The statistical analyses (or, if you wish, econometrics) that we are going to use in this text are really rather simple. Our two major objectives are also simple:

1. We want you to be able to set up a regression equation that could subsequently be estimated by using one of the readily available regression packages.
2. We want you to be able to use the output of a regression to examine those economic issues that are of interest to the manager of an enterprise.

Hence, in terms of the field of study known as econometrics, we will concentrate our attention on helping you avoid what are called *specification errors.* In simple terms, this means that we will show you how to set up an estimation equation that is appropriate for the use to which it is to be put. Specification errors—such as excluding important explanatory variables or using an inappropriate form for the equation—are serious; they can result in the estimates being *biased.*

In addition to specification errors, there are other problems that can be encountered in regression analysis. These problems, which are more difficult than the material we want to cover in this text, are reviewed briefly in the appendix to this chapter.

4.8 SUMMARY

This chapter set forth the basic principles of regression analysis: estimation and testing for statistical significance. The emphasis of the chapter was on explaining how to interpret the results of regression analysis, rather than on the mathematics of regression analysis. A mathematical derivation of the statistical techniques is presented in the appendix at the end of the chapter.

The simple linear regression model relates a dependent variable to a single explanatory variable in a linear fashion: $Y = a + bX$. The parameter a is the Y-intercept: the value of Y when X is 0. The parameter b is the slope of the regression line; it measures the rate of change in Y as X changes ($\Delta Y/\Delta X$). Because the variation in Y is affected not only by variation in X but also by various random effects, we cannot predict exactly the actual value of Y. Thus you should interpret the regression equation as giving the average or expected value of Y for any particular value of X.

Parameter estimates are obtained by choosing values of a and b that minimize the sum of the squared residuals. The residual is the difference between the actual value of Y and the fitted value of Y ($Y_i - \hat{Y}_i$). This method of estimating a and b is called the method of least-squares. The estimated regression line, $Y = \hat{a} + \hat{b}X$, is called the *sample regression line.* The sample regression line is an estimate of the true regression line.

The estimates $\hat{a}$ and $\hat{b}$ do not, in general, equal the true values of a and b. Because $\hat{a}$ and $\hat{b}$ are computed from the data in the random sample, the estimates themselves are random variables. Statisticians have shown that the distribution of values that the estimates might take is centered around the true value of the parameter. An estimator is *unbiased* if the mean value of the estimator is equal to the true value of the parameter. The method of least-squares produces unbiased estimates of a and b.

It is the randomness of the parameter estimates that necessitates testing for statistical significance. Just because the estimate $\hat{b}$ is not equal to 0 does *not* mean the true value of b is not actually equal to 0. Even when b *does* equal 0, it is still possible that the sample of Y and X values will produce a least-squares estimate $\hat{b}$ that is different from 0. It is necessary to determine statistically if there is sufficient evidence in the sample to indicate that Y is truly related to X (i.e., $b \neq 0$). This is called testing for statistical significance.

A t-test can be used to test for statistical significance of parameter estimates. To test for statistical significance of an individual parameter estimate, the researcher must first determine the level of significance for the test. The significance level of a test is the probability of finding a parameter estimate to be significantly different from 0 when, in fact, b is 0 (a Type I error). One minus the significance level is the level of confidence of the test. The choice of a significance or confidence level is in most cases rather arbitrary; lower (higher) levels of significance (confidence), other things equal, are more desirable. Traditionally, either a 0.01, 0.02, 0.05, or 0.10 level of significance is selected, which reflects the analyst's willingness to tolerate, at most, a 1, 2, 5, or 10 percent probability of finding a parameter to be significant when it is not truly significant. The appropriate significance level is determined by the analyst on the basis of the cost of making an error.

