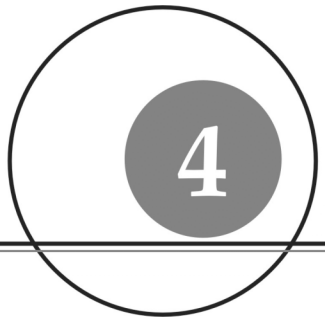


4 Basic Estimation Techniques

Learning Objectives

After reading Chapter 4 and working the problems for Chapter 4 in the textbook and in this Workbook, you should be able to:

- Set up a regression equation that can be estimated using a computerized regression routine.
- Interpret and understand how to use the computer output to investigate problems that are of interest to managers of a firm.

In order to accomplish these two goals, you must know how to:

- Specify a relation or model between a dependent variable and the appropriate independent variable(s) that can be estimated using regression techniques.
- Interpret the estimated parameters of regression equations.
- Determine whether estimated parameters are statistically significant using either a *t*-test or a *p*-value associated with the parameter estimate.
- Evaluate how well the regression equation "fits" the data by examining the R^2 statistic (also known as the coefficient of determination).
- Test for statistical significance of the whole regression equation using an *F*-test.
- Use linear regression techniques to estimate the parameters of two common nonlinear models: (1) quadratic regression models and (2) log-linear regression models.

Essential Concepts

1. A simple linear regression model relates a dependent variable Y to a single independent (or explanatory) variable X in a linear fashion

$$Y = a + bX$$

The *intercept parameter* (a) gives the value of Y at the point where the regression line crosses the Y-axis, which is the value of Y when X is zero. The *slope parameter* (b) gives the change in Y associated with a one-unit change in X ($b = \Delta Y / \Delta X$).

2. Because the variation in Y is affected not only by variation in X but also by various random effects as well, the actual value of Y cannot be predicted exactly. The regression equation is correctly interpreted as providing the average value, or the expected value, of Y for a given value of X.

3. 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*, and 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.

4. The estimates $\hat{a}$ and $\hat{b}$ do not, in general, equal the true values of a and b. Since $\hat{a}$ and $\hat{b}$ are computed using data from a random sample, the estimates themselves are random variables—the estimates would vary in value from one random sample to another random sample. 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 average value, or the expected value, of the estimator is equal to the true value of the parameter. The method of least-squares can produce unbiased estimates of a and b.

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

6. There are two ways to determine whether an estimated parameter is statistically significant. Either a t-test can be performed or the p-value for the parameter estimate can be examined.

7. To perform a t-test for significance, a 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 zero when, in fact, b is zero. This mistake is called a *Type I error*. Lower levels of significance, other things equal, are more desirable. One minus the level of significance is called the *level of confidence*.

   Once the level of significance is chosen, the t-ratio is computed as

$$t = \frac{\hat{b}}{S_{\hat{b}}}$$

   where $S_{\hat{b}}$ is the standard error of the estimate $\hat{b}$. Next, the critical value of t is found in the t-table at the end of your textbook. Choose the critical t-value with $n - k$ degrees of freedom for the desired level of significance, where n is the number of observations and k is the number of parameters being estimated. If the absolute value of the t-ratio is greater (less) than the critical t-value, then $\hat{b}$ is (is not) statistically significant.

8. 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. The p-value gives the exact level of significance for a parameter estimate, which is the probability of finding significance when none exists.

9. The coefficient of determination R^2 measures the percentage of the total variation in the dependent variable that is explained by the regression equation. The value of R^2 ranges from 0 to 1. A high R^2 indicates Y and X are highly correlated and the scatter diagram tightly fits the sample regression line.

10. The F-test is used to test for significance of the overall regression equation. The F-statistic from the computer printout is compared to the critical F-value obtained from the F-table at the end of your textbook. The critical F-value is identified by two separate degrees of freedom and the significance level. The first of the degrees of freedom is $k-1$ and the second is $n-k$. If the value for the calculated F-statistic (calculated by the computer) exceeds the critical F-value, the regression equation overall is statistically significant at the specified significance level. Alternatively, if the p-value for the F-statistic is smaller than the acceptable level of significance, the equation as a whole is statistically significant.

12. Multiple regression uses more than one explanatory variable to explain the variation in the dependent variable. The coefficient for each of the explanatory variables measures the change in Y associated with a one-unit change in that explanatory variable ($b = \Delta Y / \Delta X$).

