

- Answer 1: Using the standardized test statistic:
 Decision Rule: Reject H_0 if $F > F(0.05; 2, 22) = 3.44$
 Do not reject H_0 if $F \leq F(0.05; 2, 22) = 3.44$
 Test Statistic: $F = 64.83$
 Decision: Reject H_0
- Answer 2: Using the p value:
 Decision Rule: Reject H_0 if p value < 0.05
 Do not reject H_0 if p value ≥ 0.05
 Test Statistic: p value = 0.000
 Decision: Reject H_0
4. What conclusion can be drawn from the result of the F test for overall fit?
 Answer: At least one of the coefficients (β_1, β_2) is not equal to zero. In other words, at least one of the variables (x_1, x_2) is important in explaining the variation in y .

EXERCISES

1. **Cost Control.** Ms. Karen Ainsworth is an employee of a well-known accounting firm's management services division. She is currently on a consulting assignment to the Apex Corporation, a firm that produces corrugated paper for use in making boxes and other packing materials. Apex is consulting help to improve its cost control program, and Ms. Ainsworth is analyzing manufacturing costs to understand more fully the important influences on these costs. She has assembled monthly data on a group of variables, and she is using regression analysis to help assess how these variables are related to total manufacturing cost. The variables Ms. Ainsworth has selected to study, the data for which are contained in the file COST14 on the CD, are
- y , total manufacturing cost per month in thousands of dollars (COST)
 - x_1 , total production of paper per month in tons (PAPER)
 - x_2 , total machine hours used per month (MACHINE)
 - x_3 , total variable overhead costs per month in thousands of dollars (OVERHEAD)
 - x_4 , total direct labor hours used each month (LABOR)
- The data shown in Table 4.2 refer to the period January 2001 through March 2003. Ms. Ainsworth wants to use a cost function developed by means of
- a. What is the equation that is determined using all four explanatory variables?
 - b. Conduct the F test for overall fit of the regression. State the hypotheses to be tested, the decision rule, the test statistic, and your decision. Use a 5% level of significance. What conclusion can be drawn from the result of the test?
 - c. In the cost accounting literature, the sample regression coefficient corresponding to x_4 is regarded as an estimate of the true marginal cost of output associated with the variable x_4 . Find a point estimate of the true marginal cost associated with total machine hours per month. Also, find a 95% confidence interval estimate of the true marginal cost associated with total machine hours.
 - d. Test the hypothesis that the true marginal cost of output associated with total production of paper is 1.0. Use a 5% level of significance and a two-tailed test procedure. State the hypotheses to be tested, the decision rule, the test statistic, and your decision. What conclusion can be drawn from the result of the test?
 - e. What percentage of the variation in y has been explained by the regression?

TABLE 4.2 Data for Cost Control Exercise

COST	PAPER	MACHINE	OVERHEAD	LABOR
1102	530	218	112	325
1008	502	199	99	301
1227	616	249	126	376
1395	701	277	143	419
1710	838	363	191	682
1881	919	399	210	751
1924	939	411	216	813
1246	622	248	124	371
1255	626	259	127	383
1314	659	266	135	402
1557	740	334	181	546
1887	901	401	216	655
1204	610	238	117	351
1211	598	246	124	370
1287	646	259	127	387
1451	732	286	155	433
1828	891	389	208	878
1903	932	404	216	660
1997	964	430	233	694
1363	680	271	129	405
1421	723	286	146	426
1543	784	317	158	478
1774	841	376	199	601
1929	922	415	228	679
1317	647	260	126	378
1302	656	255	117	380
1388	704	281	142	429

Source: These data were created by Dr. Roger L. Wright, RLW Analytics, Inc., Somerville, CA, and are used (with modification) with his permission.

FIGURE 4.8 Regression Results for Cost Control Exercise.

