CHAPTER 4

MULTIPLE REGRESSION: ESTIMATION AND HYPOTHESIS TESTING

QUESTIONS

4.1. **(*a*)** It measures the change in the *mean* value of the dependent variable (Y) for a unit change in the value of an explanatory variable (X), holding the values of all other explanatory variables constant. Mathematically, it is the partial derivative of (mean) Y with respect to the given explanatory variable.

(*b*) It measures the proportion, or percentage, of the total variation in the dependent variable, $\sum (Y_i - \bar{Y})^2$, explained by *all* the explanatory variables included in the model.

(*c*) Exact *linear* relationship among the explanatory variables.

(*d*) More than one exact *linear* relationship among the explanatory variables.

(*e*) Testing the hypothesis about a single (partial) regression coefficient.

(*f*) Testing the hypothesis about two or more partial regression coefficients simultaneously.

(*g*) An R^2 value that is adjusted for degrees of freedom.

4.2. **(*a*)** (1) State the null and alternative hypotheses.

(2) Choose the level of significance.

(3) Find the t value of the coefficient under the null hypothesis, H_0.

(4) Compare this $|t|$ value with the critical value at the chosen level of significance and the given d.f.

(5) If the computed t value exceeds the critical t value, we reject the null hypothesis. Make sure that you use the appropriate one-tailed or two-tailed test.

(*b*) Here the null hypothesis is:

$$\mathrm{H}_0 : B_2 = B_3 = \ldots = B_k = 0$$

that is, all partial slopes are zero. The alternative hypothesis is that this is not so, that is, one or more partial slope coefficients are nonzero. Here, we use the ANOVA technique and the *F* test. If the computed *F* value under the null hypothesis exceeds the critical *F* value at the chosen level of significance, we reject the null hypothesis. Otherwise, we do not reject it. Make sure that the numerator and denominator d.f. are properly counted.

Note: In both (*a*) and (*b*), instead of choosing the level of significance in advance, obtain the *p* value of the estimated test statistic. If it is reasonably low, you can reject the null hypothesis.

4.3. (***a***) *True*. This is obvious from the formula relating the two R^2 s.

(***b***) *False*. Use the *F* test.

(***c***) *False*. When $R^2 = 1$, the value of *F* is infinite. But when it is zero, the *F* value is also zero.

(***d***) *True*, which can be seen from the normal and *t* distribution tables.

(***e***) *True*. It can be shown that $E(b_{12}) = B_2 + B_3 b_{32}$, where b_{32} is the slope coefficient in the regression of X_3 on X_2. From this relationship, the conclusion follows.

(***f***) *False*. It is statistically different from zero, not 1.

(***g***) *False*. We also need to know the level of significance.

(***h***) *False*. By the overall significance we mean that all partial regression coefficients are not simultaneously equal to zero, or that R^2 is different from zero.

(***i***) *Partially true*. If our concern is only with a single regression coefficient, then we use the *t* test in both cases. But if we are interested in testing the joint significance of two or more partial regression coefficients, the *t* test will not do; we will have to use the *F* test.

(***j***) *True*. This is because TSS = $\sum (Y_i - \bar{Y})^2$. We lose only one d.f. in computing the sample mean. Therefore, the d.f. are always (*n* -1).

4.4. (***a***) $\hat{\sigma}^2 = 880 / 21 = 41.9048$.

(***b***) $\hat{\sigma}^2 = 1220 / 10 = 122$.

4.5. 2.179; 2.528; -1.697; 1.960 (normal approximation).

4.6 5.05; 4.50; 1.62.

PROBLEMS

4.7. $\hat{Y}_i = -3.0 + 3.5\, X_{2i}$

$\hat{Y}_i = 4.0 - 1.3571\, X_{3i}$

$\hat{Y}_i = 2.0 + X_{2i} - X_{3i}$

(*1*) and **(*2*)** No, in both cases. As pointed out in Sec. 4.9, running a two-variable regression when a three-variable regression is called for is likely to give biased estimates of the true parameters. [See answer to question 4.3(e).] Only when $\text{cov}(X_2, X_3) = 0$ can one obtain unbiased estimates of the true parameters from the two-variable regressions. Even then, this procedure is not recommended because the standard errors can still be biased.