13. Two types of nonlinear models can be easily transformed into linear models that can be estimated using linear regression analysis. These are quadratic regression models and log-linear regression models.

    (a) *Quadratic regression models* are appropriate when the curve fitting the scatter plot is either $\cup$-shaped or $\cap$-shaped. A quadratic equation, $Y = a + bX + cX^2$, can be transformed into a linear form by computing a new variable $Z = X^2$, which is then substituted for X^2 in the regression. Then, the regression equation to be estimated is $Y = a + bX + cZ$.

    (b) *Log-linear regression models* are appropriate when the relation takes the multiplicative exponential form: $Y = aX^bZ^c$. The equation is transformed by taking natural logarithms:

$$\ln Y = \ln a + b \ln X + c \ln Z$$

    The coefficients b and c are elasticities. For example, b measures the percent change in Y that results when X changes by 1 percent.

Matching Definitions

coefficient of determination (R^2)
critical value of *t*
cross-sectional data set
degrees of freedom
dependent variable
estimates
estimators
explanatory variables
fitted or predicted value
F-statistic
hypothesis testing
intercept parameter
level of confidence
level of significance
log-linear regression model
method of least-squares
multiple regression model
parameter estimation
parameters
population regression line
p-value
quadratic regression model
random error term
regression analysis
relative frequency distribution on $\hat{b}$
residual
sample regression line
scatter diagram
slope parameter
statistically significant
time series data set
t-ratio
true (or actual) relation
t-statistic
t-test
Type I error
unbiased estimator

1. _____ The coefficients in an equation that determine the exact mathematical relationship between the variables.
2. _____ The process of finding estimates of the numerical values of the parameters in an equation.
3. _____ The technique that uses data on variables to determine a mathematical equation that describes the relationship between the variables.
4. _____ The variable whose variation is to be explained.
5. _____ Variables thought to affect the value of the dependent variable.
6. _____ The parameter that gives the value of the dependent variable (Y) when the explanatory variable (X) is zero.
7. _____ A parameter that gives the change in the dependent variable per one-unit change in one of the explanatory variables.
8. _____ The unknown relationship that exists between Y and X and is to be discovered through regression analysis.
9. _____ Captures the effects of all the minor, unpredictable factors that cannot reasonably be accounted for in the hypothesized model.
10. _____ Data on explanatory and dependent variables that is collected over time.
11. _____ Data that is collected from many different firms or industries at one point in time.
12. _____ A graph that plots the value of the dependent variable against the value of the explanatory variable.

13. _____ The equation or line that represents the exact relationship between the independent variable and explanatory variables.

14. _____ The line that best fits the sample data.

15. _____ Method of fitting a line through a scatter of data that minimizes the sum of squared distance from each sample point and the fitted point.

16. _____ Values of Y obtained by entering a value for X into the least-squares regression line.

17. _____ The observed difference between the value estimated by the regression line and the actual value.

18. _____ The formulas by which the estimates of the intercept parameter and the slope parameters are calculated.

19. _____ The values of the intercept and slope parameters that are calculated using the formulas (i.e., the estimators) for the least-squares lines.

20. _____ Sample data contain sufficient evidence that the true value of a coefficient or parameter is not zero.

21. _____ A statistical technique for making probabilistic statements about the true value of a parameter.

22. _____ Graph showing the distribution of values that $\hat{b}$ might take given the random sample of observations on Y and X.

23. _____ An estimator which, on average, gives the true value of the parameter it seeks to estimate.

24. _____ Statistical test for testing the hypothesis that the true value of a parameter is equal to zero.

25. _____ The ratio of an estimated regression parameter divided by its standard error.

26. _____ The numerical value of the t-ratio.

27. _____ The value the t-statistic must exceed in order to reject the hypothesis that the true value of the parameter is equal to zero.

28. _____ Parameter estimate is found to be statistically significant when the true parameter value is equal to zero.

29. _____ The probability of making a Type I error.

30. _____ When the true value of a parameter is zero, this gives the probability of (correctly) failing to find statistical significance.

31. _____ The number of observations in a sample minus the number of parameters being estimated.

32. _____ The exact level of significance associated with a particular test statistic, such as a *t*-statistic or an *F*-statistic.

33. _____ The fraction of the total variation that is explained by the relationship with the independent variables.

