

Chapter 6

Specification of regression variables: a preliminary skirmish

Overview

This chapter treats a variety of topics relating to the specification of the variables in a regression model. First there are the consequences for the regression coefficients, their standard errors, and R^2 of failing to include a relevant variable, and of including an irrelevant one. This leads to a discussion of the use of proxy variables to alleviate a problem of omitted variable bias. Next come F and t tests of the validity of a restriction, the use of which was advocated in Chapter 3 as a means of improving efficiency and perhaps mitigating a problem of multicollinearity. The chapter concludes by outlining the potential benefit to be derived from examining observations with large residuals after fitting a regression model.

Further material

This section contains some new material on the following topics:

1. The reparameterization of a regression model
2. The application of reparameterization to t tests of linear restrictions
3. The testing of multiple restrictions
4. Tests of zero restrictions

The reparameterization of a regression model

Suppose that you have fitted the regression model

$$Y = \beta_1 + \sum_{j=2}^k \beta_j X_j + u \quad (1)$$

and that the regression model assumptions are valid. Let the fitted model be

$$\hat{Y} = b_1 + \sum_{j=2}^k b_j X_j \quad (2)$$

as usual. Suppose that, as well as the individual parameter estimates, you are interested in some linear combination:

$$\theta = \sum_{j=1}^k \lambda_j \beta_j \quad (3)$$

To obtain a point estimate of θ , it is natural to construct the statistic $\hat{\theta} = \sum_{j=1}^k \lambda_j b_j$, and indeed, given that the

regression model assumptions are valid, it can easily be shown that this is unbiased and the most efficient estimator of θ . However you do not have information on its standard error and hence you are not able to construct confidence intervals or to perform t tests. There are three ways that you might use to obtain such information:

- (1) Some regression applications have a special command that produces it. For example, Stata has the `lincom` command.
- (2) Given the appropriate command, most regression applications will produce the variance-covariance matrix for the estimates of the parameters. This is the complete list of the estimates of their variances and

covariances, for convenience arranged in matrix form. The standard errors in the ordinary regression output are the square roots of the variances. The estimate of the variance of the estimate of θ is given by

$$s_{\theta}^2 = \sum_{j=1}^k \lambda_j^2 s_{b_j}^2 + 2 \sum_{p \neq j} \lambda_p \lambda_j s_{b_p b_j} \quad (4)$$

where $s_{b_p b_j}$ is the estimate of the covariance between b_p and b_j .

- (3) The third method is to reparameterize the model, manipulating it so that θ and its standard error are estimated directly as part of the regression output. To do this, we rewrite (3) so that one of the b parameters is expressed in terms of θ and the other b parameters. This will be illustrated with two simple examples, the general case being left as an additional exercise. Suppose the regression model is

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u \quad (5)$$

and suppose we are interested in the sum of β_2 and β_3 :

$$\theta = \beta_2 + \beta_3 \quad (6)$$

We rewrite this as

$$\beta_3 = \theta - \beta_2 \quad (7)$$

It makes no substantive difference which β we take to the left side. Substituting in the original model, we have

$$\begin{aligned} Y &= \beta_1 + \beta_2 X_2 + (\theta - \beta_2) X_3 + u \\ &= \beta_1 + \beta_2 (X_2 - X_3) + \theta X_3 + u \end{aligned} \quad (8)$$

This if we define a new variable $Z = X_2 - X_3$ and regress Y on Z and X_3 , the coefficient of Z will be an estimate of β_2 and that of X_3 will be an estimate of θ . The estimate of θ will be exactly the same as that obtained by summing the estimates of β_2 and β_3 in (5). The difference is that we obtain its standard error directly from the regression results. The estimate of β_2 and its standard error will be the same as that using (5).

Suppose instead that we were interested in the difference in β_2 and β_3 :

$$\theta = \beta_2 - \beta_3 \quad (9)$$

We rewrite this as

$$\beta_3 = \beta_2 + \theta \quad (10)$$

Substituting in the original model, we have

$$\begin{aligned} Y &= \beta_1 + \beta_2 X_2 + (\beta_2 + \theta) X_3 + u \\ &= \beta_1 + \beta_2 (X_3 + X_2) + \theta X_3 + u \end{aligned} \quad (11)$$

This if we define a new variable $Z = X_3 + X_2$ and regress Y on Z and X_3 , the coefficient of Z will be an estimate of β_2 and that of X_3 will be an estimate of θ . The estimate of θ will be exactly the same as that obtained by taking the difference of the estimates of β_2 and β_3 in (5) and we obtain its standard error directly from the regression results. The estimate of β_2 and its standard error will be the same as that using (5).

t tests of linear restrictions

An obvious application of reparameterization is its use in the testing of linear restrictions. Suppose that your hypothetical restriction is

$$\sum_{j=1}^k \lambda_j \beta_j = \alpha \quad (12)$$

where α is a scalar. Define

$$\theta = \sum_{j=1}^k \lambda_j \beta_j - \alpha \quad (13)$$

and reparameterize. θ will become the coefficient of one of the variables in the model, and a t test of $H_0: \theta = 0$ is effectively a t test of $H_0: \sum_{j=1}^k \lambda_j \beta_j = \alpha$, and hence of the restriction.

As an illustration, we will use the example discussed in Sections 3.4 and 6.4 of the text. The model relates years of schooling, S , to the cognitive ability score $ASVABC$ and years of schooling of the mother and the father, SM and SF :

$$S = \beta_1 + \beta_2 ASVABC + \beta_3 SM + \beta_4 SF + u \quad (14)$$

It was hypothesized that mother's education and father's education are equally important for educational attainment, implying the restriction $\beta_4 = \beta_3$, or

$$\beta_4 - \beta_3 = 0 \quad (15)$$

Define

$$\theta = \beta_4 - \beta_3 \quad (16)$$

and rewrite this equation as

$$\beta_4 = \beta_3 + \theta \quad (17)$$

Substitute into the original model:

$$\begin{aligned} S &= \beta_1 + \beta_2 ASVABC + \beta_3 SM + (\beta_3 + \theta)SF + u \\ &= \beta_1 + \beta_2 ASVABC + \beta_3 (SM + SF) + \theta SF + u \\ &= \beta_1 + \beta_2 ASVABC + \beta_3 SP + \theta SF + u \end{aligned} \quad (18)$$

where $SP = SM + SF$, and test the coefficient of SF , which is an estimate of θ . The output in Section 6.4 was

```
. reg S ASVABC SP SF
```

Source	SS	df	MS	Number of obs = 540		
Model	1181.36981	3	393.789935	F(3, 536) =	104.30	
Residual	2023.61353	536	3.77539837	Prob > F =	0.0000	
-----				R-squared =	0.3686	
-----				Adj R-squared =	0.3651	
Total	3204.98333	539	5.94616574	Root MSE =	1.943	

S	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
ASVABC	.1257087	.0098533	12.76	0.000	.1063528	.1450646
SP	.0492424	.0390901	1.26	0.208	-.027546	.1260309
SF	.0584401	.0617051	0.95	0.344	-.0627734	.1796536
_cons	5.370631	.4882155	11.00	0.000	4.41158	6.329681

The coefficient of SF is not significantly different from zero and so we do not reject the restriction.

