CHAPTER 15

Multiple Regression and Model Building

Learning Objectives

After mastering the material in this chapter, you will be able to:

- **LO15-1** Explain the multiple regression model and the related least squares point estimates.
- **LO15-2** Explain the assumptions behind multiple regression and calculate the standard error.
- **LO15-3** Calculate and interpret the multiple and adjusted multiple coefficients of determination.
- **LO15-4** Test the significance of a multiple regression model by using an *F*-test.
- **LO15-5** Test the significance of a single independent variable.
- **LO15-6** Find and interpret a confidence interval for a mean value of the dependent variable and a prediction interval for an individual value.
- **LO15-7** Use dummy variables to model qualitative independent variables.
- **LO15-8** Use squared and interaction variables.
- **LO15-9** Describe multicollinearity and build a multiple regression model.
- **LO15-10** Use residual analysis to check the assumptions of multiple regression.
- **LO15-11** Use a logistic model to estimate probabilities and odds ratios.

Chapter Outline

- **15.1** The Multiple Regression Model and the Least Squares Point Estimates
- **15.2** Model Assumptions and the Standard Error
- **15.3** R^2 and Adjusted R^2 (This section can be read anytime after reading Section 15.1)
- **15.4** The Overall *F*-Test
- **15.5** Testing the Significance of an Independent Variable
- **15.6** Confidence and Prediction Intervals
- **15.7** The Sales Representative Case: Evaluating Employee Performance
- **15.8** Using Dummy Variables to Model Qualitative Independent Variables
- **15.9** Using Squared and Interaction Variables
- **15.10** Model Building and the Effects of Multicollinearity
- **15.11** Residual Analysis in Multiple Regression
- **15.12** Logistic Regression

Note: After completing Section 15.7, the reader may study Sections 15.8, 15.9, 15.10, 15.11, and 15.12 in any order without loss of continuity.

Often we can more accurately describe, predict, and control a dependent variable by using a regression model that employs more than one independent variable. Such a model is called a **multiple regression model,** which is the subject of this chapter.

In order to explain the ideas of this chapter, we consider the following cases:

The Tasty Sub Shop Case: The business entrepreneur more accurately predicts the yearly revenue for a potential restaurant site by using a multiple regression model that employs as independent variables (1) the number of residents living near the site and (2) a rating of the amount of business and shopping near the site. The entrepreneur uses the more accurate predictions given by the multiple regression model to more accurately assess the profitability of the potential restaurant site.

The Sales Representative Case: A sales manager evaluates the performance of sales representatives by using a multiple regression model that predicts sales performance on the basis of five independent variables. Salespeople whose actual performance is far worse than predicted performance will get extra training to help improve their sales techniques.

15.1 The Multiple Regression Model and the Least Squares Point Estimates

LO15-1 Explain the multiple regression model and the related least squares point estimates.

Regression models that employ more than one independent variable are called **multiple regression models.** We begin our study of these models by considering the following example.

EXAMPLE 15.1 The Tasty Sub Shop Case: A Multiple Regression Model

Part 1: The data and a regression model Consider the Tasty Sub Shop problem in which the business entrepreneur wishes to predict yearly revenue for potential Tasty Sub restaurant sites. In Chapter 14 we used the number of residents, or population size x, living near a site to predict y, the yearly revenue for a Tasty Sub Shop built on the site. We now consider predicting y on the basis of the population size and a second predictor variable—the business rating. The business rating for a restaurant site reflects the amount of business and shopping near the site. This rating is expressed as a whole number between 1 and 10. Sites having only limited business and shopping nearby do not provide many potential customers—shoppers or local employees likely to eat in a Tasty Sub Shop—so they receive ratings near 1. However, sites located near substantial business and shopping activity do provide many potential customers for a Tasty Sub Shop, so they receive much higher ratings. The best possible rating for business activity is 10.

The business entrepreneur has collected data concerning yearly revenue (y), population size (x_1), and business rating (x_2) for 10 existing Tasty Sub restaurants that are built on sites similar to the site the entrepreneur is considering. These data are given in Table 15.1.

TABLE 15.1 The Tasty Sub Shop Revenue Data DS TastySub2

Restaurant	Population Size, x_1 (Thousands of Residents)	Business Rating, x_2	Yearly Revenue, y (Thousands of Dollars)
1	20.8	3	527.1
2	27.5	2	548.7
3	32.3	6	767.2
4	37.2	5	722.9
5	39.6	8	826.3
6	45.1	3	810.5
7	49.9	9	1040.5
8	55.4	5	1033.6
9	61.7	4	1090.3
10	64.6	7	1235.8

FIGURE 15.1 Plot of y (Yearly Revenue) versus x_1 (Population Size)

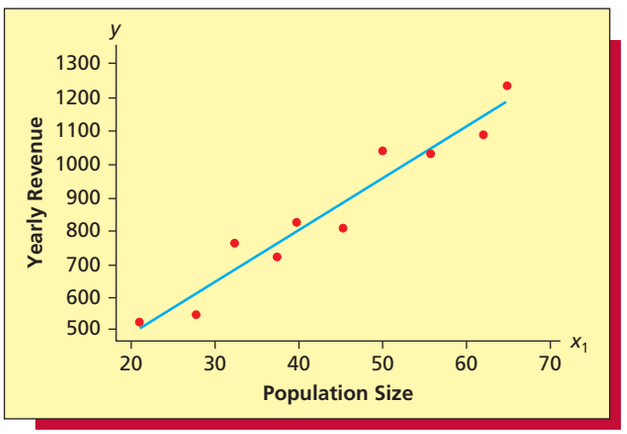


FIGURE 15.2 Plot of y (Yearly Revenue) versus x_2 (Business Rating)

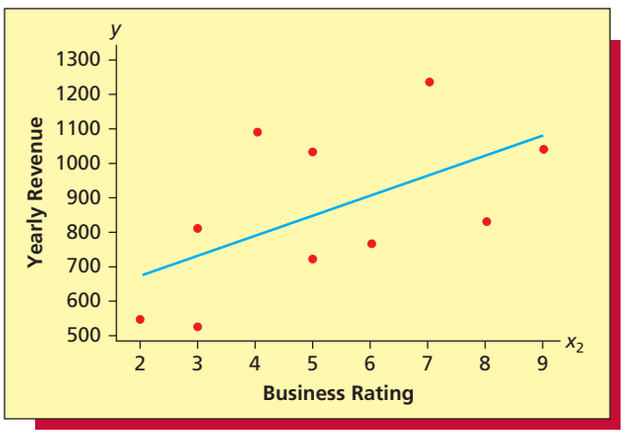


Figure 15.1 presents a scatter plot of y versus x_1. This plot shows that y tends to increase in a straight-line fashion as x_1 increases. Figure 15.2 shows a scatter plot of y versus x_2. This plot shows that y tends to increase in a straight-line fashion as x_2 increases. Together, the scatter plots in Figures 15.1 and 15.2 imply that a reasonable multiple regression model relating y (yearly revenue) to x_1 (population size) and x_2 (business rating) is

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

This model says that the values of y can be represented by a **mean level** ($\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$) that changes as x_1 and x_2 change, combined with random fluctuations (described by the **error term** ε) that cause the values of y to deviate from the mean level. Here:

1 The mean level $\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ is the mean yearly revenue for all Tasty Sub restaurants that could potentially be built near populations of size x_1 and business/shopping areas having a rating of x_2. Furthermore, the equation

$$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

is the equation of a plane—called the **plane of means**—in three-dimensional space. The plane of means is the shaded plane illustrated in Figure 15.3. Different mean yearly revenues corresponding to different population size–business rating combinations lie on the plane of means. For example, Table 15.1 tells us that restaurant 3 is built near a population of 32,300 residents and a business/shopping area having a rating of 6. It follows that

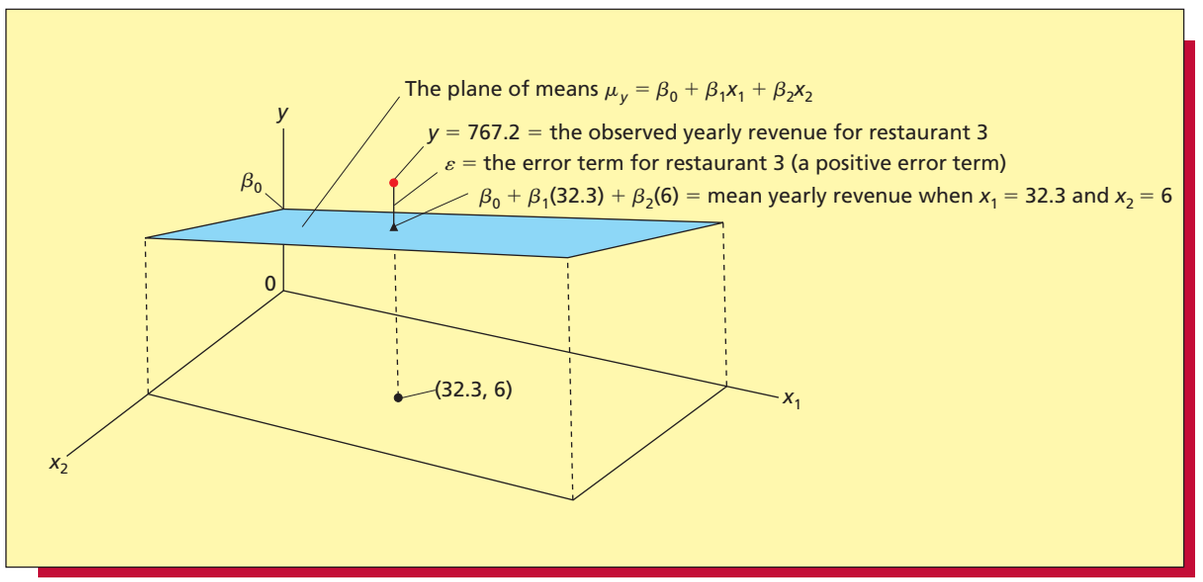
$$\beta_0 + \beta_1(32.3) + \beta_2(6)$$

is the mean yearly revenue for all Tasty Sub restaurants that could potentially be built near populations of 32,300 residents and business/shopping areas having a rating of 6.

