15 Multiple Regression Model Building

Learning Objectives

In this chapter, you learn:

- To use quadratic terms in a regression model
- To use transformed variables in a regression model
- To measure the correlation among independent variables
- To build a regression model using either the stepwise or best-subsets approach
- To avoid the pitfalls involved in developing a multiple regression model

USING STATISTICS @ WHIT-DT

As part of your job as the operations manager at WHIT-DT, your business objective is to reduce unnecessary labor expenses. Currently, the unionized graphic artists at the television station receive hourly pay for a significant number of hours during which they are idle. These hours are called *standby hours*. You have collected data concerning standby hours and four factors that you suspect are related to the excessive number of standby hours the station is currently experiencing: the total number of staff present, remote hours, Dubner hours, and total labor hours.

You plan to build a multiple regression model to help determine which factors most heavily affect standby hours. You believe that an appropriate model will help you to predict the number of future standby hours, identify the root causes of excessive numbers of standby hours, and allow you to reduce the total number of future standby hours. How do you build the model with the most appropriate mix of independent variables? Are there statistical techniques that can help you identify a "best" model without having to consider all possible models? How do you begin?

Chapter 14 discussed multiple regression models with two independent variables. This chapter extends regression analysis to models containing more than two independent variables. The chapter introduces you to various topics related to model building to help you learn to develop the best model when confronted with a set of data (such as the one described in the WHIT-DT scenario) that has many independent variables. These topics include quadratic independent variables, transformations of the dependent or independent variables, stepwise regression, and best-subsets regression.

15.1 The Quadratic Regression Model

The simple regression model discussed in Chapter 13 and the multiple regression model discussed in Chapter 14 assume that the relationship between Y and each independent variable is linear. However, in Section 13.1, several different types of nonlinear relationships between variables were introduced. One of the most common nonlinear relationships is a quadratic, or curvilinear, relationship between two variables in which Y increases (or decreases) at a changing rate for various values of X (see Figure 13.2, Panels C–E, on page 523). You can use the quadratic regression model defined in Equation (15.1) to analyze this type of relationship between X and Y.

QUADRATIC REGRESSION MODEL

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{1i}^2 + \varepsilon_i \quad \textbf{(15.1)}$$

where

$$\begin{aligned} \beta_0 &= Y \text{ intercept} \\ \beta_1 &= \text{coefficient of the linear effect on } Y \\ \beta_2 &= \text{coefficient of the quadratic effect on } Y \\ \varepsilon_i &= \text{random error in } Y \text{ for observation } i \end{aligned}$$

This **quadratic regression model** is similar to the multiple regression model with two independent variables [see Equation (14.2) on page 579] except that the second independent variable, the **quadratic term**, is the square of the first independent variable. Once again, you use the least-squares method to compute sample regression coefficients (b_0, b_1, and b_2) as estimates of the population parameters (β_0, β_1, and β_2). Equation (15.2) defines the regression equation for the quadratic model with an independent variable (X_1) and a dependent variable (Y).

QUADRATIC REGRESSION EQUATION

$$\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{1i}^2 \quad \textbf{(15.2)}$$

In Equation (15.2), the first regression coefficient, b_0, represents the Y intercept; the second regression coefficient, b_1, represents the linear effect; and the third regression coefficient, b_2, represents the quadratic effect.

Finding the Regression Coefficients and Predicting Y

To illustrate the quadratic regression model, consider a study that examined the business problem facing a concrete supplier of how adding fly ash affects the strength of concrete. (Fly ash is an inexpensive industrial waste by-product that can be used as a substitute for Portland cement, a more expensive ingredient of concrete.) Batches of concrete were prepared in which the percentage of fly ash ranged from 0% to 60%. Data were collected from a sample of 18 batches and organized and stored in **FlyAsh**. Table 15.1 summarizes the results.

TABLE 15.1
Fly Ash Percentage and Strength of 18 Batches of 28-Day-Old Concrete

Fly Ash %	Strength (psi)	Fly Ash %	Strength (psi)
0	4,779	40	5,995
0	4,706	40	5,628
0	4,350	40	5,897
20	5,189	50	5,746
20	5,140	50	5,719
20	4,976	50	5,782
30	5,110	60	4,895
30	5,685	60	5,030
30	5,618	60	4,648

By creating the scatter plot in Figure 15.1 to visualize these data, you will be better able to select the proper model for expressing the relationship between fly ash percentage and strength.

FIGURE 15.1
Scatter plot of fly ash percentage (X) and strength (Y)

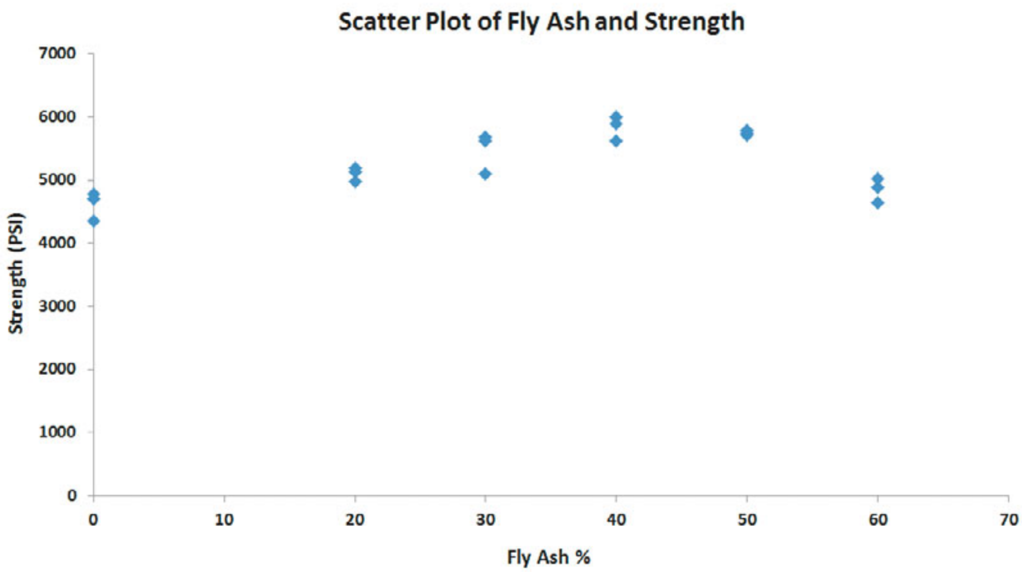


Figure 15.1 indicates an initial increase in the strength of the concrete as the percentage of fly ash increases. The strength appears to level off and then drop after achieving maximum strength at about 40% fly ash. Strength for 50% fly ash is slightly below strength at 40%, but strength at 60% fly ash is substantially below strength at 50%. Therefore, you should fit a quadratic model, not a linear model, to estimate strength based on fly ash percentage.

Figure 15.2 on page 632 shows regression results for these data.

From Figure 15.2,

$$b_0 = 4{,}486.3611 \quad b_1 = 63.0052 \quad b_2 = -0.8765$$

Therefore, the quadratic regression equation is

$$\hat{Y}_i = 4{,}486.3611 + 63.0052X_{1i} - 0.8765X_{1i}^2$$

where

$\hat{Y}_i$ = predicted strength for sample i
X_{1i} = percentage of fly ash for sample i

FIGURE 15.2

Excel and Minitab regression results for the concrete strength data

	A	B	C	D	E	F	G
1	Concrete Strength Analysis						
2							
3	*Regression Statistics*						
4	Multiple R	0.8053					
5	R Square	0.6485					
6	Adjusted R Square	0.6016					
7	Standard Error	312.1129					
8	Observations	18					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	2	2695473.4897	1347736.745	13.8351	0.0004	
13	Residual	15	1461217.0103	97414.4674			
14	Total	17	4156690.5000				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	4486.3611	174.7531	25.6726	0.0000	4113.8834	4858.8389
18	Fly Ash%	63.0052	12.3725	5.0923	0.0001	36.6338	89.3767
19	Fly Ash% ^2	-0.8765	0.1966	-4.4578	0.0005	-1.2955	-0.4574

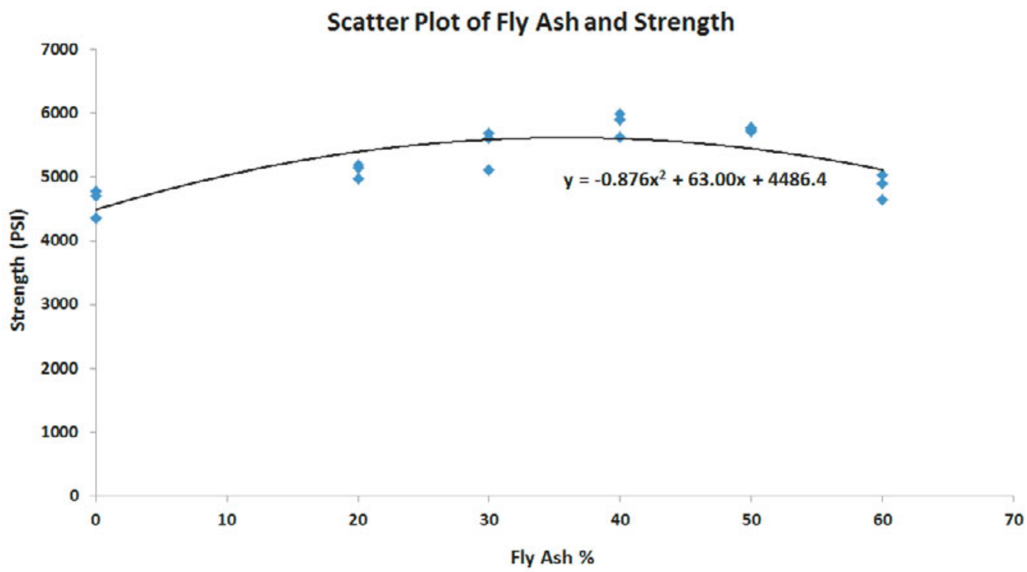
Regression Analysis: Strength versus Fly Ash%, Fly Ash%^2

```
The regression equation is
Strength = 4486 + 63.0 Fly Ash% - 0.876 Fly Ash%^2

Predictor      Coef  SE Coef      T      P
Constant     4486.4    174.8  25.67  0.000
Fly Ash%      63.01    12.37   5.09  0.000
Fly Ash%^2  -0.8765   0.1966  -4.46  0.000

S = 312.113   R-Sq = 64.8%   R-Sq(adj) = 60.2%

Analysis of Variance
Source           DF       SS       MS      F      P
Regression        2  2695473  1347737  13.84  0.000
Residual Error   15  1461217    97414
Total            17  4156690
```

Figure 15.3 is a scatter plot of this quadratic regression equation that shows the fit of the quadratic regression model to the original data.

FIGURE 15.3

Scatter plot showing the quadratic relationship between fly ash percentage and strength for the concrete data



From the quadratic regression equation and Figure 15.3, the Y intercept $(b_0 = 4{,}486.3611)$ is the predicted strength when the percentage of fly ash is 0. To interpret the coefficients b_1 and b_2, observe that after an initial increase, strength decreases as fly ash percentage increases. This nonlinear relationship is further demonstrated by predicting the strength for fly ash percentages of 20, 40, and 60. Using the quadratic regression equation,

$$\hat{Y}_i = 4{,}486.3611 + 63.0052X_{1i} - 0.8765X_{1i}^2$$

for $X_{1i} = 20$,

$$\hat{Y}_i = 4{,}486.3611 + 63.0052(20) - 0.8765(20)^2 = 5{,}395.865$$

for $X_{1i} = 40$,

$$\hat{Y}_i = 4{,}486.3611 + 63.0052(40) - 0.8765(40)^2 = 5{,}604.169$$

and for $X_{1i} = 60$,

$$\hat{Y}_i = 4{,}486.3611 + 63.0052(60) - 0.8765(60)^2 = 5{,}111.273$$

Thus, the predicted concrete strength for 40% fly ash is 208.304 psi above the predicted strength for 20% fly ash, but the predicted strength for 60% fly ash is 492.896 psi below the predicted strength for 40% fly ash.

Testing for the Significance of the Quadratic Model

After you calculate the quadratic regression equation, you can test whether there is a significant overall relationship between strength, Y, and fly ash percentage, X_1. The null and alternative hypotheses are as follows:

H_0: $\beta_1 = \beta_2 = 0$ (There is no overall relationship between X_1 and Y.)

H_1: β_1 and/or $\beta_2 \neq 0$ (There is an overall relationship between X_1 and Y.)

Equation (14.6) on page 586 defines the overall F_{STAT} test statistic used for this test:

$$F_{STAT} = \frac{MSR}{MSE}$$

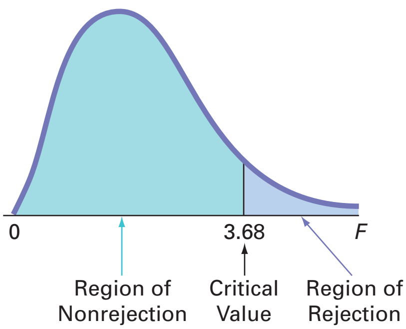
From the Figure 15.2 results on page 632,

$$F_{STAT} = \frac{MSR}{MSE} = \frac{1{,}347{,}736.745}{97{,}414.4674} = 13.8351$$

If you choose a level of significance of 0.05, from Table E.5, the critical value of the F distribution, with 2 and 15 degrees of freedom, is 3.68 (see Figure 15.4). Because $F_{STAT} = 13.8351 > 3.68$, or because the p-value $= 0.0004 < 0.05$, you reject the null hypothesis (H_0) and conclude that there is a significant overall relationship between strength and fly ash percentage.

FIGURE 15.4

Testing for the existence of the overall relationship at the 0.05 level of significance, with 2 and 15 degrees of freedom



Testing the Quadratic Effect

In using a regression model to examine a relationship between two variables, you want to find not only the most accurate model but also the simplest model that expresses that relationship. Therefore, you need to examine whether there is a significant difference between the quadratic model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{1i}^2 + \varepsilon_i$$

and the linear model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$$

In Section 14.4, you used the t test to determine whether each independent variable makes a significant contribution to the regression model. To test the significance of the contribution of the quadratic effect, you use the following null and alternative hypotheses:

H_0 : Including the quadratic effect does not significantly improve the model $(\beta_2 = 0)$.