Multiple restrictions

Multiple restrictions can be tested by multiple reparameterizations. Each one will result in one of the original parameters being dropped and replaced by a test statistic for the restriction. For example, if we had the model

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + u \quad (9)$$

and we hypothesized the restrictions

$$\beta_3 = \beta_2, \beta_4 + \beta_5 = 0 \quad (20)$$

we could define test statistics

$$\theta = \beta_3 - \beta_2, \phi = \beta_4 + \beta_5 \quad (21)$$

and substitute

$$\beta_3 = \beta_2 + \theta, \beta_5 = \phi - \beta_4 \quad (22)$$

Substituting into the original model, we have

$$\begin{aligned} Y &= \beta_1 + \beta_2 X_2 + (\beta_2 + \theta)X_3 + \beta_4 X_4 + (\phi - \beta_4)X_5 + u \\ &= \beta_1 + \beta_2(X_2 + X_3) + \beta_4(X_4 - X_5) + \theta X_3 + \phi X_5 + u \\ &= \beta_1 + \beta_2 Z + \beta_4 W + \theta X_3 + \phi X_5 + u \end{aligned} \quad (23)$$

where $Z = X_2 + X_3$ and $W = X_4 - X_5$. We would then perform t tests on the coefficients of X_3 and X_5 . We could also perform a joint test of the restrictions, hypothesizing $H_0: \theta = \phi = 0$. This would involve comparing the residual sum of squares with that when fitting the fully restricted model

$$Y = \beta_1 + \beta_2 Z + \beta_4 W + u \quad (24)$$

The test statistic would be

$$F(2, n - k) = \frac{(RSS_R - RSS_U) / 2}{RSS_U / (n - k)} \quad (25)$$

where RSS_U is the residual sum of squares in the original, unrestricted model, RSS_R is the residual sum of squares in the model with both restrictions, and k is the number of parameters in the original, unrestricted version. In general, if there were p restrictions being tested simultaneously, the test statistic would be

$$F(p, n - k) = \frac{(RSS_R - RSS_U) / p}{RSS_U / (n - k)} \quad (26)$$

Zero restrictions

You will often encounter references to zero restrictions. This just means that a particular parameter is hypothesized to be equal to zero. Taken in isolation, the appropriate test is of course the t test. It can be considered to be a special case of the t test of a restriction discussed above where there is no need for reparameterization. The test statistic is the parameter itself. Likewise the testing of multiple zero restrictions can be thought of as a special case of the testing of multiple restrictions discussed in the previous section, again with no need for reparameterization. The test of the joint explanatory power of a group of explanatory variables discussed in Section 3.5 in the text can be thought of in this way. Even the F statistic for the equation as a whole can be treated as a special case. Here the unrestricted model is

$$Y = \beta_1 + \sum_{j=2}^k \beta_j X_j + u \quad (27)$$

The restricted model is

$$Y = \beta_1 + u \quad (28)$$

since all the slope coefficients are hypothesized to be zero. If this model is fitted, the OLS estimate of β_1 is \bar{Y} , the residual in observation i is $Y_i - \bar{Y}$, and $RSS_R = \sum_{i=1}^n (Y_i - \bar{Y})^2$ which is the total sum of squares, TSS , for Y .

The F statistic is therefore

$$F(k-1, n-k) = \frac{(RSS_R - RSS_U)/(k-1)}{RSS_U/(n-k)} = \frac{(TSS - RSS_U)/(k-1)}{RSS_U/(n-k)} = \frac{ESS_U/(k-1)}{RSS_U/(n-k)} \quad (29)$$

where RSS_U and ESS_U are the residual sum of squares and the explained sum of squares for the original, unrestricted model. This is the definition of F for the equation as a whole given in equation (3.65).

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:

- derive the expression for the bias in an OLS estimator of a slope coefficient when the true model has two explanatory variables but the regression model has only one
- determine the likely direction of omitted variable bias, given data on the correlation between the explanatory variables
- explain the consequence of omitted variable bias for the standard errors of the coefficients and for t tests and F tests
- explain the consequences of including an irrelevant variable for the regression coefficients, their standard errors, and t and F tests
- explain how the regression results are affected by the substitution of a proxy variable for a missing explanatory variable
- perform an F test of a restriction, stating the null hypothesis for the test
- perform a t test of a restriction, stating the null hypothesis for the test

Additional exercises

A6.1 A researcher obtains data on household annual expenditure on books, B , and annual household income, Y , for 100 households. He hypothesizes that B is related to Y and the average cognitive ability of adults in the household, IQ , by the relationship

$$\log B = \beta_1 + \beta_2 \log Y + \beta_3 \log IQ + u \quad (A)$$

where u is a disturbance term that satisfies the regression model assumptions. He also considers the possibility that $\log B$ may be determined by $\log Y$ alone:

$$\log B = \beta_1 + \beta_2 \log Y + u \quad (B)$$

He does not have data on IQ and decides to use average years of schooling of the adults in the household, S , as a proxy in specification (A). It may be assumed that Y and S are both nonstochastic. In the sample the correlation between $\log Y$ and $\log S$ is 0.86. He performs the following regressions: (1) $\log B$ on both $\log Y$ and $\log S$, and (2) $\log B$ on $\log Y$ only, with the results shown in the table (standard errors in parentheses):

	(1)	(2)
log Y	1.10 (0.69)	2.10 (0.35)
log S	0.59 (0.35)	–
constant	–6.89 (2.28)	–3.37 (0.89)
R^2	0.29	0.27

- (a) Assuming that (A) is the correct specification, explain, with a mathematical proof, whether you would expect the coefficient of log Y to be greater in regression (2).
- (b) Assuming that (A) is the correct specification, describe the various benefits from using log S as a proxy for log IQ , as in regression (1), if log S is a good proxy.
- (c) Explain whether the low value of R^2 in regression (1) implies that log S is not a good proxy.
- (d) Assuming that (A) is the correct specification, provide an explanation of why the coefficients of log Y and log S in regression (1) are not significantly different from zero, using two-sided t tests.
- (e) Discuss whether the researcher would be justified in using one-sided t tests in regression (1).
- (f) Assuming that (B) is the correct specification, explain whether you would expect the coefficient of log Y to be lower in regression (1).
- (g) Assuming that (B) is the correct specification, explain whether the standard errors in regression (1) are valid estimates.

A6.2 *Does the omission of total household expenditure or household size give rise to omitted variable bias in your CES regressions?*

Regress $LGATPC$ (1) on both $LGEXPPC$ and $LGSIZE$, (2) on $LGEXPPC$ only, and (3) on $LGSIZE$ only. Assuming that (1) is the correct specification, analyze the likely direction of the bias in the estimate of the coefficient of $LGEXPPC$ in (2) and that of $LGSIZE$ in (3). Check whether the regression results are consistent with your analysis.

A6.3 A researcher has data on years of schooling, S , weekly earnings in dollars, W , hours worked per week, H , and hourly earnings, E (computed as W/H) for a sample of 1755 white males in the United States in the year 2000. She calculates LW , LE , and LH as the natural logarithms of W , E , and H , respectively, and fits the following regressions, with the results shown in the table below (standard errors in parentheses; RSS = residual sum of squares):

- Column 1: a regression of LE on S .
- Column 2: a regression of LW on S and LH , and
- Column 3: a regression of LE on S and LH .