Variable	Coefficient	Std Dev	T Stat	P Value
Intercept	51.72	21.70	2.38	0.026
PAPER	0.95	0.12	7.90	0.000
MACHINE	2.47	0.47	5.31	0.000
OVERHEAD	0.05	0.53	0.09	0.927
LABOR	-0.05	0.04	-1.26	0.223

Analysis of Variance		R-Sq = 99.9%		R-Sq(Adj) = 99.9%	
Source	DF	Sum of Squares	Mean Square	F Stat	P Value
Regression	4	2271423	567856	4629.17	0.000
Error	22	2699	123		
Total	26	2274122			

FIGURE 4.9 Regression Results for Salaries Exercise.

Variable	Coefficient	Std Dev	T Stat	P Value
Intercept	3179.5	383.4	8.29	0.000
EDUC	139.6	27.7	5.04	0.000
EXPER	1.5	0.7	2.13	0.036
TIME	20.6	6.2	3.35	0.001

Standard Error = 602.728 R-Sq = 30.2% R-Sq(Adj) = 27.9%

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Stat	P Value
Regression	3	13991247	4663749	12.98	0.000
Error	89	32323043	363281		
Total	92	46323290			

- f. What is the adjusted R^2 for this regression?
- g. Based on the regression equation, what actions might be taken to control costs?
2. Salaries. The file on the CD named HARRIS4 contains values of the following four variables for 93 employees of Harris Bank Chicago in 1977:
- y : beginning salary in dollars (SALARY)
 - x_1 : years of schooling at the time of hire (EDUC)
 - x_2 : number of months of previous work experience (EXPER)
 - x_3 : number of months after January 1, 1969, that the individual was hired (TIME)
- The regression results for the regression of SALARY on the three explanatory variables are shown in Figure 4.9. Use the results to help answer the following questions:
- What is the estimated regression equation relating SALARY to EDUC, EXPER, and TIME?
 - Conduct the F test for overall fit of the regression. Use a 5% level of significance. State the hypotheses to be tested, the decision rule, the test statistic, and your decision. What conclusion can be drawn from the result of the test?
 - Is education linearly related to beginning salary (after taking into account the effect of experience and time)? Perform the hypothesis test necessary to answer this question. State the hypotheses to be tested, the decision rule, the test statistic, and your decision. Use a 5% level of significance.
 - What percentage of the variation in salary has been explained by the regression?

4.4 COMPARING TWO REGRESSION MODELS

4.4.1 FULL AND REDUCED MODEL COMPARISONS USING SEPARATE REGRESSIONS

Thus far, two types of hypothesis tests for multiple regression models have been considered:

- A test of the overall fit of the regression:

$$H_0: \beta_1 = \beta_2 = \dots = \beta_K = 0$$

$$H_a: \text{At least one coefficient is not equal to zero}$$
- A test of the significance of each individual regression coefficient:

$$H_0: \beta_k = 0$$

$$H_a: \beta_k \neq 0$$

In multiple regression models, it also may be useful to test whether subsets of coefficients are equal to zero. In this section, a *partial F* test to test whether any subset of coefficients in a multiple regression equals zero is considered.

To set up this hypothesis test, consider the following regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_L x_L + \beta_{L+1} x_{L+1} + \dots + \beta_K x_K + \epsilon$$

Testing whether the variables x_{L+1}, \dots, x_K are useful in explaining any variation in y after taking account of the variation already explained by x_1, \dots, x_L can be viewed as a comparison of two regression models to determine whether it is worthwhile to include the additional variables. The two models for comparison are called the *full* and *reduced* models.

Full Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_L x_L + \beta_{L+1} x_{L+1} + \dots + \beta_K x_K + \epsilon$$

This is called the full model because all K explanatory variables of interest are included.

Reduced Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_L x_L + \epsilon$$

This is called the reduced model because the variables x_{L+1}, \dots, x_K have been removed.

The question to be answered is, "Is the full model significantly better than the reduced model at explaining the variation in y ?" This question can be formalized by setting up the following null and alternative hypotheses:

$$H_0: \beta_{L+1} = \dots = \beta_K = 0$$

$$H_a: \text{At least one of the coefficients } \beta_{L+1}, \dots, \beta_K \text{ is not equal to zero}$$

If the null hypothesis is not rejected, choose the reduced model; if the null hypothesis is rejected, at least one of x_{L+1}, \dots, x_K is contributing to the explanation of the variation in y and the full model is chosen as superior to the reduced.

