level is 3.56. The critical value of $F(15,15)$ must be a little lower, so the null hypothesis of homoscedasticity is still rejected at that significance level. The scatter diagram reveals a further problem. The relationship is clearly nonlinear.



Dairy expenditure per capita and household income per capita



Dairy expenditure per capita and the reciprocal of household income per capita

In this case taking logarithms does not linearize the relationship (try it and see), but the satiation model used in the banana consumption model in the text does. The model is now

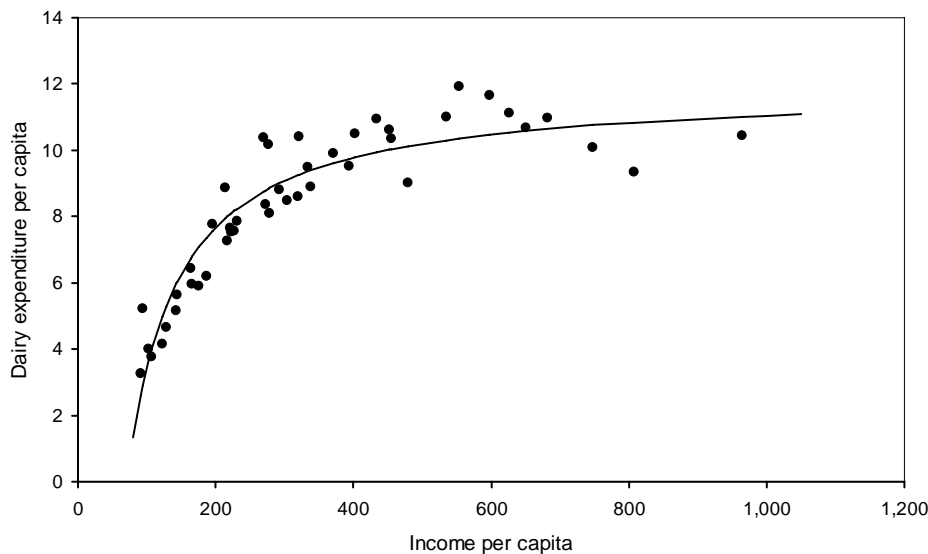
$$\frac{D}{N} = \beta_1 + \beta_2 \frac{1}{I/N} + \frac{u}{N} .$$

Sorting the observations by the reciprocal of income per capita, RSS_1 and RSS_2 are now 10.60 and 9.21, so the null hypothesis of homoscedasticity is not rejected. The regression output for the whole sample is shown (DC is dairy expenditure per capita and $ICREC$ is the reciprocal of income per capita):

```
. reg DC ICREC
```

Source	SS	df	MS			
Model	212.862218	1	212.862218	Number of obs =	45	
Residual	32.5950415	43	.758024221	F(1, 43) =	280.81	
				Prob > F =	0.0000	
				R-squared =	0.8672	
				Adj R-squared =	0.8641	
Total	245.457259	44	5.57857407	Root MSE =	.87065	

DC	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
ICREC	-843.8374	50.35598	-16.757	0.000	-945.3899	-742.2849
_cons	11.88223	.2486256	47.792	0.000	11.38082	12.38363



Dairy expenditure per capita and household income per capita

The figure shows the implicit relationship between expenditure per capita and income per capita.

- 7.6 (a) This is false. Heteroscedasticity occurs when the variance of the disturbance term is not the same for all observations.
- (b) It is true that OLS is inefficient and that the t and F tests are invalid, but ‘and this causes’ is wrong.
- (c) It is true that OLS is unbiased, but false that it is inconsistent.
- (d) This is false.
- (e) 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 higher RSS over the smaller.
The second sentence is false.
- (f) This is nonsense.
- (g) This is more nonsense.

- (h) True. A homoscedastic disturbance term in a logarithmic regression, which is responsible for proportional changes in the dependent variable, may appear to be heteroscedastic in a linear regression because the absolute changes in the dependent variable will be proportional to its size.