H_1 : Including the quadratic effect significantly improves the model $(\beta_2 \neq 0)$.

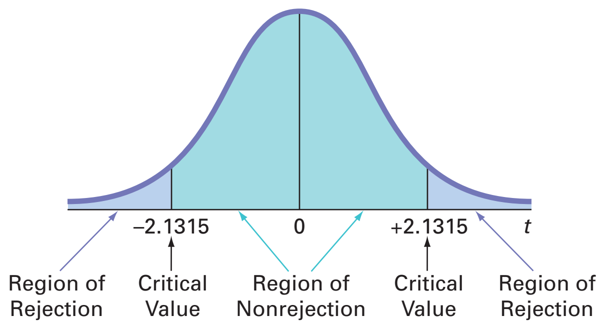
The standard error of each regression coefficient and its corresponding t_{STAT} test statistic are part of the regression results (see Figure 15.2 on page 632). Equation (14.7) on page 590 defines the t_{STAT} test statistic:

$$t_{STAT} = \frac{b_2 - \beta_2}{S_{b_2}}$$
$$= \frac{-0.8765 - 0}{0.1966} = -4.4578$$

If you select the 0.05 level of significance, then from Table E.3, the critical values for the t distribution with 15 degrees of freedom are -2.1315 and $+2.1315$ (see Figure 15.5).

FIGURE 15.5

Testing for the contribution of the quadratic effect to a regression model at the 0.05 level of significance, with 15 degrees of freedom



Because $t_{STAT} = -4.4578 < -2.1315$ or because the p-value $= 0.0005 < 0.05$, you reject H_0 and conclude that the quadratic model is significantly better than the linear model for representing the relationship between strength and fly ash percentage.

Example 15.1 provides an additional illustration of a possible quadratic effect.

EXAMPLE 15.1

Studying the Quadratic Effect in a Multiple Regression Model

A real estate developer studying the business problem of estimating the consumption of heating oil by single-family houses has decided to examine the effect of atmospheric temperature and the amount of attic insulation on heating oil consumption. Data are collected from a random sample of 15 single-family houses. The data are organized and stored in **HeatingOil**. Figure 15.6 shows the regression results for a multiple regression model using the two independent variables: atmospheric temperature and attic insulation.

FIGURE 15.6

Excel and Minitab regression results for the multiple linear regression model predicting monthly consumption of heating oil

	A	B	C	D	E	F	G
1	**Heating Oil Consumption Analysis**						
2							
3	*Regression Statistics*						
4	Multiple R	0.9827					
5	R Square	0.9656					
6	Adjusted R Square	0.9599					
7	Standard Error	26.0138					
8	Observations	15					
9							
10	**ANOVA**						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	2	228014.6263	114007.3132	168.4712	0.0000	
13	Residual	12	8120.6030	676.7169			
14	Total	14	236135.2293				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	562.1510	21.0931	26.6509	0.0000	516.1931	608.1089
18	Temperature	-5.4366	0.3362	-16.1699	0.0000	-6.1691	-4.7040
19	Insulation	-20.0123	2.3425	-8.5431	0.0000	-25.1162	-14.9084

Regression Analysis: Gallons versus Temperature, Insulation

```
The regression equation is
Gallons = 562 - 5.44 Temperature - 20.0 Insulation

Predictor        Coef   SE Coef       T      P
Constant       562.15     21.09   26.65  0.000
Temperature   -5.4366    0.3362  -16.17  0.000
Insulation    -20.012     2.343   -8.54  0.000

S = 26.0138   R-Sq = 96.6%   R-Sq(adj) = 96.0%

Analysis of Variance
Source          DF      SS      MS       F      P
Regression       2  228015  114007  168.47  0.000
Residual Error  12    8121     677
Total           14  236135
```

The residual plot for attic insulation (not shown here) contained some evidence of a quadratic effect. Thus, the real estate developer reanalyzed the data by adding a quadratic term for attic insulation to the multiple regression model. At the 0.05 level of significance, is there evidence of a significant quadratic effect for attic insulation?

SOLUTION Figure 15.7 shows the results for this regression model.

FIGURE 15.7

Excel and Minitab results for the multiple regression model with a quadratic term for attic insulation

	A	B	C	D	E	F	G
1	**Quadratic Effect for Insulation Variable?**						
2							
3	*Regression Statistics*						
4	Multiple R	0.9862					
5	R Square	0.9725					
6	Adjusted R Square	0.9650					
7	Standard Error	24.2938					
8	Observations	15					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	3	229643.1645	76547.7215	129.7006	0.0000	
13	Residual	11	6492.0649	590.1877			
14	Total	14	236135.2293				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	624.5864	42.4352	14.7186	0.0000	531.1872	717.9856
18	Temperature	-5.3626	0.3171	-16.9099	0.0000	-6.0606	-4.6646
19	Insulation	-44.5868	14.9547	-2.9815	0.0125	-77.5019	-11.6717
20	Insulation ^2	1.8667	1.1238	1.6611	0.1249	-0.6067	4.3401

Regression Analysis: Gallons versus Temperature, Insulation, ...

```
The regression equation is
Gallons = 625 - 5.36 Temperature - 44.6 Insulation + 1.87 Insulation^2

Predictor        Coef   SE Coef       T      P
Constant       624.59     42.44   14.72  0.000
Temperature   -5.3626    0.3171  -16.91  0.000
Insulation     -44.59     14.95   -2.98  0.012
Insulation^2    1.867     1.124    1.66  0.125

S = 24.2938   R-Sq = 97.3%   R-Sq(adj) = 96.5%

Analysis of Variance
Source          DF      SS     MS       F      P
Regression       3  229643  76548  129.70  0.000
Residual Error  11    6492    590
Total           14  236135
```

The multiple regression equation is

$$\hat{Y}_i = 624.5864 - 5.3626X_{1i} - 44.5868X_{2i} + 1.8667X_{2i}^2$$

To test for the significance of the quadratic effect,

H_0 : Including the quadratic effect does not significantly improve the model ($\beta_3 = 0$).

H_1 : Including the quadratic effect significantly improves the model ($\beta_3 \neq 0$).

From Figure 15.7 and Table E.3, $-2.2010 < t_{STAT} = 1.6611 < 2.2010$ (or the p-value $= 0.1249 > 0.05$). Therefore, you do not reject the null hypothesis. You conclude that there is insufficient evidence that the quadratic effect for attic insulation is different from zero. In the interest of keeping the model as simple as possible, you should use the multiple regression equation shown in Figure 15.6:

$$\hat{Y}_i = 562.1510 - 5.4366X_{1i} - 20.0123X_{2i}$$

The Coefficient of Multiple Determination

In the multiple regression model, the coefficient of multiple determination, r^2 (see Section 14.2), represents the proportion of variation in Y that is explained by variation in the independent variables. Consider the quadratic regression model you used to predict the strength of concrete using fly ash and fly ash squared. You compute r^2 by using Equation (14.4) on page 616:

$$r^2 = \frac{SSR}{SST}$$

From Figure 15.2 on page 632,

$$SSR = 2{,}695{,}473.897 \qquad SST = 4{,}156{,}690.5$$

Thus,

$$r^2 = \frac{SSR}{SST} = \frac{2{,}695{,}473.897}{4{,}156{,}690.5} = 0.6485$$

This coefficient of multiple determination indicates that 64.85% of the variation in strength is explained by the quadratic relationship between strength and the percentage of fly ash. You should also compute r^2_{adj} to account for the number of independent variables and the sample size. In the quadratic regression model, $k = 2$ because there are two independent variables, X_1 and X_1^2. Thus, using Equation (14.5) on page 585,

$$\begin{aligned} r^2_{adj} &= 1 - \left[(1 - r^2)\frac{(n - 1)}{(n - k - 1)}\right] \\ &= 1 - \left[(1 - 0.6485)\frac{17}{15}\right] \\ &= 1 - 0.3984 \\ &= 0.6016 \end{aligned}$$

Problems for Section 15.1

LEARNING THE BASICS

15.1 The following is the quadratic regression equation for a sample of $n = 25$:

$$\hat{Y}_i = 5 + 3X_{1i} + 1.5X_{1i}^2$$

a. Predict Y for $X_1 = 2$.
b. Suppose that the computed t_{STAT} test statistic for the quadratic regression coefficient is 2.35. At the 0.05 level of significance, is there evidence that the quadratic model is better than the linear model?
c. Suppose that the computed t_{STAT} test statistic for the quadratic regression coefficient is 1.17. At the 0.05 level of significance, is there evidence that the quadratic model is better than the linear model?
d. Suppose the regression coefficient for the linear effect is -3.0. Predict Y for $X_1 = 2$.

APPLYING THE CONCEPTS

15.2 Businesses actively recruit business students with well-developed higher-order cognitive skills (HOCS) such as problem identification, analytical reasoning, and content integration skills. Researchers conducted a study to see if improvement in students' HOCS was related to the students' GPA. (Data extracted from R. V. Bradley, C. S. Sankar, H. R. Clayton, V. W. Mbarika, and P. K. Raju, "A Study on the Impact of GPA on Perceived Improvement of Higher-Order Cognitive Skills," *Decision Sciences Journal of Innovative Education*, January 2007, 5(1), pp 151–168.) The researchers conducted a study in which business students were taught using the case study method. Using data collected from 300 business students, the following quadratic regression equation was derived:

$$\text{HOCS} = -3.48 + 4.53(\text{GPA}) - 0.68(\text{GPA})^2$$

where the dependent variable HOCS measured the improvement in higher-order cognitive skills, with 1 being the lowest improvement in HOCS and 5 being the highest improvement in HOCS.

a. Construct a table of predicted HOCS, using GPA equal to 2.0, 2.1, 2.2, . . . , 4.0.
b. Plot the values in the table constructed in (a), with GPA on the horizontal axis and predicted HOCS on the vertical axis.
c. Discuss the curvilinear relationship between students' GPA and their predicted improvement in HOCS.
d. The researchers reported that the model had an r^2 of 0.07 and an adjusted r^2 of 0.06. What does this tell you about the scatter of individual HOCS scores around the curvilinear relationship plotted in (b) and discussed in (c)?

15.3 A national chain of consumer electronics stores had the business objective of determining the effectiveness of newspaper advertising. To promote sales, the chain relies heavily on local newspaper advertising to support its modest exposure in nationwide television commercials. A sample of 20 cities with similar populations and monthly sales totals were assigned different newspaper advertising budgets for one month. The following table (stored in **Advertising**) summarizes the sales (in $millions) and the newspaper advertising budgets (in $thousands) observed during the study:

Sales	Newspaper Advertising	Sales	Newspaper Advertising
6.14	5	6.84	15
6.04	5	6.66	15
6.21	5	6.95	20
6.32	5	6.65	20
6.42	10	6.83	20
6.56	10	6.81	20
6.67	10	7.03	25
6.35	10	6.88	25
6.76	15	6.84	25
6.79	15	6.99	25

a. Construct a scatter plot for newspaper advertising and sales.
b. Fit a quadratic regression model and state the quadratic regression equation.
c. Predict the monthly sales for a city with newspaper advertising of $20,000.
d. Perform a residual analysis on the results and determine whether the regression assumptions are valid.
e. At the 0.05 level of significance, is there a significant quadratic relationship between monthly sales and newspaper advertising?
f. At the 0.05 level of significance, determine whether the quadratic model is a better fit than the linear model.
g. Interpret the meaning of the coefficient of multiple determination.
h. Compute the adjusted r^2.

15.4 Is the number of calories in a beer related to the number of carbohydrates and/or the percentage of alcohol in the beer? Data concerning 139 of the best-selling domestic beers in the United States are stored in **DomesticBeer**. The values for three variables are included: the number of calories per 12 ounces, the alcohol percentage, and the number of carbohydrates (in grams) per 12 ounces. (Data extracted from **www.Beer100.com**, March 18, 2010.)
a. Perform a multiple linear regression analysis, using calories as the dependent variable and percentage alcohol and number of carbohydrates as the independent variables.
b. Add quadratic terms for alcohol percentage and the number of carbohydrates.
c. Which model is better, the one in (a) or (b)?
d. Write a short summary concerning the relationship between the number of calories in a beer and the alcohol percentage and number of carbohydrates.

15.5 The per-store daily customer count (i.e., the mean number of customers in a store in one day) for a nationwide convenience store chain that operates nearly 10,000 stores has been steady, at 900, for some time. To increase the customer count, the chain is considering cutting prices for coffee beverages. The question to be determined is how much prices should be cut to increase the daily customer count without reducing the gross margin on coffee sales too much. You decide to carry out an experiment in a sample of 24 stores where customer counts have been running almost exactly at the national average of 900. In 6 of the stores, the price of a small coffee will now be $0.59, in 6 stores the price of a small coffee will now be $0.69, in 6 stores, the price of a small coffee will now be $0.79, and in 6 stores, the price of a small coffee will now be $0.89. After four weeks at the new prices, the daily customer count in the stores is determined and is stored in **CoffeeSales2**.
a. Construct a scatter plot for price and sales.
b. Fit a quadratic regression model and state the quadratic regression equation.
c. Predict the weekly sales for a small coffee priced at 79 cents.
d. Perform a residual analysis on the results and determine whether the regression model is valid.
e. At the 0.05 level of significance, is there a significant quadratic relationship between weekly sales and price?
f. At the 0.05 level of significance, determine whether the quadratic model is a better fit than the linear model.
g. Interpret the meaning of the coefficient of multiple determination.
h. Compute the adjusted r^2.
i. Compare the results of (a) through (h) to those of Problem 11.11 on page 428.

SELF Test **15.6** An agronomist designed a study in which tomatoes were grown using six different amounts of fertilizer: 0, 20, 40, 60, 80, and 100 pounds per 1,000 square feet. These fertilizer application rates were then randomly assigned to plots of land. The results including the yield of tomatoes (in pounds) are stored in **Tomato** and are listed here:

Fertilizer Application			Fertilizer Application		
Plot	Rate	Yield	Plot	Rate	Yield
1	0	6	7	60	46
2	0	9	8	60	50
3	20	19	9	80	48
4	20	24	10	80	54
5	40	32	11	100	52
6	40	38	12	100	58

a. Construct a scatter plot for fertilizer application rate and yield.
b. Fit a quadratic regression model and state the quadratic regression equation.
c. Predict the yield for a plot of land fertilized with 70 pounds per 1,000 square feet.
d. Perform a residual analysis on the results and determine whether the regression model is valid.
e. At the 0.05 level of significance, is there a significant overall relationship between the fertilizer application rate and tomato yield?
f. What is the p-value in (e)? Interpret its meaning.
g. At the 0.05 level of significance, determine whether there is a significant quadratic effect.
h. What is the p-value in (g)? Interpret its meaning.
i. Interpret the meaning of the coefficient of multiple determination.
j. Compute the adjusted r^2.