2 β_0, β_1, and β_2 are (unknown) regression parameters that relate mean yearly revenue to x_1 and x_2. Specifically:

- β_0 (the *intercept of the model*) is the mean yearly revenue for all Tasty Sub restaurants that could potentially be built near populations of zero residents and business/shopping areas having a rating of 0. This interpretation, however, is of dubious practical value, because we have not observed any Tasty Sub restaurants that are built near populations of zero residents and business/shopping areas having a rating of zero. (The lowest business rating is 1.)

FIGURE 15.3 A Geometrical Interpretation of the Regression Model Relating y to x_1 and x_2



- β_1 (the **regression parameter for the variable x_1**) is the change in mean yearly revenue that is associated with a one-unit (1000 resident) increase in the population size (x_1) when the business rating (x_2) does not change. Intuitively, β_1 is the slope of the plane of means in the x_1 direction.
- β_2 (the **regression parameter for the variable x_2**) is the change in mean yearly revenue that is associated with a one-unit increase in the business rating (x_2) when the population size (x_1) does not change. Intuitively, β_2 is the slope of the plane of means in the x_2 direction.

3 ε is an error term that describes the effect on y of all factors other than x_1 and x_2. One such factor is the skill of the owner as an operator of the restaurant under consideration. For example, Figure 15.3 shows that the error term for restaurant 3 is positive. This implies that the observed yearly revenue for restaurant 3, $y = 767.2$, is greater than the mean yearly revenue for all Tasty Sub restaurants that could potentially be built near populations of 32,300 residents and business/shopping areas having a rating of 6. In general, positive error terms cause their respective observed yearly revenues to be greater than the corresponding mean yearly revenues. On the other hand, negative error terms cause their respective observed yearly revenues to be less than the corresponding mean yearly revenues.

Part 2: The least squares point estimates If b_0, b_1, and b_2 denote point estimates of β_0, β_1, and β_2, then the point prediction of an observed yearly revenue $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ is

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2$$

which we call a *predicted yearly revenue.* Here, because the regression assumptions (to be discussed in Section 15.2) imply that the error term ε has a 50 percent chance of being positive and a 50 percent chance of being negative, we predict ε to be zero. Now, consider the 10 Tasty Sub restaurants in Table 15.1. If any particular values of b_0, b_1, and b_2 are good point estimates, they will make the predicted yearly revenue for each restaurant fairly close to the observed yearly revenue for the restaurant. This will make the restaurant's *residual*—the difference between the restaurant's observed and predicted yearly revenues—fairly small (in magnitude). We define the **least squares point estimates** to be the values of b_0, b_1, and b_2 that minimize *SSE*, the sum of squared residuals for the 10 restaurants.

FIGURE 15.4 Excel and MINITAB Outputs of a Regression Analysis of the Tasty Sub Shop Revenue Data in Table 15.1 Using the Model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

(a) The Excel output

Regression Statistics	
Multiple R	0.9905
R Square	0.9810 [8]
Adjusted R Square	0.9756 [9]
Standard Error	36.6856 [7]
Observations	10

ANOVA	df	SS	MS	F	Significance F
Regression	2	486355.7 [10]	243177.8	180.689 [13]	9.46E-07 [14]
Residual	7	9420.8 [11]	1345.835		
Total	9	495776.5 [12]			

	Coefficients	Standard Error [4]	t Stat [5]	P-value [6]	Lower 95% [19]	Upper 95% [19]
Intercept	125.289 [1]	40.9333	3.06	0.0183	28.4969	222.0807
population	14.1996 [2]	0.9100	15.60	1.07E-06	12.0478	16.3517
bus_rating	22.8107 [3]	5.7692	3.95	0.0055	9.1686	36.4527

(b) The MINITAB output

```
The regression equation is
revenue = 125 + 14.2 population + 22.8 bus_rating

Predictor          Coef      SE Coef [4]          T [5]          P [6]
Constant         125.29 [1]     40.93            3.06          0.018
population      14.1996 [2]      0.91            15.6          0.000
bus_rating       22.811 [3]     5.769            3.95          0.006
S = 36.6856 [7]              R-Sq = 98.10% [8]          R-Sq(adj) = 97.6% [9]

Analysis of Variance
Source              DF            SS              MS             F              P
Regression           2        486356 [10]      243178      180.69 [13]   0.000 [14]
Residual Error       7          9421 [11]        1346
Total                9        495777 [12]

Predicted Values for New Observations
     New Obs          Fit [15]        SE Fit [16]          95% CI [17]              95% PI [18]
           1         956.6                15          (921.0, 992.2)           (862.8, 1050.4)

Values of Predictors for New Observations
     New Obs        population         bus_rating
           1              47.3                  7
```

[1] b_0 [2] b_1 [3] b_2 [4] s_{b_j} = standard error of the estimate b_j [5] t statistics [6] p-values for t statistics [7] s = standard error
[8] R^2 [9] Adjusted R^2 [10] Explained variation [11] SSE = Unexplained variation [12] Total variation [13] F(model) statistic
[14] p-value for F(model) [15] $\hat{y}$ = point prediction when $x_1 = 47.3$ and $x_2 = 7$ [16] $s_{\hat{y}}$ = standard error of the estimate $\hat{y}$
[17] 95% confidence interval when $x_1 = 47.3$ and $x_2 = 7$ [18] 95% prediction interval when $x_1 = 47.3$ and $x_2 = 7$ [19] 95% confidence interval for β_j

The formula for the least squares point estimates of the parameters in a multiple regression model is expressed using a branch of mathematics called **matrix algebra.** This formula is presented in Bowerman, O'Connell, and Koehler (2005). In the main body of the book, we will rely on Excel and MINITAB to compute the needed estimates. For example, consider the Excel and MINITAB outputs in Figure 15.4. These outputs tell us that the least squares point estimates of β_0, β_1, and β_2 in the Tasty Sub Shop revenue model are $b_0 = 125.29$, $b_1 = 14.1996$, and $b_2 = 22.811$ (see [1], [2], and [3]). The point estimate $b_1 = 14.1996$ of β_1 says we estimate that mean yearly revenue increases by \$14,199.60 when the population size increases by 1,000 residents and the business rating does not change. The point estimate $b_2 = 22.811$ of β_2 says we estimate that mean yearly revenue increases by \$22,811 when there is a one-unit increase in the business rating and the population size does not change.

TABLE 15.2 The Point Predictions and Residuals Using the Least Squares Point Estimates, $b_0 = 125.29$, $b_1 = 14.1996$, and $b_2 = 22.811$

Restaurant	Population Size, x_1 (Thousands of Residents)	Business Rating, x_2	Yearly Revenue, y (Thousands of Dollars)	Predicted Yearly Revenue $\hat{y} = 125.29 + 14.1996x_1 + 22.811x_2$	Residual, $y - \hat{y}$
1	20.8	3	527.1	489.07	38.03
2	27.5	2	548.7	561.40	−12.70
3	32.3	6	767.2	720.80	46.40
4	37.2	5	722.9	767.57	−44.67
5	39.6	8	826.3	870.08	−43.78
6	45.1	3	810.5	834.12	−23.62
7	49.9	9	1040.7	1039.15	1.55
8	55.4	5	1033.6	1026.00	7.60
9	61.7	4	1090.3	1092.65	−2.35
10	64.6	7	1235.8	1202.26	33.54

$SSE = (38.03)^2 + (-12.70)^2 + \cdots + (33.54)^2 = 9420.8$

The equation

$$\begin{aligned}\hat{y} &= b_0 + b_1x_1 + b_2x_2 \\ &= 125.29 + 14.1996x_1 + 22.811x_2\end{aligned}$$

is called the **least squares prediction equation.** In Table 15.2 we summarize using this prediction equation to calculate the predicted yearly revenues and the residuals for the 10 observed Tasty Sub restaurants. For example, because the population size and business rating for restaurant 1 were 20.8 and 3, the predicted yearly revenue for restaurant 1 is

$$\begin{aligned}\hat{y} &= 125.29 + 14.1996(20.8) + 22.811(3) \\ &= 489.07\end{aligned}$$

It follows, because the observed yearly revenue for restaurant 1 was $y = 527.1$, that the residual for restaurant 1 is

$$y - \hat{y} = 527.1 - 489.07 = 38.03$$

If we consider all of the residuals in Table 15.2 and add their squared values, we find that **SSE**, **the sum of squared residuals,** is 9420.8. This *SSE* value is given on the Excel and MINITAB outputs in Figure 15.4 (see 11) and will be used throughout this chapter.