The correlation between S and LH is 0.06.

	(1)	(2)	(3)	(4)	(5)
Respondents	All	All	All	FT	PT
Dependent variable	<i>LE</i>	<i>LW</i>	<i>LE</i>	<i>LW</i>	<i>LW</i>
<i>S</i>	0.099 (0.006)	0.098 (0.006)	0.098 (0.006)	0.101 (0.006)	0.030 (0.049)
<i>LH</i>	–	1.190 (0.065)	0.190 (0.065)	0.980 (0.088)	0.885 (0.325)
constant	6.111 (0.082)	5.403 (0.254)	5.403 (0.254)	6.177 (0.345)	7.002 (1.093)
<i>RSS</i>	741.5	737.9	737.9	626.1	100.1
observations	1755	1755	1755	1669	86

- (a) Explain why specification (1) is a restricted version of specification (2), stating and interpreting the restriction.
- (b) Supposing the restriction to be valid, explain whether you expect the coefficient of *S* and its standard error to differ, or be similar, in specifications (1) and (2).
- (c) Supposing the restriction to be invalid, how would you expect the coefficient of *S* and its standard error to differ, or be similar, in specifications (1) and (2)?
- (d) Perform an *F* test of the restriction.
- (e) Perform a *t* test of the restriction.
- (f) Explain whether the *F* test and the *t* test could lead to different conclusions.
- (g) At a seminar, a commentator says that part-time workers tend to be paid worse than full-time workers and that their earnings functions are different. Defining full-time workers as those working at least 35 hours per week, the researcher divides the sample and fits the earnings functions for full-time workers (column 4) and part-time workers (column 5). Test whether the commentator's assertion is correct.
- (h) What are the implications of the commentator's assertion for the test of the restriction?

A6.4 A researcher has data on *HEALTH*, aggregate expenditure on health, *GDP*, aggregate gross domestic product, and *POP*, total population, for a sample of 70 countries in 1999. *HEALTH* and *GDP* are both measured in US\$ billion. *POP* is measured in million. Hypothesizing that expenditure on health per capita depends on GDP per capita, he fits the regression (standard errors in parentheses; *RSS* = residual sum of squares):

$$\log \frac{\widehat{HEALTH}}{POP} = -3.74 + 1.27 \log \frac{GDP}{POP} \quad R^2 = 0.91 \quad RSS = 14.26 \quad (1)$$

(0.10) (0.05)

He also runs the following regressions:

$$\log \widehat{HEALTH} = -3.60 + 1.27 \log GDP - 0.33 \log POP \quad R^2 = 0.95 \quad RSS = 13.90 \quad (2)$$

(0.14) (0.05) (0.07)

$$\log \frac{\widehat{HEALTH}}{POP} = -3.60 + 1.27 \log \frac{GDP}{POP} - 0.06 \log POP \quad R^2 = 0.91 \quad RSS = 13.90 \quad (3)$$

(0.14) (0.05) (0.04)

- (a) Give an interpretation of the slope coefficient in equation (1).
- (b) If a country had GDP per capita of \$1,000 in 2001, with expenditure on health per capita \$40, what is the predicted level of expenditure on health in 2002 if GDP per capita rises to \$1,010?
- (c) Give an interpretation of the coefficients of $\log GDP$ and $\log POP$ in equation (2) and comment on their plausibility.
- (d) Demonstrate that specification (1) is a restricted version of specification (2), stating the restriction.
- (e) Test the restriction, using an F test.
- (f) Demonstrate that the same restriction may be tested using a t test on the coefficient of $\log POP$ in specification (3).
- (g) In the comparison of specifications (1) and (2), the restricted version has a worse fit. Does this matter?

A6.5 *Is expenditure per capita on your CES category related to total household expenditure per capita?*

The model specified in Exercise A4.2 is a restricted version of that in Exercise A4.1. Perform an F test of the restriction. Also perform a t test of the restriction.

[A4.6: regress $LGCATPC$ on $LGEXPPC$; A4.5: regress $LGCAT$ on $LGEXP$ and $LGSIZE$.]

A6.6 Three researchers investigating the determinants of hourly earnings have the following data for a sample of 104 male workers in the United States in 2006: E , hourly earnings in dollars; S , years of schooling; NUM , score on a test of numeracy; and $VERB$, score on a test of literacy. The NUM and $VERB$ tests are marked out of 100. The correlation between them is 0.81. Defining LGE to be the natural logarithm of E , Researcher 1 fits the following regression (standard errors in parentheses; RSS = residual sum of squares):

$$\hat{LGE} = 2.02 + 0.063 S + 0.0044 NUM + 0.0026 VERB \quad RSS = 2,000$$

(1.81) (0.007) (0.0011) (0.0010)

Researcher 2 defines a new variable $SCORE$ as the average of NUM and $VERB$. She fits the regression

$$\hat{LGE} = 1.72 + 0.050 S + 0.0068 SCORE \quad RSS = 2,045$$

(1.78) (0.005) (0.0010)

Researcher 3 fits the regression

$$\hat{LGE} = 2.02 + 0.063 S + 0.0088 SCORE - 0.0018 VERB \quad RSS = 2,000$$

(1.81) (0.007) (0.0022) (0.0012)

- (a) Show that the specification of Researcher 2 is a restricted version of the specification of Researcher 1, stating the restriction.
- (b) Perform an F test of the restriction.
- (c) Show that the specification of Researcher 3 is a reparameterized version of the specification of Researcher 1 and hence perform a t test of the restriction in the specification of Researcher 2.
- (d) Explain whether the F test in (b) and the t test in (c) could have led to different results.
- (e) Perform a test of the hypothesis that the numeracy score has a greater effect on earnings than the literacy score.
- (f) Compare the regression results of the three researchers.

Answers to the starred exercises in the text

- 6.4 The table gives the results of multiple and simple regressions of $LGFDHO$, the logarithm of annual household expenditure on food eaten at home, on $LGEXP$, the logarithm of total annual household expenditure, and $LGSIZE$, the logarithm of the number of persons in the household, using a sample of 868 households in the 1995 Consumer Expenditure Survey. The correlation coefficient for $LGEXP$ and $LGSIZE$ was 0.45. Explain the variations in the regression coefficients.

	(1)	(2)	(3)
$LGEXP$	0.29 (0.02)	0.48 (0.02)	–
$LGSIZE$	0.49 (0.03)	–	0.63 (0.02)
constant	4.72 (0.22)	3.17 (0.24)	7.50 (0.02)
R^2	0.52	0.31	0.42

Answer: If the model is written as

$$LGFDHO = \beta_1 + \beta_2 LGEXP + \beta_3 LGSIZE + u,$$

the expected value of b_2 in the second regression is given by

$$E(b_2) = \beta_2 + \beta_3 \frac{\sum (LGEXP_i - \overline{LGEXP})(LGSIZE_i - \overline{LGSIZE})}{\sum (LGEXP_i - \overline{LGEXP})^2}$$

We know that the covariance is positive because the correlation is positive, and it is reasonable to suppose that β_3 is also positive, especially given the highly significant positive estimate in the first regression, and so b_2 is biased upwards. This accounts for the large increase in its size in the second regression. In the third regression,

$$E(b_3) = \beta_3 + \beta_2 \frac{\sum (LGEXP_i - \overline{LGEXP})(LGSIZE_i - \overline{LGSIZE})}{\sum (LGSIZE_i - \overline{LGSIZE})^2}$$

β_2 is certainly positive, especially given the highly significant positive estimate in the first regression, and so b_3 is also biased upwards. As a consequence, the estimate in the third regression is greater than that in the first.