Once the level of significance (or confidence) is chosen, performing a t-test is straightforward. A t-test is based on the fact that the larger the absolute value of the t-ratio, $t = \hat{b}/S_{\hat{b}}$, the more probable it is that the true value of b is not equal to 0 ($S_{\hat{b}}$ is the standard error of the parameter estimate). So first compute the t-ratio as defined. Then find the critical t-value in the t-table. Locate the critical t-value with $n - k$ degrees of freedom for the chosen level of significance. If the absolute value of the t-ratio is greater than the critical t-value, $\hat{b}$ is statistically significant at the chosen level of significance. If the absolute value is less than the critical t-value, $\hat{b}$ is not statistically significant. If $\hat{b}$ is significant at the 0.05 level of significance, either of two equivalent statements can be made: (1) The probability of incorrectly finding $\hat{b}$ to be significant is less than 5 percent or (2) you can be at least 95 percent confident that the t-test will *not* find statistical significance when there is none ($b = 0$).

An alternative method of assessing the statistical significance of parameter estimates is to treat as statistically significant only those parameter estimates whose p-values are smaller than the maximum acceptable significance level. Most regression software now calculates a p-value for each parameter estimate. The p-value gives the exact (or minimum) level of significance for a parameter estimate.

To measure how well the sample regression line fits the data, the R^2 statistic (also called the coefficient of determination) is computed. The R^2 measures the fraction of the total variation in Y that is explained by the variation in X. The value of R^2 ranges from 0 (the regression equation explains none of the variation in Y) to 1 (the regression equation explains all the variation in Y). A high R^2 indicates Y and X are highly correlated, and the scatter diagram tightly fits the sample regression line.

The F-statistic is used to test whether the equation as a whole explains a significant amount of the variation in Y. To test whether the overall equation is significant, the F-statistic is compared with the critical F-value with $k - 1$ and $n - k$ degrees of freedom and the chosen level of significance. If the value for the calculated F-statistic exceeds the critical F-value, the regression equation is statistically significant. Alternatively, if the p-value for the F-statistic is smaller than the acceptable level of significance, then the equation as a whole is statistically significant.

Multiple regression models use two or more explanatory variables to explain the variation in the dependent variable. The coefficient on each of the explanatory variables measures the degree of variation in Y associated with variation in that explanatory variable, holding all other explanatory variables constant. As in the case of simple regression, each coefficient is tested for significance by using the t-test. The degree of significance for each coefficient is given by its p-value. The F-statistic is used to test the overall equation for significance. The R^2 measures the fit of the equation.

Many economic relations of interest to managers are *nonlinear* in nature. Two types of nonlinear models are presented in this chapter: (1) quadratic regression models and (2) log-linear regression models. The quadratic regression model is appropriate when the curve fitting the scatter diagram is either U-shaped or inverted-U-shaped. The quadratic equation, $Y = a + bX + cX^2$, is transformed into a linear form by computing a new variable, $Z = X^2$, and substituting for X^2 to get a linear form: $Y = a + bX + cZ$. A second type of nonlinear model presented in this chapter is the log-linear model. In this nonlinear model, the dependent variable is related to one or more explanatory variables in a multiplicative fashion. The log-linear model for two explanatory variables takes the form $Y = aX^bZ^c$. A particularly useful feature of this specification is that the parameters b and c are elasticities. For example, b measures the percent change in Y that results when X changes by 1 percent. By taking natural logarithms, the logarithm of Y can be expressed as a linear function of the logarithms of the explanatory variables: $\ln Y = \ln a + b \ln X + c \ln Z$. Once this transformation is made, estimation and tests of statistical significance proceed as usual.

We must emphasize again that all the statistics needed for regression analysis are automatically computed when using a computerized regression routine. The purpose of this chapter was to show you how to interpret and use the regression statistics produced by the computer.

