CHAPTER 7: MULTIPLE REGRESSION ANALYSIS: THE PROBLEM OF ESTIMATION

7.1 The regression results are:

$\hat{\alpha}_1 = -3.00; \hat{\alpha}_2 = 3.50$

$\hat{\lambda}_1 = 4.00; \hat{\lambda}_2 = -1.357$

$\hat{\beta}_1 = 2.00; \hat{\beta}_2 = 1.00; \hat{\beta}_3 = -1.00$

(a) No. Given that model (3) is the true model, $\hat{\alpha}_2$ is a biased estimator of β_2.

(b) No. $\hat{\lambda}_3$ is a biased estimator of β_3, for the same reason as in (*a*).

The lesson here is that misspecifying an equation can lead to biased estimation of the parameters of the true model.

7.2 Using the formulas given in the text, the regression results are as follows:

$$\hat{Y}_i = 53.1612 + 0.727 X_{2i} + 2.736 X_{3i}$$

$$se \qquad (0.049) \quad (0.849) \quad R^2 = 0.9988; \ \bar{R}^2 = 0.9986$$

7.3 Omitting the observation subscript *i* for convenience, recall that

$$\hat{\beta}_2 = \frac{(\sum yx_2)(\sum x_3^2) - (\sum yx_3)(\sum x_2x_3)}{(\sum x_2^2)(\sum x_3^2) - (\sum x_2x_3)^2}$$

$$= \frac{(\sum yx_2) - (\sum yx_3)(\sum x_2x_3)/(\sum x_3^2)}{(\sum x_2^2) - (\sum x_2x_3)^2/(\sum x_3^2)}$$

$$= \frac{(\sum yx_2) - (\sum yx_3)b_{23}}{(\sum x_2^2) - b_{23}(\sum x_2x_3)}, \text{ using } b_{23} = \frac{(\sum x_2x_3)}{(\sum x_3^2)}$$

$$= \frac{\sum y(x_2 - b_{23}x_3)}{\sum x_2(x_2 - b_{23}x_3)}$$

7.4 Since we are told that is, $u_i \sim N(0,4)$, generate, say, 25 observations from a normal distribution with these parameters. Most computer packages do this routinely. From these 25 observations, compute the sample variance as $S^2 = \dfrac{\sum(X_i - \bar{X})^2}{24}$, where X_i = the observed value of u_i in the sample of 25 observations. Repeat this exercise, say, 99 more times, for a total of 100 experiments. In all there will be 100 values of S^2. Take the average of these 100 S^2 values. This average value should be close to $\sigma^2 = 4$. Sometimes you may need more than 100 samples for the approximation to be good.

7.5 From Eq. (7.11.7) from the text, we have

$$R^2 = r_{13}^2 + (1 - r_{13}^2) r_{12.3}^2 .$$

Therefore,

$$r_{12.3}^2 = \frac{R^2 - r_{13}^2}{1 - r_{13}^2}$$

This is the coefficient of partial determination and may be interpreted as describing the proportion of the variation in the dependent variable not explained by explanatory variable X_3, but has been explained by the addition of the explanatory variable X_2 to the model.

7.6 The given equation can be written as:

$$X_1 = (-\alpha_2/\alpha_1)X_2 + (-\alpha_3/\alpha_1)X_3, or$$
$$X_2 = (-\alpha_1/\alpha_2)X_1 + (-\alpha_3/\alpha_2)X_3, or$$
$$X_3 = (-\alpha_1/\alpha_3)X_1 + (-\alpha_2/\alpha_3)X_2$$

Therefore, the partial regression coefficients would be as follows:

$$\beta_{12.3} = -(\alpha_2/\alpha_1); \beta_{13.2} = -(\alpha_3/\alpha_1)$$
$$\beta_{21.3} = -(\alpha_1/\alpha_2); \beta_{23.1} = -(\alpha_2/\alpha_3)$$
$$\beta_{31.2} = -(\alpha_1/\alpha_3); \beta_{32.1} = -(\alpha_2/\alpha_3)$$

Recalling Question 3.6, it follows:

$$r_{12.3} = \sqrt{(\beta_{12.3})(\beta_{21.3})} = \sqrt{\frac{(-\alpha_2)(-\alpha_1)}{(\alpha_1)(\alpha_2)}} = \sqrt{1} = \pm 1$$

7.7 (*a*) No. An r-value cannot exceed 1 in absolute value. Plugging the given data in Eq. (7.11.2), the reader can should verify that: $r_{12.3} = 2.295$, which is logically impossible.