Part 3: Estimating means and predicting individual values The least squares prediction equation is the equation of a plane—called the **least squares plane**—in three-dimensional space. The least squares plane is the estimate of the plane of means. It follows that the point on the least squares plane corresponding to the population size x_1 and the business rating x_2

$$\begin{aligned}\hat{y} &= b_0 + b_1x_1 + b_2x_2 \\ &= 125.29 + 14.1996x_1 + 22.811x_2\end{aligned}$$

is the point estimate of $\beta_0 + \beta_1x_1 + \beta_2x_2$, the mean yearly revenue for all Tasty Sub restaurants that could potentially be built near populations of size x_1 and business/shopping areas having a rating of x_2. In addition, because we predict the error term to be 0, $\hat{y}$ is also the point prediction of $y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \varepsilon$, the yearly revenue for a single Tasty Sub restaurant that is built near a population of size x_1 and a business/shopping area having a rating of x_2.

For example, suppose that one of the business entrepreneur's potential restaurant sites is near a population of 47,300 residents and a business/shopping area having a rating of 7. It follows that

$$\begin{aligned}\hat{y} &= 125.29 + 14.1996(47.3) + 22.811(7) \\ &= 956.6 \text{ (that is, \$956.600)}\end{aligned}$$

is

1. The **point estimate** of the mean yearly revenue for all Tasty Sub restaurants that could potentially be built near populations of 47,300 residents and business/shopping areas having a rating of 7, and
2. The **point prediction** of the yearly revenue for a single Tasty Sub restaurant that is built near a population of 47,300 residents and a business/shopping area having a rating of 7.

Notice that $\hat{y} = 956.6$ is given at the bottom of the MINITAB output in Figure 15.4 on page 558 (see 15). Moreover, recall that the yearly rent and other fixed costs for the entrepreneur's potential restaurant will be \$257,550 and that (according to Tasty Sub corporate headquarters) the yearly food and other variable costs for the restaurant will be 60 percent of the yearly revenue. Because we predict that the yearly revenue for the restaurant will be \$956,600, it follows that we predict that the yearly total operating cost for the restaurant will be \$257,550 + .6(\$956,600) = \$831,510. In addition, if we subtract this predicted yearly operating cost from the predicted yearly revenue of \$956,600, we predict that the yearly profit for the restaurant will be \$125,090. Of course, these predictions are point predictions. In Section 15.6 we will predict the restaurant's yearly revenue and profit *with confidence*.

BI

The Tasty Sub Shop revenue model expresses the dependent variable as a function of two independent variables. In general, we can use a multiple regression model to express a dependent variable as a function of any number of independent variables. For example, in the past, natural gas utilities serving the Cincinnati, Ohio, area have predicted daily natural gas consumption by using four independent (predictor) variables—average temperature, average wind velocity, average sunlight, and change in average temperature from the previous day. The general form of a multiple regression model expresses the dependent variable y as a function of k independent variables $x_1, x_2, \ldots, x_k$. We express this general form in the following box.

The Multiple Regression Model

The **multiple regression model relating y to $x_1, x_2, \ldots, x_k$** is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

Here

1. $\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$ is the mean value of the dependent variable y when the values of the independent variables are $x_1, x_2, \ldots, x_k$.
2. $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ are (unknown) **regression parameters** relating the mean value of y to $x_1, x_2, \ldots, x_k$.
3. ε is an **error term** that describes the effects on y of all factors other than the values of the independent variables $x_1, x_2, \ldots, x_k$.

If $b_0, b_1, b_2, \ldots, b_k$ denote point estimates of $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$, then

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$$

is the **point estimate of the mean value of the dependent variable** when the values of the independent variables are $x_1, x_2, \ldots, x_k$. In addition, because we predict the error term ε to be 0, $\hat{y}$ is also the **point prediction of an individual value of the dependent variable** when the values of the independent variables are $x_1, x_2, \ldots, x_k$. Now, assume that we have obtained n observations, where each observation consists of an observed value of the dependent variable y and corresponding observed values of the independent variables $x_1, x_2, \ldots, x_k$. For the ith observation, let y_i and $\hat{y}_i$ denote the observed and predicted values of the dependent variable, and define the residual to be $e_i = y_i - \hat{y}_i$. It then follows that the **least squares point estimates** are the values of $b_0, b_1, b_2, \ldots, b_k$ that minimize the sum of squared residuals:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

As illustrated in Example 15.1, we use Excel and MINITAB to find the least squares point estimates.

To conclude this section, consider an arbitrary independent variable, which we will denote as x_j, in a multiple regression model. We can then interpret the parameter β_j to be the change in the mean value of the dependent variable that is associated with a one-unit increase in x_j when the other independent variables in the model do not change. This interpretation is based, however, on the assumption that x_j can increase by one unit without the other independent variables in the model changing. In some situations (as we will see) this assumption is not reasonable.

Exercises for Section 15.1

CONCEPTS

15.1 In the multiple regression model, what sum of squared deviations do the least squares point estimates minimize?

15.2 When using the multiple regression model, how do we obtain a point estimate of the mean value of the dependent variable and a point prediction of an individual value of the dependent variable?

METHODS AND APPLICATIONS

15.3 THE NATURAL GAS CONSUMPTION CASE GasCon2

Consider the situation in which a gas company wishes to predict weekly natural gas consumption for its city. In the exercises of Chapter 14, we used the single predictor variable x, average hourly temperature, to predict y, weekly natural gas consumption. We now consider predicting y on the basis of average hourly temperature and a second predictor variable—the chill index. The chill index for a given average hourly temperature expresses the combined effects of all other major weather-related factors that influence natural gas consumption, such as wind velocity, sunlight, cloud cover, and the passage of weather fronts. The chill index is expressed as a whole number between 0 and 30. A weekly chill index near 0 indicates that, given the average hourly temperature during the week, all other major weather-related factors will only slightly increase weekly natural gas consumption. A weekly chill index near 30 indicates that, given the average hourly temperature during the week, other weather-related factors will greatly increase weekly natural gas consumption. The natural gas company has collected data concerning weekly natural gas consumption (y, in MMcF), average hourly temperature (x_1, in degrees Fahrenheit), and the chill index (x_2) for the last eight weeks. The data are given in Table 15.3, and scatter plots of y versus x_1 and y versus x_2 are given below the data. Moreover, Figure 15.5 on the next page gives Excel and MINITAB outputs of a regression analysis of these data using the model

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

a Using the Excel or MINITAB output (depending on the package used in your class), find (on the output) b_1 and b_2, the least squares point estimates of β_1 and β_2, and report their values. Then interpret b_1 and b_2.

b Calculate a point estimate of the mean natural gas consumption for all weeks that have an average hourly temperature of 40 and a chill index of 10, and a point prediction of the amount of natural gas consumed in a single week that has an average hourly temperature of 40 and a chill index of 10. Find this point estimate (prediction), which is given at the bottom of the MINITAB output, and verify that it equals (within rounding) your calculated value.

15.4 THE REAL ESTATE SALES PRICE CASE RealEst2

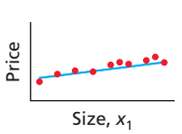
A real estate agency collects the data in Table 15.4 concerning

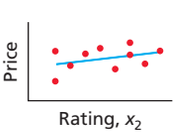
y = sales price of a house (in thousands of dollars)

x_1 = home size (in hundreds of square feet)

x_2 = rating (an overall "niceness rating" for the house expressed on a scale from 1 [worst] to 10 [best], and provided by the real estate agency)

Scatter plots of y versus x_1 and y versus x_2 are as follows:

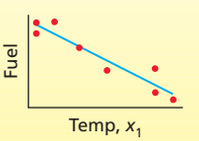




connect

TABLE 15.3
The Natural Gas Consumption Data
GasCon2

y	x_1	x_2
12.4	28.0	18
11.7	28.0	14
12.4	32.5	24
10.8	39.0	22
9.4	45.9	8
9.5	57.8	16
8.0	58.1	1
7.5	62.5	0



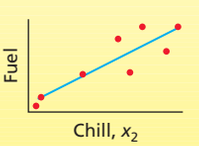


TABLE 15.4
The Real Estate Sales Price Data
RealEst2

y	x_1	x_2
180	23	5
98.1	11	2
173.1	20	9
136.5	17	3
141	15	8
165.9	21	4
193.5	24	7
127.8	13	6
163.5	19	7
172.5	25	2

Source: R. L. Andrews and J. T. Ferguson, "Integrating Judgement with a Regression Appraisal," *The Real Estate Appraiser and Analyst* 52, no. 2 (1986). Reprinted by permission.