To test the hypotheses (that is, to compare the full and reduced models), an F statistic is used. The F statistic can be written:

$$F = \frac{(SSE_R - SSE_F)/(K - L)}{SSE_F/(n - K - 1)}$$

where the subscript F stands for full model and the subscript R stands for reduced model.

Now consider what is being computed in the F statistic. If the full and reduced models are estimated, the regression output includes the error sum of squares for each of these regressions. In the F statistic, SSE_F refers to the error sum of squares from the full model output using all K explanatory variables. SSE_R refers to the error sum of squares from the reduced model output using only L explanatory variables. Recall that the error sum of squares represents the variation in y unexplained by the

FIGURE 4.14
MINITAB Regression of SALES on ADV, BONUS, MKTSHR, and COMPET.

The regression equation is
 $SALES = -594 + 2.51 ADV + 1.91 BONUS + 2.65 MKTSHR - 0.121 COMPET$

Predictor	Coef	SE Coef	T	P
Constant	-593.5	259.2	-2.29	0.033
ADV	2.5131	0.3143	8.00	0.000
BONUS	1.9059	0.7424	2.57	0.018
MKTSHR	2.651	4.636	0.57	0.574
COMPET	-0.1207	0.3718	-0.32	0.749

S = 93.7697 R-Sq = 85.9% R-Sq(adj) = 83.1%

Analysis of Variance		DF	SS	MS	F	P
Source	Regression	4	1073119	268280	30.51	0.000
	Residual Error	20	175855	8793		
	Total	24	1248974			

Source	DF	Seq SS	Fit	SE Fit	Residual	SE Resid
ADV	1	1012408	39.4	-187.0	-2.20R	
BONUS	1	55389				
MKTSHR	1	4394				
COMPET	1	927				

Unusual Observations
 Obs ADV SALES Fit SE Fit Residual SE Resid
 20 525 1159.3 1346.2 39.4 -187.0 -2.20R
 R denotes an observation with a large standardized residual.

The null hypothesis cannot be rejected. The variables x_3 and x_4 do not significantly improve the model's ability to explain sales.

EXERCISES

3. **Cost Control (continued).** Consider again the cost data from Exercises 4.1 and the regression results in Figure 4.8. Consider this output to be for the full model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + e$$

where y , x_1 , x_2 , x_3 , and x_4 were defined in the first exercise.

Now consider the reduced model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$$

Conduct the test to compare these two models. State the hypotheses to be tested, the decision

rule, the test statistic, and your decision. What conclusion can be drawn from the result of the test? The regression results for the reduced model can be found in Figure 4.15. Use a 5% level of significance.

4. **Salaries (continued).** Consider again the salary data from Exercise 4.2 and the regression results in Figure 4.9. Consider this output to be for the full model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + e$$

where y , x_1 , x_2 , and x_3 were defined in Exercise 4.2.

FIGURE 4.15
Regression Results for the Reduced Model in the Cost Control Exercise.

Variable	Coefficient	Std Dev	T Stat	P Value
Intercept	59.4318	19.6388	3.03	0.006
PAPER	0.9489	0.1101	8.62	0.000
MACHINE	2.3864	0.2101	11.36	0.000

Standard Error = 10.9835 R-Sq = 99.9% R-Sq(adj) = 99.9%

FIGURE 4.16
Regression Results for the Reduced Model in the Salaries Exercise.

Analysis of Variance		DF	Sum of Squares	Mean Square	F Stat	P Value
Source	Regression	2	2271227	1135613	9413.48	0.000
	Error	24	2895	121		
	Total	26	2274122			

Variable	Coefficient	Std Dev	T Stat	P Value
Intercept	3818.56	377.44	10.12	0.000
EDUC	128.09	29.70	4.31	0.000

Standard Error = 650.112 R-Sq = 17.0% R-Sq(adj) = 16.1%

Now consider the reduced model:

$$y = \beta_0 + \beta_1 x_1 + e$$

Conduct the test to compare these two models. State the hypotheses to be tested, the decision

rule, the test statistic, and your decision. What conclusion can be drawn from the result of the test? The regression results for the reduced model can be found in Figure 4.16. Use a 5% level of significance.