15.7 An auditor for a county government would like to develop a model to predict county taxes, based on the age of single-family houses. She selects a random sample of 19 single-family houses, and the results are stored in **Taxes**.

a. Construct a scatter plot of age and county taxes.

b. Fit a quadratic regression model and state the quadratic regression equation.

c. Predict the county taxes for a house that is 20 years old.

d. Perform a residual analysis on the results and determine whether the regression model is valid.

e. At the 0.05 level of significance, is there a significant overall relationship between age and county taxes?

f. What is the p-value in (e)? Interpret its meaning.

g. At the 0.05 level of significance, determine whether the quadratic model is superior to the linear model.

h. What is the p-value in (g)? Interpret its meaning.

i. Interpret the meaning of the coefficient of multiple determination.

j. Compute the adjusted r^2.

15.2 Using Transformations in Regression Models

This section introduces regression models in which the independent variable, the dependent variable, or both are transformed in order to either overcome violations of the assumptions of regression or to make a model whose form is not linear into a linear model. Among the many transformations available (see reference 1) are the square-root transformation and transformations involving the common logarithm (base 10) and the natural logarithm (base e).[1]

[1]For more information on logarithms, see Appendix Section A.3.

The Square-Root Transformation

The **square-root transformation** is often used to overcome violations of the equal-variance assumption as well as to transform a model whose form is not linear into a linear model. Equation (15.3) shows a regression model that uses a square-root transformation of the independent variable.

REGRESSION MODEL WITH A SQUARE-ROOT TRANSFORMATION

$$Y_i = \beta_0 + \beta_1 \sqrt{X_{1i}} + \varepsilon_i \tag{15.3}$$

Example 15.2 illustrates the use of a square-root transformation.

EXAMPLE 15.2

Using the Square-Root Transformation

Given the following values for Y and X, use a square-root transformation for the X variable:

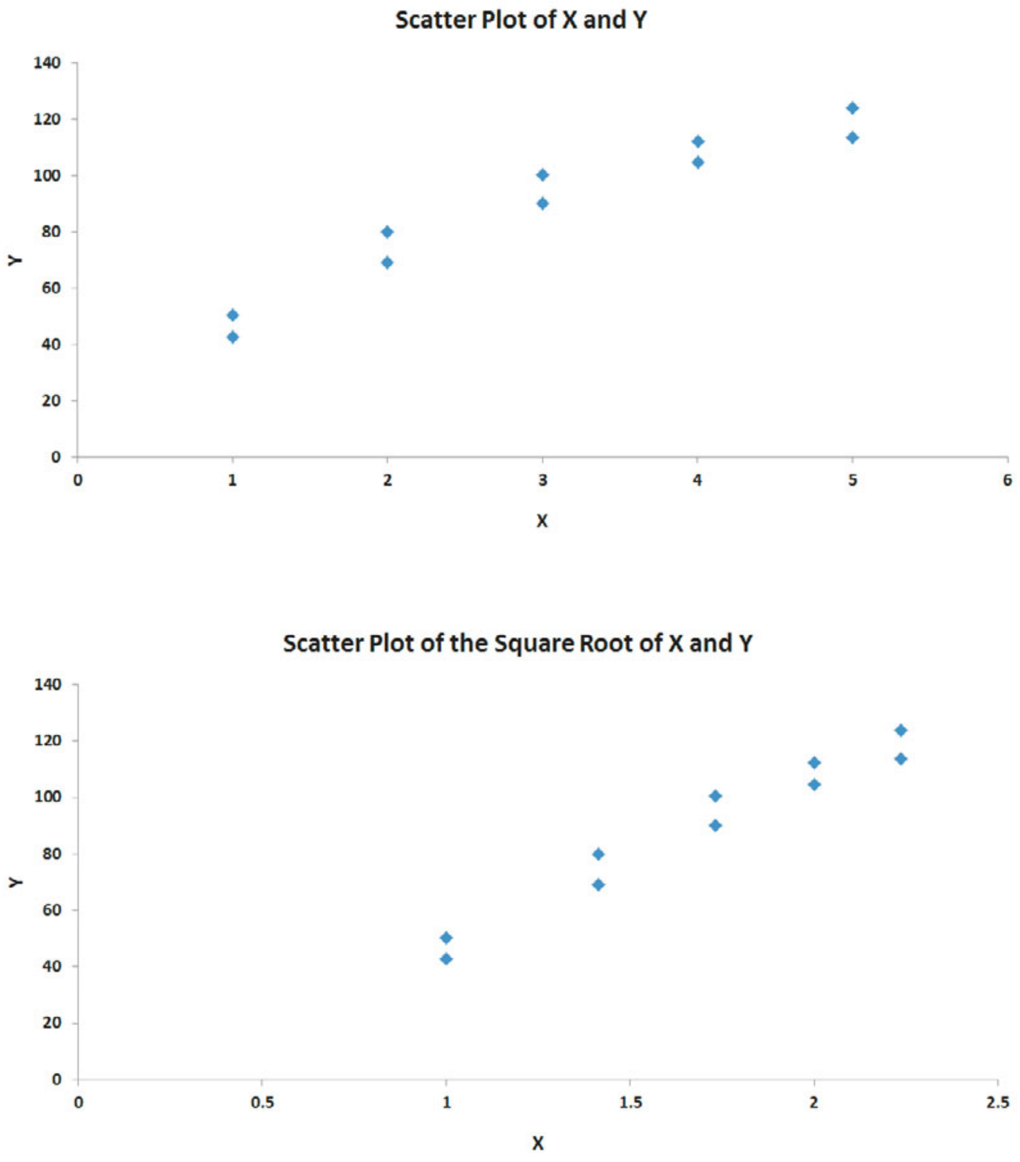
Y	X	Y	X
42.7	1	100.4	3
50.4	1	104.7	4
69.1	2	112.3	4
79.8	2	113.6	5
90.0	3	123.9	5

Construct a scatter plot for X and Y and for the square root of X and Y.

SOLUTION Figure 15.8 displays both scatter plots.

FIGURE 15.8

Example 15.2 scatter plots of X and Y and the square root of X and Y



You can see that the square-root transformation has transformed a nonlinear relationship into a linear relationship.

The Log Transformation

The **logarithmic transformation** is often used to overcome violations to the equal-variance assumption. You can also use the logarithmic transformation to change a nonlinear model into a linear model. Equation (15.4) shows a multiplicative model.

ORIGINAL MULTIPLICATIVE MODEL

$$Y_i = \beta_0 X_{1i}^{\beta_1} X_{2i}^{\beta_2} \varepsilon_i \tag{15.4}$$

By taking base 10 logarithms of both the dependent and independent variables, you can transform Equation (15.4) to the model shown in Equation (15.5).

TRANSFORMED MULTIPLICATIVE MODEL

$$\begin{aligned}\log Y_i &= \log(\beta_0 X_{1i}^{\beta_1} X_{2i}^{\beta_2} \varepsilon_i) \\ &= \log\beta_0 + \log(X_{1i}^{\beta_1}) + \log(X_{2i}^{\beta_2}) + \log\varepsilon_i \\ &= \log\beta_0 + \beta_1 \log X_{1i} + \beta_2 \log X_{2i} + \log\varepsilon_i\end{aligned} \tag{15.5}$$

Thus, Equation (15.5) is linear in the logarithms. Similarly, you can transform the exponential model shown in Equation (15.6) to a linear form by taking the natural logarithm of both sides of the equation. Equation (15.7) is the transformed model.

ORIGINAL EXPONENTIAL MODEL

$$Y_i = e^{\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}} \varepsilon_i \tag{15.6}$$

TRANSFORMED EXPONENTIAL MODEL

$$\begin{aligned}\ln Y_i &= \ln(e^{\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}} \varepsilon_i) \\ &= \ln(e^{\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}}) + \ln\varepsilon_i \\ &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ln\varepsilon_i\end{aligned} \tag{15.7}$$

Example 15.3 illustrates the use of a natural log transformation.

EXAMPLE 15.3

Using the Natural Log Transformation

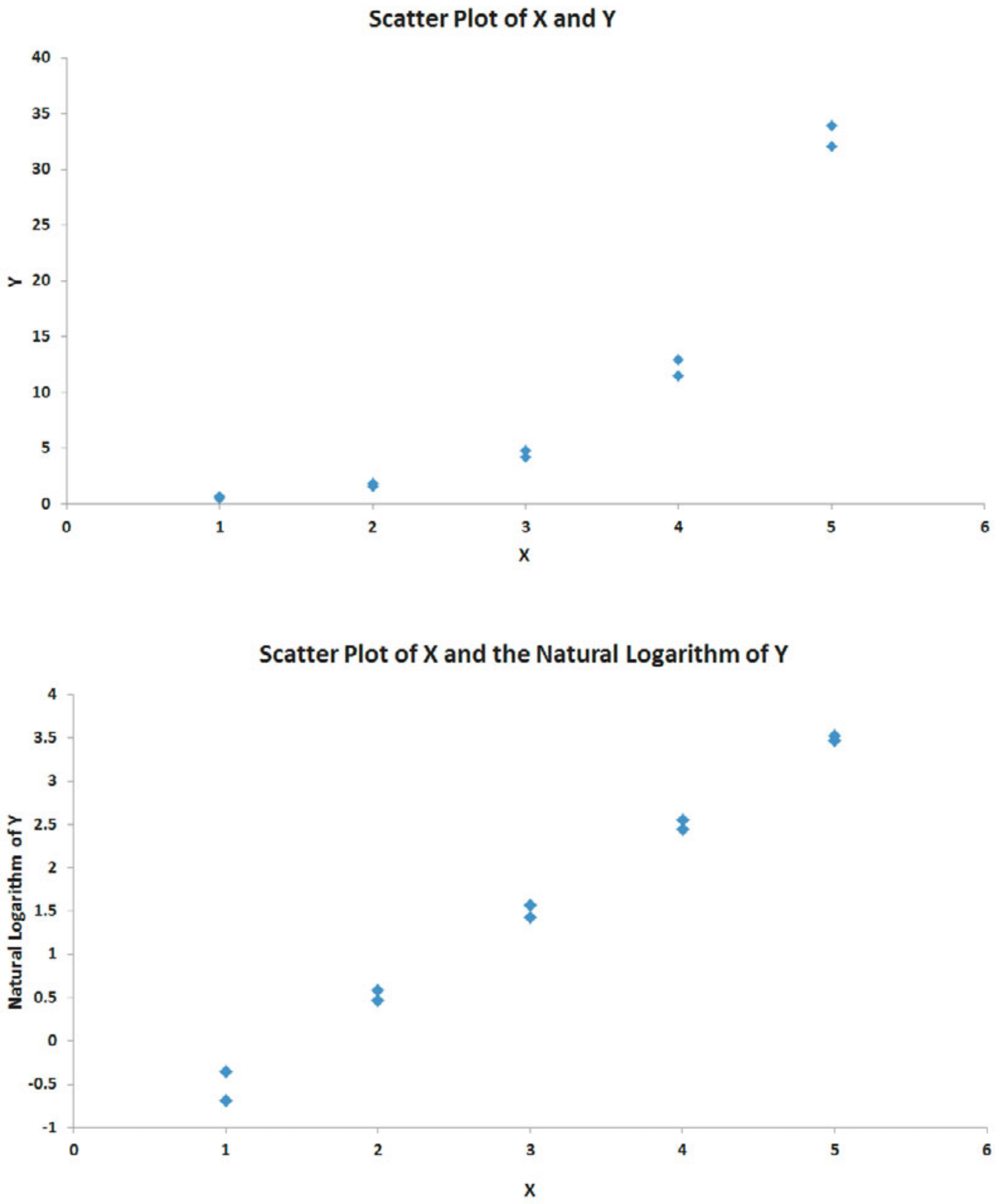
Given the following values for Y and X, use a natural logarithm transformation for the Y variable:

Y	X	Y	X
0.7	1	4.8	3
0.5	1	12.9	4
1.6	2	11.5	4
1.8	2	32.1	5
4.2	3	33.9	5

Construct a scatter plot for X and Y and for X and the natural logarithm of Y.

SOLUTION Figure 15.9 displays both scatter plots. The plots show that the natural logarithm transformation has transformed a nonlinear relationship into a linear relationship.

FIGURE 15.9
Example 15.3 scatter plots of X and Y and X and the natural logarithm of Y



Problems for Section 15.2

LEARNING THE BASICS

15.8 Consider the following regression equation:

$$\log \hat{Y}_i = \log 3.07 + 0.9 \log X_{1i} + 1.41 \log X_{2i}$$

a. Predict the value of Y when $X_1 = 8.5$ and $X_2 = 5.2$.
b. Interpret the meaning of the regression coefficients b_0, b_1, and b_2.

15.9 Consider the following regression equation:

$$\ln \hat{Y}_i = 4.62 + 0.5X_{1i} + 0.7X_{2i}$$

a. Predict the value of Y when $X_1 = 8.5$ and $X_2 = 5.2$.
b. Interpret the meaning of the regression coefficients b_0, b_1, and b_2.

APPLYING THE CONCEPTS

SELF Test **15.10** Using the data of Problem 15.4 on page 637, stored in **DomesticBeer**, perform a square-root transformation on each of the independent variables (percentage alcohol and number of carbohydrates). Using calories as the dependent variable and the transformed independent variables, perform a multiple regression analysis.
a. State the regression equation.
b. Perform a residual analysis of the results and determine whether the regression model is valid.
c. At the 0.05 level of significance, is there a significant relationship between calories and the square root of the

percentage of alcohol and the square root of the number of carbohydrates?

d. Interpret the meaning of the coefficient of determination, r^2, in this problem.

e. Compute the adjusted r^2.

f. Compare your results with those in Problem 15.4. Which model is better? Why?