FIGURE 15.5 Excel and MINITAB Outputs of a Regression Analysis of the Natural Gas Consumption Data Using the Model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

(a) The Excel output

Regression Statistics	
Multiple R	0.9867
R Square	0.9736
Adjusted R Square	0.9631
Standard Error	0.3671
Observations	8

ANOVA	df	SS	MS	F	Significance F
Regression	2	24.8750	12.4375	92.3031	0.0001
Residual	5	0.6737	0.1347		
Total	7	25.5488			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	13.1087	0.8557	15.3193	2.15E-05	10.9091	15.3084
TEMP	-0.0900	0.0141	-6.3942	0.0014	-0.1262	-0.0538
CHILL	0.0825	0.0220	3.7493	0.0133	0.0259	0.1391

(b) The MINITAB output

```
The regression equation is
FuelCons = 13.1 - 0.0900 Temp + 0.0825 Chill

Predictor           Coef         SE Coef             T             P
Constant         13.1087          0.8557         15.32         0.000
Temp            -0.09001         0.01408         -6.39         0.001
Chill            0.08249         0.02200          3.75         0.013

s = 0.367078        R-Sq = 97.4%         R-Sq(adj) = 96.3%

Analysis of Variance
Source               DF             SS             MS             F             P
Regression            2         24.875         12.438         92.30         0.000
Residual Error        5          0.674          0.135
Total                 7         25.549

Values of Predictors for New Obs   Predicted Values for New Observations
New Obs Temp Chill                 New Obs      Fit      SE Fit         95% CI          95% PI
      1 40.0  10.0                       1   10.333       0.170   (9.895, 10.771) (9.293, 11.374)
```

FIGURE 15.6 MINITAB Output of a Regression Analysis of the Real Estate Sales Price Data Using the Model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

```
The regression equation is
SalesPrice = 29.3 + 5.61 HomeSize + 3.83 Rating

Predictor     Coef  SE Coef      T      P
Constant    29.347    4.891   6.00  0.001
HomeSize    5.6128   0.2285  24.56  0.000
Rating      3.8344   0.4332   8.85  0.000

S = 3.24164   R-Sq = 99.0%   R-Sq(adj) = 98.7%

Analysis of Variance
Source           DF      SS      MS       F      P
Regression        2  7374.0  3687.0  350.87  0.000
Residual Error    7    73.6    10.5
Total             9  7447.5

Values of Predictors for New Obs   Predicted Values for New Observations
New Obs  HomeSize  Rating          New Obs     Fit   SE Fit       95% CI            95% PI
      1      20.0    8.00                1  172.28     1.57  (168.56, 175.99)  (163.76, 180.80)
```

The agency wishes to develop a regression model that can be used to predict the sales prices of future houses it will list. Figure 15.6 gives the MINITAB output of a regression analysis of the real estate sales price data in Table 15.4 using the model

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

a Using the MINITAB output, identify and interpret b_1 and b_2, the least squares point estimates of β_1 and β_2.

b Calculate a point estimate of the mean sales price of all houses having 2,000 square feet and a rating of 8, and a point prediction of the sales price of a single house having 2,000 square feet and a rating of 8. Find this point estimate (prediction), which is given at the bottom of the MINITAB output, and verify that it equals (within rounding) your calculated value.

15.5 THE FRESH DETERGENT CASE DS Fresh2

Enterprise Industries produces Fresh, a brand of liquid laundry detergent. In order to manage its inventory more effectively and make revenue projections, the company would like to better predict demand for Fresh. To develop a prediction model, the company has gathered data concerning demand for Fresh over the last 30 sales periods (each sales period is defined to be a four-week period). The demand data are presented in Table 15.5. Here, for each sales period,

y = the demand for the large size bottle of Fresh (in hundreds of thousands of bottles) in the sales period

x_1 = the price (in dollars) of Fresh as offered by Enterprise Industries in the sales period

x_2 = the average industry price (in dollars) of competitors' similar detergents in the sales period

x_3 = Enterprise Industries' advertising expenditure (in hundreds of thousands of dollars) to promote Fresh in the sales period

Figure 15.7 on the next page gives the Excel output of a regression analysis of the Fresh Detergent demand data in Table 15.5 using the model

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

a Find (on the output) and report the values of b_1, b_2, and b_3, the least squares point estimates of β_1, β_2, and β_3. Interpret b_1, b_2, and b_3.

TABLE 15.5 Historical Data Concerning Demand for Fresh Detergent DS Fresh2

Sales Period	Price for Fresh, x_1	Average Industry Price, x_2	Advertising Expenditure for Fresh, x_3	Demand for Fresh, y	Sales Period	Price for Fresh, x_1	Average Industry Price, x_2	Advertising Expenditure for Fresh, x_3	Demand for Fresh, y
1	3.85	3.80	5.50	7.38	16	3.80	4.10	6.80	8.87
2	3.75	4.00	6.75	8.51	17	3.70	4.20	7.10	9.26
3	3.70	4.30	7.25	9.52	18	3.80	4.30	7.00	9.00
4	3.70	3.70	5.50	7.50	19	3.70	4.10	6.80	8.75
5	3.60	3.85	7.00	9.33	20	3.80	3.75	6.50	7.95
6	3.60	3.80	6.50	8.28	21	3.80	3.75	6.25	7.65
7	3.60	3.75	6.75	8.75	22	3.75	3.65	6.00	7.27
8	3.80	3.85	5.25	7.87	23	3.70	3.90	6.50	8.00
9	3.80	3.65	5.25	7.10	24	3.55	3.65	7.00	8.50
10	3.85	4.00	6.00	8.00	25	3.60	4.10	6.80	8.75
11	3.90	4.10	6.50	7.89	26	3.65	4.25	6.80	9.21
12	3.90	4.00	6.25	8.15	27	3.70	3.65	6.50	8.27
13	3.70	4.10	7.00	9.10	28	3.75	3.75	5.75	7.67
14	3.75	4.20	6.90	8.86	29	3.80	3.85	5.80	7.93
15	3.75	4.10	6.80	8.90	30	3.70	4.25	6.80	9.26

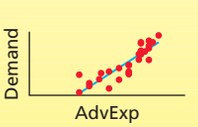


FIGURE 15.7 Excel Output of a Regression Analysis of the Fresh Detergent Demand Data Using the Model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$

(a) The Excel output

Regression Statistics	
Multiple R	0.9453
R Square	0.8936
Adjusted R Square	0.8813
Standard Error	0.2347
Observations	30

ANOVA	df	SS	MS	F	Significance F
Regression	3	12.0268	4.0089	72.797	8.883E-13
Residual	26	1.4318	0.0551		
Total	29	13.4586			

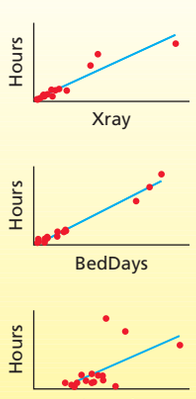
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	7.5891	2.4450	3.1039	0.0046	2.5633	12.6149
Price (X1)	-2.3577	0.6379	-3.6958	0.0010	-3.6690	-1.0464
IndPrice (X2)	1.6122	0.2954	5.4586	0.0000	1.0051	2.2193
AdvExp (X3)	0.5012	0.1259	3.9814	0.0005	0.2424	0.7599

(b) Prediction using an Excel add-in (MegaStat)

Predicted values for: Demand (y)

				95% Confidence Interval		95% Prediction Interval		
Price (x1)	IndPrice (x2)	AdvExp (x3)	Predicted	lower	upper	lower	upper	Leverage
3.7	3.9	6.5	8.4107	8.3143	8.5070	7.9188	8.9025	0.040

TABLE 15.6 Hospital Labor Needs Data HospLab



Hospital	Monthly X-Ray Exposures, x_1	Monthly Occupied Bed Days, x_2	Average Length of Stay, x_3	Monthly Labor Hours Required, y
1	2,463	472.92	4.45	566.52
2	2,048	1,339.75	6.92	696.82
3	3,940	620.25	4.28	1,033.15
4	6,505	568.33	3.90	1,603.62
5	5,723	1,497.60	5.50	1,611.37
6	11,520	1,365.83	4.60	1,613.27
7	5,779	1,687.00	5.62	1,854.17
8	5,969	1,639.92	5.15	2,160.55
9	8,461	2,872.33	6.18	2,305.58
10	20,106	3,655.08	6.15	3,503.93
11	13,313	2,912.00	5.88	3,571.89
12	10,771	3,921.00	4.88	3,741.40
13	15,543	3,865.67	5.50	4,026.52
14	34,703	12,446.33	10.78	11,732.17
15	39,204	14,098.40	7.05	15,414.94
16	86,533	15,524.00	6.35	18,854.45

Source: *Procedures and Analysis for Staffing Standards Development Regression Analysis Handbook* (San Diego, CA: Navy Manpower and Material Analysis Center, 1979).

b Consider the demand for Fresh Detergent in a future sales period when Enterprise Industries' price for Fresh will be $x_1 = 3.70$, the average price of competitors' similar detergents will be $x_2 = 3.90$ and Enterprise Industries' advertising expenditure for Fresh will be $x_3 = 6.50$. The point prediction of this demand is given at the bottom of the Excel add-in output. Report this point prediction and show (within rounding) how it has been calculated.