4.5 PREDICTION WITH A MULTIPLE REGRESSION EQUATION

As with simple regression, one of the possible goals of fitting a multiple regression equation is using it to predict values of the dependent variable. The two cases considered here are the same as in simple regression.

4.5.1 ESTIMATING THE CONDITIONAL MEAN OF Y GIVEN X_1, X_2, \dots, X_k

In this case, the goal is to estimate the point on the regression surface for specific values of the explanatory variables. For example, in the Meddicorp example (Example 4.1), consider the population regression equation

$$\mu_{y|x_1, x_2} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

FIGURE 4.23
Regression Results for
Wheat Export Exercise.

Variable	Coefficient	Std Dev	T Stat	P Value	
Intercept	3361.932	633.194	5.31	0.000	
EXCHANGE	1.869	4.223	0.44	0.659	
PRICE	-2413.837	846.480	-2.85	0.005	
Standard Error = 798.260					
R-Sq = 8.58 R-Sq(Adj) = 7.48					
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Stat	P Value
Regression	2	8117338	4058669	6.37	0.002
Error	132	8412922	637219		
Total	134	9230260			

- hypothesis test to answer this question and use a 5% level of significance. State the hypotheses to be tested, the decision rule, the test statistic, and your decision. What conclusion can be drawn from the result of the test?
- What percentage of the variation in the dependent variable has been explained by the regression?
 - Construct a 95% confidence interval estimate for the population regression coefficient of PRICE.
 - What is the value of the R^2 adjusted for degrees of freedom? What, if any, is the advantage of this number over the coefficient of determination? (Source: Data are from D. A. Beales and R. A. Bibbitt, "Forecasting Wheat Exports: Do Exchange Rates Really Matter?" *Journal of Business and Economic Statistics*, 5, 1987, pp. 397-406. Copyright 1987 by the American Statistical Association. Used with permission. All rights reserved.)
9. **Mortgage Rates.** The regression in Figure 4.24 is an attempt to develop an equation to forecast mortgage rates. The explanatory variables include prime rate and the one through six-period lagged values of prime rate. The forecaster initially believed that prime rate in some of the past time periods and possibly the current time period may have an effect on mortgage rates. After examining the regression, the forecaster notes the small t statistics (large p values) for all the variables included in the regression and concludes that the regression is basically worthless: None of the variables included—current or lagged—are helpful in predicting mortgage

rates. Do you agree or disagree? Justify your position.

10. **Dividends.** A random sample of 42 firms was chosen from the S&P 500 firms listed in the Spring 2003 Special Issue of *Business Week* (The Business Week Fifty Best Performers). The indicated dividend yield (DIVYIELD), the earnings per share (EPS), and the stock price (PRICE) were recorded for these 42 firms. These data are available on the CD in a file named DIV4. Run a regression using DIVYIELD as the dependent variable and EPS and PRICE as the independent variables. Use the output to answer the following questions:
- What is the sample regression equation relating DIVYIELD to PRICE and EPS?
 - What percentage of the variation of DIVYIELD has been explained by the regression?
 - Test the overall fit of the regression. Use a 10% level of significance. State the hypotheses to be tested, the decision rule, the test statistic, and your decision.
 - What conclusion can be drawn from the test result?
 - Is it necessary to test each coefficient individually to see if either PRICE or EPS is related to DIVYIELD? Why or why not? (Source: Copyright 2003, *Business Week*. Visit us at our Web site at www.businessweek.com for additional information.)
11. **Fuel Consumption.** The data file FUELCON4 on the CD contains the following variables for all 50 states plus the District of Columbia:

FIGURE 4.24
Mortgage Rate and
Prime Rate Regression.