- 6.7 A social scientist thinks that the level of activity in the shadow economy, Y , depends either positively on the level of the tax burden, X , or negatively on the level of government expenditure to discourage shadow economy activity, Z . Y might also depend on both X and Z . International cross-section data on Y , X , and Z , all measured in US \$ million, are obtained for a sample of 30 industrialized countries and a second sample of 30 developing countries. The social scientist regresses (1) $\log Y$ on both $\log X$ and $\log Z$, (2) $\log Y$ on $\log X$ alone, and (3) $\log Y$ on $\log Z$ alone, for each sample, with the following results (standard errors in parentheses):

	<i>Industrialized Countries</i>			<i>Developing Countries</i>		
	(1)	(2)	(3)	(1)	(2)	(3)
log X	0.699 (0.154)	0.201 (0.112)	–	0.806 (0.137)	0.727 (0.090)	–
log Z	–0.646 (0.162)	–	–0.053 (0.124)	–0.091 (0.117)	–	0.427 (0.116)
constant	–1.137 (0.863)	–1.065 (1.069)	1.230 (0.896)	–1.122 (0.873)	–1.024 (0.858)	2.824 (0.835)
R ²	0.44	0.10	0.01	0.71	0.70	0.33

X was positively correlated with Z in both samples. Having carried out the appropriate statistical tests, write a short report advising the social scientist how to interpret these results.

Answer: One way to organize an answer to this exercise is, for each sample, to consider the evidence for and against each of the three specifications in turn. The *t* statistics for the slope coefficients are given in the following table. * indicates significance at the 5 percent level, ** at the 1 percent level, and *** at the 0.1 percent level, using one-sided tests. (Justification for one-sided tests: one may rule out a negative coefficient for X and a positive one for Y.)

	<i>Industrialized Countries</i>			<i>Developing Countries</i>		
	(1)	(2)	(3)	(1)	(2)	(3)
log X	4.54***	1.79*	–	5.88***	8.08***	–
log Z	–3.99***	–	–0.43	–0.78	–	3.68***

Industrialized countries:

The first specification is clearly the only satisfactory one for this sample, given the *t* statistics. Writing the model as

$$\log Y = \beta_1 + \beta_2 \log X + \beta_3 \log Z + u,$$

in the second specification

$$E(b_2) = \beta_2 + \beta_3 \frac{\sum (\log X_i - \overline{\log X})(\log Z_i - \overline{\log Z})}{\sum (\log X_i - \overline{\log X})^2}$$

Anticipating that β_3 is negative, and knowing that X and Z are positively correlated, the bias term should be negative. The estimate of β_2 is indeed lower in the second specification. In the third specification,

$$E(b_3) = \beta_3 + \beta_2 \frac{\sum (\log X_i - \overline{\log X})(\log Z_i - \overline{\log Z})}{\sum (\log Z_i - \overline{\log Z})^2}$$

and the bias should be positive, assuming β_2 is positive. b_3 is indeed less negative than in the first specification.

Note that the sum of the R^2 statistics for the second and third specifications is less than R^2 in the first. This is because the bias terms undermine the apparent explanatory power of X and Z in the second and third specifications. In the third specification, the bias term virtually neutralizes the true effect and R^2 is very low indeed.

Developing countries:

In principle the first specification is acceptable. The failure of the coefficient of Z to be significant might be due to a combination of a weak effect of Z and a relatively small sample.

The second specification is also acceptable since the coefficient of Z and its t statistic in the first specification are very low. Because the t statistic of Z is low, R^2 is virtually unaffected when it is omitted.

The third specification is untenable because it cannot account for the highly significant coefficient of X in the first. The omitted variable bias is now so large that it overwhelms the negative effect of Z with the result that the estimated coefficient is positive.

- 6.10 A researcher has data on output per worker, Y , and capital per worker, K , both measured in thousands of dollars, for 50 firms in the textiles industry in 2001. She hypothesizes that output per worker depends on capital per worker and perhaps also the technological sophistication of the firm, $TECH$:

$$Y = \beta_1 + \beta_2 K + \beta_3 TECH + u$$

where u is a disturbance term. She is unable to measure $TECH$ and decides to use expenditure per worker on research and development in 2001, $R\&D$, as a proxy for it. She fits the following regressions (standard errors in parentheses):

$$\hat{Y} = 1.02 + 0.32 K \quad R^2=0.749$$

(0.45) (0.04)

$$\hat{Y} = 0.34 + 0.29K + 0.05 R\&D \quad R^2=0.750$$

(0.61) (0.22) (0.15)

The correlation coefficient for K and $R\&D$ is 0.92. Discuss these regression results

1. assuming that Y does depend on both K and $TECH$,
2. assuming that Y depends only on K .

Answer: If Y depends on both K and $TECH$, the first specification is subject to omitted variable bias, with the expected value of b_2 being given by

$$E(b_2) = \beta_2 + \beta_3 \frac{\sum (K_i - \bar{K})(TECH_i - \overline{TECH})}{\sum (K_i - \bar{K})^2}$$

Since K and $R\&D$ have a high positive correlation, it is reasonable to assume that K and $TECH$ are positively correlated. It is also reasonable to assume that β_3 is positive. Hence one would expect b_2 to be upwards biased. It is indeed greater than in the second equation, but not by much. The second specification is clearly subject to multicollinearity, with the consequence that, although the estimated coefficients remain unbiased, they are erratic, this being reflected in large standard errors. The large variance of the estimate of the coefficient of K means that much of the difference between it and the estimate in the first specification is likely to be purely random, and this could account for the fact that the omitted variable bias appears to be so small.

If Y depends only on K , the inclusion of $R\&D$ in the second specification gives rise to inefficiency. Since the standard errors in both equations remain valid, they can be compared and it is evident that the loss of efficiency is severe. As expected in this case, the coefficient of $R\&D$ is not significantly different from zero and the increase in R^2 in the second specification is minimal.