34. _____ A statistic used to test whether the overall regression equation is statistically significant.

35. _____ A regression having more than one explanatory variable to explain the variation in the dependent variable.

36. _____ A regression that fits a $\cup$ - or $\cap$ -shaped pattern of data.

37. _____ A multiplicative nonlinear regression model in which the slope parameters are elasticities.

Study Problems

1. A simple linear regression equation relates *G* and *H* as follows:

$$G = a + bH$$

   a. The explanatory variable is _____, and the dependent variable is _____.

   b. The slope parameter is _____, and the intercept parameter is _____.

   c. When *H* is zero, *G* equals _____.

   d. For each one unit increase in *H*, the change in *G* is _____ units.

2. Using the statistical tables in your textbook, find the values of the appropriate test statistic in the following two situations:

   a. Testing for statistical significance (at the 10 percent level of significance) of the individual regression coefficients in the model

$$Y = a + bX + cZ + dR$$

   which is estimated using a time-series sample containing monthly observations over a two-year period.

   b. Testing the statistical significance (at the 95 percent level of confidence) of the overall regression equation

$$Z = a + bY + cX$$

   which is estimated using cross-section data on 21 firms.

3. The linear regression equation in question 1 above is estimated using 22 observations on *G* and *H*. The least-squares estimate of *b* is –210.4, and the standard error of the estimate is 80.92. Perform a *t*-test for statistical significance at the 2 percent level of significance.

   a. There are _____ degrees of freedom for this *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 0.017. The *p*-value gives the probability of rejecting the hypothesis that _____ ($b = 0, b \neq 0$) when b is truly equal to _____. The exact level of significance for $\hat{b}$ is _____ percent.

4. Thirty 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 shown at the top of the next page:

DEPENDENT VARIABLE:	Y	R-SQUARE	F-RATIO	P-VALUE ON F	
OBSERVATIONS:	30	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. The equation of the sample regression line is _____.

b. Test the intercept and slope estimates for statistical significance at the 5 percent significance level. The critical *t*-value is _____. The parameter estimate for *a* is _____, which _____ (is, is not) statistically significant. The parameter estimate for *b* is _____, which _____ (is, is not) statistically significant.

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

d. Test the overall equation for statistical significance at the 5 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 500, the fitted (or predicted) value of *Y* is _____.

f. The fraction of the total variation in *Y* explained by the regression is _____ percent.

5. A manager wishes to determine the relation between a firm's sales and its level of advertising in the newspaper. The manager believes sales (*S*) and advertising expenditures (*A*) are related in a nonlinear way:

$$S = a + bA + cA^2 + dA^3$$

Explain how to transform this nonlinear model into a linear regression model.

6. Suppose Y is related to X, W, and Z in the following nonlinear way:

$$Y = aX^{b}W^{c}Z^{d}$$

a. This nonlinear relation can be transformed into the linear regression model

_____ .

The computer output from the regression analysis is shown below.

DEPENDENT VARIABLE:	LNY	R-SQUARE	F-RATIO	P-VALUE ON F	
OBSERVATIONS:	25	0.7360	19.52	0.0001	
VARIABLE		PARAMETER ESTIMATE	STANDARD ERROR	T-RATIO	P-VALUE
INTERCEPT		3.1781	1.1010	2.89	0.0088
LNX		–2.173	0.6555	–3.32	0.0033
LNW		1.250	1.780	0.70	0.4902
LNZ		–0.8415	0.1525	–5.52	0.0001

b. At the 99 percent level of confidence, perform t-tests for statistical significance of $\hat{b}$, $\hat{c}$, and $\hat{d}$.

c. This regression leaves _____ percent of the variation in the dependent variable unexplained.

d. The estimated value of a is _____.

e. If $X = 10$, $W = 5$, and $Z = 2$, the expected value of Y is _____.

f. If Z decreases by 10 percent (all other things constant), Y will _____ (increase, decrease) by _____ percent.

g. If W decreases by 12 percent (all other things constant), Y will _____ (increase, decrease) by _____ percent.