4.8. **(*a*)**, **(*b*)**, and **(*c*)** $\hat{Y}_i = 53.1600 + 0.7266\, X_{2i} + 2.7363\, X_{3i}$

$$\begin{aligned} \text{se} &= (13.0261) \quad (0.0487) \quad (0.8486) \\ t &= (4.0810) \quad (14.9199) \quad (3.2245) \end{aligned}$$

$R^2 = 0.9988;\ \overline{R}^2 = 0.9986$

(*d*) For 12 d.f. the two-tailed 5% critical t value is 2.179

95% CI for B_2: $0.7266 \pm 2.179\ (0.0487)$ or $0.6205 \le B_2 \le 0.8327$

95% CI for B_3: $2.7363 \pm 2.179\ (0.8486)$ or $0.8872 \le B_3 \le 4.5854$

(*e*) The null hypothesis that each partial slope coefficient is zero can be easily rejected at the 5% level of significance, since the confidence intervals established in (d) do not include the zero value.

(*f*) This hypothesis is equivalent to the hypothesis that $R^2 = 0$. It can be tested using the R^2 variant of the F test:

$$F = \frac{0.9988/2}{(1-0.9988)/12} = 4{,}994$$

This F value is obviously highly significant, leading to the rejection of the null of the null hypothesis. Set up the ANOVA table as indicated in the text.

4.9. (***a***) 15 (***b***) 77 (***c***) 2 and 12, respectively

(***d***) $R^2 = 0.9988$; $\overline{R}^2 = 0.9986$

(***e***) $F = \dfrac{65{,}965/2}{77/12} = 5{,}140.13$. This F value is highly significant, leading to the rejection of the null hypothesis

(***f***) No. We need the results of the two-variable regression models.

4.10. Follow Table 4-3 in the text.

4.11. (***a***) *Ceteris paribus,* if the BTU rating of an air conditioner goes up by a unit, the average price of the air conditioner goes up by about 2.3 cents. Other partial slope coefficients should be interpreted similarly. The intercept value has no viable economic meaning in the present case.

(***b***) Yes. *A priori*, each X variable is expected to have a positive impact on the price.

(***c***) For 15 d.f. the 5% one-tailed critical t value is 1.753. The observed t value of 0.023 / 0.005 = 4.6 exceeds this critical t value. Hence, we reject the null hypothesis.

(***d***) $H_0: R^2 = 0$ and $H_1: R^2 > 0$. Using the F test, we obtain

$$F = \frac{0.84/3}{0.16/15} = 26.25$$

This F value is significant beyond the 0.01 level of significance. So, reject the null hypothesis.

4.12. (***a***) The MPC is 0.93.

(***b***) $t = \dfrac{0.93 - 1}{0.003734} = -18.7465$

For 73 d.f. this t value is highly significant. Hence reject the null hypothesis that the true MPC is unity (*Note*: The se is obtained as 0.93 / 249.06 = 0.003734).

(***c***) Since expenditure on items such as automobiles, washers and dryers, etc., is often financed, the cost of borrowing becomes an important

determinant of consumption expenditure. Therefore, the interest rate, representing the cost of borrowing, is expected to have a negative impact on consumption expenditure.

(*d*) Yes. The *t* value is -3.09, which is significant at about the 0.01 level of significance (two-tailed test).

(*e*) $F = \dfrac{0.9996/2}{(1-0.9996)/73} = 91{,}213.5$

This *F* value is obviously very high, leading to the rejection of the null hypothesis that $R^2 = 0$. (*Note*: The *F* value reported by the authors is different because of rounding.)

(*f*) $\text{se}(b_1) = 3.2913$; $\text{se}(b_2) = 0.003734$; $\text{se}(b_3) = 0.6764$.