TECHNICAL PROBLEMS

1. A simple linear regression equation relates R and W as follows:

$$R = a + bW$$

   a. The explanatory variable is ______ , and the dependent variable is ______.
   b. The slope parameter is ______ , and the intercept parameter is ______.
   c. When W is zero, R equals ______.
   d. For each one unit increase in W, the change in R is ______ units.
2. Regression analysis is often referred to as least-squares regression. Why is this name appropriate?
3. Regression analysis involves estimating the values of parameters and testing the estimated parameters for significance. Why must parameter estimates be tested for statistical significance?
4. Evaluate the following statements:
   a. "The smaller the standard error of the estimate, $S_{\hat{b}}$, the more accurate the parameter estimate."
   b. "If $\hat{b}$ is an unbiased estimate of b, then $\hat{b}$ equals b."
   c. "The more precise the estimate of $\hat{b}$ (i.e., the smaller the standard error of $\hat{b}$), the higher the t-ratio."
5. The linear regression in problem 1 is estimated using 26 observations on R and W. The least-squares estimate of b is 40.495, and the standard error of the estimate is 16.250. Perform a t-test for statistical significance at the 5 percent level of significance.
   a. There are ______ degrees of freedom for the t-test.
   b. The value of the t-statistic is ______. The critical t-value for the test is ______.
   c. Is $\hat{b}$ statistically significant? Explain.
   d. The p-value for the t-statistic is ______. (*Hint:* In this problem, the t-table provides the answer.) The p-value gives the probability of rejecting the hypothesis that ______ ($b = 0$, $b \neq 0$) when b is truly equal to ______. The confidence level for the test is ______ percent.
   e. What does it mean to say an estimated parameter is statistically significant at the 5 percent significance level?
   f. What does it mean to say an estimated parameter is statistically significant at the 95 percent confidence level?
   g. How does the level of significance differ from the level of confidence?
6. Ten data points on Y and X are employed to estimate the parameters in the linear relation $Y = a + bX$. The computer output from the regression analysis is the following:

DEPENDENT VARIABLE: Y	R-SQUARE	F-RATIO	P-VALUE ON F
OBSERVATIONS: 10	0.5223	8.747	0.0187

VARIABLE	PARAMETER ESTIMATE	STANDARD ERROR	T-RATIO	P-VALUE
INTERCEPT	800.0	189.125	4.23	0.0029
X	−2.50	0.850	−2.94	0.0187

a. What is the equation of the sample regression line?

b. Test the intercept and slope estimates for statistical significance at the 1 percent significance level. Explain how you performed this test, and present your results.

c. Interpret the p-values for the parameter estimates.

d. Test the overall equation for statistical significance at the 1 percent significance level. Explain how you performed this test, and present your results. Interpret the p-value for the F-statistic.

e. If X equals 140, what is the fitted (or predicted) value of Y?

f. What fraction of the total variation in Y is explained by the regression?

7. A simple linear regression equation, $Y = a + bX$, is estimated by a computer program, which produces the following output:

DEPENDENT VARIABLE: Y	R-SQUARE	F-RATIO	P-VALUE ON F
OBSERVATIONS: 25	0.7482	68.351	0.0106

VARIABLE	PARAMETER ESTIMATE	STANDARD ERROR	T-RATIO	P-VALUE
INTERCEPT	325.24	125.09	2.60	0.0160
X	0.8057	0.2898	2.78	0.0106

a. How many degrees of freedom does this regression analysis have?

b. What is the critical value of t at the 5 percent level of significance?

c. Test to see if the estimates of a and b are statistically significant.

d. Discuss the p-values for the estimates of a and b.

e. How much of the total variation in Y is explained by this regression equation? How much of the total variation in Y is unexplained by this regression equation?

f. What is the critical value of the F-statistic at a 5 percent level of significance? Is the overall regression equation statistically significant?

g. If X equals 100, what value do you expect Y will take? If X equals 0?

8. Evaluate each of the following statements:

a. "In a multiple regression model, the coefficients on the explanatory variables measure the percent of the total variation in the dependent variable Y explained by that explanatory variable."

b. "The more degrees of freedom in the regression, the more likely it is that a given t-ratio exceeds the critical t-value."

c. "The coefficient of determination (R^2) can be exactly equal to one only when the sample regression line passes through each and every data point."