6.13 The first regression shows the result of regressing *LGFDHO*, the logarithm of annual household expenditure on food eaten at home, on *LGEXP*, the logarithm of total annual household expenditure, and *LGSIZE*, the logarithm of the number of persons in the household, using a sample of 868 households in the 1995 Consumer Expenditure Survey. In the second regression, *LGFDHOPC*, the logarithm of food expenditure per capita (*FDHO/SIZE*), is regressed on *LGEXPPC*, the logarithm of total expenditure per capita (*EXP/SIZE*). In the third regression *LGFDHOPC* is regressed on *LGEXPPC* and *LGSIZE*.

```
. reg LGFDHO LGEXP LGSIZE
```

Source	SS	df	MS	Number of obs =	868
Model	138.776549	2	69.3882747	F(2, 865) =	460.92
Residual	130.219231	865	.150542464	Prob > F =	0.0000
				R-squared =	0.5159
				Adj R-squared =	0.5148
Total	268.995781	867	.310260416	Root MSE =	.388

LGFDHO	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
LGEXP	.2866813	.0226824	12.639	0.000	.2421622 .3312003
LGSIZE	.4854698	.0255476	19.003	0.000	.4353272 .5356124
_cons	4.720269	.2209996	21.359	0.000	4.286511 5.154027

```
. reg LGFDHOPC LGEXPPC
```

Source	SS	df	MS	Number of obs =	868
Model	51.4364364	1	51.4364364	F(1, 866) =	313.04
Residual	142.293973	866	.164311747	Prob > F =	0.0000
				R-squared =	0.2655
				Adj R-squared =	0.2647
Total	193.73041	867	.223449146	Root MSE =	.40535

LGFDHOPC	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
LGEXPPC	.376283	.0212674	17.693	0.000	.3345414 .4180246
_cons	3.700667	.1978925	18.700	0.000	3.312262 4.089072

```
. reg LGFDHOPC LGEXPPC LGSIZE
```

Source	SS	df	MS	Number of obs =	868
Model	63.5111811	2	31.7555905	F(2, 865) =	210.94
Residual	130.219229	865	.150542461	Prob > F =	0.0000
				R-squared =	0.3278
				Adj R-squared =	0.3263
Total	193.73041	867	.223449146	Root MSE =	.388

LGFDHOPC	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
LGEXPPC	.2866813	.0226824	12.639	0.000	.2421622 .3312004
LGSIZE	-.2278489	.0254412	-8.956	0.000	-.2777826 -.1779152
_cons	4.720269	.2209996	21.359	0.000	4.286511 5.154027

1. Explain why the second model is a restricted version of the first, stating the restriction.
2. Perform an *F* test of the restriction.
3. Perform a *t* test of the restriction.
4. Summarize your conclusions from the analysis of the regression results.

Answer: Write the first specification as

$$LGF DHO = \beta_1 + \beta_2 LGEXP + \beta_3 LGSIZE + u$$

Then the restriction implicit in the second specification is $\beta_3 = 1 - \beta_2$, for then

$$LGF DHO = \beta_1 + \beta_2 LGEXP + (1 - \beta_2)LGSIZE + u$$

$$LGF DHO - LGSIZE = \beta_1 + \beta_2(LGEXP - LGSIZE) + u$$

$$\log \frac{FDHO}{SIZE} = \beta_1 + \beta_2 \log \frac{EXP}{SIZE} + u$$

$$LGF DHO PC = \beta_1 + \beta_2 LGEXP PC + u,$$

the last equation being the second specification. The F statistic for the null hypothesis $H_0: \beta_3 = 1 - \beta_2$ is

$$F(1,865) = \frac{(142.29 - 130.22)/1}{130.22/865} = 80.2 .$$

The critical value of $F(1,865)$ at the 0.1 percent level is 10.9, and hence the restriction is rejected at that significance level. This is not a surprising result, given that the estimates of β_2 and β_3 in the unrestricted specification were 0.29 and 0.49, respectively, their sum being well short of 1, as implied by the restriction.

Alternatively, we could use the t test approach. The restriction may be written $\beta_2 + \beta_3 - 1 = 0$ and hence our test statistic is $\theta = \beta_2 + \beta_3 - 1$. From this we obtain $\beta_3 = \theta - \beta_2 + 1$. Substituting for β_3 , the unrestricted version may be rewritten

$$LGF DHO = \beta_1 + \beta_2 LGEXP + (\theta - \beta_2 + 1)LGSIZE + u$$

Hence

$$LGF DHO - LGSIZE = \beta_1 + \beta_2(LGEXP - LGSIZE) + \theta LGSIZE + u,$$

that is,

$$LGF DHO PC = \beta_1 + \beta_2 LGEXP PC + \theta LGSIZE + u.$$

We use a t test to see if the coefficient of $LGSIZE$ is significantly different from zero. If it is not, we can drop the $LGSIZE$ term and we conclude that the restricted specification is an adequate representation of the data. If it is, we have to stay with the unrestricted specification. From the output for the third regression, we see that t is -8.96 and hence the null hypothesis $H_0: \beta_2 + \beta_3 - 1 = 0$ is rejected (critical value of t at the 0.1 percent level is 3.31). Note that the t statistic is the square root of the F statistic and the critical value of t at the 0.1 percent level is the square root of the critical value of F .

Answers to the additional exercises

- A6.1 (a) To simplify the algebra, throughout this answer $\log B$, $\log Y$, and $\log IQ$ will be written as B , Y , and IQ , it being understood that these are logarithms.

$$\begin{aligned}
 b_2 &= \frac{\sum (B_i - \bar{B})(Y_i - \bar{Y})}{\sum (Y_i - \bar{Y})^2} = \frac{\sum (\beta_1 + \beta_2 Y_i + \beta_3 IQ_i + u_i - \beta_1 - \beta_2 \bar{Y} - \beta_3 \bar{IQ} - \bar{u})(Y_i - \bar{Y})}{\sum (Y_i - \bar{Y})^2} \\
 &= \frac{\sum (\beta_2 Y_i - \beta_2 \bar{Y})(Y_i - \bar{Y}) + \sum (\beta_3 IQ_i - \beta_3 \bar{IQ})(Y_i - \bar{Y}) + \sum (u_i - \bar{u})(Y_i - \bar{Y})}{\sum (Y_i - \bar{Y})^2} \\
 &= \beta_2 + \beta_3 \frac{\sum (IQ_i - \bar{IQ})(Y_i - \bar{Y})}{\sum (Y_i - \bar{Y})^2} + \frac{\sum (u_i - \bar{u})(Y_i - \bar{Y})}{\sum (Y_i - \bar{Y})^2}
 \end{aligned}$$

Hence

$$\begin{aligned}
 E(b_2) &= \beta_2 + \beta_3 \frac{\sum (IQ_i - \bar{IQ})(Y_i - \bar{Y})}{\sum (Y_i - \bar{Y})^2} + \frac{1}{\sum (Y_i - \bar{Y})^2} E(\sum (u_i - \bar{u})(Y_i - \bar{Y})) \\
 &= \beta_2 + \beta_3 \frac{\sum (IQ_i - \bar{IQ})(Y_i - \bar{Y})}{\sum (Y_i - \bar{Y})^2} + \frac{1}{\sum (Y_i - \bar{Y})^2} \sum E((u_i - \bar{u})(Y_i - \bar{Y})) \\
 &= \beta_2 + \beta_3 \frac{\sum (IQ_i - \bar{IQ})(Y_i - \bar{Y})}{\sum (Y_i - \bar{Y})^2} + \frac{1}{\sum (Y_i - \bar{Y})^2} \sum (Y_i - \bar{Y}) E(u_i - \bar{u}) \\
 &= \beta_2 + \beta_3 \frac{\sum (IQ_i - \bar{IQ})(Y_i - \bar{Y})}{\sum (Y_i - \bar{Y})^2}
 \end{aligned}$$

assuming that Y and IQ are nonstochastic. Thus b_2 is biased, the direction of the bias depending on the signs of β_3 and $\sum (IQ_i - \bar{IQ})(Y_i - \bar{Y})$. We would expect the former to be positive and we expect the latter to be positive since we are told that the correlation between S and Y is positive and S is a proxy for IQ . So we would expect an upward bias in regression (2).