7. A multiple regression model, $Y = a + bX + cX^2$, is estimated by creating a new variable named "X2" that equals X^2. A computer package produces the following output:

DEPENDENT VARIABLE:	Y	R-SQUARE	F-RATIO	P-VALUE ON F	
OBSERVATIONS:	27	0.8766	85.25	0.0001	
VARIABLE		PARAMETER ESTIMATE	STANDARD ERROR	T-RATIO	P-VALUE
INTERCEPT		8000.00	3524.0	2.27	0.0325
X		–12.00	4.50	–2.67	0.0135
X2		0.005	0.002	2.50	0.0197

a. The regression has _____ degrees of freedom.

b. Test to see if the estimates of a, b, and c are statistically significant at the 5 percent significance level.

c. The exact levels of significance of $\hat{a}, \hat{b}$, and $\hat{c}$ are _____, _____, and _____, respectively.

d. _____ percent of the total variation in Y is explained by the regression. _____ percent of the variation in Y is unexplained by the regression.

e. The critical value of the F-statistic at the 5 percent level of significance is _____. Is the overall regression equation statistically significant at the 5 percent level? The exact level of significance of the equation as a whole is _____ percent.

f. If X is equal to 1,200, then Y = _____.

Computer Problem

1. Use the following 12 observations on Y and X and a computer regression package, such as the Student Edition of Statistix 8, to work this computer problem.

Observation	X	Y
1	15	75
2	20	150
3	30	125
4	35	250
5	40	200
6	50	225
7	55	300
8	60	200
9	70	250
10	75	175
11	80	225
12	90	175

Run the appropriate regression to estimate the parameters of the linear model: $Y = a + bX$.

a. At the 5 percent level of significance, does X play a significant role in explaining the variation in Y? Explain.

b. What percentage of the variation in Y is explained by variation in the explanatory variable X? What percentage of the variation in Y is unexplained by this model?

c. Using the computer software package, plot a scatter diagram of the data. (Plot Y on the vertical axis and X on the horizontal axis.) By looking at the scatter diagram, can you see why the R^2 for the linear model is low or high? Explain.

Now run the appropriate regression to estimate the parameters of the curvilinear model: $Y = a + bX + cX^2$.

d. Does the curvilinear model seem more appropriate for this data? Explain carefully using the regression results for the curvilinear model.

Multiple Choice / True-False

1. Using regression analysis for a linear equation $Y = a + bX$, the objective is to
   a. estimate the parameters a and b.
   b. fit a straight line through the data scatter in such a way that the sum of the squared errors is minimized.
   c. estimate the variables Y and X.
   d. both a and b.
   e. all of the above.

2. In a linear regression equation of the form $Y = a + bX$, the slope parameter b shows
   a. $\Delta X / \Delta Y$.
   b. $\Delta X / \Delta b$.
   c. $\Delta Y / \Delta b$.
   d. $\Delta Y / \Delta X$.
   e. none of the above

3. In a linear regression equation of the form $Y = a + bX$, the intercept parameter a shows
   a. the value of Y when X is zero.
   b. the value of X when Y is zero.
   c. the amount that Y changes when X changes by one unit.
   d. the amount that X changes when Y changes by one unit.

In questions 4 – 10, use the following estimation results for $Y = a + bX$:

DEPENDENT VARIABLE:	Y	R-SQUARE	F-RATIO	P-VALUE ON F	
OBSERVATIONS:	22	0.4815	18.57	0.0003	
VARIABLE		PARAMETER ESTIMATE	STANDARD ERROR	T-RATIO	P-VALUE
INTERCEPT		276.320	105.060	2.63	0.0160
X		–24.291	5.636	–4 .31	0.0003

4. What is the critical value of the t-statistic at the 1 percent level of significance?
   a. 1.725
   b. 1.717
   d. 2.819
   e. 2.845

5. Which of the following statements is true?
   a. Since 5.1885 > 2.819, $\hat{b}$ is statistically significant.
   b. Since 4.31 > 2.819, $\hat{b}$ is statistically significant.
   c. Since 5.1885 > 2.845, $\hat{b}$ is statistically significant.
   d. Since 4.31 > 2.845, $\hat{b}$ is statistically significant.

6. Given the *t*-ratio calculated for $\hat{b}$, what would be the lowest level of significance that would allow the hypothesis $b = 0$ to be rejected in favor of the alternative hypothesis $b \neq 0$?
   a. 0.03 percent
   b. 1.0 percent
   c. 1.6 percent
   d. 48.57 percent
   e. 51.43 percent

7. What is the critical value for *F* at the 1 percent level of significance?
   a. 4.35
   b. 5.93
   c. 7.94
   d. 8.10

8. Which of the following is true?
   a. Since 4.35 < 21.177, the regression equation is statistically significant.
   b. Since 4.35 < 21.177, the regression equation is statistically significant.
   c. Since 21.177 > 8.10, the regression equation is statistically significant.
   d. Since 21.177 > 7.94, the regression equation is statistically significant.