9. A multiple regression model, $R = a + bW + cX + dZ$, is estimated by a computer package, which produces the following output:

DEPENDENT VARIABLE: R	R-SQUARE	F-RATIO	P-VALUE ON F
OBSERVATIONS: 34	0.3179	4.660	0.00865

VARIABLE	PARAMETER ESTIMATE	STANDARD ERROR	T-RATIO	P-VALUE
INTERCEPT	12.6	8.34	1.51	0.1413
W	22.0	3.61	6.09	0.0001
X	−4.1	1.65	−2.48	0.0188
Z	16.3	4.45	3.66	0.0010

a. How many degrees of freedom does this regression analysis have?

b. What is the critical value of t at the 2 percent level of significance?

c. Test to see if the estimates of a, b, c, and d are statistically significant at the 2 percent significance level. What are the exact levels of significance for each of the parameter estimates?

d. How much of the total variation in R is explained by this regression equation? How much of the total variation in R is unexplained by this regression equation?

e. What is the critical value of the F-statistic at the 1 percent level of significance? Is the overall regression equation statistically significant at the 1 percent level of significance? What is the exact level of significance for the F-statistic?

f. If W equals 10, X equals 5, and Z equals 30, what value do you predict R will take? If W, X, and Z are all equal to 0?

10. Eighteen data points on M and X are used to estimate the quadratic regression model $M = a + bX + cX^2$. A new variable, Z, is created to transform the regression into a linear form. The computer output from this regression is

DEPENDENT VARIABLE: M	R-SQUARE	F-RATIO	P-VALUE ON F
OBSERVATIONS: 18	0.6713	15.32	0.0002

VARIABLE	PARAMETER ESTIMATE	STANDARD ERROR	T-RATIO	P-VALUE
INTERCEPT	290.0630	53.991	5.37	0.0001
X	−5.8401	2.1973	−2.66	0.0179
Z	0.07126	0.01967	3.62	0.0025

a. What is the variable Z equal to?

b. Write the estimated quadratic relation between M and X.

c. Test each of the three estimated parameters for statistical significance at the 2 percent level of significance. Show how you performed these tests and present the results.

d. Interpret the p-value for $\hat{c}$.

e. What is the predicted value of M when X is 300?

11. Suppose Y is related to R and S in the following nonlinear way:

$$Y = aR^bS^c$$

a. How can this nonlinear equation be transformed into a linear form that can be analyzed by using multiple regression analysis?

Sixty-three observations are used to obtain the following regression results:

DEPENDENT VARIABLE: LNY		R-SQUARE	F-RATIO	P-VALUE ON F
OBSERVATIONS: 63		0.8151	132.22	0.0001

VARIABLE	PARAMETER ESTIMATE	STANDARD ERROR	T-RATIO	P-VALUE
INTERCEPT	−1.386	0.83	−1.67	0.1002
LNR	0.452	0.175	2.58	0.0123
LNS	0.30	0.098	3.06	0.0033

b. Test each estimated coefficient for statistical significance at the 5 percent level of significance. What are the exact significance levels for each of the estimated coefficients?

c. Test the overall equation for statistical significance at the 5 percent level of significance. Interpret the p-value on the F-statistic.

d. How well does this nonlinear model fit the data?

e. Using the estimated value of the intercept, compute an estimate of a.

f. If $R = 200$ and $S = 1{,}500$, compute the expected value of Y.

g. What is the estimated elasticity of R? Of S?

APPLIED PROBLEMS

1. The director of marketing at Vanguard Corporation believes that sales of the company's Bright Side laundry detergent (S) are related to Vanguard's own advertising expenditure (A), as well as the combined advertising expenditures of its three biggest rival detergents (R). The marketing director collects 36 weekly observations on S, A, and R to estimate the following multiple regression equation:

$$S = a + bA + cR$$

where S, A, and R are measured in dollars per week. Vanguard's marketing director is comfortable using parameter estimates that are statistically significant at the 10 percent level or better.

a. What sign does the marketing director expect a, b, and c to have?

b. Interpret the coefficients a, b, and c.