15.6 THE HOSPITAL LABOR NEEDS CASE HospLab

Table 15.6 presents data concerning the need for labor in 16 U.S. Navy hospitals. Here, y = monthly labor hours required; x_1 = monthly X-ray exposures; x_2 = monthly occupied bed

FIGURE 15.8 Excel Output of a Regression Analysis of the Hospital Labor Needs Data Using the Model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$

(a) The Excel output

Regression Statistics

Multiple R	0.9981
R Square	0.9961
Adjusted R Square	0.9952
Standard Error	387.1598
Observations	16

ANOVA	df	SS	MS	F	Significance F
Regression	3	462327889.4	154109296.5	1028.1309	9.92E-15
Residual	12	1798712.2	149892.7		
Total	15	464126601.6			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	1946.8020	504.1819	3.8613	0.0023	848.2840	3045.3201
XRay (x1)	0.0386	0.0130	2.9579	0.0120	0.0102	0.0670
BedDays (x2)	1.0394	0.0676	15.3857	2.91E-09	0.8922	1.1866
LengthStay (x3)	-413.7578	98.5983	-4.1964	0.0012	-628.5850	-198.9306

(b) Prediction Using an Excel add-in (MegaStat)

Predicted values for: LaborHours

				95% Confidence Interval		95% Prediction Interval		
XRay (x1)	BedDays (x2)	LengthStay (x3)	Predicted	lower	upper	lower	upper	Leverage
56194	14077.88	6.89	15,896.2473	15,378.0313	16,414.4632	14,906.2361	16,886.2584	0.3774

days (a hospital has one occupied bed day if one bed is occupied for an entire day); and x_3 = average length of patients' stay (in days). Figure 15.8 gives the Excel output of a regression analysis of the data using the model

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

Note that the variables x_1, x_2, and x_3 are denoted as XRay, BedDays, and LengthStay on the output.

a Find (on the output) and report the values of b_1, b_2, and b_3, the least squares point estimates of β_1, β_2, and β_3. Interpret b_1, b_2, and b_3. Note that the negative value of b_3 ($= -413.7578$) might say that, when XRay and BedDays stay constant, an increase in LengthStay implies less patient turnover and thus fewer start-up hours needed for the initial care of new patients.

b Consider a questionable hospital for which XRay = 56,194, BedDays = 14,077.88, and LengthStay = 6.89. A point prediction of the labor hours corresponding to this combination of values of the independent variables is given on the Excel add-in output. Report this point prediction and show (within rounding) how it has been calculated.

c If the actual number of labor hours used by the questionable hospital was $y = 17{,}207.31$, how does this y value compare with the point prediction?

15.2 Model Assumptions and the Standard Error

LO15-2 Explain the assumptions behind multiple regression and calculate the standard error.

Model assumptions In order to perform hypothesis tests and set up various types of intervals when using the multiple regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

we need to make certain assumptions about the error term ε. At any given combination of values of $x_1, x_2, \ldots, x_k$, there is a population of error term values that could potentially occur. These error term values describe the different potential effects on y of all factors other than the combination of values of $x_1, x_2, \ldots, x_k$. Therefore, these error term values explain the variation in the y values that could be observed at the combination of values of $x_1, x_2, \ldots, x_k$. We make the following four assumptions about the potential error term values.

Assumptions for the Multiple Regression Model

1. At any given combination of values of $x_1, x_2, \ldots, x_k$, the population of potential error term values has a mean equal to 0.
2. **Constant variance assumption:** At any given combination of values of $x_1, x_2, \ldots, x_k$, the population of potential error term values has a variance that does not depend on the combination of values of $x_1, x_2, \ldots, x_k$. That is, the different populations of potential error term values corresponding to different combinations of values of $x_1, x_2, \ldots, x_k$ have equal variances. We denote the constant variance as σ^2.
3. **Normality assumption:** At any given combination of values of $x_1, x_2, \ldots, x_k$, the population of potential error term values has a **normal distribution.**
4. **Independence assumption:** Any one value of the error term ε is **statistically independent** of any other value of ε. That is, the value of the error term ε corresponding to an observed value of y is statistically independent of the error term corresponding to any other observed value of y.

Taken together, the first three assumptions say that, at any given combination of values of $x_1, x_2, \ldots, x_k$, the population of potential error term values is normally distributed with mean 0 and a variance σ^2 that does not depend on the combination of values of $x_1, x_2, \ldots, x_k$. Because the potential error term values cause the variation in the potential y values, the first three assumptions imply that, at any given combination of values of $x_1, x_2, \ldots, x_k$, the population of y values that could be observed is normally distributed with mean $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$ and a variance σ^2 that does not depend on the combination of values of $x_1, x_2, \ldots, x_k$. Furthermore, the independence assumption says that, when time series data are utilized in a regression study, there are no patterns in the error term values. In Section 15.11 we show how to check the validity of the regression assumptions. That section can be read at any time after Section 15.7. As in simple linear regression, only pronounced departures from the assumptions must be remedied.

The mean square error and the standard error To present statistical inference formulas in later sections, we need to be able to compute point estimates of σ^2 and σ (the constant variance and standard deviation of the different error term populations). We show how to do this in the following box:

The Mean Square Error and the Standard Error

Suppose that the multiple regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

utilizes k independent variables and thus has $(k + 1)$ parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$. Then, if the regression assumptions are satisfied, and if *SSE* denotes the sum of squared residuals for the model:

1. A point estimate of σ^2 is the **mean square error**

$$s^2 = \frac{SSE}{n - (k + 1)}$$

2. A point estimate of σ is the **standard error**

$$s = \sqrt{\frac{SSE}{n - (k + 1)}}$$

In order to explain these point estimates, recall that σ^2 is the variance of the population of y values (for given values of $x_1, x_2, \ldots, x_k$) around the mean value μ_y. Because $\hat{y}$ is the point estimate of this mean, it seems natural to use $SSE = \Sigma(y_i - \hat{y}_i)^2$ to help construct a point estimate of σ^2. We divide *SSE* by $n - (k + 1)$ because it can be proven that doing so makes the resulting s^2 an unbiased point estimate of σ^2. We call $n - (k + 1)$ the **number of degrees of freedom** associated with *SSE*.

We will see in Section 15.6 that if a particular regression model gives a small standard error, then the model will give short prediction intervals and thus accurate predictions of individual y values. For example, Table 15.2 (page 559) shows that *SSE* for the Tasty Sub Shop revenue model

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

is 9420.8. Because this model utilizes $k = 2$ independent variables and thus has $k + 1 = 3$ parameters (β_0, β_1, and β_2), a point estimate of σ^2 is the mean square error

$$s^2 = \frac{SSE}{n - (k + 1)} = \frac{9420.8}{10 - 3} = \frac{9420.8}{7} = 1345.835$$

and a point estimate of σ is the standard error $s = \sqrt{1345.835} = 36.6856$. Note that $SSE = 9420.8$, $s^2 = 1345.835$, and $s = 36.6856$ are given on the Excel and MINITAB outputs in Figure 15.4 (page 558). Also note that the s of 36.6856 for the two independent variable model is less than the s of 61.7052 for the simple linear regression model that uses only the population size to predict yearly revenue (see Example 14.3, page 502).

15.3 R^2 and Adjusted R^2

LO15-3 Calculate and interpret the multiple and adjusted multiple coefficients of determination.

The multiple coefficient of determination, R^2 In this section we discuss several ways to assess the utility of a multiple regression model. We first discuss a quantity called the **multiple coefficient of determination,** which is denoted R^2. The formulas for R^2 and several other related quantities are given in the following box:

The Multiple Coefficient of Determination, R^2

For the multiple regression model:

1. **Total variation** $= \sum (y_i - \bar{y})^2$
2. **Explained variation** $= \sum (\hat{y}_i - \bar{y})^2$
3. **Unexplained variation** $= \sum (y_i - \hat{y}_i)^2$
4. **Total variation = Explained variation + Unexplained variation**
5. The **multiple coefficient of determination** is
$$R^2 = \frac{\text{Explained variation}}{\text{Total variation}}$$
6. R^2 is the proportion of the total variation in the n observed values of the dependent variable that is explained by the overall regression model.
7. **Multiple correlation coefficient** $= R = \sqrt{R^2}$

As an example, consider the Tasty Sub Shop revenue model

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

and the following MINITAB output:

```
S = 36.6856   R-Sq = 98.10%   R-Sq(adj) = 97.6%

Analysis of Variance
Source           DF      SS      MS       F      P
Regression        2  486356  243178  180.69  0.000
Residual Error    7    9421    1346
Total             9  495777
```

This output tells us that the total variation (SS Total), explained variation (SS Regression), and unexplained variation (SS Residual Error) for the model are, respectively, 495,777, 486,356, and 9,421. The output also tells us that the multiple coefficient of determination is

$$R^2 = \frac{\text{Explained variation}}{\text{Total variation}} = \frac{486{,}356}{495{,}777} = .981 \quad (98.1\% \text{ on the output})$$

which implies that the multiple correlation coefficient is $R = \sqrt{.981} = .9905$. The value of $R^2 = .981$ says that the two independent variable Tasty Sub Shop revenue model explains 98.1 percent of the total variation in the 10 observed yearly revenues. Note this R^2 value is larger than the r^2 of .939 for the simple linear regression model that uses only the population size to predict yearly revenue. Also note that the quantities given on the MINITAB output are given on the following Excel output.

Regression Statistics					
Multiple R	0.9905				
R Square	0.9810				
Adjusted R Square	0.9756				
Standard Error	36.6856				
Observations	10				
ANOVA	**df**	**SS**	**MS**	**F**	**Significance F**
Regression	2	486355.7	243177.8	180.689	9.46E-07
Residual	7	9420.8	1345.835		
Total	9	495776.5			

Adjusted R^2 Even if the independent variables in a regression model are unrelated to the dependent variable, they will make R^2 somewhat greater than 0. To avoid overestimating the importance of the independent variables, many analysts recommend calculating an *adjusted* multiple coefficient of determination.

