8.5. Refer to Equations (4.21), (4.22), (4.25), and (4.27). Let $x_{3i} = 2x_{2i}$. Show why it is impossible to estimate these equations.

8.6. What are the theoretical consequences of imperfect multicollinearity?

8.7. What are the practical consequences of imperfect multicollinearity?

8.8. What is meant by the variance inflation factor (VIF)? From the formula (8.14), can you tell the least possible and the highest possible value of the VIF?

8.9. Fill in the gaps in the following sentences:

a. In cases of near multicollinearity, the standard errors of regression coefficients tend to be _______ and the t ratios tend to be _______.

b. In cases of perfect multicollinearity, OLS estimators are _______ and their variances are _______.

c. Ceteris paribus, the higher the VIF is, the higher the _______ of OLS estimators will be.

8.10. State with reasons whether the following statements are true or false:

a. Despite perfect multicollinearity, OLS estimators are best linear unbiased estimators (BLUE).

b. In cases of high multicollinearity, it is not possible to assess the individual significance of one or more partial regression coefficients.

c. If an auxiliary regression shows that a particular R_i^2 is high, there is definite evidence of high collinearity.

d. High pairwise correlations do not necessarily suggest that there is high multicollinearity.

e. Multicollinearity is harmless if the objective of the analysis is prediction only.

8.11. In data involving economic time series such as unemployment, money supply, interest rate, or consumption expenditure, multicollinearity is usually suspected. Why?

8.12. Consider the following model:

$$Y_t = B_1 + B_2X_t + B_3X_{t-1} + B_4X_{t-2} + B_3X_{t-3} + u_t$$

where Y = the consumption
X = the income
t = the time

This model states that consumption expenditure at time t is a linear function of income not only at time t but also of income in three previous time periods. Such models are called *distributed lag models* and represent what are called *dynamic models* (i.e., models involving change over time).

a. Would you expect multicollinearity in such models and why?

b. If multicollinearity is suspected, how would you get rid of it?

PROBLEMS

8.13. Consider the following set of hypothetical data:

Y:	−10	−8	−6	−4	−2	0	2	4	6	8	10
X_2:	1	2	3	4	5	6	7	8	9	10	11
X_3:	1	3	5	7	9	11	13	15	17	19	21

Suppose you want to do a multiple regression of Y on X_2 and X_3.

a. Can you estimate the parameters of this model? Why or why not?

b. If not, which parameter or combination of parameters can you estimate?

8.14. You are given the annual data in Table 8-5 for the United States for the period 1971 to 1986. Consider the following aggregate demand function for passenger cars:

$$\ln Y_i = B_1 + B_2 \ln X_{2t} + B_3 \ln X_{3t} + B_4 \ln X_{4t} + B_5 \ln X_{5t} + B_6 \ln X_{6t} + u_t$$

where ln = the natural log

a. What is the rationale for the introduction of both price indexes X_2 and X_3?

b. What might be the rationale for the introduction of the "employed civilian labor force" (X_6) in the demand function?

c. How would you interpret the various partial slope coefficients?

d. Obtain OLS estimates of the preceding model.

8.15. Continue with Problem 8.14. Is there multicollinearity in the previous problem? How do you know?

8.16. If there is collinearity in Problem 8.14, estimate the various auxiliary regressions and find out which of the X variables are highly collinear.

TABLE 8-5 DEMAND FOR NEW PASSENGER CARS IN THE UNITED STATES, 1971 TO 1986

Year	Y	X_2	X_3	X_4	X_5	X_6
1971	10227	112.0	121.3	776.8	4.89	79367
1972	10872	111.0	125.3	839.6	4.55	82153
1973	11350	111.1	133.1	949.8	7.38	85064
1974	8775	117.5	147.7	1038.4	8.61	86794
1975	8539	127.6	161.2	1142.8	6.16	85846
1976	9994	135.7	170.5	1252.6	5.22	88752
1977	11046	142.9	181.5	1379.3	5.50	92017
1978	11164	153.8	195.4	1551.2	7.78	96048
1979	10559	166.0	217.4	1729.3	10.25	98824
1980	8979	179.3	246.8	1918.0	11.28	99303
1981	8535	190.2	272.4	2127.6	13.73	100397
1982	7980	197.6	289.1	2261.4	11.20	99526
1983	9179	202.6	298.4	2428.1	8.69	100834
1984	10394	208.5	311.1	2670.6	9.65	105005
1985	11039	215.2	322.2	2841.1	7.75	107150
1986	11450	224.4	328.4	3022.1	6.31	109597

Notes: Y = New passenger cars sold (thousands), seasonally unadjusted.
X_2 = New cars Consumer Price Index (1967 = 100), seasonally unadjusted.
X_3 = Consumer Price Index, all items, all urban consumers (1967 = 100), seasonally unadjusted.
X_4 = Personal disposable income (PDI) ($, in billions), unadjusted for seasonal variation.
X_5 = Interest rate (percent), finance company paper placed directly.
X_6 = Employed civilian labor force (thousands), unadjusted for seasonal variation.