15.11 Using the data of Problem 15.4 on page 637, stored in DomesticBeer, perform a natural logarithmic transformation of the dependent variable (calories). Using the transformed dependent variable and the percentage of alcohol and the number of carbohydrates as the independent variables, perform a multiple regression analysis.

a. State the regression equation.

b. Perform a residual analysis of the results and determine whether the regression assumptions are valid.

c. At the 0.05 level of significance, is there a significant relationship between the natural logarithm of calories and the percentage of alcohol and the number of carbohydrates?

d. Interpret the meaning of the coefficient of determination, r^2, in this problem.

e. Compute the adjusted r^2.

f. Compare your results with those in Problems 15.4 and 15.10. Which model is best? Why?

15.12 Using the data of Problem 15.6 on page 637, stored in Tomato, perform a natural logarithm transformation of the dependent variable (yield). Using the transformed dependent variable and the fertilizer application rate as the independent variable, perform a regression analysis.

a. State the regression equation.

b. Predict the yield when 55 pounds of fertilizer is applied per 1,000 square feet.

c. Perform a residual analysis of the results and determine whether the regression assumptions are valid.

d. At the 0.05 level of significance, is there a significant relationship between the natural logarithm of yield and the fertilizer application rate?

e. Interpret the meaning of the coefficient of determination, r^2, in this problem.

f. Compute the adjusted r^2.

g. Compare your results with those in Problem 15.6. Which model is better? Why?

15.13 Using the data of Problem 15.6 on page 637, stored in Tomato, perform a square-root transformation of the independent variable (fertilizer application rate). Using yield as the dependent variable and the transformed independent variable, perform a regression analysis.

a. State the regression equation.

b. Predict the yield when 55 pounds of fertilizer is applied per 1,000 square feet.

c. Perform a residual analysis of the results and determine whether the regression model is valid.

d. At the 0.05 level of significance, is there a significant relationship between yield and the square root of the fertilizer application rate?

e. Interpret the meaning of the coefficient of determination, r^2, in this problem.

f. Compute the adjusted r^2.

g. Compare your results with those of Problems 15.6 and 15.12. Which model is best? Why?

h. How much fertilizer should you apply in order to grow the most tomatoes?

15.3 Collinearity

One important problem in the application of multiple regression analysis involves the possible **collinearity** of the independent variables. This condition refers to situations in which two or more of the independent variables are highly correlated with each other. In such situations, collinear variables do not provide unique information, and it becomes difficult to separate the effects of such variables on the dependent variable. When collinearity exists, the values of the regression coefficients for the correlated variables may fluctuate drastically, depending on which independent variables are included in the model.

One method of measuring collinearity is to determine the **variance inflationary factor (*VIF*)** for each independent variable. Equation (15.8) defines VIF_j, the variance inflationary factor for variable j.

VARIANCE INFLATIONARY FACTOR

$$VIF_j = \frac{1}{1 - R_j^2} \quad \textbf{(15.8)}$$

where

R_j^2 is the coefficient of multiple determination for a regression model, using variable X_j as the dependent variable and all other X variables as independent variables.

If there are only two independent variables, R_1^2 is the coefficient of determination between X_1 and X_2. It is identical to R_2^2, which is the coefficient of determination between X_2 and X_1. If there are three independent variables, then R_1^2 is the coefficient of multiple determination of X_1 with X_2 and X_3; R_2^2 is the coefficient of multiple determination of X_2 with X_1 and X_3; and R_3^2 is the coefficient of multiple determination of X_3 with X_1 and X_2.

If a set of independent variables is uncorrelated, each VIF_j is equal to 1. If the set is highly correlated, then a VIF_j might even exceed 10. Marquardt (see reference 2) suggests that if VIF_j is greater than 10, there is too much correlation between the variable X_j and the other independent variables. However, other statisticians suggest a more conservative criterion. Snee (see reference 5) recommends using alternatives to least-squares regression if the maximum VIF_j exceeds 5.

You need to proceed with extreme caution when using a multiple regression model that has one or more large *VIF* values. You can use the model to predict values of the dependent variable *only* in the case where the values of the independent variables used in the prediction are in the relevant range of the values in the data set. However, you cannot extrapolate to values of the independent variables not observed in the sample data. And because the independent variables contain overlapping information, you should always avoid interpreting the regression coefficient estimates separately because there is no way to accurately estimate the individual effects of the independent variables. One solution to the problem is to delete the variable with the largest *VIF* value. The reduced model (i.e., the model with the independent variable with the largest *VIF* value deleted) is often free of collinearity problems. If you determine that all the independent variables are needed in the model, you can use methods discussed in reference 1.

In the OmniPower sales data (see Section 14.1), the correlation between the two independent variables, price and promotional expenditure, is -0.0968. Because there are only two independent variables in the model, from Equation (15.8) on page 642:

$$\begin{aligned} VIF_1 = VIF_2 &= \frac{1}{1 - (-0.0968)^2} \\ &= 1.009 \end{aligned}$$

Thus, you can conclude that you should not be concerned with collinearity for the OmniPower sales data.

In models containing quadratic and interaction terms, collinearity is usually present. The linear and quadratic terms of an independent variable are usually highly correlated with each other, and an interaction term is often correlated with one or both of the independent variables making up the interaction. Thus, you cannot interpret individual parameter estimates separately. You need to interpret the linear and quadratic parameter estimates together in order to understand the nonlinear relationship. Likewise, you need to interpret an interaction parameter estimate in conjunction with the two parameter estimates associated with the variables comprising the interaction. In summary, large *VIF*s in quadratic or interaction models do not necessarily mean that the model is not a good one. They do, however, require you to carefully interpret the parameter estimates.

Problems for Section 15.3

LEARNING THE BASICS

15.14 If the coefficient of determination between two independent variables is 0.20, what is the *VIF*?

15.15 If the coefficient of determination between two independent variables is 0.50, what is the *VIF*?

APPLYING THE CONCEPTS

SELF Test **15.16** Refer to Problem 14.4 on page 583. Perform a multiple regression analysis using the data in **WareCost** and determine the *VIF* for each independent variable in the model. Is there reason to suspect the existence of collinearity?

15.17 Refer to Problem 14.5 on page 583. Perform a multiple regression analysis using the data in **Auto2010** and determine the *VIF* for each independent variable in the model. Is there reason to suspect the existence of collinearity?

15.18 Refer to Problem 14.6 on page 583. Perform a multiple regression analysis using the data in **Advertise** and determine the *VIF* for each independent variable in the model. Is there reason to suspect the existence of collinearity?

15.19 Refer to Problem 14.7 on page 584. Perform a multiple regression analysis using the data in **Standby** and determine the *VIF* for each independent variable in the model. Is there reason to suspect the existence of collinearity?

15.20 Refer to Problem 14.8 on page 584. Perform a multiple regression analysis using the data in **GlenCove** and determine the *VIF* for each independent variable in the model. Is there reason to suspect the existence of collinearity?

15.4 Model Building

This chapter and Chapter 14 have introduced you to many different topics in regression analysis, including quadratic terms, dummy variables, and interaction terms. In this section, you learn a structured approach to building the most appropriate regression model. As you will see, successful model building incorporates many of the topics you have studied so far.

To begin, refer to the WHIT-DT scenario introduced on page 629, in which four independent variables (total staff present, remote hours, Dubner hours, and total labor hours) are considered in the business problem that involves developing a regression model to predict standby hours of unionized graphic artists. Data are collected over a period of 26 weeks and organized and stored in **Standby**. Table 15.2 summarizes the data.

TABLE 15.2

Predicting Standby Hours Based on Total Staff Present, Remote Hours, Dubner Hours, and Total Labor Hours

Week	Standby Hours	Total Staff Present	Remote Hours	Dubner Hours	Total Labor Hours
1	245	338	414	323	2,001
2	177	333	598	340	2,030
3	271	358	656	340	2,226
4	211	372	631	352	2,154
5	196	339	528	380	2,078
6	135	289	409	339	2,080
7	195	334	382	331	2,073
8	118	293	399	311	1,758
9	116	325	343	328	1,624
10	147	311	338	353	1,889
11	154	304	353	518	1,988
12	146	312	289	440	2,049
13	115	283	388	276	1,796
14	161	307	402	207	1,720
15	274	322	151	287	2,056
16	245	335	228	290	1,890
17	201	350	271	355	2,187
18	183	339	440	300	2,032
19	237	327	475	284	1,856
20	175	328	347	337	2,068
21	152	319	449	279	1,813
22	188	325	336	244	1,808
23	188	322	267	253	1,834
24	197	317	235	272	1,973
25	261	315	164	223	1,839
26	232	331	270	272	1,935

To develop a model to predict the dependent variable, standby hours in the WHIT-DT scenario, you need to be guided by a general problem-solving strategy or *heuristic*. One heuristic appropriate for building regression models uses the principle of parsimony.

Parsimony guides you to select the regression model with the fewest independent variables that can predict the dependent variable adequately. Regression models with fewer independent variables are easier to interpret, particularly because they are less likely to be affected by collinearity problems (described in Section 15.3).

The selection of an appropriate model when many independent variables are under consideration involves complexities that are not present with a model that has only two independent variables. The evaluation of all possible regression models is more computationally complex. And, although you can quantitatively evaluate competing models, there may not be a *uniquely* best model but several *equally appropriate* models.

To begin analyzing the standby-hours data, you compute the variance inflationary factors [see Equation (15.8) on page 642] to measure the amount of collinearity among the independent variables. The values for the four *VIF*s for this model appear in Figure 15.10, along with the results for the model that uses the four independent variables.

FIGURE 15.10

Excel and Minitab regression results for predicting standby hours based on four independent variables (Excel results contain additional worksheets for Durbin-Watson statistic and *VIF* inset)

Variance Inflationary Factor (*VIF*) Calculations

	Regression Model			
	Total Staff and all other X	Remote and all other X	Dubner and all other X	Total Labor and all other X
R Square	0.4143	0.1891	0.3147	0.4998
VIF	1.7074	1.2333	1.4592	1.9993

Standby Hours Analysis

Regression Statistics	
Multiple R	0.7894
R Square	0.6231
Adjusted R Square	0.5513
Standard Error	31.8350
Observations	26

Durbin-Watson Calculations

Sum of Squared Difference of Residuals	47241.6126
Sum of Squared Residuals	21282.8217
Durbin-Watson Statistic	2.2197

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	35181.7937	8795.4484	8.6786	0.0003
Residual	21	21282.8217	1013.4677		
Total	25	56464.6154			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-330.8318	110.8954	-2.9833	0.0071	-561.4514	-100.2123
Total Staff	1.2456	0.4121	3.0229	0.0065	0.3887	2.1026
Remote	-0.1184	0.0543	-2.1798	0.0408	-0.2314	-0.0054
Dubner	-0.2971	0.1179	-2.5189	0.0199	-0.5423	-0.0518
Total Labor	0.1305	0.0593	2.2004	0.0391	0.0072	0.2539

Regression Analysis: Standby versus Total Staff, ...

```
The regression equation is
Standby = - 331 + 1.25 Total Staff - 0.118 Remote
          - 0.297 Dubner + 0.131 Total Labor

Predictor        Coef   SE Coef      T      P    VIF
Constant       -330.8     110.9  -2.98  0.007
Total Staff    1.2456    0.4121   3.02  0.006  1.707
Remote       -0.11842   0.05432  -2.18  0.041  1.233
Dubner        -0.2971    0.1179  -2.52  0.020  1.459
Total Labor   0.13053   0.05932   2.20  0.039  1.999

S = 31.8350   R-Sq = 62.3%   R-Sq(adj) = 55.1%

Analysis of Variance
Source          DF     SS    MS     F      P
Regression       4  35182  8795  8.68  0.000
Residual Error  21  21283  1013
Total           25  56465

Source       DF  Seq SS
Total Staff   1   20667
Remote        1    6995
Dubner        1    2612
Total Labor   1    4907

Durbin-Watson statistic = 2.21971
```

Observe that all the *VIF* values in Figure 15.10 are relatively small, ranging from a high of 1.999 for the total labor hours to a low of 1.233 for remote hours. Thus, on the basis of the criteria developed by Snee that all *VIF* values should be less than 5.0 (see reference 5), there is little evidence of collinearity among the set of independent variables.

The Stepwise Regression Approach to Model Building

You continue your analysis of the standby-hours data by attempting to determine whether a subset of all independent variables yields an adequate and appropriate model. The first approach described here is **stepwise regression**, which attempts to find the "best" regression model without examining all possible models.

The first step of stepwise regression is to find the best model that uses one independent variable. The next step is to find the best of the remaining independent variables to add to the model selected in the first step. An important feature of the stepwise approach is that an independent variable that has entered into the model at an early stage may subsequently be removed after other independent variables are considered. Thus, in stepwise regression, variables are either added to or deleted from the regression model at each step of the model-building process. The t test for the slope (see Section 14.4) or the partial F_{STAT} test statistic (see Section 14.5) is used to determine whether variables are added or deleted. The stepwise procedure terminates with the selection of a best-fitting model when no additional variables can be added to or deleted from the last model evaluated. Figure 15.11 shows the Excel (using PHStat2) and Minitab stepwise regression results for the standby-hours data.

FIGURE 15.11

Excel (PHStat2) and Minitab stepwise regression results for the standby-hours data

	A	B	C	D	E	F	G	H
1	Stepwise Analysis for Standby Hours							
2	Table of Results for General Stepwise							
3								
4	Total Staff entered.							
5								
6			df	SS	MS	F	Significance F	
7		Regression	1	20667.3980	20667.3980	13.8563	0.0011	
8		Residual	24	35797.2174	1491.5507			
9		Total	25	56464.6154				
10								
11			Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
12		Intercept	-272.3816	124.2402	-2.1924	0.0383	-528.8008	-15.9625
13		Total Staff	1.4241	0.3826	3.7224	0.0011	0.6345	2.2136
14								
15								
16	Remote entered.							
17								
18			df	SS	MS	F	Significance F	
19		Regression	2	27662.5429	13831.2714	11.0450	0.0004	
20		Residual	23	28802.0725	1252.2640			
21		Total	25	56464.6154				
22								
23			Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
24		Intercept	-330.6748	116.4802	-2.8389	0.0093	-571.6322	-89.7175
25		Total Staff	1.7649	0.3790	4.6562	0.0001	0.9808	2.5490
26		Remote	-0.1390	0.0588	-2.3635	0.0269	-0.2606	-0.0173
27								
28								
29	No other variables could be entered into the model. Stepwise ends.							