Variable	Coefficient	Std Dev	T Stat	P Value	
Intercept	5.3188	0.5203	10.22	0.000	
PRMR1	0.6266	0.5047	1.24	0.223	
PRMR2	-0.1148	0.8281	-0.14	0.890	
PRMR3	-0.2286	0.8286	-0.28	0.783	
PRMR4	-0.2137	0.8295	-0.38	0.706	
PRMR5	-0.2468	0.8286	-0.30	0.767	
PRMR6	-0.0368	0.8262	-0.04	0.965	
PRMR6	0.6845	0.4947	1.38	0.170	
Standard Error = 0.7587					
R-Sq = 29.7% R-Sq(Adj) = 23.7%					
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Stat	P Value
Regression	7	19.9288	2.8470	4.95	0.000
Error	82	47.1977	0.5756		
Total	89	67.1265			

- FUELCON: Per capita fuel consumption in gallons
 DRIVERS: The ratio of licensed drivers to private and commercial motor vehicles registered
 HWYMILES: The number of miles of federally funded highways
 GASTAX: The tax per gallon of gasoline in cents
 INCOME: The average household income in dollars
- Run the regression with FUELCON as the dependent variable and the other four variables as independent variables. Use the output to help answer the following questions:
- What is the estimated regression equation?
 - Test the overall fit of the regression. Use a 5% level of significance. Be sure to state your decision rule, test statistic value, and your decision. What conclusion can you draw from the result of the test?
 - What percentage of the variation in FUELCON has been explained by the regression?
 - Are there any variables that appear to be unnecessary in the regression? Justify your answer.
12. **Pricing Communications Nodes.** The cost of adding a new communications node at a location not currently included on the network was of concern to a major Fort Worth manufacturing company. To try to predict the price of new communications nodes, data were obtained on a sample of existing nodes. The installation cost (COST) and the number of ports (NUMPORTS) available for access in each existing node were readily available. Data on two additional characteristics of communications nodes were also obtained: bandwidth (BANDWIDTH) and port speed (PORTSPEED). These data are shown in Table 4.6 and are available on the CD in a file named COMMNODE4.
- The network administrator wants to develop a method of estimating the cost of new nodes in a quick and fairly accurate manner. You have been asked to help in this project. Using the data available, develop an equation to help in the

TABLE 4.6 Data for Communications Modes Exercise

Cost	Number of Ports	Bandwidth	Port Speed
52,388	68	58	653
51,761	52	179	499
50,221	44	123	422
36,095	32	38	307
27,500	16	29	154
57,088	56	141	538
54,475	56	141	538
33,969	28	48	269
31,309	24	29	230
23,444	24	10	115
24,269	12	56	499
53,479	52	131	192
33,543	20	38	230
33,056	24	29	230

pricing of new communications nodes. Justify your choice of equation.

Start with all three explanatory variables in the equation. Do you encounter any problems when you estimate this equation? Request variance inflation factors for the variables included in the regression. What do the VIFs tell you?

13. Prime Rate. The file named PRIME4 on the CD contains monthly prime rates for the time period from January 1988 through December 2002.

(These data are from the web site www.fedreserve.gov and are from the Federal Reserve Bank of St. Louis.) Develop an extrapolative model to forecast the prime rate for each month in 2003. Find the actual rates for each month in 2003 and compare them to your forecasts. How well did your model do? (How will you measure the accuracy of your forecasts?)

14. Graduation Rates. Kiplinger's *Personal Finance* provides information on the best public and private college values. Some of the variables included in this issue are as follows. All are based on the most recent available data.

GRADRATE4	the percentage of students who earned a bachelor's degree in four years (expressed as a percentage)
ADMINSRATE	admission rate expressed as a percentage
SFACRATIO	student faculty ratio
AVGDEBT	average debt at graduation

This information is included in a file named COLLEGE4 on the CD for 195 schools. Only schools listed are included. (Using the graduation rate as the dependent variable, use the techniques discussed in this chapter to develop an equation to explain four year graduation rates.)