Source: Business Statistics, 1986, a Supplement to the *Current Survey of Business,* U.S. Department of Commerce.

8.17. Continuing with the preceding problem, if there is severe collinearity, which variable would you drop and why? If you drop one or more X variables, what type of error are you likely to commit?

8.18. After eliminating one or more X variables, what is your final demand function for passenger cars? In what ways is this "final" model better than the initial model that includes all X variables?

8.19. What other variables do you think might better explain the demand for automobiles in the United States?

8.20. In a study of the production function of the United Kingdom bricks, pottery, glass, and cement industry for the period 1961 to 1981, R. Leighton Thomas obtained the following results:[21]

$$1.\ \log Q = -5.04 + 0.887 \log K + 0.893 \log H$$
$$\text{se} = (1.40)\quad (0.087)\quad (0.137) \qquad R^2 = 0.878$$

$$2.\ \log Q = -8.57 + 0.0272t + 0.460 \log K + 1.285 \log H$$
$$\text{se} = (2.99)\quad (0.0204)\quad (0.333)\quad (0.324) \qquad R^2 = 0.889$$

where Q = the index of production at constant factor cost
K = the gross capital stock at 1975 replacement cost
H = hours worked
t = the time trend, a proxy for technology

The figures in parentheses are the estimated standard errors.

a. Interpret both regressions.
b. In regression (1) verify that each partial slope coefficient is statistically significant at the 5% level.
c. In regression (2) verify that the coefficients of t and $\log K$ are individually insignificant at the 5% level.
d. What might account for the insignificance of $\log K$ variable in Model 2?
e. If you were told that the correlation coefficient between t and $\log K$ is 0.980, what conclusion would you draw?
f. Even if t and $\log K$ are individually insignificant in Model 2, would you accept or reject the hypothesis that in Model 2 all partial slopes are simultaneously equal to zero? Which test would you use?
g. In Model 1, what are the returns to scale?

8.21. Establish Eqs. (8.12) and (8.13). (*Hint:* Find out the coefficient of correlation between X_2 and X_3, say, r_{23}^2.)

8.22. You are given the hypothetical data in Table 8-6 on weekly consumption expenditure (Y), weekly income (X_2), and wealth (X_3), all in dollars.

a. Do an OLS regression of Y on X_2 and X_3.
b. Is there collinearity in this regression? How do you know?
c. Do separate regressions of Y on X_2 and Y on X_3. What do these regressions reveal?
d. Regress X_3 on X_2. What does this regression reveal?
e. If there is severe collinearity, would you drop one of the X variables? Why or why not?

[21]See R. Leighton Thomas, *Introductory Econometrics: Theory and Applications*, Longman, London, 1985, pp. 244–246.

TABLE 8-6 HYPOTHETICAL DATA ON CONSUMPTION EXPENDITURE (Y), WEEKLY INCOME (X_2), AND WEALTH (X_3)

Y	X_2	X_3
70	80	810
65	100	1009
90	120	1273
95	140	1425
110	160	1633
115	180	1876
120	200	2252
140	220	2201
155	240	2435
150	260	2686

8.23. Utilizing the data given in Table 8-1, estimate Eq. (8.20) and compare your results.

8.24. Check that all R^2 values in Table 8-4 are statistically significant.

8.25. Refer to Problem 7.19 and the data given in Table 7-9. How would your answer to this problem change knowing what you now know about multicollinearity? Present the necessary regression results.

8.26. Refer to Problem 2.16. Suppose you regress ASP on GPA, GMAT, acceptance rate (%), tuition, and recruiter rating. A priori, would you face the multicollinearity problem? If so, how would you resolve it? Show all the necessary regression results.

8.27. Based on the quarterly data for the U.K. for the period 1990-1Q to 1998-2Q, the following results were obtained by Asteriou and Hall.[22] The dependent variable in these regressions is Log(IM) = logarithm of imports (t ratios in parentheses).