The regression output from the computer is as follows:

DEPENDENT VARIABLE: S	R-SQUARE	F-RATIO	P-VALUE ON F	
OBSERVATIONS: 36	0.2247	4.781	0.0150	
VARIABLE	PARAMETER ESTIMATE	STANDARD ERROR	T-RATIO	P-VALUE
INTERCEPT	175086.0	63821.0	2.74	0.0098
A	0.8550	0.3250	2.63	0.0128
R	−0.284	0.164	−1.73	0.0927

c. Does Vanguard's advertising expenditure have a statistically significant effect on the sales of Bright Side detergent? Explain, using the appropriate *p*-value.

d. Does advertising by its three largest rivals affect sales of Bright Side detergent in a statistically significant way? Explain, using the appropriate *p*-value.

e. What fraction of the total variation in sales of Bright Side remains unexplained? What can the marketing director do to increase the explanatory power of the sales equation? What other explanatory variables might be added to this equation?

f. What is the expected level of sales each week when Vanguard spends $40,000 per week and the combined advertising expenditures for the three rivals are $100,000 per week?

2. In his analysis of California's Proposition 103 (see Illustration 4.3), Benjamin Zycher notes that one of the most important provisions of this proposition is eliminating the practice by insurance companies of basing premiums (in part) on the geographic location of drivers. Prohibiting the use of geographic location to assess the risk of a driver creates a substantial implicit subsidy from low-loss counties to high-loss counties, such as Los Angeles, Orange, and San Francisco counties. Zycher hypothesizes that the percent of voters favoring Proposition 103 in a given county (V) is inversely related to the (average) percentage change in auto premiums (P) that the proposition confers upon the drivers of that county.

   The data in the table below were presented by Zycher to support his contention that V and P are inversely related:

County	Percent for Proposition 103 (V)	Change in average premium (P)
Los Angeles	62.8	−21.4
Orange	51.7	−8.2
San Francisco	65.2	−0.9
Alameda	58.9	+8.0
Marin	53.5	+9.1
Santa Clara	51.0	+11.8
San Mateo	52.8	+12.6
Santa Cruz	54.2	+13.0
Ventura	44.8	+1.4
San Diego	44.1	+10.7
Monterey	41.6	+15.3
Sacramento	39.3	+16.0
Tulare	28.7	+23.3
Sutter	32.3	+37.1
Lassen	29.9	+46.5
Siskiyou	29.9	+49.8
Modoc	23.2	+57.6

Sources: California Department of Insurance and Office of the California Secretary of State.

Using the data in the table, we estimated the regression equation

$$V = a + bP$$

to see if voting behavior is related to the change in auto insurance premiums in a statistically significant way. Here is the regression output from the computer:

DEPENDENT VARIABLE: V		R-SQUARE	F-RATIO	P-VALUE ON F
OBSERVATIONS: 17		0.7399	42.674	0.0001

VARIABLE	PARAMETER ESTIMATE	STANDARD ERROR	T-RATIO	P-VALUE
INTERCEPT	53.682	2.112	25.42	0.0001
P	−0.528	0.081	−6.52	0.0001

a. Does this regression equation provide evidence of a statistically significant relation between voter support for Proposition 103 in a county and changes in average auto premiums affected by Proposition 103 in that county? Perform an F-test at the 95 percent level of confidence.

b. Test the intercept estimate for significance at the 95 percent confidence level. If Proposition 103 has no impact on auto insurance premiums in any given county, what percent of voters do you expect will vote for the proposition?

c. Test the slope estimate for significance at the 95 percent confidence level. If P increases by 10 percent, by what percent does the vote for Proposition 103 decline?

3. A security analyst specializing in the stocks of the motion picture industry wishes to examine the relation between the number of movie theater tickets sold in December and the annual level of earnings in the motion picture industry. Time-series data for the last 15 years are used to estimate the regression model

$$E = a + bN$$

where E is total earnings of the motion picture industry measured in dollars per year and N is the number of tickets sold in December. The regression output is as follows:

DEPENDENT VARIABLE: E	R-SQUARE	F-RATIO	P-VALUE ON F
OBSERVATIONS: 15	0.8311	63.96	0.0001

VARIABLE	PARAMETER ESTIMATE	STANDARD ERROR	T-RATIO	P-VALUE
INTERCEPT	25042000.0	20131000.0	1.24	0.2369
N	32.31	8.54	3.78	0.0023

a. How well do movie ticket sales in December explain the level of earnings for the entire year? Present statistical evidence to support your answer.

b. On average, what effect does a 100,000-ticket increase in December sales have on the annual earnings in the movie industry?

c. Sales of movie tickets in December are expected to be approximately 950,000. According to this regression analysis, what do you expect earnings for the year to be?