(b) Yes. Following the same procedure as in (*a*), the reader will find that $r_{12.3} = 0.397$, which is possible.

(c) Yes, again it can be shown that $r_{12.3} = 0.880$, which is possible.

7.8 If you leave out the years of experience (X_3) from the model, the coefficient of education (X_2) will be biased, the nature of the bias depending on the correlation between X_2 and X_3. The standard error, the residual sum of squares, and R^2 will all be affected as a result of this omission. This is an instance of the *omitted variable bias.*

7.9 The slope coefficients in the double-log models give direct estimates of the (constant) elasticity of the left-hand side variable with respect to the right hand side variable. Here:

$$\frac{\partial \ln Y}{\partial \ln X_2} = \frac{\partial Y / Y}{\partial X_2 / X_2} = \beta_2 \text{, and}$$

$$\frac{\partial \ln Y}{\partial \ln X_3} = \frac{\partial Y / Y}{\partial X_3 / X_3} = \beta_3$$

7.10 (*a*) & (*b*) If you multiply X_2 by 2, you can verify from Equations (7.4.7) and (7.4.8), that the slopes remain unaffected. On the other hand, if you multiply Y by 2, the slopes as well as the intercept coefficients and their standard errors are all multiplied by 2. Always keep in mind the units in which the regressand and regressors are measured.

7.11 From (7.11.5) we know that

$$R^2 = \frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_3}{1 - r_{23}^2}.$$

Therefore, when $r_{23} = 0$, that is, no correlation between variables X_2 and X_3,

$R^2 = r^2_{12} + r^2_{13}$, that is, the multiple coefficient of determination is the sum of the coefficients of determination in the regression of Y on X_2 and that of Y on X_3.

7.12 (*a*) Rewrite Model B as:

$$Y_t = \beta_1 + (1+\beta_2)X_{2t} + \beta_3 X_{3t} + u_t$$

$$= \beta_1 + \beta_2^* X_{2t} + \beta_3 X_{3t} + u_t \text{, where } \beta_2^* = (1+\beta_2)$$

Therefore, the two models are similar.Yes, the intercepts in the models are the same.

(*b*)The OLS estimates of the slope coefficient of X_3 in the two models will be the same.

(*c*) $\beta_2^* = (1+\beta_2) = \alpha_2$

(d) No, because the regressands in the two models are different.

7.13 (*a*) Using OLS, we obtain:

$$\hat{\alpha}_2 = \frac{\sum y_i x_i}{\sum x_i^2} = \frac{\sum (x_i - z_i)(x_i)}{\sum x_i^2}$$

$$= \frac{\sum x_i^2}{\sum x_i^2} - \frac{\sum z_i x_i}{\sum x_i^2}$$

$$= 1 - \hat{\beta}_2$$

That is, the slope in the regression of savings on income (i.e., the marginal propensity to save) is one minus the slope in the regression of consumption on income. (i.e., the marginal propensity to consume). Put differently, the sum of the two marginal propensities is 1, as it should be in view of the identity that total income is equal to total consumption expenditure and total savings. Incidentally, note that $\hat{\alpha}_1 = -\hat{\beta}_1$

*(b)*Yes. The RSS for the consumption function is:

$$\sum (Y_i - \hat{\alpha}_1 - \hat{\alpha}_2 X_i)^2$$

Now substitute (X_i-Y_i) for Z_i, $\hat{\alpha}_1 = -\hat{\beta}_1$ *and* $\hat{\alpha}_2 = (1 - \hat{\beta}_2)$ and verify that the two RSS are the same.

*(c)*No, since the two regressands are not the same.