Explanatory variable	Model 1	Model 2	Model 3
Intercept	0.6318	0.2139	0.6857
	(1.8348)	(0.5967)	(1.8500)
Log(GDP)	1.9269	1.9697	2.0938
	(11.4117)	(12.5619)	(12.1322)
Log(CPI)	0.2742	1.0254	—
	(1.9961)	(3.1706)	0.1195
Log(PPI)	—	−0.7706	0.1195
		(−2.5248)	(0.8787)
Adjusted-R^2	0.9638	0.9692	0.9602

Notes: GDP = gross domestic product
CPI = Consumer Price Index
PPI = producer price index

[22]See Dimitrios Asteriou and Stephen Hall, *Applied Econometrics: A Modern Approach*, Palgrave/Macmillan, New York, 2007, Chapter 6. Note that these results are summarized from various tables given in that chapter.

a. Interpret each equation.
b. In Model 1, which drops Log(PPI), the coefficient of Log(CPI) is positive and significant at about the 5% level. Does this make economic sense?
c. In Model 3, which drops Log(CPI), the coefficient of Log(PPI) is positive but statistically insignificant. Does this make economic sense?
d. Model 2 includes the logs of both price variables and their coefficients are individually statistically significant. However, the coefficient of Log(CPI) is positive and that of Log(PPI) is negative. How would you rationalize this result?
e. Do you think multicollinearity is the reason why some of these results are conflicting? Justify your answer.
f. If you were told that the correlation between PPI and CPI is 0.9819, would that suggest that there is a multicollinearity problem?
g. Of the three models given above, which would you choose and why?

8.28. Table 8-7 on the textbook's Web site gives data on imports, GDP, and the Consumer Price Index (CPI) for the United States over the period 1975–2005. You are asked to consider the following model:

$$\ln Imports_t = \beta_1 + \beta_2 \ln GDP_t + \beta_3 \ln CPI_t + u_t$$

a. Estimate the parameters of this model using the data given in the table.
b. Do you suspect that there is multicollinearity in the data?
c. Regress:
(1) $\ln Imports_t = A_1 + A_2 \ln GDP_t$
(2) $\ln Imports_t = B_1 + B_2 \ln CPI_t$
(3) $\ln GDP_t = C_1 + C_2 \ln CPI_t$

On the basis of these regressions, what can you say about the nature of multicollinearity in the data?
d. Suppose there is multicollinearity in the data but $\hat{\beta}_2$ and $\hat{\beta}_3$ are individually significant at the 5% level and the overall F test is also significant. In this case, should we worry about the collinearity problem?

8.29. Table 8-8 on the textbook's Web site gives data on new passenger cars sold in the United States as a function of several variables.
a. Develop a suitable linear or log-linear model to estimate a demand function for automobiles in the United States.
b. If you decide to include all the regressors given in the table as explanatory variables, do you expect to face the multicollinearity problem? Why?
c. If you do expect to face the multicollinearity problem, how will you go about resolving the problem? State your assumptions clearly and show all calculations.

8.30. As cheese ages, several chemical processes take place that determine the taste of the final product. Table 8-9 on the textbook's Web site contains data on the concentrations of various chemicals in 30 samples of mature cheddar cheese and a subjective measure of taste for each sample. The variables *Acetic* and H2S are the natural logarithm of the concentration of acetic acid and hydrogen sulfide, respectively. The variable *lactic* has not been log-transformed.
a. Draw a scatterplot of the four variables.
b. Do a bivariate regression of taste on acetic and H2S and interpret your results.

of a given MSA, there were fewer unmarried females per 100 unmarried males in the MSA. In other words, based on this sample regression function, it appeared to the student that if he wanted to live in an MSA in which marriage was relatively more prevalent, he would need to relocate to an MSA with a lower median income (which likely also indicates a lower cost-of-living).

Based on these findings, our former student was well on his way to completing a successful empirical research project.

LOOKING AHEAD TO CHAPTER 7

While this chapter has introduced multiple linear regression analysis, there are still a number of issues that need to be addressed if we are to get the maximum value out of this very valuable tool. First, suppose that we are interested in determining the linear relationship between the dependent variable and a quality, such as a person's sex, rather than a quantity, such as years of work experience. Second, suppose that we suspect that the true relationship between the dependent variable and a given independent variable is actually quadratic rather than linear. Finally, suppose that we are interested in estimating the elasticity between the dependent variable and a given independent variable, which is defined as the percentage change in the dependent variable that results from a given percentage change in the independent variable. Chapter 7 introduces methods for dealing with such concerns in multiple linear regression analysis.