9. R^2 tells us
   a. the amount of variation in *Y* that is unexplained.
   b. the percent of the variation in *X* that is explained.
   c. that 51.43 percent of the total variation in *Y* is explained by the regression.
   d. that 48.57 percent of the total variation in *Y* is explained by the regression.

10. If $X = 40$, then $Y =$ _____.
   a. –8 40.32
   b. –6 95.32
   c. 1,478.32
   d. 1,845.32
   e. 1,945.32

Questions 11 – 13 cover topics presented in the Appendix to Chapter 4 in the textbook.

11. Multicollinearity will most likely be a problem when
   a. time-series data are used.
   b. cross-section data are used.
   c. the explanatory variables are not independent.
   d. endogenous variables are not independent.

12. Autocorrelation will most likely be a problem when
    a. time-series data are used.
    b. cross-section data are used.
    c. the explanatory variables are not independent.
    d. endogenous variables are not independent.

13. Heteroscedasticity will most likely be a problem when
    a. time-series data are used.
    b. cross-section data are used.
    c. the explanatory variables are not independent.
    d. endogenous variables are not independent.

14. T F Generally, each and every data point in the sample lies on the sample regression line.

15. T F In the model $Y = a + bX$, a high R^2 tells us that the variation in Y is caused by the variation in X.

16. T F The amount of energy consumed each month at a specific factory is an example of a time-series data set.

17. T F If the regression equation is statistically significant, then all of the individual parameters are statistically significant.

18. T F If $\hat{b}$ is statistically significant at a 10 percent significance level, then $\hat{b}$ is far enough from zero that there is only a 10 percent probability that the true value of b equals zero.

19. T F The method of least-squares produces unbiased estimates of parameters; therefore, $\hat{b} = b$ when the least-squares technique is employed to estimate the parameters of a regression model.

20. T F The p-value gives the probability of making a Type I error, which is the probability of finding significance when none exists.

Answers

MATCHING DEFINITIONS

1. parameters
2. parameter estimation
3. regression analysis
4. dependent variable
5. explanatory variables
6. intercept parameter
7. slope parameter
8. true (or actual) relation
9. random error term
10. time series data set
11. cross-sectional data set
12. scatter diagram
13. population regression line
14. sample regression line
15. method of least-squares
16. fitted or predicted value
17. residual
18. estimators
19. estimates
20. statistically significant
21. hypothesis testing
22. relative frequency distribution on $\hat{b}$
23. unbiased estimator
24. *t*-test
25. *t*-ratio
26. *t*-statistic
27. critical value of *t*
28. Type I error
29. level of significance
30. level of confidence
31. degrees of freedom
32. *p*-value
33. coefficient of determination (R^2)
34. *F*-statistic
35. multiple regression model
36. quadratic regression model
37. log-linear regression model

STUDY PROBLEMS

1. a. *H*; *G*
   b. *b*; *a*
   c. *a*
   d. *b*

2. a. $t_{20} = 1.725$ (at the 10 percent significance level with 24 – 4 = 20 degrees of freedom)
   b. $F_{k-1,n-k} = F_{2,18} = 3.55$

3. a. 20
   b. –2.60 = –210.4 / 80.92; $t_{\text{critical}} = 2.528$

c. The estimated value of b is (just) statistically significant because the calculated value of the t-ratio, in absolute value, is greater than the critical value from the t-table: 2.60 > 2.528. There is no more than a 2% chance that the true value of b is zero when the t-ratio is –2.60.

d. $b = 0$; zero; 1.7 (There is exactly a 1.8% chance that the true value of b is zero when the t-ratio is –2.60.)