- (b) The use of S as a proxy for IQ will alleviate the problem of omitted variable bias. In particular, comparing the results of regression (1) with those that would have been obtained if B had been regressed on Y and IQ :
- the coefficient of Y will be approximately the same
 - its standard error will be approximately the same
 - the t statistic for S will be approximately equal to that of IQ
 - R^2 will be approximately the same
- (c) Not necessarily. It could be that S is a poor proxy for IQ , but it could also be that the original model had low explanatory power.
- (d) The high correlation between Y and S has given rise to multicollinearity, the standard errors being so large that the coefficients are not significantly different from zero.
- (e) Yes. It is reasonable to suppose that expenditure on books should not be negatively influenced by either income or cognitive ability. (Note that one should *not* say that it is reasonable to suppose that expenditure on books is positively influenced by them. This rules out the null hypothesis.)
- (f) No. It would be randomly higher or lower, if S is an irrelevant variable.
- (g) Yes. The inclusion of an irrelevant variable in general does not invalidate the standard errors. It causes them to be larger than those in the correct specification.

A6.2 The output below gives the results of a simple regression of $LGCATPC$ on $LGSIZE$. See Exercise A5.2 for the simple regression of $LGCATPC$ on $LGEXPPC$ and Exercise A5.3 for the multiple regression of $LGCATPC$ on $LGEXPPC$ and $LGSIZE$.

```
. reg LGFDHOPC LGSIZE
```

Source	SS	df	MS			
Model	39.4632274	1	39.4632274	Number of obs =	868	
Residual	154.267181	866	.178137622	F(1, 866) =	221.53	
Total	193.730408	867	.223449145	Prob > F =	0.0000	
				R-squared =	0.2037	
				Adj R-squared =	0.2028	
				Root MSE =	.42206	

LGFDHOPC	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
LGSIZE	-.3696776	.0248373	-14.88	0.000	-.418426	-.3209293
_cons	7.498327	.0249932	300.01	0.000	7.449272	7.547381

If the model is written as

$$LGFDHOPC = \beta_1 + \beta_2 LGEXPPC + \beta_3 LGSIZE + u,$$

the expected value of b_2 in the second regression is given by

$$E(b_2) = \beta_2 + \beta_3 \frac{\sum (LGEXPPC_i - \overline{LGEXPPC})(LGSIZE_i - \overline{LGSIZE})}{\sum (LGEXPPC_i - \overline{LGEXPPC})^2}.$$

We know that the numerator of the second factor in the bias term is negative because the correlation is negative:

```
. cor LGEXPPC LGSIZE if FDHO>0
(obs=868)
```

	LGEXPPC	LGSIZE
LGEXPPC	1.0000	
LGSIZE	-0.4411	1.0000

It is reasonable to suppose that economies of scale will cause β_3 to be negative, and the highly significant negative estimate in the multiple regression provides empirical support, so b_2 is biased upwards. This accounts for the increase in its size in the second regression. In the third regression,

$$E(b_3) = \beta_3 + \beta_2 \frac{\sum (LGEXPPC_i - \overline{LGEXPPC})(LGSIZE_i - \overline{LGSIZE})}{\sum (LGSIZE_i - \overline{LGSIZE})^2}$$

β_2 is certainly positive, especially given the highly significant positive estimate in the first regression, and so b_3 is biased downwards. As a consequence, the estimate in the third regression is lower than that in the first.

Similar results are obtained for the other categories of expenditure. The correlation between *LGEXPPC* and *LGSIZE* varies because the missing observations are different for different categories, but it is always at least -0.4 .

Omitted variable bias, dependent variable <i>LGCATPC</i>					
		<i>Multiple regression</i>		<i>Simple regressions</i>	
	<i>n</i>	<i>LGEXPPC</i>	<i>LGSIZE</i>	<i>LGEXPPC</i>	<i>LGSIZE</i>
<i>FDHO</i>	868	0.2867	-0.2278	0.3763	-0.3697
<i>FDAW</i>	827	1.4164	0.2230	1.3203	-0.5293
<i>HOUS</i>	867	1.0384	-0.1566	1.1006	-0.6731
<i>TELE</i>	858	0.4923	-0.3537	0.6312	-0.5955
<i>DOM</i>	454	0.8786	0.2084	0.7977	-0.1564
<i>TEXT</i>	482	0.9543	-0.1565	1.0196	-0.6386
<i>FURN</i>	329	0.6539	-0.4622	0.8560	-0.8142
<i>MAPP</i>	244	0.5136	-0.4789	0.7572	-0.8007
<i>SAPP</i>	467	0.7223	-0.5076	0.9481	-0.8840
<i>CLOT</i>	847	1.1138	0.3502	0.9669	-0.2236
<i>FOOT</i>	686	0.6992	-0.0813	0.7339	-0.4515
<i>GASO</i>	797	0.6770	-0.0785	0.7107	-0.4491
<i>TRIP</i>	309	1.0563	-0.3570	1.2434	-0.9050
<i>LOCT</i>	172	-0.0141	-0.5429	0.1993	-0.5367
<i>HEAL</i>	821	0.6612	-0.5121	0.8629	-0.8229
<i>ENT</i>	824	1.4679	0.3771	1.3069	-0.4213
<i>FEES</i>	676	1.7907	0.4286	1.5884	-0.6093
<i>TOYS</i>	592	0.9522	0.0054	0.9497	-0.5498
<i>READ</i>	764	0.9652	-0.4313	1.1532	-0.9210
<i>EDUC</i>	288	1.2243	-0.1707	1.2953	-0.9835
<i>TOB</i>	368	0.4329	-0.5379	0.6646	-0.7917

- A6.3 (a) First note that, since $E = W/H$, $LE = \log(W/H) = LW - LH$.
Write specification (2) as

$$LW = \beta_1 + \beta_2 S + \beta_3 LH + u$$

If one imposes the restriction $\beta_3 = 1$, the model becomes specification (1):

$$LW - LH = \beta_1 + \beta_2 S + u$$

The restriction implies that weekly earnings are proportional to hours worked, controlling for schooling..

- (b) If the restriction is valid, the coefficient of S should be similar in the restricted specification (1) and the unrestricted specification (2). Both estimates will be unbiased, but that in specification (1) will be more efficient. The gain in efficiency in specification (1) should be reflected in a smaller standard error. However, the gain will be small, given the low correlation.
- (c) The estimate of the coefficient of S would be biased. The standard error in specification (1) would be invalid and so a comparison with the standard error in specification (2) would be illegitimate.
- (d) The null and alternative hypotheses are $H_0: \beta_3 = 1$ and $H_0: \beta_3 \neq 1$.

$$F(1,1752) = \frac{(741.5 - 737.9)/1}{737.9/1752} = 8.55 \tag{3}$$

The critical value of $F(1,1000)$ at the 1 percent level is 6.66. The critical value of $F(1,1752)$ must be lower. Thus we reject the restriction at the 1 percent level. (The critical value at the 0.1 percent level is about 10.8.)