4.13. Use the *F* test: $F = \dfrac{0.96/2}{(1-0.96)/16} = 192$

For 2 and 16 d.f., this *F* value is highly significant. Hence, reject the null hypothesis that X_2 and X_3 have no influence on *Y*. The *F* test assumes that the error term is distributed normally.

4.14. **(*a*)** CM is expected to be negatively related to FLR and PGNP but positively related to TFR.

(*b*) The *EViews* regression results are:

Dependent Variable: CM
Sample: 1 64

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	263.8635	12.22499	21.58395	0.0000
FLR	-2.390496	0.213263	-11.20917	0.0000
R-squared	0.669590			

(c) The regression output is:

(Regression output is shown in the following page)

Dependent Variable: CM Sample: 1 64				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	263.6416	11.59318	22.74109	0.0000
FLR	-2.231586	0.209947	-10.62927	0.0000
PGNP	-0.005647	0.002003	-2.818703	0.0065
R-squared	0.707665			

(*d*) Adding the variable TFR, we obtain:

Dependent Variable: CM Sample: 1 64				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	168.3067	32.89165	5.117003	0.0000
FLR	-1.768029	0.248017	-7.128663	0.0000
PGNP	-0.005511	0.001878	-2.934275	0.0047
TFR	12.86864	4.190533	3.070883	0.0032
R-squared	0.747372			

The ANOVA table is straightforward. Set it up using the R^2 value.

(*e*) The model in (*d*) seems to be better in that all the variables have the expected the expected signs, each variable is individually statistically significant since the *p* values are very low, and the overall R^2 value is fairly high for cross-sectional data.

(*f*) In each case we will be committing a specification error, namely, the error of omitting a relevant variable(s). As a result, the coefficients of the incorrectly estimated model are likely to be inconsistent, a topic explored in Chapter 7.

(*g*) To answer this question, we use Equation (4.56). In the present case the unrestricted coefficient of determination, R_{ur}^2 (i.e., model (*d*)) is 0.7474 (approx.) and the restricted coefficient of determination R_r^2 (i.e., model (*b*) is 0.6696 (approx.). The number of restrictions here is 2 because model (*b*) excludes 2 variables (PGNP and TFR). Using Equation (4.56), we get:

$$F = \frac{(0.7474 - 0.6696)/2}{(1 - 0.7474)/(64 - 4)} = \frac{0.03890}{0.00421} = 9.2399$$

For 2 numerator and 60 denominator d.f., the computed F value is highly significant (the 1 percent critical value is 4.98), suggesting that both PGNP and TFR belong in the model.

4.15 The adjusted R^2 values are shown in the last column of the following table:

Value of R^2	n	k	$\overline{R}^2$
0.83	50	6	0.8107
0.55	18	9	0.1500
0.33	16	12	-1.5125
0.12	1,200	32	0.0966

These calculations show that the $\overline{R}^2$ value depends on the sample size as well as on the number of explanatory variables in the model. If the sample size is rather small and if the number of explanatory variables is relatively large, the $\overline{R}^2$ can be substantially smaller than the (unadjusted) R^2, as the second example shows so clearly, or even negative, as in the third example.

4.16. Using formula (4.50), we obtain:

$$F = \frac{0.689/4}{(1-0.689)/15} = 8.3079$$

For 4 and 15 d.f., this F value is significant beyond the 0.01 level. Therefore, we can reject the hypothesis that $R^2 = 0$.

4.17. This is straightforward, but use the R^2 version of the ANOVA table.

4.18 **(*a*)** As a first pass, consider the following results obtained from *EViews*.

The dependent variable is average starting pay (ASP).

Note: In this regression output, we present the adjusted R^2 for the first time.

(Regression output is shown in the following page)

Dependent Variable: ASP Sample: 1 49				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-53580.30	37064.295	-1.445604	0.1555
GMAT	-14.25427	74.522067	-0.191276	0.8492
GPA	6616.380	12078.929	0.547762	0.5867
PCTEMPLOYED	49411.20	15115.723	3.268861	0.0021
TUITION	0.871869	0.1822566	4.783745	0.0000
RATING	20444.504	4875.2060	4.193567	0.0001
R-squared	0.85858			
Adjusted R-squared	0.84213			

As these results suggest, GPA, tuition and recruiter perception have statistically significant positive impact on average starting salaries at the 0.1% or lower level of significance. The percentage of employed graduates also has a positive effect, indicating that higher demand for the graduates of a particular school translates into a higher salary. The R^2 value is reasonably high.