4. The manager of Collins Import Autos believes the number of cars sold in a day (Q) depends on two factors: (1) the number of hours the dealership is open (H) and (2) the number of salespersons working that day (S). After collecting data for two months (53 days), the manager estimates the following log-linear model:

$$Q = aH^bS^c$$

a. Explain how to transform this log-linear model into a linear form that can be estimated using multiple regression analysis.

The computer output for the multiple regression analysis is shown below:

DEPENDENT VARIABLE: LNQ	R-SQUARE	F-RATIO	P-VALUE ON F
OBSERVATIONS: 53	0.5452	29.97	0.0001

VARIABLE	PARAMETER ESTIMATE	STANDARD ERROR	T-RATIO	P-VALUE
INTERCEPT	0.9162	0.2413	3.80	0.0004
LNH	0.3517	0.1021	3.44	0.0012
LNS	0.2550	0.0785	3.25	0.0021

b. How do you interpret coefficients b and c? If the dealership increases the number of salespersons by 20 percent, what will be the percentage increase in daily sales?

c. Test the overall model for statistical significance at the 5 percent significance level.

d. What percent of the total variation in daily auto sales is explained by this equation? What could you suggest to increase this percentage?

e. Test the intercept for statistical significance at the 5 percent level of significance. If H and S both equal 0, are sales expected to be 0? Explain why or why not.

f. Test the estimated coefficient b for statistical significance. If the dealership decreases its hours of operation by 10 percent, what is the expected impact on daily sales?

STATISTICAL APPENDIX Further Development

Least-Squares Parameter Estimation

Consider the sample regression line

(1) $$\hat{Y}_i = \hat{a} + \hat{b}X_i$$

where $\hat{a}$ and $\hat{b}$ are estimates of the population parameters a and b. The goal of least-squares regression is to find estimates of $\hat{a}$ and $\hat{b}$ that minimize the sum of the squared residuals (denoted *ESS*) for the sample of n observations on Y and X:

(2) $$ESS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{a} - \hat{b}X_i)^2$$

Notice that *ESS* is a function of the estimates $\hat{a}$ and $\hat{b}$. To find the values of the estimates that minimize *ESS*, partially differentiate equation (2) with respect to $\hat{a}$ and $\hat{b}$, and then set the two partial derivatives equal to 0:

(3) $$\frac{\partial ESS}{\partial \hat{a}} = -2\,\Sigma\,(Y_i - \hat{a} - \hat{b}X_i) = 0$$

(4) $$\frac{\partial ESS}{\partial \hat{b}} = -2\,\Sigma\,X_i\,(Y_i - \hat{a} - \hat{b}X_i) = 0$$

Multiplying and rearranging terms:

(3a) $$\Sigma Y_i = n\hat{a} + \hat{b}\Sigma\,X_i$$

(4a) $$\Sigma Y_i X_i = \hat{a}\Sigma\,X_i + \hat{b}\Sigma\,X_i^2$$

Now solve (3a) and (4a) simultaneously for $\hat{a}$ and $\hat{b}$. To do this, multiply (3a) by ΣX_i and (4a) by n:

(3b) $$\Sigma X_i \Sigma Y_i = n\hat{a}\Sigma X_i + \hat{b}(\Sigma\,X_i)^2$$

(4b) $$n\,\Sigma Y_i \Sigma X_i = n\hat{a}\Sigma X_i + \hat{b}n\,\Sigma\,X_i^2$$

Next, subtract (3b) from (4b):