- (e) The restriction is so simple that it can be tested with no reparameterization: a simple t test on the coefficient of LH in specification (2), $H_0: \beta_3 = 1$.

Alternatively, mechanically following the standard procedure, we rewrite the restriction as $\beta_3 - 1 = 0$. The test statistic will be

$$\theta = \beta_3 - 1$$

and so

$$\beta_3 = \theta + 1$$

Substituting this into the unrestricted specification, the latter may be rewritten

$$LW = \beta_1 + \beta_2 S + (\theta + 1)LH + u$$

Hence

$$LW - LH = \beta_1 + \beta_2 S + \theta LH + u$$

This is regression specification (3) and the restriction may be tested with a t test on the coefficient of LH , the null hypothesis being $H_0: \theta = \beta_3 - 1 = 0$. The t statistic is 2.92, which is significant at the 1 percent level, implying that the restriction should be rejected.

- (f) The test must lead to the same conclusion since the F statistic is the square of the t statistic and the critical value of F is the square of the critical value of t .
- (g) The appropriate test is a Chow test. The test statistic under the null hypothesis of no difference in the earnings functions is

$$F(3,1749) = \frac{(737.9 - 626.1 - 100.1)/3}{(626.1 + 100.1)/1749} = 9.39$$

The critical value of $F(3,1000)$ at the 0.1 percent level is 5.46. Hence we reject the null hypothesis and conclude that the commentator is correct.

- (h) The elasticity of LH is now not significantly different from 1 for either full-time or part-time workers, so the restriction is no longer rejected.

- A6.4 (a) The elasticity of expenditure on health per capita with respect to GDP per capita is 1.27. A one percent increase in GDP per capita leads to a 1.27 percent increase in health expenditure per capita.
- (b) GDP per capita increases by 1 percent, so expenditure on health per capita will increase by 1.27 percent. 1.27 percent of \$40 is \$0.51, so the predicted level of expenditure on health in 2002 is \$40.51.
- (c) The elasticity of expenditure on health with respect to GDP, controlling for population, is 1.27. This seems plausible since health expenditure should at least keep pace with GDP.

An increase in population, holding GDP constant, has two effects: a direct effect, which may be expected to be positive and probably with an elasticity about 1, and an indirect income effect attributable to the fact that an increase in population, holding GDP constant, means a reduction in GDP per capita. Since the income elasticity is greater than 1, the total effect should be negative, and accordingly the elasticity, -0.33 , seems plausible.

- (d) Write the first specification

$$\log \frac{HEALTH}{POP} = \beta_1 + \beta_2 \log \frac{GDP}{POP} + u.$$

It may be rewritten as

$$\log HEALTH - \log POP = \beta_1 + \beta_2 \log GDP - \beta_2 \log POP + u$$

and this in turn may be rewritten

$$\log HEALTH = \beta_1 + \beta_2 \log GDP + (1 - \beta_2) \log POP + u$$

This is a restricted version of the more general specification

$$\log HEALTH = \beta_1 + \beta_2 \log GDP + \beta_3 \log POP + u$$

with the restriction $\beta_3 = 1 - \beta_2$.

- (e) $F(1,67) = \frac{(14.26 - 13.90)/1}{13.90/67} = 1.74$. The null hypothesis is $H_0: \beta_3 = 1 - \beta_2$. The critical value of $F(1,67)$ at the 5 percent significance level is about 3.99. Hence we do not reject the restriction.
- (f) The restriction may be rewritten as $\beta_2 + \beta_3 - 1 = 0$. The test statistic will therefore be $\theta = \beta_2 + \beta_3 - 1$. Rewriting this relationship as

$$\beta_3 = \theta - \beta_2 + 1$$

and substituting for β_3 in the unrestricted specification, one has

$$\log HEALTH = \beta_1 + \beta_2 \log GDP + (\theta - \beta_2 + 1) \log POP + u$$

This may be rewritten

$$\log \frac{HEALTH}{POP} = \beta_1 + \beta_2 \log \frac{GDP}{POP} + \theta \log POP + u$$

A t test on the coefficient of $\log POP$ is thus a test of the restriction. If the coefficient is not significantly different from zero, we could drop the term and simplify the model to the restricted version. Regression (3) fits this specification. The t statistic is -1.5 , subject to rounding error, and hence one does not reject the restriction.

- (g) No, the deterioration in the fit, as indicated by the increase in RSS , does not necessarily matter. A restriction is imposed with the objective of obtaining more precise point estimates of the coefficients and it is accepted that there will be some deterioration in the overall fit. However, if the restriction is valid, imposing it should not lead to a significantly worse fit. This is the point of the F test.

A6.5

```
. reg LGFDHO LGEXP LGSIZE
```

Source	SS	df	MS			
Model	138.776549	2	69.3882747	Number of obs =	868	
Residual	130.219231	865	.150542464	F(2, 865) =	460.92	
Total	268.995781	867	.310260416	Prob > F =	0.0000	
				R-squared =	0.5159	
				Adj R-squared =	0.5148	
				Root MSE =	.388	

LGFDHO	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
LGEXP	.2866813	.0226824	12.639	0.000	.2421622	.3312003
LGSIZE	.4854698	.0255476	19.003	0.000	.4353272	.5356124
_cons	4.720269	.2209996	21.359	0.000	4.286511	5.154027

Dougherty: Introduction to Econometrics 3e
Study Guide

. reg LGFDHOPC LGEXPPC

Source	SS	df	MS			
Model	51.4364364	1	51.4364364	Number of obs =	868	
Residual	142.293973	866	.164311747	F(1, 866) =	313.04	
Total	193.73041	867	.223449146	Prob > F =	0.0000	
				R-squared =	0.2655	
				Adj R-squared =	0.2647	
				Root MSE =	.40535	

LGFDHOPC	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
LGEXPPC	.376283	.0212674	17.693	0.000	.3345414 .4180246
_cons	3.700667	.1978925	18.700	0.000	3.312262 4.089072

Write the first specification as

$$LGFDHO = \beta_1 + \beta_2 LGEXP + \beta_3 LGSIZE + u$$

Then the restriction implicit in the second specification is $\beta_3 = 1 - \beta_2$, for then

$$LGFDHO = \beta_1 + \beta_2 LGEXP + (1 - \beta_2) LGSIZE + u$$

$$LGFDHO - LGSIZE = \beta_1 + \beta_2 (LGEXP - LGSIZE) + u$$

$$\log \frac{FDHO}{SIZE} = \beta_1 + \beta_2 \log \frac{EXP}{SIZE} + u$$

$$LGFDHOPC = \beta_1 + \beta_2 LGEXPPC + u,$$

the last equation being the second specification. The F statistic for the null hypothesis $H_0: \beta_3 = 1 - \beta_2$ is

$$F(1,865) = \frac{(142.29 - 130.22)/1}{130.22/865} = 80.2.$$

The critical value of $F(1,865)$ at the 0.1 percent level is 10.9, and hence the restriction is rejected at that significance level. This is not a surprising result, given that the estimates of β_2 and β_3 in the unrestricted specification were 0.29 and 0.49, respectively, their sum being well short of 1, as implied by the restriction.