7.14 (*a*) As discussed in Sec. 6.9, to use the classical normal linear regression model (CNLRM), we must assume that

$$\ln u_i \sim N(0, \sigma^2)$$

After estimating the Cobb-Douglas model, obtain the residuals and subject them to normality test, such as the Jarque-Bera test.

(b) No. As discussed in Sec. 6.9,

$$u_i \sim \log\text{-}normal[e^{\sigma^2/2}, e^{\sigma^2}(e^{\sigma^2} - 1)]$$

7.15 (*a*) The normal equations would be:

$$\sum Y_i X_{2i} = \beta_2 \sum X_{2i}^2 + \beta_3 \sum X_{2i} X_{3i}$$

$$\sum Y_i X_{3i} = \beta_2 \sum X_{2i} X_{3i} + \beta_3 \sum X_{3i}^2$$

(*b*) No, for the same reason as the two-variable case.

(*c*) Yes, these conditions still hold.

(b) It will depend on the underlying theory.

(c) This is a straightforward generalization of the normal equations given above.

Empirical Exercises

7.16 (*a*) **Linear Model:**

$$\hat{Y}_t = 10816.04 - 2227.704X_{2i} + 1251.141X_{3i} + 6.283X_{4i} - 197.399X_{5i}$$
$$se \quad (5988.348)(920.538) \quad (1157021) \quad (29.919) \quad (101.156)$$
$$R^2 = 0.835$$

In this model the slope coefficients measure the rate of change of Y with respect to the relevant variable.

(*b*) **Log-Linear Model**

$$\ln \hat{Y}_t = 0.627 - 1.274 \ln X_{2i} + 0.937 \ln X_{3i} + 1.713 \ln X_{4i} - 0.182 \ln X_{5i}$$
$$se \quad (6.148)\ (0.527) \quad (0.659) \quad (1.201) \quad (0.128)$$
$$R^2 = 0.778$$

In this model all the partial slope coefficients are partial elasticities of Y with respect to the relevant variable.

(c) The own-price elasticity is expected to be negative, the cross price elasticity is expected to be positive for substitute goods and negative for complimentary goods, and the income elasticity is expected to be positive, since roses are a normal good.

(d) The general formula for elasticity for linear equation is:

$$Elasticity = \frac{\partial Y}{\partial X_i}\frac{\bar{X}_i}{\bar{Y}}$$

, where Xi is the relevant regressor.
That is for a linear model, the elasticity can be computed at the mean values.

(e) Both models give similar results. One advantage of the log-linear model is that the slope coefficients give direct estimates of the (constant) elasticity of the relevant variable with respect to the regressor under consideration. But keep in mind that the R^2s of the two models are not directly comparable.

7.17 (*a*) A priori, all the variables seem relevant to explain wildcat activity. With the exception of the trend variable, all the slope coefficients are expected to be positive; trend may be positive or negative.

(*b*) The estimated model is:

$$\hat{Y}_i = -37.186 + 2.775X_{2i} + 24.152X_{3i} - 0.011X_{4i} - 0.213X_{5i}$$
$$se = (12.877) \quad (0.57) \quad (5.587) \quad (0.008) \quad (0.259)$$

$$R^2 = 0.656;\ \bar{R}^2 = 0.603$$

(c) Price per barrel and domestic output variables are statistically significant at the 5% level and have the expected signs. The other variables are not statistically different from zero.

(d) The log-linear model may be another specification. Besides giving direct estimates of the elasticities, it may capture nonlinearities (in the variables), if any.

7.18 (*a*) The regression results are:

$$\begin{aligned}\hat{Y}_i &= 19.443 + 0.018X_{2i} - 0.284X_{3i} + 1.343X_{4i} + 6.332X_{5i}\\ se &= (3.406)\quad (0.006)\quad (0.457)\quad (0.259)\quad (3.024)\end{aligned}$$

$$R^2 = 0.978;\ \bar{R}^2 = 0.972;\ \text{modified } R^2 = 0.734$$

(*b*) A priori, all the slope coefficients are expected to be positive. Except the coefficient for US military sales, all the other variables have the expected signs and are statistically significant at the 5% level.

(*c*) Overall federal outlays and some form of trend variable may be valuable.