(5) $$n\,\Sigma Y_i X_i - \Sigma X_i \Sigma Y_i = \hat{b}[n\,\Sigma\,X^2 - (\Sigma\,X_i)^2]$$

and it follows that

(6) $$\hat{b} = \frac{n\Sigma X_i Y_i - \Sigma X_i \Sigma Y_i}{n\Sigma X_i^2 - (\Sigma X_i)^2}$$

where X and Y are the sample means of X and Y. After finding $\hat{b}$, $\hat{a}$ is calculated as

(7) $$\hat{a} = Y - \hat{b}X$$

To illustrate how the computer uses equations (6) and (7) to estimate a and b, Table 4A.1 shows the computations that will be done for you by a computer. Table 4A.1 calculates the parameter estimates for the sample regression line that best fits the data in Table 4.2.

Derivation of the Coefficient of Determination (R^2)

The coefficient of determination, denoted R^2, measures how well the overall equation explains the variation in the dependent variable Y. The total variation in Y can be attributed to one of two things: variation due to changes in the explanatory variable(s) or variation due to random influences. R^2 is derived by decomposing the variation in Y into these two component parts.

Statisticians measure the variation in the dependent variable as $Y_i - \overline{Y}$, the variation in Y about the sample mean ($\overline{Y}$). To motivate why the sample mean is used, recall that if Y_i does not vary at all in the sample (i.e., Y_i is constant for all observations), Y_i equals $\overline{Y}$ for every observation in the sample. The sample mean, then, provides a point of reference about which the variation in Y_i can be measured. The amount by which the value of Y predicted by the regression ($\hat{Y}_i$) deviates from the sample mean ($\overline{Y}$) is referred to as the explained variation and is denoted

TABLE 4A.1
Computing the Least-Squares Estimates $\hat{a}$ and $\hat{b}$ for the Sample of Seven Travel Agencies

Sales (Y_i)	Advertising expenditure (X_i)	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})$	$(X_i - \bar{X})(Y_i - \bar{Y})$
15,000	2,000	−3,143	9,878,551	−22,143	69,591,837
30,000	2,000	−3,143	9,878,551	−7,143	22,448,980
30,000	5,000	−143	20,408	−7,143	1,020,408
25,000	3,000	−2,143	4,591,837	−12,143	26,020,408
55,000	9,000	3,857	14,877,551	17,857	68,877,551
45,000	8,000	2,857	8,163,265	7,857	22,448,980
60,000	7,000	1,857	3,448,980	22,857	42,448,980
Σ 260,000	36,000		50,857,143		252,857,144

$\bar{Y} = 260{,}000/7 = 37{,}143.$
$\bar{X} = 36{,}000/7 = 5{,}143.$
$\hat{b} = \Sigma(X_i - \bar{X})(Y_i - \bar{Y})/\Sigma(X_i - \bar{X})^2 = 252{,}857{,}144/50{,}857{,}143 = 4.9719.$
$\hat{a} = \bar{Y} - \hat{b}\bar{X} = 37{,}143 - (4.9719 \times 5{,}143) = 11{,}573.$

$\hat{Y}_i - \bar{Y}$. The unexplained variation in Y_i is the residual amount, $Y_i - \hat{Y}_i$. Figure 4A.1 shows how the variation in Y_i is decomposed for one particular observation in the sample.

The total variation in Y_i in the sample is computed by squaring the total variation in Y_i and summing across all observations in the sample:

$$\text{Total variation} = \Sigma(Y_i - \bar{Y}_i)^2$$

Given the decomposition discussed above, the total variation in Y can also be expressed as the sum of the explained and unexplained variation in Y:

$$\begin{array}{ccccc} \text{Total} & & \text{Total} & & \text{Total} \\ \text{variation} & = & \text{explained} & + & \text{unexplained} \\ \text{in } Y & & \text{variation} & & \text{variation} \\ \Sigma(Y_i - \bar{Y})^2 & = & \Sigma(\hat{Y}_i - \bar{Y})^2 & + & \Sigma(Y_i - \hat{Y}_i)^2 \end{array}$$

The coefficient of determination measures the fraction of total variation explained by the regression:

$$R^2 = \frac{\text{Total explained variation}}{\text{Total variation in } Y} = \frac{\Sigma(\hat{Y}_i - \bar{Y})^2}{\Sigma(Y_i - \bar{Y}_i)^2}$$

Thus it follows that R^2 can vary from 0 to 1 in value.