4. a. $\hat{Y} = 800.0 - 2.50X$

b. $t_{critical} = 2.048$; 800.0; is; –2.50; is

c. The p-value gives the probability of committing a Type I error; that is rejecting the hypothesis that a parameter's true value is zero when the parameter value really is zero. For the intercept parameter estimate, there is only a 0.29% chance that $a = 0$, given the t-ratio of 4.23. For the slope parameter estimate, there is only a 1.87% chance that $b = 0$, given the t-ratio of –2.94.

d. The critical value of the F-statistic is $F_{k-1, n-k} = F_{1, 28} = 4.20$. Since the calculated F-statistic, 8.747, exceeds the critical value of F, the equation is statistically significant. The p-value for the F-ratio indicates there is only a 1.82% chance the equation is *not* truly significant when the F-ratio is as large as 8.747.

e. $-450 = 800.0 + (-2.50)(500) = 800.0 - 1{,}250$

f. 52.23%

5. Two new variables must be computed and substituted for A^2 and A^3. Let $X = A^2$ and $Z = A^3$ so that the nonlinear relation can be written in linear form as: $S = a + bA + cX + dZ$.

6. a. $\ln Y = \ln a + b \ln X + c \ln W + d \ln Z$

b. The critical value of the t-statistic for $n - k = 25 - 4 = 21$ degrees of freedom and a 1 percent significance level is 2.831. When t-ratios are negative, their absolute values are used in their t-tests. Since $|t_{\hat{b}}|$ and $|t_{\hat{d}}|$ both exceed 2.831, the estimates of b and d are statistically significant. Since $t_{\hat{c}}$ is less than 2.831, the estimate of c is *not* statistically significant.

c. 26.4% of the variation in ln Y is unexplained.

d. The intercept estimate provides an estimate for log a, (i.e., 3.1781 is an estimate of the *natural log* of a.) Therefore, $\hat{a}$ is found by taking the antilog: $\hat{a} = e^{3.1781} = 24.00$. If the intercept estimate (of ln a) is statistically significant, then the estimate of a (24.00) is also statistically significant. Since the t-statistic for the intercept is 2.89 (= 3.1781/1.1010) exceeds 2.831 (barely), the estimate of a ($\hat{a} = 24$) is statistically significant.

e. When $X = 10$, $W = 5$, and $Z = 2$, $Y = 0.6724$ [$= 24(10)^{-2.1730}(5)^{1.2500}(2)^{-0.8415}$]

f. increase; 8.415 ($= 10 \times 0.8415$)

g. decrease; 15 ($= 12 \times 1.2500$)

7. a. $24 = 27 - 3$

b. The critical value of the t-statistic is 2.064 at the 5% level of significance with 24 degrees of freedom. Since all three calculated t-ratios (2.27, –2.67, and 2.50) exceed the critical t-value, all three parameter estimates are statistically significant at the 5% level of significance.

c. 3.25%; 1.35%; 1.97%

d. 87.66%; 12.34%

e. 3.40; The equation is statistically significant at the 5% level. The p-value for the F-statistic indicates the exact level of significance is much better than 5%: there is less than a 0.01% chance the equation as a whole is not statistically significant.

f. 800 ($= 8{,}000 - 12 \times 1{,}200 + 0.005 \times 1{,}200^2$)

COMPUTER PROBLEM

Your regression printout for the linear model $Y = a + bX$ should look like this:

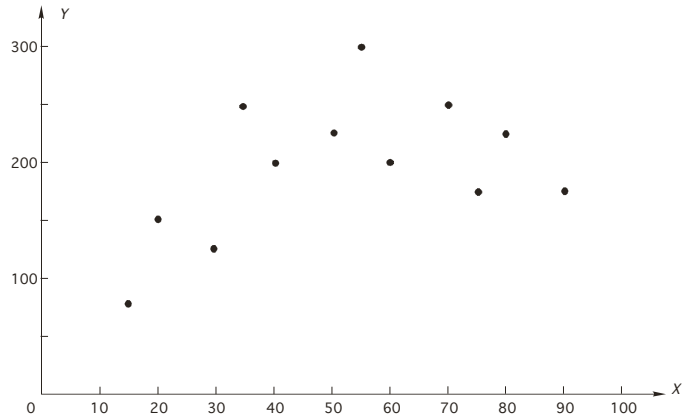
DEP. VARIABLE:	Y	R-SQUARE	F-RATIO	P-VALUE ON F
OBS:	12	0.2026	2.54	0.1420

VARIABLE	PARAMETER ESTIMATE	STD. ERROR	T-RATIO	P-VALUE
INTERCEPT	137.242	40.2942	3.41	0.0067
X	1.13402	0.71138	1.59	0.1420