7.19 (*a*) Model (5) seems to be the best as it includes all the economically relevant variables, including the composite real price of chicken substitutes, which should help alleviate the multicollinearity problem that may exist in model (4) between the price of beef and price of pork. Model (1) contains no substitute good information, and models (2) and (3) have limited substitute good information.

(*b*) The coefficient of ln X_2 represents income elasticity; the coefficient of ln X_3 represents own-price elasticity.

(*c*) Model (2) considers only pork as a substitute good, while model(4) considers both pork and beef.

(d) There may be a problem of multicollinearity between the price of beef and the price of pork.

(e) Yes. This might alleviate the problem of multicollinearity.

(f) They should be substitute goods because they compete with chicken as a food consumption product.

(g) The regression results of Model (5) are as follows:

$$\ln \hat{Y}_t = 2.030 + 0.481 \ln X_{2t} - 0.351 \ln X_{3t} - 0.061 \ln X_{6t}$$
$$se = (0.119)\ (0.068)\qquad (0.079)\qquad (0.130)$$
$$R^2 = 0.980;\ \bar{R}^2 = 0.977;\ \text{modified } R^2 = 0.810$$

The income elasticity and own-price elasticity have the correct signs.

(h) The consequence of estimating model (2) would be that the estimators are likely to be biased due to model misspecification. This topic is discussed in detail in Chap. 13.

7.20 (*a*) *Ceteris paribus*, on average, a 1% increase in the unemployment rate leads to a 0.34% increase in the quite rate, a 1% increase in the percentage of employees under 25 leads to a 1.22% increase in the quite rate, and 1% increase in the relative manufacturing employment leads to 1.22 % increase in the quite rate, a 1% increase in the percentage of women employees leads to a 0.80 % increase in the quite rate, and that over the time period under study, the quite rate declined at the rate of 0.54% per year.

(b) Yes, quite rate and the unemployment rate are expected to be negatively related.

(*c*) As more people under the age of 25 are hired, the quite rate is expected to go up because of turnover among younger workers.

(*d*) The decline rate is 0.54%. As working conditions and pensions benefits have increased over time, the quit rate has probably declined.

(*e*) No. Low is a relative term.

(*f*) Since the *t* values are given, we can easily compute the standard errors. Under the null hypothesis that the true β_i is zero, we have the relationship:

$$t = \frac{\hat{\beta}_i}{se(\hat{\beta}_i)} \Rightarrow se(\hat{\beta}_i) = \frac{\hat{\beta}_i}{t}$$

7.21 *(a)* The regression results are as follows:

$$\widehat{\ln M_2} = 1.2394 + 0.5243 \ln RGDP - 0.0255 \ln Tbrate$$
$$se = (0.6244)\ \ (0.1445)\qquad (0.0513)\quad R^2 = 0.7292$$

The regression results using the long-term (30 year bond) rate are as follows:

$$\hat{\ln M}_{2t} = 1.4145 + 0.4946 \ln RGDP_t - 0.0516 \ln LTRATE_t$$
$$se = (1.3174)\quad (0.2686) \qquad\qquad (0.1501) \quad R^2 = 0.7270$$

The income elasticites (0.5243 or 0.4946) and the interest rate elasticities (-0.0255 or –0.0516) are not vastly different, but as we will see in Chapter 8, regression using the short-term interest (TBrate) gives better statistical results.

(*b*) The ratio, M/GDP is known in the literature as the **Cambridge k.** It represents the proportion of the income that people wish to hold in the form of money. This ratio is sensitive to interest rate, as the latter represents the cost of holding money, which generally does not yield much interest income. The regression results are as follows:

$$\ln\left(\frac{M_2}{GDP}\right)_t = 3.4785 - 0.1719 \ln TBrate_t$$
$$se = (0.0780)\quad (0.0409) \qquad r^2 = 0.5095$$

$$\ln\left(\frac{M_2}{GDP}\right)_t = 3.8318 - 0.3123 \ln LTRATE_t$$
$$se \quad (0.1157)\quad (0.0532) \qquad r^2 = 0.6692$$

Since these are both bi-variate regressions, the reader can check that the Cambride k is statistically inversely related to the interest rate, as per prior expectations. Numerically, it is more sensitive to the long-term rate than the short-term rate. Since the dependent variable in the two models is the same, we can see that the r^2 value using the long-term interest rate as the regressor gives a much better fit.