Some Additional Problems in Regression Analysis

Multicollinearity

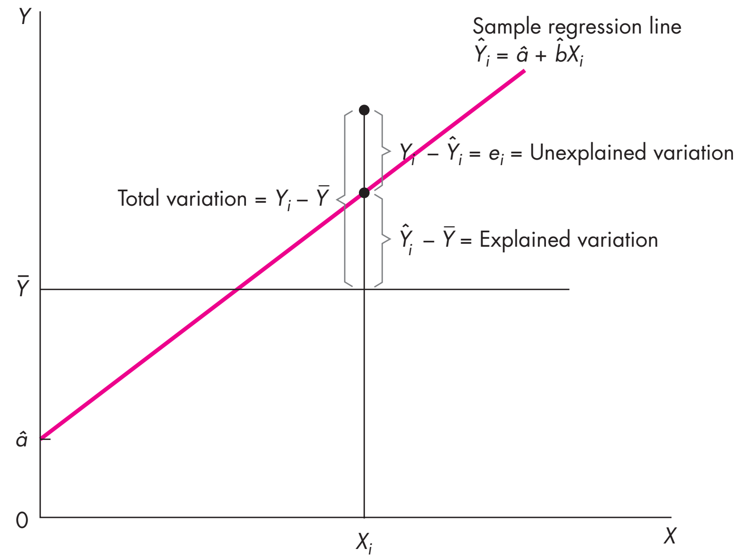
When using regression analysis, we assume that the explanatory (right-hand side) variables are linearly independent of one another. If this assumption is violated, we have the problem of *multicollinearity.* Under normal circumstances, multicollinearity will result in the estimated standard errors being larger than their true values. This means, then, that if multicollinearity exists, finding statistical significance will be more difficult. More specifically, if moderate multicollinearity is present, the estimate of the coefficient, $\hat{b}$, will be unbiased but the estimated standard error, $S_{\hat{b}}$, will be increased. Thus the t-coefficient, $t = \hat{b}/S_{\hat{b}}$, will be reduced, and it will be more difficult to find statistical significance.

Multicollinearity is not unusual. The question is what to do about it. As a general rule, the answer is *nothing.* To illustrate, consider the following function that denotes some true relation:

$$Y = a + bX + cZ$$

If X and Z are not linearly independent—if X and Z are collinear—the standard errors of the estimates for b and c will be increased. Shouldn't we just drop one? Not in the normal instance. If Z is an important explanatory variable, the exclusion of Z would be a *specification error* and would result in biased estimates of the coefficients—a much more severe problem.

FIGURE 4A.1
Decomposition of Total Variation in *Y*



Heteroscedasticity

The problem of *heteroscedasticity* is encountered when the variance of the error term is not constant. It can be encountered when there exists some relation between the error term and one or more of the explanatory variables—for example, when there exists a positive relation between X and the errors (i.e., large errors are associated with large values of X).

In such a case, the estimated parameters are still unbiased, but the standard errors of the coefficients are biased, so the calculated t-ratios are unreliable. This problem, most normally encountered in cross-section studies, sometimes can be corrected by performing a transformation on the data or equation. Otherwise, it becomes necessary to employ a technique called weighted least-squares estimation.

Autocorrelation

The problem of *autocorrelation*, associated with time-series data, occurs when the errors are not independent over time. For example, it could be the case that a high error in one period tends to promote a high error in the following period.

With autocorrelation (sometimes referred to as *serial correlation*) the estimated parameters are unbiased, but the standard errors are again biased, resulting in unreliability of the calculated t-ratios. Tests for determining if autocorrelation is present (most notably the Durbin-Watson test) are included in most of the available regression packages. Furthermore, most packages also include techniques for estimating an equation in the presence of autocorrelation.