Problems

6.1 Suppose that you have a data set that has student-level information on college GPA, hours studied, SAT score, and high school GPA and that you obtain the following correlation matrix.

	College GPA	Hours Studied	SAT Score	HS GPA
College GPA	1			
Hours Studied	0.7817	1		
SAT score	0.4433	0.1848	1	
HS GPA	0.3957	0.3275	0.1376	1

a. What information does this matrix tell you about these data? Be as specific as possible.

b. What information is not available in this correlation matrix?

6.2 Suppose you run the following three regressions with intercepts:

- A simple regression of *IQ* on *education* to obtain the slope coefficient $\hat{\delta}_1$.
- A simple regression of *log(wage)* on *education* to obtain the slope coefficient $\tilde{\beta}_1$.
- A multiple regression of *log(wage)* on *education* and *IQ* to obtain the slope coefficients $\hat{\beta}_1$ and $\hat{\beta}_2$, respectively.

a. Under what two conditions are $\tilde{\beta}_1$ and $\hat{\beta}_1$ equal? Put the conditions in terms of a subset of $\hat{\delta}_1$, $\tilde{\beta}_1$, $\hat{\beta}_1$, and $\hat{\beta}_2$.

b. Draw two Venn diagrams that illustrate your answer to part (a).

c. Draw a Venn diagram to explain why $\tilde{\beta}_1$ and $\hat{\beta}_1$ are generally not equal.

6.3 Suppose you are interested in estimating the ceteris paribus relationship between y and x_1. For this purpose, you can collect data on two control variables, x_2 and x_3. Let $\tilde{\beta}_1$ be the simple linear regression estimate from y on x_1, and let $\hat{\beta}_1$ be the multiple regression estimate from y on x_1, x_2, and x_3.

a. Assume that x_1 is not correlated with x_2 but that x_2 and x_3 have a large partial effect on y. Would you expect $\tilde{\beta}_1$ and $\hat{\beta}_1$ to be similar or very different? Explain, and draw a Venn diagram to illustrate your answer.

b. If x_1 is highly correlated with x_2 and x_3 but x_2 and x_3 have no partial effect on y, would you expect $\tilde{\beta}_1$ and $\hat{\beta}_1$ to be similar or very different? Explain, and draw a Venn diagram to illustrate your answer.

6.4 I know that y depends linearly on x, but I am not sure whether or not it also depends on another variable z. A friend suggests that I should regress y on x first, calculate the residuals, and then see whether they are correlated with z. What advice would you offer your friend?

6.5 Medicorp Company, which sells medical supplies to health care providers, is reviewing its new bonus program. The program provides bonuses to salespeople based on performance. Medicorp management wants to know whether bonuses paid in 1995 had an effect on 1996 sales. You have been asked to help them develop a regression model to explain sales. You have a random sample of 50 of Medicorp's sales territories containing the following information on each territory:

ADV	Amount spent on advertising in 1996 ($1,000)
BONUS	Total amount of bonuses paid in 1995 ($1,000)
RELPRICE	Medicorp's average price relative to the industry average
COMPET	Largest competitor's sales ($1,000)
SALES	Medicorp's sales in 1996 ($1,000)

The partial Excel results are shown here:

SSExplained = 1,237,904

MSUnexplained = 24,665

ANOVA

	df	SS	MS	F	Significance F
Regression	____	____	____	____	0.0001
Residual	____	____	____		
Total	____	____			

	Coefficients	Standard Error	*t* Stat	*p*-value
Intercept	889.1	162.1		0.0001
ADV	10.8	1.8		0.0001
BONUS	11.6	3.02		0.0003
RELPRICE	−122	71.65		0.0955
COMPET	−0.18	0.15		0.2364

a. Complete the ANOVA table.

b. Find R^2. Explain what the number tells you about this regression model.

c. What should R^2 be if all variations in SALES have been explained by our model?

d. Test whether SALES is related to any of the X's (the overall significance of the model). Use a significance level of 0.01.

e. Test whether SALES is related to BONUS. Use a significance level of 0.05.

f. Explain in the context of this problem what the numbers for the following coefficients mean: ADV, BONUS, RELPRICE.

g. Find a 95 confidence interval for BONUS. What parameter does this confidence interval bracket?

h. Management expects the following in 1997 for the Raleigh sales territory:

Amount spent on advertising: 60 ($1,000)

Amount paid in bonuses in 1997: 45 ($1,000)

Average price relative to the industry average: 0.7

Largest competitor's sales: 100 ($1,000)

Forecast 1997 sales for the Raleigh sales territory.