Adjusted R^2

The **adjusted multiple coefficient of determination (adjusted R^2)** is

$$\overline{R}^2 = \left(R^2 - \frac{k}{n-1}\right)\left(\frac{n-1}{n-(k+1)}\right)$$

where R^2 is the multiple coefficient of determination, n is the number of observations, and k is the number of independent variables in the model under consideration.

To briefly explain this formula, note that it can be shown that subtracting $k/(n-1)$ from R^2 helps avoid overestimating the importance of the k independent variables. Furthermore, multiplying $[R^2 - (k/(n-1))]$ by $(n-1)/(n-(k+1))$ makes $\overline{R}^2$ equal to 1 when R^2 equals 1.

As an example, consider the Tasty Sub Shop revenue model

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

Because we have seen that $R^2 = .981$, it follows that

$$\begin{aligned}\overline{R}^2 &= \left(R^2 - \frac{k}{n-1}\right)\left(\frac{n-1}{n-(k+1)}\right)\\ &= \left(.981 - \frac{2}{10-1}\right)\left(\frac{10-1}{10-(2+1)}\right)\\ &= .9756\end{aligned}$$

which is given on the MINITAB and Excel outputs.

If R^2 is less than $k/(n-1)$, which can happen, then $\overline{R}^2$ will be negative. In this case, statistical software systems set $\overline{R}^2$ equal to 0. Historically, R^2 and $\overline{R}^2$ have been popular measures of model utility—possibly because they are unitless and between 0 and 1. In general, we desire R^2 and $\overline{R}^2$ to be near 1. However, sometimes even if a regression model has an R^2 and an $\overline{R}^2$ that are near 1, the model is still not able to predict accurately. We will discuss assessing a model's ability to predict accurately, as well as using R^2 and $\overline{R}^2$ to help choose a regression model, as we proceed through the rest of this chapter.

15.4 The Overall *F*-Test

LO15-4 Test the significance of a multiple regression model by using an *F*-test.

Another way to assess the utility of a regression model is to test the significance of the regression relationship between y and $x_1, x_2, \ldots, x_k$. For the multiple regression model, we test the null hypothesis H_0: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$, which says that **none of the independent variables $x_1, x_2, \ldots, x_k$ is significantly related to y (the regression relationship is not significant),** versus the alternative hypothesis H_a: At least one of $\beta_1, \beta_2, \ldots, \beta_k$ does not equal 0, which says that **at least one of the independent variables is significantly related to y (the regression relationship is significant).** If we can reject H_0 at level of significance α, we say that **the multiple regression model is significant at level of significance α.** We carry out the test as follows:

An *F*-Test for the Multiple Regression Model

Suppose that the regression assumptions hold and that the multiple regression model has $(k + 1)$ parameters, and consider testing

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

versus

H_a: At least one of $\beta_1, \beta_2, \ldots, \beta_k$ does not equal 0.

We define the **overall *F* statistic** to be

$$F(\text{model}) = \frac{(\text{Explained variation})/k}{(\text{Unexplained variation})/[n - (k + 1)]}$$

Also define the p-value related to F(model) to be the area under the curve of the F distribution (having k and $[n - (k + 1)]$ degrees of freedom) to the right of F(model). Then, we can reject H_0 in favor of H_a at level of significance α if either of the following equivalent conditions holds:

1. $F(\text{model}) > F_\alpha$
2. $p\text{-value} < \alpha$

Here the point F_α is based on k numerator and $n - (k + 1)$ denominator degrees of freedom.

Condition 1 is intuitively reasonable because a large value of F(model) would be caused by an explained variation that is large relative to the unexplained variation. This would occur if at least one independent variable in the regression model significantly affects y, which would imply that H_0 is false and H_a is true.

EXAMPLE 15.2 The Tasty Sub Shop Case: The Overall *F*-Test

Consider the Tasty Sub Shop revenue model

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

and the following MINITAB output

```
Analysis of Variance
Source           DF      SS      MS       F       P
Regression        2  486356  243178  180.69   0.000
Residual Error    7    9421    1346
Total             9  495777
```

This output tells us that the explained and unexplained variations for this model are, respectively, 486,356 and 9,421. It follows, because there are $k = 2$ independent variables, that

$$\begin{aligned} F(\text{model}) &= \frac{(\text{Explained variation})/k}{(\text{Unexplained variation})/[n - (k + 1)]} \\ &= \frac{486{,}356/2}{9421/[10 - (2 + 1)]} = \frac{243{,}178}{1345.8} \\ &= 180.69 \end{aligned}$$

Note that this overall F statistic is given on the MINITAB output and is also given on the following Excel output:

ANOVA	df	SS	MS	F	Significance F
Regression	2	486355.7	243177.8	180.689	9.46E-07
Residual	7	9420.8	1345.835		
Total	9	495776.5			

The p-value related to F(model) is the area to the right of 180.69 under the curve of the F distribution having $k = 2$ numerator and $n - (k + 1) = 10 - 3 = 7$ denominator degrees of freedom. Both the MINITAB and Excel outputs say this p-value is less than .001.

If we wish to test the significance of the regression model at level of significance $\alpha = .05$, we use the critical value $F_{.05}$ based on 2 numerator and 7 denominator degrees of freedom. Using Table A.7 (page 796), we find that $F_{.05} = 4.74$. Because $F(\text{model}) = 180.69 > F_{.05} = 4.74$, we can reject H_0 in favor of H_a at level of significance .05. Alternatively, because the p-value is smaller than .05, .01, and .001, we can reject H_0 at level of significance .05, .01, and .001. Therefore, we have extremely strong evidence that the Tasty Sub Shop revenue model is significant. That is, we have extremely strong evidence that at least one of the independent variables x_1 and x_2 in the model is significantly related to y.

If the overall F-test tells us that at least one independent variable in a regression model is significant, we next attempt to decide which independent variables are significant. In the next section we discuss one way to do this.

Exercises for Sections 15.2, 15.3, and 15.4

connect

CONCEPTS

15.7 What is estimated by the mean square error, and what is estimated by the standard error?

15.8 **a** What do R^2 and $\overline{R}^2$ measure? **b** How do R^2 and $\overline{R}^2$ differ?

15.9 What is the purpose of the overall F-test?

METHODS AND APPLICATIONS

In Exercises 15.10 to 15.13 we give Excel and MINITAB outputs of regression analyses of the data sets related to four case studies introduced in Section 15.1. Above each output we give the regression model and the number of observations, n, used to perform the regression analysis under consideration. Using the appropriate model, sample size n, and output:

1. Report SSE, s^2, and s as shown on the output. Calculate s^2 from SSE and other numbers.
2. Report the total variation, unexplained variation, and explained variation as shown on the output.
3. Report R^2 and $\overline{R}^2$ as shown on the output. Interpret R^2 and $\overline{R}^2$. Show how $\overline{R}^2$ has been calculated from R^2 and other numbers.
4. Calculate the F(model) statistic by using the explained and unexplained variations (as shown on the output) and other relevant quantities. Find F(model) on the output to check your answer (within rounding).
5. Use the F(model) statistic and the appropriate critical value to test the significance of the linear regression model under consideration by setting α equal to .05.
6. Use the F(model) statistic and the appropriate critical value to test the significance of the linear regression model under consideration by setting α equal to .01.
7. Find the p-value related to F(model) on the output. Using the p-value, test the significance of the linear regression model by setting $\alpha = .10, .05, .01$, and .001. What do you conclude?

15.10 THE NATURAL GAS CONSUMPTION CASE GasCon2

Model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ Sample size: $n = 8$

The output follows on the next page:

```
S = 0.367078   R-Sq = 97.4%   R-Sq(adj) = 96.3%

Analysis of Variance
Source           DF      SS      MS      F      P
Regression        2  24.875  12.438  92.30  0.000
Residual Error    5   0.674   0.135
Total             7  25.549
```

15.11 THE REAL ESTATE SALES PRICE CASE RealEst2

Model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ Sample size: $n = 10$

```
S = 3.24164   R-Sq = 99.0%   R-Sq(adj) = 98.7%

Analysis of Variance
Source           DF      SS      MS      F      P
Regression        2  7374.0  3687.0  350.87  0.000
Residual Error    7    73.6    10.5
Total             9  7447.5
```

15.12 THE FRESH DETERGENT CASE Fresh2

Model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$ Sample size: $n = 30$

Regression Statistics	
Multiple R	0.9453
R Square	0.8936
Adjusted R Square	0.8813
Standard Error	0.2347
Observations	30

ANOVA	df	SS	MS	F	Significance F
Regression	3	12.0268	4.0089	72.7973	0.0000
Residual	26	1.4318	0.0551		
Total	29	13.4586			

15.13 THE HOSPITAL LABOR NEEDS CASE HospLab

Model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$ Sample size: $n = 16$

Regression Statistics	
Multiple R	0.9981
R Square	0.9961
Adjusted R Square	0.9952
Standard Error	387.1598
Observations	16

ANOVA	df	SS	MS	F	Significance F
Regression	3	462327889.4	154109296.5	1028.1309	9.92E-15
Residual	12	1798712.2	149892.7		
Total	15	464126601.6			

15.5 Testing the Significance of an Independent Variable

LO15-5 Test the significance of a single independent variable.