1. a. No. The *t*-ratio for $\hat{b}$, 1.59, is less than the critical *t* (= 2.228) for a 5 percent level of significance and 10 ($= n - k = 12–10$) degrees of freedom. The *p*-value for $\hat{b}$ reveals that the exact level of significance is 14.20 percent.

   b. 20.26% of the variation in *Y* is explained by *X*, which is the only explanatory variable in the model. This leaves 79.74% of the variation in *Y* unexplained.

   c. Your scatter diagram should look like the following:



As you can see, a straight line fits the scatter of data points rather poorly, and this explains the low R^2 and *F*-statistic in the linear model.

Your regression printout for the curvilinear model $Y = a + bX + cX^2$ should look like:

DEP. VARIABLE:	Y	R-SQUARE	F-RATIO	P-VALUE ON F
OBS:	12	0.6379	7.93	0.0103

VARIABLE	PARAMETER ESTIMATE	STD. ERROR	T-RATIO	P-VALUE
INTERCEPT	–31.774	58.8211	–0.54	0.6021
X	9.31115	2.53695	3.67	0.0052
X2	–0.07900	0.02402	–3.29	0.0094

Note that a new variable, *X2*, was created by squaring *X* (*X2* = X^2).

   d. Yes. For the curvilinear regression, the *F*-ratio and R^2 are higher than the linear model, which indicates a better fit. Looking at the scatter diagram, you can also see a curvature to the data pattern. More important, however, X and X2 are both significant explanatory variables as $\hat{b}$ and $\hat{c}$ are significant at better than the 1 percent level, as indicated by their *p*-values. The lack of significance for the intercept estimate, $\hat{a}$, indicates the true curvilinear relation may pass through the origin (i.e., the true value of *a* is zero.)

MULTIPLE CHOICE / TRUE-FALSE

1. d Least-squares estimation is equivalent to fitting a line through a scatter of data points.

2. d The slope parameter b gives the rate of change in the dependent variable as the independent variable changes.

3. a The intercept parameter a gives the value of the dependent variable when the line crosses the axis on which the dependent variable is plotted.

4. d From the t-table at the end of your textbook, the critical value of t with $n - k = 22 - 2 = 20$ degrees of freedom and a 99% confidence level is 2.845.

5. d Since $\left|t_{\hat{b}}\right| > t_{\text{critical}}$, $\hat{b}$ is statistically significant.

6. a The p-value for $\hat{b}$ is 0.0003, which gives the lowest level of significance for which the hypothesis $b = 0$ can be rejected.

7. d For the 99 percent level of confidence, $F_{\text{n–k, k–1}} = F_{20,1} = 8.10$.

8. c Since $F_{\text{calculated}} > F_{\text{critical}}$, the regression equation is statistically significant.

9. c 51.43% of variation in Y is explained by the equation, while 48.57% is unexplained.

10. b –695.32 (= 276.32 –24.291×40)

11. c Multicollinearity arises when explanatory variables are correlated with one another.

12. a Autocorrelation is most common in time series data because random effects tend to "carry-over" from one time period to the next.

13. b Heteroscedasticity is most common in cross-section data.

14. F The sample regression line is the best-fitting line, but it cannot pass through every data point in a scatter diagram (unless all points lie perfectly on a straight line, which is an extremely unlikely occurrence).

15. F A high degree of correlation does not imply *causality*. Two variables may be highly correlated even though changes in one variable do not cause changes in the other.

16. T This is indeed a time-series data set. A cross-section data set would contain energy consumption for all factories in a given month.

17. F It is possible for the equation as a whole to be significant without all of the individual parameters being significant.

18. T The level of significance is the probability that b is actually zero when $\hat{b}$ is found to be statistically significant.

19. F Unbiasedness only means that if you collected many random samples and calculated $\hat{b}$ the estimates *on average* would be equal to the true value.