Stepwise Regression: Standby versus Total Staff, Remote, ...

```
 Alpha-to-Enter: 0.05  Alpha-to-Remove: 0.05

Response is Standby on 4 predictors, with N = 26

Step              1       2
Constant     -272.4  -330.7

Total Staff    1.42    1.76
T-Value        3.72    4.66
P-Value       0.001   0.000

Remote               -0.139
T-Value               -2.36
P-Value               0.027

S              38.6    35.4
R-Sq          36.60   48.99
R-Sq(adj)     33.96   44.56
Mallows Cp     13.3     8.4
```

Figure 15.11 contains an Excel worksheet created by PHStat2. Although manually creating stepwise results in Excel is not impossible to do, the decision making inherent in adding and deleting variables and the need to cut and paste or delete partial regression results in order to report results makes relying on an add-in such as PHStat2 the only practical choice.

For this example, a significance level of 0.05 is used to enter a variable into the model or to delete a variable from the model. The first variable entered into the model is total staff, the variable that correlates most highly with the dependent variable standby hours. Because the p-value of 0.0011 is less than 0.05, total staff is included in the regression model.

The next step involves selecting a second independent variable for the model. The second variable chosen is one that makes the largest contribution to the model, given that the first variable has been selected. For this model, the second variable is remote hours. Because the p-value of 0.0269 for remote hours is less than 0.05, remote hours is included in the regression model.

After the remote hours variable is entered into the model, the stepwise procedure determines whether total staff is still an important contributing variable or whether it can be eliminated from the model. Because the p-value of 0.0001 for total staff is less than 0.05, total staff remains in the regression model.

The next step involves selecting a third independent variable for the model. Because none of the other variables meets the 0.05 criterion for entry into the model, the stepwise procedure terminates with a model that includes total staff present and the number of remote hours.

This stepwise regression approach to model building was originally developed more than four decades ago, when regression computations on computers were time-consuming and costly. Although stepwise regression limited the evaluation of alternative models, the method was deemed a good trade-off between evaluation and cost.

Given the ability of today's computers to perform regression computations at very low cost and high speed, stepwise regression has been superseded to some extent by the best-subsets approach, discussed next, which evaluates a larger set of alternative models. Stepwise regression is not obsolete, however. Today, many businesses use stepwise regression as part of the research technique called **data mining** (see Online Section 15.7), which tries to identify significant statistical relationships in very large data sets that contain extremely large numbers of variables.

The Best-Subsets Approach to Model Building

The **best-subsets approach** evaluates all possible regression models for a given set of independent variables. Figure 15.12 presents best-subsets regression results of all possible regression models for the standby-hours data.

FIGURE 15.12

Excel and Minitab best-subsets regression results for the standby-hours data

	A	B	C	D	E	F
1	**Best-Subsets Analysis for Standby Hours**					
2						
3	Intermediate Calculations					
4	*R*2T	0.6231				
5	1 - *R*2T	0.3769				
6	n	26				
7	T	5				
8	n - T	21				
9						
10	Model	Cp	k+1	R Square	Adj. R Square	Std. Error
11	X1	13.3215	2	0.3660	0.3396	38.6206
12	X1X2	8.4193	3	0.4899	0.4456	35.3873
13	X1X2X3	7.8418	4	0.5362	0.4729	34.5029
14	X1X2X3X4	5.0000	5	0.6231	0.5513	31.8350
15	X1X2X4	9.3449	4	0.5092	0.4423	35.4921
16	X1X3	10.6486	3	0.4499	0.4021	36.7490
17	X1X3X4	7.7517	4	0.5378	0.4748	34.4426
18	X1X4	14.7982	3	0.3754	0.3211	39.1579
19	X2	33.2078	2	0.0091	-0.0322	48.2836
20	X2X3	32.3067	3	0.0612	-0.0205	48.0087
21	X2X3X4	12.1381	4	0.4591	0.3853	37.2608
22	X2X4	23.2481	3	0.2238	0.1563	43.6540
23	X3	30.3884	2	0.0597	0.0205	47.0345
24	X3X4	11.8231	3	0.4288	0.3791	37.4466
25	X4	24.1846	2	0.1710	0.1365	44.1619

Best Subsets Regression: Standby versus Total Staff, Remote, ...

```
Response is Standby
                                                 T     T
                                                 o     o
                                                 t     t
                                                 a     a
                                                 l     l
                                                   R D
                                                 S e u L
                                                 t m b a
                                                 a o n b
                                 Mallows         f t e o
Vars  R-Sq  R-Sq(adj)                 Cp       S f e r r
   1  36.6       34.0       13.3  38.621 X
   1  17.1       13.7       24.2  44.162       X
   1   6.0        2.1       30.4  47.035     X
   2  49.0       44.6        8.4  35.387 X X
   2  45.0       40.2       10.6  36.749 X   X
   2  42.9       37.9       11.8  37.447     X X
   3  53.8       47.5        7.8  34.443 X   X X
   3  53.6       47.3        7.8  34.503 X X X
   3  50.9       44.2        9.3  35.492 X X   X
   4  62.3       55.1        5.0  31.835 X X X X
```

A criterion often used in model building is the adjusted r^2, which adjusts the r^2 of each model to account for the number of independent variables in the model as well as for the sample size (see Section 14.2). Because model building requires you to compare models with different numbers of independent variables, the adjusted r^2 is more appropriate than r^2. Referring to Figure 15.12, you see that the adjusted r^2 reaches a maximum value of 0.5513 when all four independent variables plus the intercept term (for a total of five estimated parameters) are included in the model.

A second criterion often used in the evaluation of competing models is the C_p statistic developed by Mallows (see reference 1). The $\boldsymbol{C_p}$ **statistic**, defined in Equation (15.9), measures the differences between a fitted regression model and a *true* model, along with random error.

C_p STATISTIC

$$C_p = \frac{(1 - R_k^2)(n - T)}{1 - R_T^2} - [n - 2(k + 1)] \qquad \textbf{(15.9)}$$

where

k = number of independent variables included in a regression model
T = total number of parameters (including the intercept) to be estimated in the full regression model
R_k^2 = coefficient of multiple determination for a regression model that has k independent variables
R_T^2 = coefficient of multiple determination for a full regression model that contains all T estimated parameters

Using Equation (15.9) to compute C_p for the model containing total staff and remote hours,

$$n = 26 \qquad k = 2 \qquad T = 4 + 1 = 5 \qquad R_k^2 = 0.4899 \qquad R_T^2 = 0.6231$$

so that

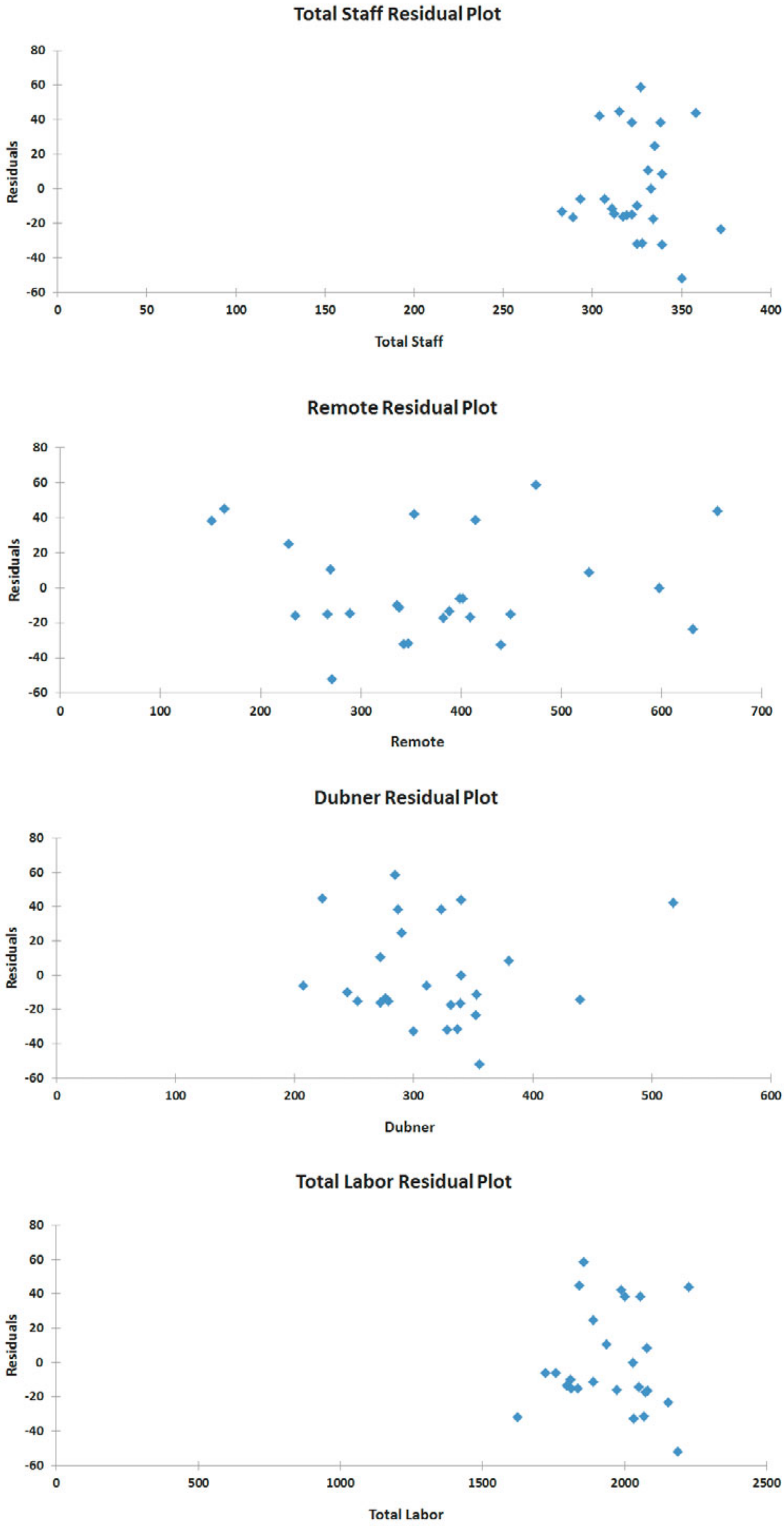
$$C_p = \frac{(1 - 0.4899)(26 - 5)}{1 - 0.6231} - [26 - 2(2 + 1)]$$
$$= 8.4193$$

When a regression model with k independent variables contains only random differences from a *true* model, the mean value of C_p is $k + 1$, the number of parameters. Thus, in evaluating many alternative regression models, the goal is to find models whose C_p is close to or less than $k + 1$. In Figure 15.12, you see that only the model with all four independent variables considered contains a C_p value close to or below $k + 1$. Therefore, using the C_p criterion, you should choose that model.

Although it is not the case here, the C_p statistic often provides several alternative models for you to evaluate in greater depth. Moreover, the best model or models using the C_p criterion might differ from the model selected using the adjusted r^2 and/or the model selected using the stepwise procedure. (Note here that the model selected using stepwise regression has a C_p value of 8.4193, which is substantially above the suggested criterion of $k + 1 = 3$ for that model.) Remember that there may not be a uniquely best model, but there may be several equally appropriate models. Final model selection often involves using subjective criteria, such as parsimony, interpretability, and departure from model assumptions (as evaluated by residual analysis).

When you have finished selecting the independent variables to include in the model, you should perform a residual analysis to evaluate the regression assumptions, and because the data were collected in time order, you also need to compute the Durbin-Watson statistic to determine whether there is autocorrelation in the residuals (see Section 13.6). From Figure 15.10 on page 645, you see that the Durbin-Watson statistic, D, is 2.2197. Because D is greater than 2.0, there is no indication of positive correlation in the residuals. Figure 15.13 presents the plots used in the residual analysis.

FIGURE 15.13
Residual plots for the standby-hours data



None of the residual plots versus the total staff, the remote hours, the Dubner hours, and the total labor hours reveal apparent patterns. In addition, a histogram of the residuals (not shown here) indicates only moderate departure from normality, and a plot of the residuals versus the predicted values of Y (also not shown here) does not show evidence of unequal variance. Thus, from Figure 15.10 on page 645, the regression equation is

$$\hat{Y}_i = -330.8318 + 1.2456X_{1i} - 0.1184X_{2i} - 0.2971X_{3i} + 0.1305X_{4i}$$

Example 15.4 presents a situation in which there are several alternative models in which the C_p statistic is close to or less than $k + 1$.

EXAMPLE 15.4

Choosing Among Alternative Regression Models

Table 15.3 shows results from a best-subsets regression analysis of a regression model with seven independent variables. Determine which regression model you would choose as the *best* model.

SOLUTION From Table 15.3, you need to determine which models have C_p values that are less than or close to $k + 1$. Two models meet this criterion. The model with six independent variables $(X_1, X_2, X_3, X_4, X_5, X_6)$ has a C_p value of 6.8, which is less than $k + 1 = 6 + 1 = 7$, and the full model with seven independent variables $(X_1, X_2, X_3, X_4, X_5, X_6, X_7)$ has a C_p value of 8.0. One way you can choose among the two models is to select the model with the largest adjusted r^2—that is, the model with six independent variables. Another way to select a final model is to determine whether the models contain a subset of variables that are common. Then you test whether the contribution of the additional variables is significant. In this case, because the models differ only by the inclusion of variable X_7 in the full model, you test whether variable X_7 makes a significant contribution to the regression model, given that the variables X_1, X_2, X_3, X_4, X_5, and X_6 are already included in the model. If the contribution is statistically significant, then you should include variable X_7 in the regression model. If variable X_7 does not make a statistically significant contribution, you should not include it in the model.