Consider the multiple regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

In order to gain information about which independent variables significantly affect y, we can test the significance of a single independent variable. We arbitrarily refer to this variable as x_j and assume that it is multiplied by the parameter β_j. For example, if $j = 1$, we are testing the significance of x_1, which is multiplied by β_1; if $j = 2$, we are testing the significance of x_2, which is

multiplied by β_2. To test the significance of x_j, we test the null hypothesis H_0: $\beta_j = 0$. We usually test H_0 versus the alternative hypothesis H_a: $\beta_j \neq 0$. **It is reasonable to conclude that x_j is significantly related to y in the regression model under consideration if H_0 can be rejected in favor of H_a at a small level of significance.** Here the phrase *in the regression model under consideration* is very important. This is because it can be shown that whether x_j is significantly related to y in a particular regression model can depend on what other independent variables are included in the model. This issue will be discussed in detail in Section 15.10.

Testing the significance of x_j in a multiple regression model is similar to testing the significance of the slope in the simple linear regression model (recall we test H_0: $\beta_1 = 0$ in simple regression). It can be proved that, if the regression assumptions hold, the population of all possible values of the least squares point estimate b_j is normally distributed with mean β_j and standard deviation σ_{b_j}. The point estimate of σ_{b_j} is called the **standard error of the estimate b_j** and is denoted s_{b_j}. The formula for s_{b_j} involves matrix algebra and is discussed in Bowerman, O'Connell, and Koehler (2005). In our discussion here, we will rely on Excel and MINITAB to compute s_{b_j}. It can be shown that, if the regression assumptions hold, then the population of all possible values of

$$\frac{b_j - \beta_j}{s_{b_j}}$$

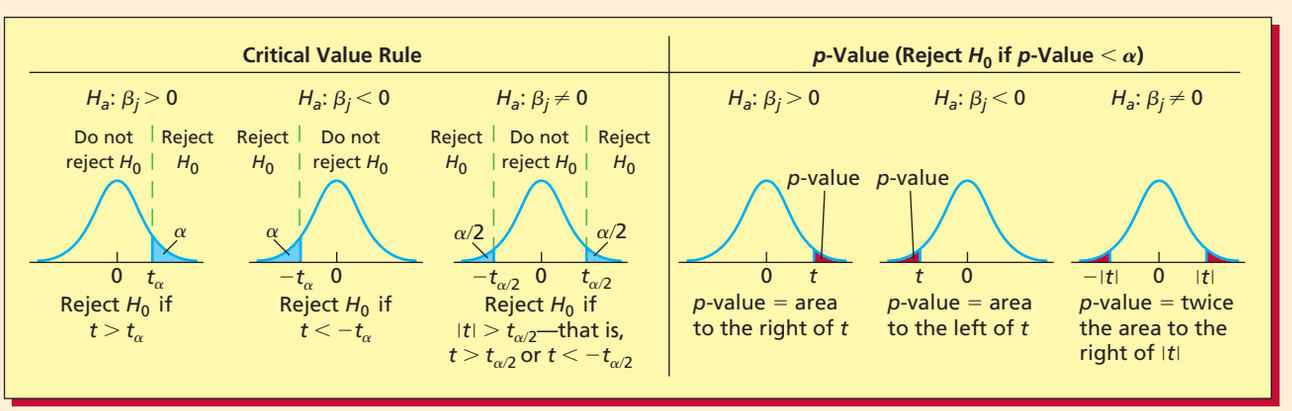
has a t distribution with $n - (k + 1)$ degrees of freedom. It follows that, if the null hypothesis H_0: $\beta_j = 0$ is true, then the population of all possible values of the test statistic

$$t = \frac{b_j}{s_{b_j}}$$

has a t distribution with $n - (k + 1)$ degrees of freedom. Therefore, we can test the significance of x_j as follows:

Testing the Significance of the Independent Variable x_j

Null Hypothesis $H_0: \beta_j = 0$ **Test Statistic** $t = \dfrac{b_j}{s_{b_j}}$ $df = n - (k + 1)$ **Assumptions** The regression assumptions



As in testing H_0: $\beta_1 = 0$ in simple linear regression, we usually use the two-sided alternative hypothesis H_a: $\beta_j \neq 0$. Excel and MINITAB present the results for the two-sided test.

It is customary to test the significance of each and every independent variable in a regression model. Generally speaking,

1. If we can reject H_0: $\beta_j = 0$ at the .05 level of significance, we have strong evidence that the independent variable x_j is significantly related to y in the regression model.
2. If we can reject H_0: $\beta_j = 0$ at the .01 level of significance, we have very strong evidence that x_j is significantly related to y in the regression model.
3. The smaller the significance level α at which H_0 can be rejected, the stronger is the evidence that x_j is significantly related to y in the regression model.

TABLE 15.7 *t* Statistics and *p*-Values for Testing the Significance of the Intercept, x_1, and x_2 in the Tasty Sub Shop Revenue Model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

(a) Calculation of the *t* statistics

Independent Variable	Null Hypothesis	b_j	s_{b_j}	$t = \frac{b_j}{s_{b_j}}$	*p*-Value
Intercept	H_0: $\beta_0 = 0$	$b_0 = 125.29$	$s_{b_0} = 40.93$	$t = \frac{b_0}{s_{b_0}} = \frac{125.29}{40.93} = 3.06$	.0183
x_1	H_0: $\beta_1 = 0$	$b_1 = 14.1996$	$s_{b_1} = 0.91$	$t = \frac{b_1}{s_{b_1}} = \frac{14.1996}{.91} = 15.6$	$< .001$
x_2	H_0: $\beta_2 = 0$	$b_2 = 22.811$	$s_{b_2} = 5.769$	$t = \frac{b_2}{s_{b_2}} = \frac{22.811}{5.769} = 3.95$	.0055

(b) The MINITAB output

Predictor	Coef	SE Coef	T	P
Constant	125.29	40.93	3.06	0.018
population	14.1996	0.91	15.6	0.000
bus_rating	22.811	5.769	3.95	0.006

(c) The Excel output

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	125.289	40.9333	3.06	0.0183	28.4969	222.0807
population	14.1996	0.9100	15.60	1.07E-06	12.0478	16.3515
bus_rating	22.8107	5.7692	3.95	0.0055	9.1686	36.4527

EXAMPLE 15.3 The Tasty Sub Shop Case: *t* statistics and related *p*-values

Again consider the Tasty Sub Shop revenue model

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

Table 15.7(a) summarizes the calculation of the t statistics and related p-values for testing the significance of the intercept and each of the independent variables x_1 and x_2. Here the values of b_j, s_{b_j}, t, and the p-value have been obtained from the MINITAB and Excel outputs of Table 15.7(b) and (c). If we wish to carry out tests at the .05 level of significance, we use the critical value $t_{.05/2} = t_{.025} = 2.365$, which is based on $n - (k + 1) = 10 - 3 = 7$ degrees of freedom. Looking at Table 15.7 (a), we see that

1. For the intercept, $|t| = 3.06 > 2.365$.
2. For x_1, $|t| = 15.6 > 2.365$.
3. For x_2, $|t| = 3.95 > 2.365$.

Because in each case $|t| > t_{.025}$, we reject each of the null hypotheses in Table 15.7(a) at the .05 level of significance. Furthermore, because the p-value related to x_1 is less than .001, we can reject H_0: $\beta_1 = 0$ at the .001 level of significance. Also, because the p-value related to x_2 is less than .01, we can reject H_0: $\beta_2 = 0$ at the .01 level of significance. On the basis of these results, we have extremely strong evidence that in the above model x_1 (population size) is significantly related to y. We also have very strong evidence that in this model x_2 (business rating) is significantly related to y.

We next consider how to calculate a confidence interval for a regression parameter.

A Confidence Interval for the Regression Parameter β_j

If the regression assumptions hold, a **100(1 − α) percent confidence interval for β_j** is

$$[b_j \pm t_{\alpha/2} s_{b_j}]$$

Here $t_{\alpha/2}$ is based on $n - (k + 1)$ degrees of freedom.

EXAMPLE 15.4 The Tasty Sub Shop Case: A Confidence Interval for β_1

Consider the Tasty Sub Shop revenue model

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

The MINITAB and Excel outputs in Table 15.7 tell us that $b_1 = 14.1996$ and $s_{b_1} = .91$. It follows, because $t_{.025}$ based on $n - (k + 1) = 10 - 3 = 7$ degrees of freedom equals 2.365, that a 95 percent confidence interval for β_1 is (see the Excel output)

$$[b_1 \pm t_{.025} s_{b_1}] = [14.1996 \pm 2.365(.91)]$$
$$= [12.048, 16.352]$$

This interval says we are 95 percent confident that, if the population size increases by 1,000 residents and the business rating does not change, then mean yearly revenue will increase by between \$12,048 and \$16,352. Furthermore, because this 95 percent confidence interval does not contain 0, we can reject H_0: $\beta_1 = 0$ in favor of H_a: $\beta_1 \neq 0$ at the .05 level of significance.

Exercises for Section 15.5

CONCEPTS

connect

15.14 What do we conclude about x_j if we can reject H_0: $\beta_j = 0$ in favor of H_a: $\beta_j \neq 0$ by setting

a α equal to .05?

b α equal to .01?