(*b*) Since GPA and GMAT are likely to be collinear, if we introduce them both in the model, as in (*a*), we would not expect both the variables to be individually statistically significant. This is borne out by the results given in (*a*).

(c) If the tuition variable is a proxy for the quality of education, higher tuition may well have a positive impact on ASP, *ceteris paribus*. The results in (*a*) may support such a hypothesis.

(*d*) Regressing GMAT on GPA, we obtain the following *EViews* output:

Dependent Variable: GMAT Sample: 1 65				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	208.8593	115.8665	1.802585	0.0763
GPA	134.7269	34.31539	3.926135	0.0002
R-squared	0.19912			
Adjusted R-squared	0.18620			

From these results it seems that GMAT and GPA are collinear.

(***e***) The *Excel* Analysis of Variance output is as follows (*EViews* does not automatically provide an ANOVA table in regressions):

Source of variation	**SS**	**df**	**MSS**	***F***	***p*-value**
Regression	10376408086	5	2075281617	52.21057924	3.576E-17
Residual	1709176777	43	39748297.13		
Total	12085584863	48			

Note: In the source of variation, Regression is ESS, Residual is RSS, and Total is TSS.

Since the *p* value of the estimated *F* value is so virtually zero, we can conclude that collectively all the slope coefficients are not equal to zero, multicollinearity among some variables notwithstanding.

(***f***) Following the format of Table 4.3, we obtain:

Source of variation	***SS***	***df***	***MSS***	***F***	***p*-value**
Regression	$0.8586\,(\sum y_i^2)$	5	$\frac{0.8586(\sum y_i^2)}{5}$	52.21	0.0000
Residual	$(1-0.8586)\,(\sum y_i^2)$	43	$\frac{(1-0.8586)(\sum y_i^2)}{43}$		
Total	$\sum y_i^2$	48			

Note: $\sum y_i^2 =$ 550,977,068,808.00

The conclusion is the same as before.

4.19. (***a***) It seems that way, because a straight line *reasonably* fits the residuals. There may be a slight departure from normality, but that doesn't typically have a large impact on the regression results.

(***b***) No, it is not significant: The *p* value of obtaining the Anderson-Darling A^2 value of 0.468 or greater is about 23 percent. This supports the conclusion in (*a*) that the error term is normally distributed. See the discussion on normal probability plots in Chapter 3.

(*c*) The mean value is zero and the variance is 0.2575 (Divide the sum of the squared residuals by $n - 3 = 25$, since there are $n = 28$ observations in Table 1-1, on which the regression is based). Any minor differences between your regression output and the regression shown in the book are due to rounding.

4.20. Here are the raw data for calculations:

Dependent variable	**Explanatory variable(s)**	**RSS**
Auction price	None	4,803,756.7
Auction price	Age	2,245,713.7
Auction price	Number of bidders	4,059,311.8
Auction price	Age, number of bidders	525,462.2

In all the cases the total sum of squares is 4,803,756.7.

Note: The RSS can easily be obtained from the *EViews* regression outputs for the above regressions.

We compare the first model that has no explanatory variables since price is regressed only on the intercept ($\text{RSS}_r = 4{,}803{,}756.7$) with the model with all the explanatory variables ($\text{RSS}_{ur} = 525{,}462.2$). Applying the F formula given in this question, we obtain:

$$F = \frac{(4{,}803{,}756.7 - 525{,}462.2)/2}{(525{,}462.2)/(32-3)} = \frac{2{,}139{,}147.25}{18{,}119.38} \approx 118.058$$

This F value is about the same as in Equation (4.57), save the rounding errors.