Summarizing the results of the test for all the categories, we have:

Restriction rejected at the 1 percent level: *FDHO, FDAW, HOUS, TELE, FURN, MAPP, SAPP, CLOT, HEAL, ENT, FEES, READ, TOB.*

Restriction rejected at the 5 percent level: *TRIP, LOCT.*

Restriction not rejected at the 5 percent level: *DOM, TEXT, FOOT, GASO, TOYS, EDUC.*

	<i>n</i>	<i>RSS restricted</i>	<i>RSS unrestricted</i>	<i>F</i>	<i>t</i>
<i>FDHO</i>	868	142.29	130.22	80.18	-8.96
<i>FDAW</i>	827	608.05	597.61	14.39	3.79
<i>HOUS</i>	867	502.08	496.41	9.87	-3.14
<i>TELE</i>	858	380.59	351.81	69.94	-8.36
<i>DOM</i>	454	1325.21	1319.71	1.88	1.37
<i>TEXT</i>	482	560.37	557.55	2.42	-1.56
<i>FURN</i>	329	697.33	681.45	7.60	-2.76
<i>MAPP</i>	244	291.76	280.41	9.75	-3.12
<i>SAPP</i>	467	522.31	493.39	27.20	-5.22
<i>CLOT</i>	847	686.45	659.59	34.37	5.86
<i>FOOT</i>	686	589.34	588.21	1.31	-1.14
<i>GASO</i>	797	366.92	365.73	2.58	-1.60
<i>TRIP</i>	309	527.42	517.96	5.59	-2.36
<i>LOCT</i>	172	450.92	433.51	6.79	-2.60
<i>HEAL</i>	821	1351.63	1294.03	36.41	-6.03
<i>ENT</i>	824	754.86	725.85	32.81	5.73
<i>FEES</i>	676	1145.09	1117.00	16.92	4.11
<i>TOYS</i>	592	809.01	809.01	0.00	0.05
<i>READ</i>	764	897.63	861.92	31.53	-5.61
<i>EDUC</i>	288	828.35	826.85	0.52	-0.72
<i>TOB</i>	368	385.63	360.58	25.36	-5.04

For the t test, we first rewrite the restriction as $\beta_2 + \beta_3 - 1 = 0$. The test statistic is therefore $\theta = \beta_2 + \beta_3 - 1$. This allows us to write $\beta_3 = \theta - \beta_2 + 1$. Substituting for β_3 , the unrestricted version becomes

$$LGF\text{DHO} = \beta_1 + \beta_2 LG\text{EXP} + (\theta - \beta_2 + 1)LG\text{SIZE} + u$$

Hence the unrestricted version may be rewritten

$$LGF\text{DHO} - LG\text{SIZE} = \beta_1 + \beta_2(LG\text{EXP} - LG\text{SIZE}) + \theta LG\text{SIZE} + u$$

that is,

$$LGF\text{DHOPC} = \beta_1 + \beta_2 LG\text{EXPPC} + \theta LG\text{SIZE} + u.$$

We use a t test to see if the coefficient of $LG\text{SIZE}$ is significantly different from 0. If it is not, we can drop the $LG\text{SIZE}$ term and we conclude that the restricted specification is an adequate representation of the data. If it is, we have to stay with the unrestricted specification.

From the output for the third regression, we see that t is -8.96 and hence the null hypothesis $H_0: \beta_2 + \beta_3 - 1 = 0$ is rejected (critical value of t at the 0.1 percent level is 3.31). Note that the t statistic is the square root of the F statistic and the critical value of t at the 0.1 percent level is the square root of the critical value of F . The results for the other categories are likewise identical to those for the F test..

. reg LGFDHOPC LGEXPPC LGSIZE

Source	SS	df	MS			
Model	63.5111811	2	31.7555905	Number of obs =	868	
Residual	130.219229	865	.150542461	F(2, 865) =	210.94	
				Prob > F =	0.0000	
				R-squared =	0.3278	
				Adj R-squared =	0.3263	
Total	193.73041	867	.223449146	Root MSE =	.388	

LGFDHOPC	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
LGEXPPC	.2866813	.0226824	12.639	0.000	.2421622	.3312004
LGSIZE	-.2278489	.0254412	-8.956	0.000	-.2777826	-.1779152
_cons	4.720269	.2209996	21.359	0.000	4.286511	5.154027

A6.6 (a) Let the model be written

$$LGE = \beta_1 + \beta_2 S + \beta_3 NUM + \beta_4 VERB + u$$

The restriction is $\beta_4 = \beta_3$ since *NUM* and *VERB* are given equal weights in the construction of *SCORE*. Using the restriction, the model can be rewritten

$$\begin{aligned} LGE &= \beta_1 + \beta_2 S + \beta_3 (NUM + VERB) + u \\ &= \beta_1 + \beta_2 S + 2\beta_3 SCORE + u \end{aligned}$$

(b) The null and alternative hypotheses are $H_0: \beta_4 = \beta_3$ and $H_1: \beta_4 \neq \beta_3$. The *F* statistic is

$$F(1,100) = \frac{(2045 - 2000)/1}{2000/100} = 2.25.$$

The critical value of $F(1,100)$ is 3.94 at the 5 percent level. Hence we do not reject the restriction at the 5 percent level.

(c) The restriction may be rewritten $\beta_4 - \beta_3 = 0$. The test statistic is therefore $\theta = \beta_4 - \beta_3$. Hence $\beta_4 = \theta + \beta_3$. Substituting for β_4 in the unrestricted model, one has

$$\begin{aligned} LGE &= \beta_1 + \beta_2 S + \beta_3 NUM + (\theta + \beta_3) VERB + u \\ &= \beta_1 + \beta_2 S + \beta_3 (NUM + VERB) + \theta VERB + u \\ &= \beta_1 + \beta_2 S + 2\beta_3 SCORE + \theta VERB + u \end{aligned}$$

This is the specification of Researcher 3. To test the hypothesis that the restriction is valid, we perform a *t* test on the coefficient of *VERB*. The *t* statistic is -1.5 , so we do not reject the restriction at the 5 percent level.

(d) No, the *F* test and the *t* test must give the same result because the *F* statistic must be the square of the *t* statistic and the critical value of *F* must be the square of the critical value of *t* for any given significance level. Note that this assumes a two-sided *t* test. If one is in a position to perform a one-sided test, the *t* test would be more powerful.

(e) One should perform a one-sided *t* test on the coefficient of *VERB* in regression 3 with the null hypothesis $H_0: \theta = 0$ and the alternative hypothesis $H_1: \theta < 0$. The null hypothesis is not rejected and hence one concludes that there is no significant difference.

(f) The regression results of Researchers 1 and 3 are equivalent, the only difference being that the coefficient of *VERB* provides a direct estimate of β_4 in the specification of Researcher 1 and $(\beta_4 - \beta_3)$ in the specification of Researcher 3. Assuming the restriction is valid, there is a large gain in efficiency in the estimation of β_3 in specification (2) because its standard error is effectively 0.0005, as opposed to 0.0011 in specifications (1) and (3).