(Source: Used by permission from the November and December 2003 issues of Kiplinger's *Personal Finance* Copyright © 2003 The Kiplinger Washington Editors, Inc. Visit our website at www.kiplinger.com for further information.)

15. Absenteeism. The ABX Company is interested in conducting a study of the factors that affect absenteeism among its production employees.

After reviewing the literature on absenteeism and interviewing several production supervisors and a number of employees, the researcher in charge of the project defined the variables shown in Figure 4.25. Then a sample of 77 employees was randomly selected, and the data contained in the file on the CD named ABSENT4 was collected. The dependent variable is absenteeism. The other variables are considered possible explanatory variables.

Use the procedures discussed in Chapters 3 and 4 to identify factors that may be related to absenteeism. Write down your final model and justify your choice of variables in the model. Check to see if your choice of variables and the coefficient estimates make intuitive sense. How much variation in absenteeism has been explained? What does this tell you? Does your model give you some sense of which

FIGURE 4.25 Absenteeism Study Variables.

Variable	Description
1. Absentism (ABSENT):	The number of distinct occasions that the worker was absent during 2003. Each occasion consists of one or more consecutive days of absence. An index ranging from 0 to 100.
2. Job Complexity (COMPLX):	Base hourly pay rate in dollars.
3. Base Pay (PAV):	Number of complete years with the company on December 31, 2003.
4. Seniority (SENIOR):	Employee's age on December 31, 2003.
5. Age (AGE):	Determined by employee response to the question: "How many individuals other than yourself depend on you for most of their financial support?"
6. Dependents (DEPEND):	

Source: These data were created by Dr. Roger L. Wright, R.W. Analytics, Inc., Sonoma, CA, and are used (with modification) with his permission.

employees might be absent most often? If so, which ones? What might be done to reduce absenteeism?

16. Fanfare. Fanfare International, Inc. designs, distributes, and markets ceiling fans and lighting fixtures. The company's product line includes 120 basic models of ceiling fans and 138 compatible fan light kits and table lamps. These products are marketed to over 1000 lighting showrooms and electrical wholesalers that supply the remodeling and new construction markets. The product line is distributed by a sales organization of 58 independent sales representatives.

In the summer of 1994, Fanfare decided it needed to develop forecasts of future sales to help determine future sales force needs, capital expenditures, and so on. The data file named FAN4 on the CD contains data on the following variables:

SALES	= total monthly sales in thousands of dollars
ADEX	= advertising expense in thousands of dollars
MTRATE	= mortgage rate for 30-year loans (%)
HSTARTS	= housing starts in thousands of units

The data are monthly and cover the period from July 1990 through May 1994. (Note: These data have been modified as requested by the company to provide confidentiality.)

As a consultant to Fanfare, your job is to find a causal regression model to forecast future sales. Use the techniques discussed in Chapters 3 and 4

to help you decide which variables you should include in the equation and which should be omitted. Justify your choices. How well do you believe the equation you developed will do at forecasting future sales? What additional analyses might you use to examine forecasting ability?

Now use the techniques discussed in Chapters 3 and 4 to build an extrapolative model to forecast sales. Generate forecasts from both the causal model and the extrapolative model. What are the benefits and drawbacks of each of the models? How could you compare the forecasting ability of the two models?

17. Major League Baseball. What factor is most important in building a winning baseball team?

Some might argue for a high batting average. Or it might be a team that hits for power as measured by the number of home runs. On the other hand, many believe that it is quality pitching as measured by the earned run average of the team's pitchers. The file MLB4 on the CD contains data on the following variables for the 30 major league baseball teams during the 2002 season:

WINS	= number of games won for each team
HR	= number of home runs hit by each team
BA	= average batting average for each team
ERA	= earned run average for each team

Using WINS as the dependent variable, use scatterplots and regression to investigate the relationship of the other variables to WINS. Use the variables to build a multiple regression

model to explain WINS. Interpret what your model tells you about a successful baseball team.

(Source: Courtesy of the *For Worth Star-Telegram*.)