TABLE 15.3

Partial Results from Best-Subsets Regression

Number of Variables	r^2	Adjusted r^2	C_p	Variables Included
1	0.121	0.119	113.9	X_4
1	0.093	0.090	130.4	X_1
1	0.083	0.080	136.2	X_3
2	0.214	0.210	62.1	X_3, X_4
2	0.191	0.186	75.6	X_1, X_3
2	0.181	0.177	81.0	X_1, X_4
3	0.285	0.280	22.6	X_1, X_3, X_4
3	0.268	0.263	32.4	X_3, X_4, X_5
3	0.240	0.234	49.0	X_2, X_3, X_4
4	0.308	0.301	11.3	X_1, X_2, X_3, X_4
4	0.304	0.297	14.0	X_1, X_3, X_4, X_6
4	0.296	0.289	18.3	X_1, X_3, X_4, X_5
5	0.317	0.308	8.2	X_1, X_2, X_3, X_4, X_5
5	0.315	0.306	9.6	X_1, X_2, X_3, X_4, X_6
5	0.313	0.304	10.7	X_1, X_3, X_4, X_5, X_6
6	0.323	0.313	6.8	$X_1, X_2, X_3, X_4, X_5, X_6$
6	0.319	0.309	9.0	$X_1, X_2, X_3, X_4, X_5, X_7$
6	0.317	0.306	10.4	$X_1, X_2, X_3, X_4, X_6, X_7$
7	0.324	0.312	8.0	$X_1, X_2, X_3, X_4, X_5, X_6, X_7$

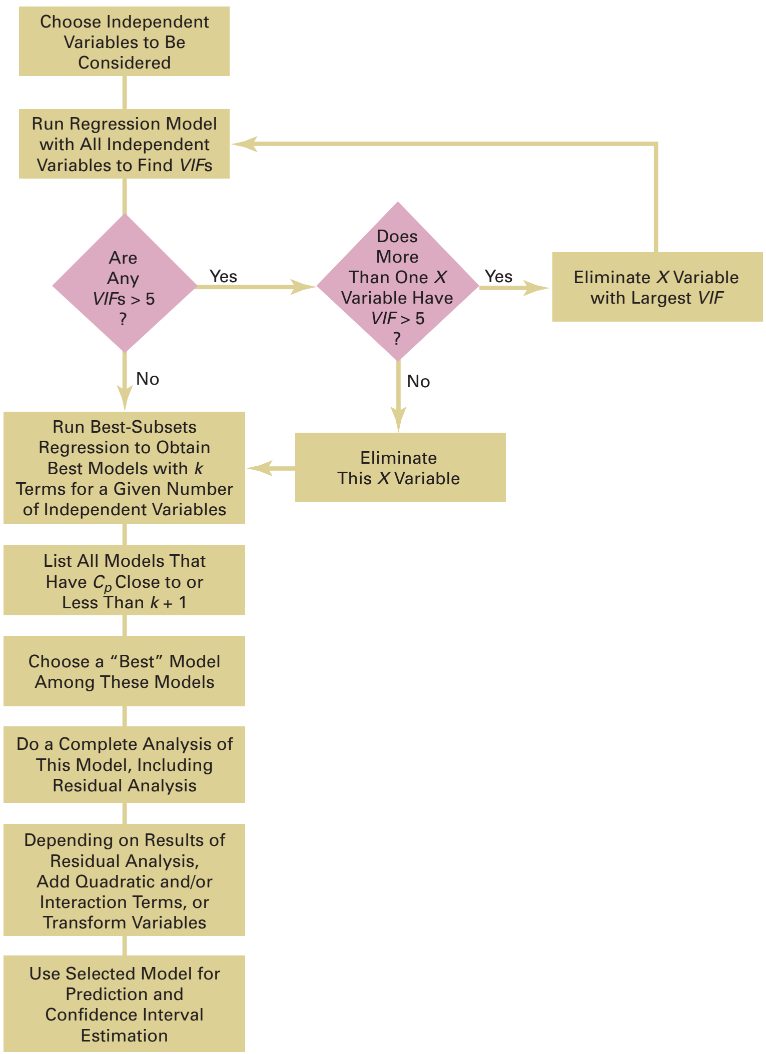
Exhibit 15.1 summarizes the steps involved in model building.

EXHIBIT 15.1
Steps Involved in Model Building

1. Compile a list of all independent variables under consideration.
2. Fit a regression model that includes all the independent variables under consideration and determine the *VIF* for each independent variable. Three possible results can occur:
   a. None of the independent variables has a $VIF > 5$; in this case, proceed to step 3.
   b. One of the independent variables has a $VIF > 5$; in this case, eliminate that independent variable and proceed to step 3.
   c. More than one of the independent variables has a $VIF > 5$; in this case, eliminate the independent variable that has the highest *VIF* and repeat step 2.
3. Perform a best-subsets regression with the remaining independent variables and determine the C_p statistic and/or the adjusted r^2 for each model.
4. List all models that have C_p close to or less than $k + 1$ and/or a high adjusted r^2.
5. From the models listed in step 4, choose a best model.
6. Perform a complete analysis of the model chosen, including a residual analysis.
7. Depending on the results of the residual analysis, add quadratic and/or interaction terms, transform variables, and reanalyze the data.
8. Use the selected model for prediction and inference.

Figure 15.14 represents a roadmap for the steps involved in model building.

FIGURE 15.14
Roadmap for model building



Model Validation

The final step in the model-building process is to validate the selected regression model. This step involves checking the model against data that were not part of the sample analyzed. The following are several ways of validating a regression model:

- Collect new data and compare the results.
- Compare the results of the regression model to previous results.
- If the data set is large, split the data into two parts and cross-validate the results.

Perhaps the best way of validating a regression model is by collecting new data. If the results with new data are consistent with the selected regression model, you have strong reason to believe that the fitted regression model is applicable in a wide set of circumstances.

If it is not possible to collect new data, you can use one of the two other approaches. In one approach, you compare your regression coefficients and predictions to previous results. If the data set is large, you can use **cross-validation**. First, you split the data into two parts. Then you use the first part of the data to develop the regression model. You then use the second part of the data to evaluate the predictive ability of the regression model.

Problems for Section 15.4

LEARNING THE BASICS

15.21 You are considering four independent variables for inclusion in a regression model. You select a sample of $n = 30$, with the following results:

1. The model that includes independent variables A and B has a C_p value equal to 4.6.
2. The model that includes independent variables A and C has a C_p value equal to 2.4.
3. The model that includes independent variables A, B, and C has a C_p value equal to 2.7.

a. Which models meet the criterion for further consideration? Explain.

b. How would you compare the model that contains independent variables A, B, and C to the model that contains independent variables A and B? Explain.

15.22 You are considering six independent variables for inclusion in a regression model. You select a sample of $n = 40$, with the following results:

$$k = 2 \quad T = 6 + 1 = 7 \quad R_k^2 = 0.274 \quad R_T^2 = 0.653$$

a. Compute the C_p value for this two-independent-variable model.

b. Based on your answer to (a), does this model meet the criterion for further consideration as the best model? Explain.

APPLYING THE CONCEPTS

15.23 In Problems 13.85 through 13.89 on page 568, you constructed simple linear regression models to investigate the relationship between demographic information and monthly sales for a chain of sporting goods stores using the data in **Sporting**. Develop the most appropriate multiple regression model to predict a store's monthly sales. Be sure to include a thorough residual analysis. In addition, provide a detailed explanation of the results, including a comparison of the most appropriate multiple regression model to the best simple linear regression model.

15.24 You need to develop a model to predict the selling price of houses in a small city, based on assessed value, time in months since the house was reassessed, and whether the house is new $(0 = \text{no}, 1 = \text{yes})$. A sample of 30 recently sold single-family houses that were reassessed at full value one year prior to the study is selected and the results are stored in **House1**. Develop the most appropriate multiple regression model to predict selling price. Be sure to perform a thorough residual analysis. In addition, provide a detailed explanation of the results.

15.25 The human resources (HR) director for a large company that produces highly technical industrial instrumentation devices has the business objective of improving recruiting decisions concerning sales managers.

The company has 45 sales regions, each headed by a sales manager. Many of the sales managers have degrees in electrical engineering and, due to the technical nature of the product line, several company officials believe that only applicants with degrees in electrical engineering should be considered. At the time of their application, candidates are asked to take the Strong-Campbell Interest Inventory Test and the Wonderlic Personnel Test. Due to the time and money involved with the testing, some discussion has taken place about dropping one or both of the tests. To start, the HR director gathered information on each of the 45 current sales managers, including years of selling experience, electrical engineering background, and the scores from both the Wonderlic and Strong-Campbell tests. The HR director has decided to use regression modeling to predict a dependent variable of "sales index" score, which is the ratio of the regions' actual sales divided by the target sales. The target values are constructed each year by upper management, in consultation with the sales managers, and are based on past performance and market potential within each region. The file **Managers** contains information on the 45 current sales managers. The following variables are included:

Sales—Ratio of yearly sales divided by the target sales value for that region. The target values were mutually agreed-upon "realistic expectations."

Wonder—Score from the Wonderlic Personnel Test. The higher the score, the higher the applicant's perceived ability to manage.

SC—Score on the Strong-Campbell Interest Inventory Test. The higher the score, the higher the applicant's perceived interest in sales.

Experience—Number of years of selling experience prior to becoming a sales manager.

Engineer—Dummy variable that equals 1 if the sales manager has a degree in electrical engineering and 0 otherwise.

a. Develop the most appropriate regression model to predict sales.

b. Do you think that the company should continue administering both the Wonderlic and Strong-Campbell tests? Explain.

c. Do the data support the argument that electrical engineers outperform the other sales managers? Would you support the idea to hire only electrical engineers? Explain.

d. How important is prior selling experience in this case? Explain.

e. Discuss in detail how the HR director should incorporate the regression model you developed into the recruiting process.

15.5 Pitfalls in Multiple Regression and Ethical Issues

Pitfalls in Multiple Regression

Model building is an art as well as a science. Different individuals may not always agree on the best multiple regression model. To try to construct a best regression model, you should use the process described in Exhibit 15.1 on page 651. In doing so, you must avoid certain pitfalls that can interfere with the development of a useful model. Section 13.9 discussed pitfalls in simple linear regression and strategies for avoiding them. Now that you have studied a variety of multiple regression models, you need to take some additional precautions. To avoid pitfalls in multiple regression, you also need to

- Interpret the regression coefficient for a particular independent variable from a perspective in which the values of all other independent variables are held constant.
- Evaluate residual plots for each independent variable.
- Evaluate interaction and quadratic terms.
- Compute the *VIF* for each independent variable before determining which independent variables to include in the model.
- Examine several alternative models, using best-subsets regression.
- Validate the model before implementing it.

Ethical Issues

Ethical issues arise when a user who wants to make predictions manipulates the development process of the multiple regression model. The key here is intent. In addition to the situations discussed in Section 13.9, unethical behavior occurs when someone uses multiple regression analysis and *willfully fails* to remove from consideration independent variables that exhibit a high collinearity with other independent variables or *willfully fails* to use methods other than least-squares regression when the assumptions necessary for least-squares regression are seriously violated.

15.6 *Online Topic:* Influence Analysis

Influence analysis measures the influence of individual observations on a regression model. To study this topic, read the Section 15.6 online topic file that is available on this book's companion website. (See Appendix C to learn how to access the online topic files.)

15.7 *Online Topic:* Analytics and Data Mining

Analytics and data mining are methods that are used with very large data sets to present summary results and to discern patterns that may exist. To study this topic, read the Section 15.7 online topic file that is available on this book's companion website. (See Appendix C to learn how to access the online topic files.)

USING STATISTICS @ WHIT-DT Revisited

In the Using Statistics scenario, you were the operations manager of WHIT-DT, looking for ways to reduce labor expenses. You needed to determine which variables have an effect on standby hours, the time during which unionized graphic artists are idle but are getting paid. You have collected data concerning standby hours and the total number of staff present, remote hours, Dubner hours, and total labor hours over a period of 26 weeks.

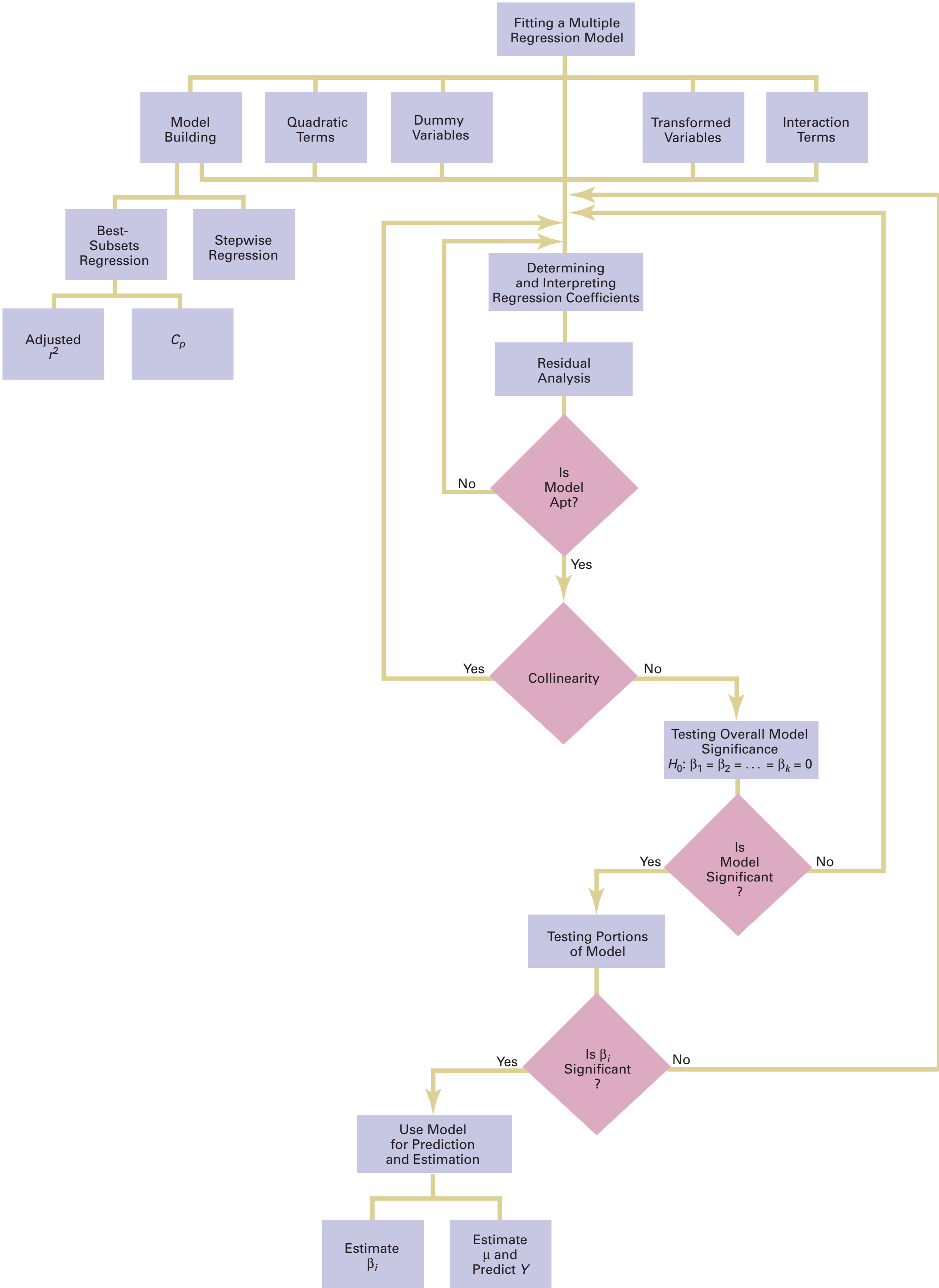
You performed a multiple regression analysis on the data. The coefficient of multiple determination indicated that 62.31% of the variation in standby hours can be explained by variation in the total number of staff present, remote hours, Dubner hours, and total labor hours. The model indicated that standby hours are estimated to increase by 1.2456 hours for each additional staff hour holding constant the other independent variables; to decrease by 0.1184 hour for each additional remote hour holding constant the other independent variables; to decrease by 0.2974 hour for each additional Dubner hour holding constant the other independent variables; and to increase by 0.1305 hour for each additional labor hour holding constant the other independent variables. Each of the four independent variables had a significant effect on standby hours holding constant the other independent variables. This regression model enables you to predict standby hours based on the total number of staff present, remote hours, Dubner hours, and total labor hours. It also enables you to investigate how changing each of these four independent variables could affect standby hours.