6.6 Suppose you are interested in predicting the price of a laptop computer based on its various features.

Price: the price of the laptop in dollars

Speed: speed of the laptop in megahertz

Charge: time (in minutes) the battery takes to charge

The partial Excel results are shown here:

SSUnexplained = 1698387.21

MSExplained = 554446.062

ANOVA

	df	SS	MS	F	Significance F
Regression	____	____	____	____	0.00479
Residual	____	____	____		
Total	21	____			

	Coefficients	Standard Error	t Stat	p-value
Intercept	1500.6007	1085.6889		
Speed	10.1506	14.5019		
Charge	1.6284	4.4568		

a. Complete the ANOVA table.

b. Find R^2. Explain what the number tells you about this regression model.

c. Test whether risk is related to any of the X's (the overall significance of the model). Use a significance level of 0.05.

d. Test whether price is related to speed. Use a significance level of 0.05. What does this result imply?

e. Explain in the context of this problem what the numbers for the following coefficients mean: speed and charge.

f. Find a 95 confidence interval for speed. What parameter does this confidence interval bracket?

g. What is the standard error of the regression?

h. Write out the regression equation.

i. Estimate the price of a laptop with a speed of 33 megahertz and the charge lasts 305 minutes.

6.7 Management proposed the following regression model to predict sales at a fast-food outlet.

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \varepsilon_i$$

where $x_{1,i}$ = Number of competitors within one mile

$x_{2,i}$ = Population within one mile (1000s)

$x_{3,i}$ = Population beyond one mile but within five miles (1000s)

y_i = Sales ($1000s)

The following sample regression function was developed after 20 outlets were surveyed.

$$\hat{y}_i = 10.1 - 4.2x_{1,i} + 6.8x_{2,1} + 2.3x_{3,1}$$

a. How does the number of competitors within one mile affect sales?

b. How does the population within 1 mile affect sales?

c. How does the population beyond 1 mile but within 5 miles affect sales?

d. Predict sales for a store with two competitors, a population of 8,000 within 1 mile, and population beyond 1 mile but within 5 miles of 13,000.

6.8 A 10-year study conducted by the American Heart Association provided data on how age, blood pressure, and smoking relate to the risk of strokes. Assume that the regression results pertain to data on the following.

Risk: the probability that the patient will have a stroke over the next 10-year period

Age: age of the person

Pressure: blood pressure

Family: number of strokes the person's parents have had

The partial Excel results are shown here:

ANOVA					
	df	SS	MS	F	Significance F
Regression	3	3660.74	1220.25	36.82	2.06E-07
Residual	16	530.21	33.14		
Total	19	4190.95			

	Coefficients	Standard Error	*t* Stat	*p*-value
Intercept	−91.76	15.22		
Age	1.08	0.17		
Pressure	0.25	0.05		
Family	8.74	3		

a. What is the R^2 of the regression model? What does it mean?

b. What is the standard error of the regression? What does it mean?

c. Explain in the context of this problem what the numbers for the following coefficients mean: Age, Pressure, Family.

d. Test the overall significance of the regression model at a 5 percent significance level.

e. Is pressure statistically significant in explaining risk? Test this hypothesis at a 10 percent significance level.

f. Find a 95 percent confidence interval for age. What does this interval mean, and what parameter does it bracket?

g. Estimate the risk of a heart attack for a person who is 70 years old, with a blood pressure of 165, and whose parents have had one stroke.

h. Can you see a potential problem with predicting using this regression model?

6.9 You are trying to determine how the price of a home in Florida is related to the square footage of the house, the number of bathrooms, how nice the house is, and whether the house has a pool. Because you have taken Econ 301, you propose the following model

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,1} + \beta_3 x_{3,1} + \varepsilon_i$$

where x_1 = Square footage

x_2 = Number of bathrooms

x_3 = A niceness rating expressed as an integer from 1 to 7

y = Price of the house ($1000s)

The following estimated regression equation was developed after a sample of 80 homes was taken.

$$\hat{y}_i = 24.9760 + 0.0526 \times x_{1,i} + 10.0430 \times x_{2,1} + 10.0420 \times x_{3,1} + \varepsilon_i$$