18. **NBA.** The following data were obtained from the *For Worth Star-Telegram* and refer to the 2002–2003 National Basketball Association (NBA) season. The data are included in a file on the CD named NBA4 for all 29 NBA teams:

wins (WINS)
 field goals attempted (FGA)
 field goals made (FGM)
 field goals attempted for opponents (FGAOP)
 field goals made for opponents (FGMOP)
 three-point field goals attempted (TFGA)
 three-point field goals made (TFGM)
 three-point field goals attempted for opponents (TFGAOP)
 three-point field goals made for opponents (TFGMOP)
 offensive rebounds (OFFREB)
 total rebounds (TOREB)
 offensive rebounds for opponents (OFFREHOP)
 total rebounds for opponents (TOREHOP)
 assists (ASST)
 assists for opponents (ASSTOP)
 steals (STL)
 steals for opponents (STLOP)
 blocked shots (BLK)
 blocked shots for opponents (BLKOP)

The dependent variable is the number of wins (WINS) for the season. The other variables are to be considered possible explanatory variables.

You have been hired by your favorite NBA team to try and determine what factors might be important

in helping to achieve a winning season. Using the available data, determine which combination of the variables provides the best explanation of what makes a winning team.

Write a report with your results. Your report should consist of a letter/executive summary of your results for team management and a technical section with a description and justification of your regression equation. In the technical section you will want to discuss aspects such as your regression equation, the choice of variables, the strength of the relationship, and the practical usefulness of the results. What do your results tell you about winning teams?

(Source: Courtesy of the *For Worth Star-Telegram*.)

19. **Multicollinearity.** Consider the following regression model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + e_i$$

Multicollinearity is suspected to exist among the four explanatory variables used in this regression. The analyst using the regression computes pairwise correlations between the four explanatory variables and the dependent variable. The following correlation matrix results:

	y	x ₁	x ₂	x ₃	x ₄
x ₁	0.4	1.0			
x ₂	0.3	0.2	1.0		
x ₃	0.6	0.3	0.4	1.0	
x ₄	0.7	0.3	0.5	0.3	1.0

Based on these correlations, the analyst concludes that there will be no problems with multicollinearity. Do these correlations provide sufficient evidence to conclude that multicollinearity will not be a problem? Justify your answer.

USING THE COMPUTER

The Using the Computer section in each chapter describes how to perform the computer analyses in the chapter using Excel, MINITAB, and SAS. For further detail on Excel, MINITAB, and SAS, see Appendix C.

EXCEL

Multiple Regression

TOOLS: DATA ANALYSIS: REGRESSION

Figure 4.26 shows the Excel Regression dialog box. Regression is accessed in Excel by clicking on Tools and then Data Analysis. The Regression option is chosen from the Data Analysis menu. Put the range of the y variable in "Input Y Range." Put the range of the x variables in "Input X Range." Note that all x variables to be used in a multiple regression must be in adjacent columns. To accommodate this restriction, variables often must be moved around.

Click "Labels" if the variables have labels in the first row. Typically, "Constant in Zero" is not an option that is used. This option forces the constant or y intercept in a regression to be zero. It is seldom a good idea to use this option and can make interpretation of the regression results difficult.

Excel produces 95% confidence interval estimates of the population regression coefficients by default. If another level is required, click the "Confidence Level" box and insert the desired level.

Click the output option desired. The Residuals and Normal Probability options are discussed in Chapter 6.

Variance Inflation Factors

Excel does not provide options for automatically producing the variance inflation factors. Although these could be computed by creating formulas, this option is not discussed in this text.

A regression add-in is available that does compute a number of additional statistics. See the Excel section in Appendix C for more information on the add-in.

Creating a Lagged Variable

One way to create a lagged variable in Excel is simply to copy the necessary portion of the column to be lagged and paste it in the appropriate position. Figure 4.27 shows an example. Once the column is copied, the initial values with no matches are not used when running any regressions.

FIGURE 4.26 Excel Regression Dialog Box.