(*c*) The answer is given in Exercise 8.29

7.22 The results of fitting the Cobb-Douglas production function, obtained from *EViews3* are as follows:

Dependent Variable: LOG(OUTPUT)
.
.
Sample: 1961 1987
Included observations: 27

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-11.93660	3.211064	-3.717335	0.0011
LOG(LABOR)	2.328402	0.599490	3.883972	0.0007
LOG(CAPITAL)	0.139810	0.165391	0.845330	0.4063

R-squared	0.971395	Mean dependent var	4.493912
Adjusted R-squared	0.969011	S.D. dependent var	0.461432
S.E. of regression	0.081229	Akaike info criterion	-2.078645
Sum squared resid	0.158356	Schwarz criterion	-1.934663
Log likelihood	31.06171	F-statistic	407.5017
Durbin-Watson stat	0.373792	Prob(F-statistic)	0.000000

(*a*) The estimated output/labor and output/capital elasticities are positive, as one would expect. But as we will see in the next chapter, the results do not make economic sense in that the capital input has no bearing on output, which, if true, would be very surprising. As we will see, perhaps collinearity may be the problem with the data.

(b) The regression results are as follows:

Dependent Variable: LOG(PRODUCTIVITY)
.
Sample: 1961 1987
Included observations: 27

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-1.155956	0.074217	-15.57533	0.0000
LOG(CLRATIO)	0.680756	0.044535	15.28571	0.0000

R-squared	0.903345	Mean dependent var	-2.254332
Adjusted R-squared	0.899479	S.D. dependent var	0.304336
S.E. of regression	0.096490	Akaike info criterion	-1.767569
Sum squared resid	0.232758	Schwarz criterion	-1.671581
Log likelihood	25.86218	F-statistic	233.6528
Durbin-Watson stat	0.263803	Prob(F-statistic)	0.000000

The elasticity of output/labor ratio (i.e., labor productivity) with respect to capital/labor ratio is about 0.68, meaning that if the latter increases by 1%, labor productivity, on average, goes up by about 0.68%. A key characteristic of developed economies is a relatively high capital/labor ratio.

7.23 This is a class exercise. Note that your answer will depend on the number of replications you carry out. The larger the number of replications, the closer the approximation.

7.24 *(a)*

$$C_t = -20.6327 + 0.7340 Y_d + 0.0360 Wealth - 5.5212\, Interest$$
$$t = (-1.6085) \quad (53.3762) \quad (14.4882) \quad (2.3067) \qquad R^2 = 0.9994$$

(b) The three independent variables are statistically significant at the 5% level. It seems that increases in Income (Y_d) and Wealth are related to increases in Consumption, whereas an increase in the Interest rate corresponds to a decrease in the Consumption level. This makes sense.

7.25 *(a)* Using a transformed time index (where *t* = 1 for the first observation on 1/3/95 and *t* = 260 on 12/20/99), the linear regression model is:

$$Close_t = -4.6941 + 0.5805\, t$$
$$t = (-0.6822) \quad (12.7005) \qquad R^2 = 0.3847$$

Although the independent variable *time* is statistically significant at the 5% (and even the 1%) level, the R^2 value isn't very strong. This is not surprising given the curved appearance of the graph.

(b)

$$Close_t = 72.6825 - 1.1915\, t + 0.0068 t^2$$
$$t = (8.9214) \quad (-8.2661) \quad (12.6937) \qquad R^2 = 0.6218$$

Yes, this model fits better than the one in part a. Both *time* variables are significant and the R^2 value has gone up dramatically.

(c)

$$Close_t = -10.8543 + 2.6128\, t - 0.0296 t^2 + 0.00009 t^3$$
$$t = (-1.415) \quad (10.2865) \quad (-13.0938) \quad (16.3256) \qquad R^2 = 0.8147$$

This cubic model fits the best of the three. All three *time* variables are significant and the R^2 value is the highest by almost 20%.