4.21 **(a)** We compare the model that does not include *population* as an explanatory variable ($\text{RSS}_r = 74{,}658{,}917.2$) with the model that does include it ($\text{RSS}_{ur} = 43{,}364{,}140$). Applying the F formula given in this question, we obtain:

$$F = \frac{(74{,}658{,}917.2 - 43{,}364{,}140)/1}{(43{,}364{,}140)/(38-3)} = \frac{31{,}294{,}777.2}{1{,}238{,}975.43} \approx 25.2586$$

Since this F statistic is so large, with a p-value close to 0, we can be assured that there is a significant difference in the two models, indicating that population does have a significant impact on the model.

(b) The new regression results are as follows:

$$\begin{aligned} percapEduc_i &= -67.166 + 0.0584\, percapGDP \\ se &= (46.257) \quad (0.004) \\ t &= (-1.452) \quad (16.019) \\ p-value &= (0.155) \quad\quad (0.000) \end{aligned}$$

$R^2 = 0.8770;\ \bar{R}^2 = 0.8736;\ F = 256.611;\ p-value \text{ of } F = 0.000$

This model seems to explain a bit less of the variability in per capita Education, but on the whole this model seems similar to the one presented in example 4.5. Also, in both models the intercept term does not seem to be significant.

4.22 (a) The regression results are as follows:

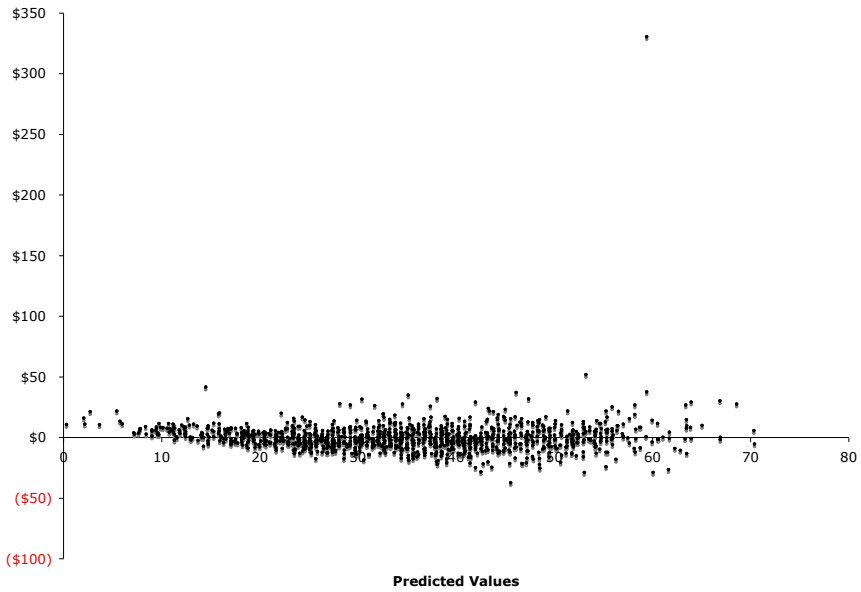
$$\begin{aligned} \text{Price}_i &= -40.3952 + 1.1928 Food + 1.7536 Decor + 1.1135 Service \\ se &= (2.6458) \quad (0.1594) \quad\quad (0.1004) \quad\quad (0.1929) \\ t &= (-15.2672) \quad (7.4838) \quad\quad (17.4722) \quad\quad (5.7716) \\ p-value &= (0.000) \quad\quad (0.000) \quad\quad (0.000) \quad\quad (0.000) \end{aligned}$$

$R^2 = 0.4929;\ \bar{R}^2 = 0.4916;\ F = 401.0392;\ p-value \text{ of } F = 0.000$

Since all the p-values are essentially 0.000, all 3 independent variables are statistically significant for predicting price.

(b) No, the normal probability plot appears to be fine.

(c) Residual plot is as follows:



Since there appears to be a highly significant outlier (a Japanese restaurant called Urasawa), it is a little hard to tell about the shape of the graph. It does appear, however, that there may be some significant heteroscedasticity.