SUMMARY

In this chapter, various multiple regression topics were considered (see Figure 15.15), including quadratic regression models, transformations, collinearity, and model building.

FIGURE 15.15

Roadmap for multiple regression



KEY EQUATIONS

Quadratic Regression Model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{1i}^2 + \varepsilon_i \quad \textbf{(15.1)}$$

Quadratic Regression Equation

$$\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{1i}^2 \quad \textbf{(15.2)}$$

Regression Model with a Square-Root Transformation

$$Y_i = \beta_0 + \beta_1 \sqrt{X_{1i}} + \varepsilon_i \quad \textbf{(15.3)}$$

Original Multiplicative Model

$$Y_i = \beta_0 X_{1i}^{\beta_1} X_{2i}^{\beta_2} \varepsilon_i \quad \textbf{(15.4)}$$

Transformed Multiplicative Model

$$\begin{aligned} \log Y_i &= \log(\beta_0 X_{1i}^{\beta_1} X_{2i}^{\beta_2} \varepsilon_i) \quad \textbf{(15.5)} \\ &= \log\beta_0 + \log(X_{1i}^{\beta_1}) + \log(X_{2i}^{\beta_2}) + \log \varepsilon_i \\ &= \log\beta_0 + \beta_1 \log X_{1i} + \beta_2 \log X_{2i} + \log \varepsilon_i \end{aligned}$$

Original Exponential Model

$$Y_i = e^{\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}} \varepsilon_i \quad \textbf{(15.6)}$$

Transformed Exponential Model

$$\begin{aligned} \ln Y_i &= \ln(e^{\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}} \varepsilon_i) \quad \textbf{(15.7)} \\ &= \ln(e^{\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}}) + \ln \varepsilon_i \\ &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ln \varepsilon_i \end{aligned}$$

Variance Inflationary Factor

$$VIF_j = \frac{1}{1 - R_j^2} \quad \textbf{(15.8)}$$

C_p Statistic

$$C_p = \frac{(1 - R_k^2)(n - T)}{1 - R_T^2} - [n - 2(k + 1)] \quad \textbf{(15.9)}$$

KEY TERMS

CHAPTER REVIEW PROBLEMS

CHECKING YOUR UNDERSTANDING

15.26 How can you evaluate whether collinearity exists in a multiple regression model?

15.27 What is the difference between stepwise regression and best-subsets regression?

15.28 How do you choose among models according to the C_p statistic in best-subsets regression?

APPLYING THE CONCEPTS

15.29 Crazy Dave has expanded his analysis, presented in Problem 14.73 on page 619, of which variables are important in predicting a team's wins in a given baseball season. He has collected data in **BB2009** related to wins, ERA, saves, runs scored, hits allowed, walks allowed, and errors for the 2009 season.

a. Develop the most appropriate multiple regression model to predict a team's wins. Be sure to include a thorough residual analysis. In addition, provide a detailed explanation of the results.

b. Develop the most appropriate multiple regression model to predict a team's ERA on the basis of hits allowed, walks allowed, errors, and saves. Be sure to include a thorough residual analysis. In addition, provide a detailed explanation of the results.

15.30 Professional basketball has truly become a sport that generates interest among fans around the world. More and more players come from outside the United States to play in the National Basketball Association (NBA). Many factors could impact the number of wins achieved by each NBA team. In addition to the number of wins, the file **NBA2010** contains team statistics for points per game (for team, opponent, and the difference between team and opponent), field goal (shots made) percentage (for team, opponent, and the difference between team and opponent), steals per game (for team, opponent, and the difference between team and opponent), rebounds per game (for team, opponent, and the difference between team and opponent).

a. Consider team points per game, opponent points per game, team field goal percentage, opponent field goal percentage, steals per game, and rebounds per game as independent variables for possible inclusion in the multiple regression model. Develop the most appropriate multiple regression model to predict the number of wins.

b. Consider the difference between team points and opponent points per game, the difference between team field goal percentage and opponent field goal percentage, the difference in team and opponent steals, and the difference between team and opponent rebounds per game as independent variables for possible inclusion in the multiple regression model. Develop the most appropriate multiple regression model to predict the number of wins.

c. Compare the results of (a) and (b). Which model is better for predicting the number of wins? Explain.

15.31 Hemlock Farms is a community located in the Pocono Mountains area of eastern Pennsylvania. The file **HemlockFarms** contains information on homes that were recently for sale. The variables included were

List Price—Asking price of the house
Hot Tub—Whether the house has a hot tub, with 0 = No and 1 = Yes
Lake View—Whether the house has a lake view, with 0 = No and 1 = Yes
Bathrooms—Number of bathrooms
Bedrooms—Number of bedrooms
Loft/Den—Whether the house has a loft or den, with 0 = No and 1 = Yes
Finished basement—Whether the house has a finished basement, with 0 = No and 1 = Yes
Acres—Number of acres for the property

Develop the most appropriate multiple regression model to predict the asking price. Be sure to perform a thorough residual analysis. In addition, provide a detailed explanation of your results.

15.32 Nassau County is located approximately 25 miles east of New York City. Data in **GlenCove** are from a sample of 30 single-family homes located in Glen Cove. Variables included are the appraised value, land area of the property (acres), interior size of the house (square feet), age (years), number of rooms, number of bathrooms, and number of cars that can be parked in the garage.

a. Develop the most appropriate multiple regression model to predict appraised value.

b. Compare the results in (a) with those of Problems 15.33 (a) and 15.34 (a).

15.33 Data similar to those in Problem 15.32 are available for homes located in Roslyn (approximately 8 miles from Glen Cove) and are stored in **Roslyn**.

a. Perform an analysis similar to that of Problem 15.32.

b. Compare the results in (a) with those of Problems 15.32 (a) and 15.34 (a).

15.34 Data similar to Problem 15.32 are available for homes located in Freeport (located approximately 20 miles from Roslyn) and are stored in **Freeport**.

a. Perform an analysis similar to that of Problem 15.32.

b. Compare the results in (a) with those of Problems 15.32 (a) and 15.33 (a).

15.35 You are a real estate broker who wants to compare property values in Glen Cove and Roslyn (which are located approximately 8 miles apart). Use the data in **GCRoslyn**. Make sure to include the dummy variable for location (Glen Cove or Roslyn) in the regression model.

a. Develop the most appropriate multiple regression model to predict appraised value.

b. What conclusions can you reach concerning the differences in appraised value between Glen Cove and Roslyn?

15.36 You are a real estate broker who wants to compare property values in Glen Cove, Freeport, and Roslyn. Use the data in **GCFreeRoslyn**.

a. Develop the most appropriate multiple regression model to predict appraised value.

b. What conclusions can you reach concerning the differences in appraised value between Glen Cove, Freeport, and Roslyn?

15.37 Over the past 30 years, public awareness and concern about air pollution have escalated dramatically. Venturi scrubbers are used for the removal of submicron particulate matter from smoke stacks. An experiment was conducted to determine the effect of air flow rate, water flow rate (liters/minute), recirculating water flow rate (liters/minute), and orifice size (mm) in the air side of the pneumatic nozzle on the performance of the scrubber, as measured by the number of transfer units. The results are stored in **Scrubber**.

Develop the most appropriate multiple regression model to predict the number of transfer units. Be sure to perform a thorough residual analysis. In addition, provide a detailed explanation of your results.

Source: Data extracted from D. A. Marshall, R. J. Sumner, and C. A. Shook, "Removal of SiO_2 Particles with an Ejector Venturi Scrubber," *Environmental Progress*, 14 (1995), 28–32.

15.38 A recent article (J. Conklin, "It's a Marathon, Not a Sprint," *Quality Progress*, June 2009, pp. 46–49) discussed a metal deposition process in which a piece of metal is placed in an acid bath and an alloy is layered on top of it. The key quality characteristic is the thickness of the alloy layer. The file **Thickness** contains the following variables:

Thickness—Thickness of the alloy layer
Catalyst—Catalyst concentration in the acid bath
pH—pH level of the acid bath
Pressure—Pressure in the tank holding the acid bath
Temp—Temperature in the tank holding the acid bath
Voltage—Voltage applied to the tank holding the acid bath

Develop the most appropriate multiple regression model to predict the thickness of the alloy layer. Be sure to perform a thorough residual analysis. The article suggests that there is a significant interaction between the pressure and the temperature in the tank. Do you agree?

15.39 A headline in *The New York Times* on March 4, 1990, read: "Wine equation puts some noses out of joint." The article explained that Professor Orley Ashenfelter, a Princeton University economist, had developed a multiple regression model to predict the quality of French Bordeaux, based on the amount of winter rain, the average temperature during the growing season, and the harvest rain. The multiple regression equation is

$$Q = -12.145 + 0.00117WR + 0.6164TMP - 0.00386HR$$

where

Q = logarithmic index of quality
WR = winter rain (October through March), in millimeters
TMP = average temperature during the growing season (April through September), in degrees Celsius
HR = harvest rain (August to September), in millimeters

You are at a cocktail party, sipping a glass of wine, when one of your friends mentions to you that she has read the article. She asks you to explain the meaning of the coefficients in the equation and also asks you about analyses that might have been done and were not included in the article. What is your reply?

REPORT WRITING EXERCISE

15.40 In Problem 15.23 on page 652, you developed a multiple regression model to predict monthly sales at sporting goods stores for the data stored in **Sporting**. Now write a report based on the model you developed. Append all appropriate charts and statistical information to your report.

TEAM PROJECT

15.41 The file **Bond Funds** contains information regarding eight variables from a sample of 184 bond mutual funds:

Type—Type of bonds comprising the bond mutual fund (intermediate government or short-term corporate)
Assets—In millions of dollars
Fees—Sales charges (no or yes)
Expense ratio—Ratio of expenses to net assets in percentage
Return 2009—Twelve-month return in 2009
Three-year return—Annualized return, 2007–2009
Five-year return—Annualized return, 2005–2009
Risk—Risk-of-loss factor of the bond mutual fund (below average, average, or above average)

Develop regression models to predict the 2009 return, the three-year return, and the five-year return, based on fees, expense ratio, type, and risk. (For the purpose of this analysis, combine below-average risk and average risk into one category.) Be sure to perform a thorough residual analysis. In addition, provide a detailed explanation of your results. Append all appropriate charts and statistical information to your report.

THE MOUNTAIN STATES POTATO COMPANY

Mountain States Potato Company sells a by-product of its potato-processing operation, called a filter cake, to area feedlots as cattle feed. The business problem faced by the feedlot owners is that the cattle are not gaining weight as quickly as they once were. The feedlot owners believe that the root cause of the problem is that the percentage of solids in the filter cake is too low.

Historically, the percentage of solids in the filter cakes ran slightly above 12%. Lately, however, the solids are running in the 11% range. What is actually affecting the solids is a mystery, but something has to be done quickly. Individuals involved in the process were asked to identify variables that might affect the percentage of solids. This review turned up the six variables (in addition to the percentage of solids) listed in the table on page 659. Data collected by monitoring the process several times daily for 20 days are stored in **Potato**.

Variable	Comments
SOLIDS	Percentage of solids in the filter cake.
PH	Acidity. This measure of acidity indicates bacterial action in the clarifier and is controlled by the amount of downtime in the system. As bacterial action progresses, organic acids are produced that can be measured using pH.
LOWER	Pressure of the vacuum line below the fluid line on the rotating drum.
UPPER	Pressure of the vacuum line above the fluid line on the rotating drum.
THICK	Filter cake thickness, measured on the drum.
VARIDRIV	Setting used to control the drum speed. May differ from DRUMSPD due to mechanical inefficiencies.
DRUMSPD	Speed at which the drum is rotating when collecting the filter cake. Measured with a stopwatch.

1. Thoroughly analyze the data and develop a regression model to predict the percentage of solids.

2. Write an executive summary concerning your findings to the president of the Mountain States Potato Company. Include specific recommendations on how to get the percentage of solids back above 12%.

DIGITAL CASE

Apply your knowledge of multiple regression model building in this Digital Case, which extends the Chapter 14 OmniFoods Using Statistics scenario.

Still concerned about ensuring a successful test marketing of its OmniPower energy bars, the marketing department of OmniFoods has contacted Connect2Coupons (C2C), another merchandising consultancy. C2C suggests that earlier analysis done by In-Store Placements Group (ISPG) was faulty because it did not use the correct type of data. C2C claims that its Internet-based viral marketing will have an even greater effect on OmniPower energy bar sales, as new data from the same 34-store sample will show. In response, ISPG says its earlier claims are valid and has reported to the OmniFoods marketing department that it can discern no simple relationship between C2C's viral marketing and increased OmniPower sales.