20. T This is the definition of p-values.

Homework Exercises

1. Thirty 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

DEPENDENT VARIABLE:	Y	R-SQUARE	F-RATIO	P-VALUE ON F	
OBSERVATIONS:	30	0.3301	13.79	0.0009	
VARIABLE		PARAMETER ESTIMATE	STANDARD ERROR	T-RATIO	P-VALUE
INTERCEPT		93.54	46.210	2.02	0.0526
X		–3.25	0.875	–3.71	0.0009

   a. The equation of the sample regression line is $\hat{Y}$ = _____.

   b. There are _____ degrees of freedom for the t-test. At the 1% level of significance, the critical t-value for the test is _____.

   c. At the 1% level of significance, $\hat{a}$ _____ (is, is not) significant, and $\hat{b}$ _____ (is, is not) significant.

   d. At the 2% level of significance, the critical t-value for a t-test is _____. At the 2% level of significance, $\hat{a}$ _____ (is, is not) significant, and $\hat{b}$ _____ (is, is not) significant.

   e. The p-value for $\hat{b}$ indicates that the exact level of significance is _____ percent, which is the probability of _____ _____.

   f. At the 1% level of significance, the critical value of the F-statistic is _____. The model as a whole _____(is, is not) significant at the 1% level.

   g. If X equals 500, the fitted (or predicted) value of Y is _____.

   h. The percentage of the total variation in Y *not* explained by the regression is _____ percent.

   i. Explain why it is necessary to assess the statistical significance of the parameter estimates.

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

$$Y = aR^bS^c$$

a. In order to estimate the parameters a, b, and c, the equation must be transformed into the form: _____.

Twenty-six observations are used to obtain the following regression results:

DEPENDENT VARIABLE:	LNY	R-SQUARE	F-RATIO	P-VALUE ON F	
OBSERVATIONS:	26	0.3647	4.21	0.0170	
VARIABLE		PARAMETER ESTIMATE	STANDARD ERROR	T-RATIO	P-VALUE
INTERCEPT		2.9957	0.3545	8.45	0.0001
LNR		2.34	0.87	2.69	0.0134
LNS		–0.687	0.334	–2.06	0.0517

b. There are _____ degrees of freedom for the t-test. At the 1% level of significance, the critical t-value for the test is _____.

c. At the 1% level of significance, $\hat{a}$ _____ (is, is not) significant, $\hat{b}$ _____ (is, is not) significant, and $\hat{c}$ _____ (is, is not) significant.

d. The estimated value of a is _____.

e. The p-value for $\hat{b}$ indicates that the exact level of significance is _____ percent, which is the probability of _____ _____.

f. At the 1% level of significance, the critical value of the F-statistic is _____. The model as a whole _____ (is, is not) significant at the 1% level.

g. If $R = 12$ and $S = 30$, the fitted (or predicted) value of Y is _____.

h. The percentage of the total variation in the dependent variable *not* explained by the regression is _____ percent.

i. If R increases by 14%, Y will increase by _____ percent.

j. A 6.87% increase in Y will occur if S _____ (increases, decreases) by _____ percent.

3. ***COMPUTER EXERCISE***

Use a computer regression package, such as the Student Edition of Statistix 8, to work this computer exercise.

In Illustration 4.2 on page 140 of the textbook, a linear regression model was employed to determine whether the auto insurance premiums (P) can be adequately explained differences in costs across counties in California. Two variables were used to measure costs: (1) the number of claims per thousand insured vehicles (N), and the average dollar amount of each bodily injury claim (C).

a. Estimate the following log-linear regression model using the data provided in Illustration 4.2.

$$P = aN^{b}C^{c}$$

Note: The bodily injury claim data in Illustration 4.2 are stored in a Statistix 8 data file, which is included on the Statistix 8 disk that came with your textbook.

b. In order to estimate the parameters a, b, and c, the equation must be transformed into the form: _____.

c. The estimated log-linear model of bodily injury claims is

$$\hat{P} = \_\_\_\_\_ N^{\_\_\_\_\_} C^{\_\_\_\_\_}$$

d. There are _____ degrees of freedom for the t-test. At the 2% level of significance, the critical t-value for the test is _____.

e. At the 2% level of significance, $\hat{a}$ _____ (is, is not) significant, $\hat{b}$ _____ (is, is not) significant, and $\hat{c}$ _____ (is, is not) significant.

f. If N increases by 6%, auto premiums (P) are predicted to increase by _____ percent. Show your work here:

g. According to the estimated equation, auto premiums would be expected to rise by 10% if the average amount of each claim (C) _____ (increases, decreases) by _____ percent. Show your work here:

h. Compare the F-ratios and R^2s of the linear and log-linear models. Which model does a better job overall of explaining the variation in auto premiums (P)? Briefly explain.