Open **OmniPowerForum15.pdf** to review all the claims made in a private online forum and chat hosted on the OmniFoods corporate website. Then answer the following:

1. Which of the claims are true? False? True but misleading? Support your answer by performing an appropriate statistical analysis.

2. If the grocery store chain allowed OmniFoods to use an unlimited number of sales techniques, which techniques should it use? Explain.

3. If the grocery store chain allowed OmniFoods to use only one sales technique, which technique should it use? Explain.

REFERENCES

1. Kutner, M., C. Nachtsheim, J. Neter, and W. Li, *Applied Linear Statistical Models*, 5th ed. (New York: McGraw-Hill/Irwin, 2005).
2. Marquardt, D. W., "You Should Standardize the Predictor Variables in Your Regression Models," discussion of "A Critique of Some Ridge Regression Methods," by G. Smith and F. Campbell, *Journal of the American Statistical Association*, 75 (1980), 87–91.
3. *Microsoft Excel 2010* (Redmond, WA: Microsoft Corp., 2010).
4. *Minitab Release 16* (State College, PA: Minitab, Inc., 2010).
5. Snee, R. D., "Some Aspects of Nonorthogonal Data Analysis, Part I. Developing Prediction Equations," *Journal of Quality Technology*, 5 (1973), 67–79.

CHAPTER 15 EXCEL GUIDE

EG15.1 The QUADRATIC REGRESSION MODEL

To the worksheet that contains your regression data, add a new column of formulas that computes the square of one of the independent variables to create a quadratic term. For example, to create a quadratic term for the Section 15.1 fly ash analysis, open to the **DATA worksheet** of the **FlyAsh workbook**. That worksheet contains the independent variable **FlyAsh%** in column A and the dependent variable **Strength** in column B. While the quadratic term **FlyAsh%^2** could be created in any column, a good practice is to place independent variables in contiguous columns. (You must follow this practice if you are using the Analysis ToolPak Regression procedure.) To do so, first select column B (**Strength**), right-click, and click **Insert** from the shortcut menu to add a new column B. (Strength becomes column C.) Enter the label **FlyAsh%^2** in cell B1 and then enter the formula **=A2^2** in cell **B2**. Copy this formula down the column through all the data rows.

To perform a regression analysis using this new variable, apply the Section EG14.1 instructions on page 622.

EG15.2 USING TRANSFORMATIONS in REGRESSION MODELS

The Square-Root Transformation

To the worksheet that contains your regression data, add a new column of formulas that computes the square root of one of the independent variables to create a square-root transformation. For example, to create a square root transformation in a blank column D for an independent variable in a column C, enter the formula **=SQRT(C2)** in cell D2 of that worksheet and copy the formula down through all data rows. If the rightmost column in the worksheet contains the dependent variable, first select that column, right-click, and click **Insert** from the shortcut menu and place the transformation in that new column.

The Log Transformation

To the worksheet that contains your regression data, add a new column of formulas that compute the common (base 10) logarithm or natural logarithm (base *e*) of one of the independent variables to create a log transformation. For example, to create a common logarithm transformation in a blank column D for an independent variable in a column C, enter the formula **=LOG(C2)** in cell D2 of that worksheet and copy the formula down through all data rows. To create a natural logarithm transformation in a blank column D for an independent variable in a column C, enter the formula **=LN(C2)** in cell D2 of that worksheet and copy the formula down through all data rows.

If the dependent variable appears in a column to the immediate right of the independent variable being transformed, first select the dependent variable column, right-click, and click **Insert** from the shortcut menu and then place the transformation of the independent variable in that new column.

EG15.3 COLLINEARITY

PHStat2 To compute the variance inflationary factor, use the Section EG14.1 “Interpreting the Regression Coefficients” *PHStat2* instructions on page 622 but modify step 6 by checking **Variance Inflationary Factor** (*VIF*) before you click **OK**. The *VIF* will appear in cell B9 of the regression results worksheet, immediately following the Regression Statistics area.

In-Depth Excel To compute the variance inflationary factor, first use the Section EG14.1 “Interpreting the Regression Coefficients” *In-Depth Excel* instructions on page 622 to create regression results worksheets for every combination of independent variables in which one serves as the dependent variable. Then, in each of the regression results worksheets, enter the label *VIF* in cell **A9** and enter the formula **=1/(1 - B5)** in cell **B9** to compute the *VIF*.

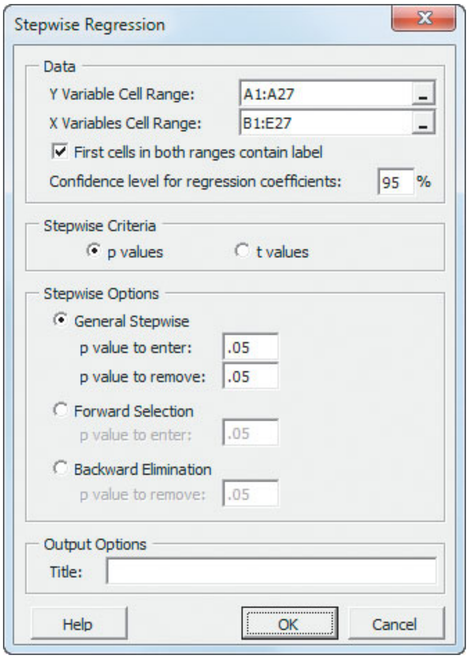
EG15.4 MODEL BUILDING

The Stepwise Regression Approach to Model Building

PHStat2 Use **Stepwise Regression** to use the stepwise regression approach to model building. For example, to create the Figure 15.11 stepwise analysis of the standby-hours data on page 646, open to the **DATA worksheet** of the **Standby workbook.** Select **PHStat → Regression → Stepwise Regression**. In the procedure’s dialog box (shown on page 661):

1. Enter **A1:A27** as the **Y Variable Cell Range**.
2. Enter **B1:E27** as the **X Variables Cell Range**.
3. Check **First cells in both ranges contain label**.
4. Enter **95** as the **Confidence level for regression coefficients**.
5. Click **p values** as the **Stepwise Criteria**.

6. Click **General Stepwise** and keep the pair of **.05** values as the **p value to enter** and the **p value to remove**.
7. Enter a **Title** and click **OK**.

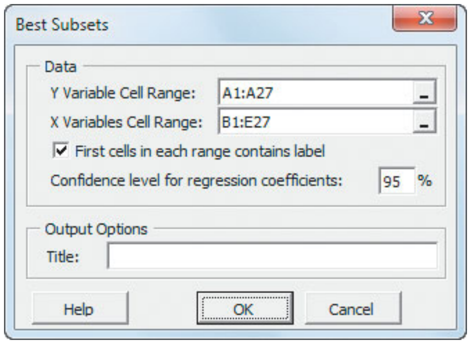


This procedure may take a noticeable amount of time to create its results. The procedure finishes when the statement "Stepwise ends" (as shown in row 29 in the Figure 15.11 Excel results on page 646) is added to the stepwise regression results worksheet.

The Best-Subsets Approach to Model Building

PHStat2 Use **Best Subsets** to use a best-subsets approach to model building. For example, to create the Figure 15.12 best subsets analysis of the standby-hours data on page 647, open to the **DATA worksheet** of the **Standby workbook**. Select **PHStat ➔ Regression ➔ Best Subsets**. In the procedure's dialog box (shown below):

1. Enter **A1:A27** as the **Y Variable Cell Range**.
2. Enter **B1:E27** as the **X Variables Cell Range**.
3. Check **First cells in each range contains label**.
4. Enter **95** as the **Confidence level for regression coefficients**.
5. Enter a **Title** and click **OK**.



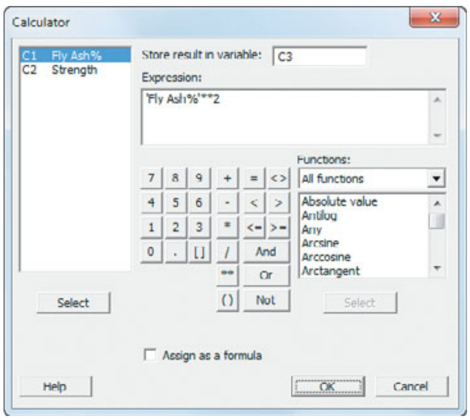
This procedure creates many regression results worksheets (seen as a flickering in the Excel windows) as it evaluates each subset of independent variables.

CHAPTER 15 MINITAB GUIDE

MG15.1 The QUADRATIC REGRESSION MODEL

Use **Calculator** to compute the square of one of the independent variables to create a quadratic term. For example, to create a quadratic term for the Section 15.1 fly ash analysis, open to the **FlyAsh worksheet**. Select **Calc ➔ Calculator**. In the Calculator dialog box (shown in the right column):

1. Enter **C3** in the **Store result in variable** box and press **Tab**.
2. Double-click **C1 Fly Ash%** in the variables list to add **'Fly Ash%'** to the **Expression** box.
3. Click ** and then **2** on the simulated calculator keypad to add ****2** to the **Expression** box.
4. Click **OK**.
5. Enter **Fly Ash%^2** as the name for column **C3**.



To perform a regression analysis using this new variable, see Section MG14.1 on page 625.

MG15.2 USING TRANSFORMATIONS in REGRESSION MODELS

Use **Calculator** to transform a variable. Open to the worksheet that contains your regression data. Select **Calc ➔ Calculator**. In the Calculator dialog box:

1. Enter the name of the empty column that will contain the transformed values in the **Store result in variable** box and press **Tab**.
2. Select **All functions** from the **Functions** drop-down list.
3. In the list of functions, select one of these choices: **Square root**, **Log base 10**, or **Natural log (log base e)**. Selecting these choices enters **SQRT(number)**, **LOGTEN(number)**, or **LN(number)**, respectively, in the **Expression** box.
4. Double-click the name of the variable to be transformed in the variables list to replace **number** with the variable name in the **Expression** box.
5. Click **OK**.
6. Enter a column name for the transformed values.

To perform a regression analysis using this new variable, see Section MG14.1 on page 625.

MG15.3 COLLINEARITY

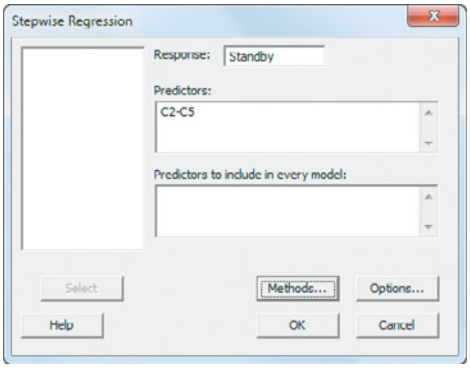
To compute the variance inflationary factor, modify the Section MG14.1 "Interpreting the Regression Coefficients" instructions on page 625. In step 15, check **Variance inflation factors** while clearing the other **Display** and **Lack of Fit Test** check boxes in the Regression - Options dialog box.

MG15.4 MODEL BUILDING

The Stepwise Regression Approach to Model Building

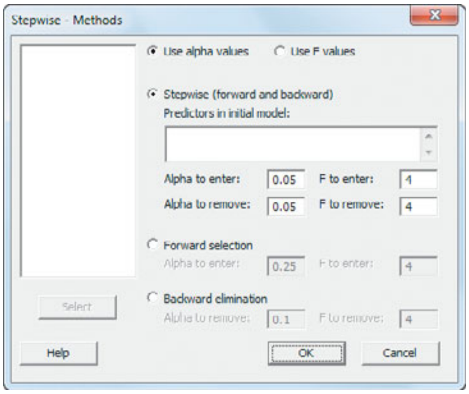
Use **Stepwise** to use the stepwise regression approach to model building. For example, to create the Figure 15.11 stepwise analysis of the standby-hours data on page 646, open to the **Standby worksheet**. Select **Stat ➔ Regression ➔ Stepwise**. In the Stepwise Regression dialog box (shown in the right column):

1. Double-click **C1 Standby** in the variables list to add **Standby** in the **Response** box.
2. Enter **C2-C5** in the **Predictors** box. (Entering **C2-C5** is a shortcut way of referring to the four variables in columns 2 through 5. This shortcut avoids having to double-click the name of each of these variables in order to add them to the Predictors box.)
3. Click **Methods**.



In the Stepwise-Methods dialog box (shown below):

4. Click **Use alpha values**.
5. Click **Stepwise**.
6. Enter **0.05** in the **Alpha to enter** box and **0.05** in the **Alpha to remove** box.
7. Click **OK**.



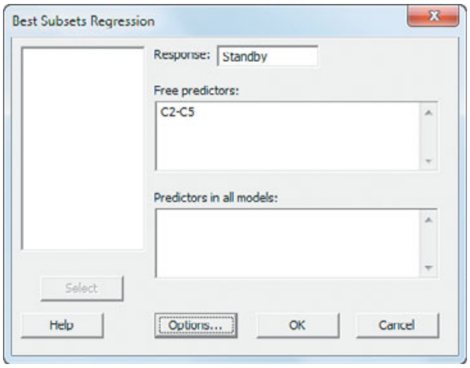
8. Back in the Stepwise Regression dialog box, click **OK**.

The Best-Subsets Approach to Model Building

Use **Best Subsets** to use a best-subsets approach to model building. For example, to create the Figure 15.12 stepwise analysis of the standby-hours data on page 647, open to the **Standby worksheet**. Select **Stat ➔ Regression ➔ Best Subsets**. In the Best Subsets Regression dialog box (shown on page 663):

1. Double-click **C1 Standby** in the variables list to add **Standby** in the **Response** box.
2. Enter **C2-C5** in the **Free Predictors** box. (Entering **C2-C5** is a shortcut way of referring to the four variables in columns 2 through 5 as explained in the previous set of instructions.)

3. Click **Options**.



In the Best Subsets Regression - Options dialog box (shown in the right column):

4. Enter **1** in the **Minimum box** and keep the **Maximum** box empty.
5. Enter **3** in the **Models of each size to print** box.
6. Check **Fit intercept**
7. Click **OK**.
8. Back in the Best Subsets Regression dialog box, click **OK**.