15.15 Give an example of a practical application of the confidence interval for β_j.

METHODS AND APPLICATIONS

In Exercises 15.16 through 15.19 we refer to Excel and MINITAB outputs of regression analyses of the data sets related to four case studies introduced in Section 15.1. The outputs are given in Figure 15.9. Using the appropriate output, do the following for **each parameter** β_j in the model under consideration:

1. Find b_j, s_{b_j}, and the t statistic for testing H_0: $\beta_j = 0$ on the output and report their values. Show how t has been calculated by using b_j and s_{b_j}.
2. Using the t statistic and appropriate critical values, test H_0: $\beta_j = 0$ versus H_a: $\beta_j \neq 0$ by setting α equal to .05. Which independent variables are significantly related to y in the model with $\alpha = .05$?
3. Using the t statistic and appropriate critical values, test H_0: $\beta_j = 0$ versus H_a: $\beta_j \neq 0$ by setting α equal to .01. Which independent variables are significantly related to y in the model with $\alpha = .01$?
4. Find the p-value for testing H_0: $\beta_j = 0$ versus H_a: $\beta_j \neq 0$ on the output. Using the p-value, determine whether we can reject H_0 by setting α equal to .10, .05, .01, and .001. What do you conclude about the significance of the independent variables in the model?
5. Calculate the 95 percent confidence interval for β_j. Discuss one practical application of this interval.
6. Calculate the 99 percent confidence interval for β_j.

15.16 THE NATURAL GAS CONSUMPTION CASE DS GasCon2

Use the MINITAB output in Figure 15.9(a) to do (1) through (6) for each of β_0, β_1, and β_2.

FIGURE 15.9 ***t* Statistics and *p*-Values for Four Case Studies**

(a) MINITAB output for the natural gas consumption case (sample size : $n = 8$)

```
Predictor       Coef   SE Coef       T       P
Constant     13.1087    0.8557   15.32   0.000
Temp        -0.09001   0.01408   -6.39   0.001
Chill        0.08249   0.02200    3.75   0.013
```

(b) MINITAB output for the real estate sales price case (sample size: $n = 10$)

```
Predictor     Coef   SE Coef       T       P
Constant    29.347     4.891    6.00   0.001
HomeSize    5.6128    0.2285   24.56   0.000
Rating      3.8344    0.4332    8.85   0.000
```

(c) Excel output for the Fresh detergent case (sample size: $n = 30$)

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	7.5891	2.4450	3.1039	0.0046	2.5633	12.6149
Price (x1)	-2.3577	0.6379	-3.6958	0.0010	-3.6690	-1.0464
IndPrice (x2)	1.6122	0.2954	5.4586	0.0000	1.0051	2.2193
AdvExp (x3)	0.5012	0.1259	3.9814	0.0005	0.2424	0.7599

(d) Excel output for the hospital labor needs case (sample size: $n = 16$)

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	1946.8020	504.1819	3.8613	0.0023	848.2840	3045.3201
XRay (x1)	0.0386	0.0130	2.9579	0.0120	0.0102	0.0670
BedDays (x2)	1.0394	0.0676	15.3857	2.91E-09	0.8922	1.1866
LengthStay (x3)	-413.7578	98.5983	-4.1964	0.0012	-628.5850	-198.9306

15.17 THE REAL ESTATE SALES PRICE CASE DS RealEst2

Use the MINITAB output in Figure 15.9(b) to do (1) through (6) for each of β_0, β_1, and β_2.

15.18 THE FRESH DETERGENT CASE DS Fresh2

Use the Excel output in Figure 15.9(c) to do (1) through (6) for each of β_0, β_1, β_2, and β_3.

15.19 THE HOSPITAL LABOR NEEDS CASE DS HospLab

Use the Excel output in Figure 15.9(d) to do (1) through (6) for each of β_0, β_1, β_2, and β_3.

15.6 Confidence and Prediction Intervals

LO15-6 Find and interpret a confidence interval for a mean value and a prediction interval for an individual value.

In this section we show how to use the multiple regression model to find a **confidence interval for a mean value of *y*** and a **prediction interval for an individual value of *y*.** We first present an example of these intervals, and we then discuss (in an optional technical note) the formulas used to compute the intervals.

EXAMPLE 15.5 The Tasty Sub Shop Case: Estimating Mean Revenue and Predicting Revenue and Profit

In the Tasty Sub Shop problem, recall that one of the business entrepreneur's potential sites is near a population of 47,300 residents and a business/shopping area having a rating of 7. Also, recall that

$$\begin{aligned}\hat{y} &= b_0 + b_1x_1 + b_2x_2 \\ &= 125.29 + 14.1996\,(47.3) + 22.811(7) \\ &= 956.6 \text{ (that is, \$956,600)}\end{aligned}$$

is:

1 The **point estimate** of the mean yearly revenue for all Tasty Sub restaurants that could potentially be built near populations of 47,300 residents and business/shopping areas having a rating of 7, and
2 The **point prediction** of the yearly revenue for a single Tasty Sub restaurant that is built near a population of 47,300 residents and a business/shopping area having a rating of 7.

This point estimate and prediction are given at the bottom of the MINITAB output in Figure 15.4, which we repeat here as follows:

New Obs	Fit	SE Fit	95% CI	95% PI
1	956.6	15	(921.0, 992.2)	(862.8, 1050.4)

In addition to giving $\hat{y} = 956.6$, the MINITAB output also gives a 95 percent confidence interval and a 95 percent prediction interval. The 95 percent confidence interval, [921.0, 992.2], says that we are 95 percent confident that the mean yearly revenue for all Tasty Sub restaurants that could potentially be built near populations of 47,300 residents and business/shopping areas having a rating of 7 is between \$921,000 and \$992,200. The 95 percent prediction interval, [862.8, 1050.4], says that we are 95 percent confident that the yearly revenue for a single Tasty Sub restaurant that is built near a population of 47,300 residents and a business/shopping area having a rating of 7 will be between \$862,800 and \$1,050,400.

Now, recall that the yearly rent and other fixed costs for the entrepreneur's potential restaurant will be \$257,550 and that (according to Tasty Sub corporate headquarters) the yearly food and other variable costs for the restaurant will be 60 percent of the yearly revenue. Using the lower end of the 95 percent prediction interval [862.8, 1050.4], we predict that (1) the restaurant's yearly operating cost will be \$257,550 + .6(862,800) = \$775,230 and (2) the restaurant's yearly profit will be \$862,800 − \$775,230 = \$87,570. Using the upper end of the 95 percent prediction interval [862.8, 1050.4], we predict that (1) the restaurant's yearly operating cost will be \$257,550 + .6(1,050,400) = \$887,790 and (2) the restaurant's yearly profit will be \$1,050,400 − \$887,790 = \$162,610. Combining the two predicted profits, it follows that we are 95 percent confident that the potential restaurant's yearly profit will be between \$87,570 and \$162,610. If the entrepreneur decides that this is an acceptable range of potential yearly profits, then the entrepreneur might decide to purchase a Tasty Sub franchise for the potential restaurant site.

BI

A technical note (optional) In general

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k$$

is the **point estimate of the mean value of the dependent variable *y*** when the values of the independent variables are $x_1, x_2, \ldots, x_k$ and is the **point prediction of an individual value of the dependent variable *y*** when the values of the independent variables are $x_1, x_2, \ldots, x_k$. Furthermore:

A Confidence Interval and a Prediction Interval

If the regression assumptions hold,

1 A **100(1 − α) percent confidence interval for the mean value of *y*** when the values of the independent variables are $x_1, x_2, \ldots, x_k$ is

$$[\hat{y} \pm t_{\alpha/2}s\sqrt{\text{distance value}}]$$

2 A **100(1 − α) percent prediction interval for an individual value of *y*** when the values of the independent variables are $x_1, x_2, \ldots, x_k$ is

$$[\hat{y} \pm t_{\alpha/2}s\sqrt{1 + \text{distance value}}]$$

Here $t_{\alpha/2}$ is based on $n - (k + 1)$ degrees of freedom and s is the standard error (see page 566). Furthermore, the formula for the **distance value** (also sometimes called the **leverage value**) involves matrix algebra and is given in Bowerman, O'Connell, and Koehler (2005). In practice, we can obtain the distance value from the outputs of statistical software packages (such as MINITAB and an Excel add-in).

Intuitively, the **distance value** is a measure of the distance of the combination of values x_1, x_2, . . . , x_k from the center of the observed data. The farther that this combination is from the center of the observed data, the larger is the distance value, and thus the longer are both the confidence interval and the prediction interval.

MINITAB gives $s_{\hat{y}} = s\sqrt{\text{distance value}}$ under the heading "SE Fit." Because the MINITAB output also gives s, the distance value can be found by calculating $(s_{\hat{y}}/s)^2$. For example, the MINITAB output in Figure 15.4 (page 558) tells us that $\hat{y} = 956.6$ (see "Fit") and $s_{\hat{y}} = 15$ (see "SE Fit"). Therefore, because s for the two-variable Tasty Sub Shop revenue model equals 36.6856 (see Figure 15.4), the distance value equals $(15/36.6856)^2 = .1671826$. It follows that the 95 percent confidence and prediction intervals given on the MINITAB output of Figure 15.4 have been calculated (within rounding) as follows:

$$[\hat{y} \pm t_{.025}\, s\sqrt{\text{distance value}}]$$
$$= [956.6 \pm 2.365(36.6856)\sqrt{.1671826}]$$
$$= [956.6 \pm 35.47]$$
$$= [921.1, 992.1]$$

$$[\hat{y} \pm t_{.025}\, s\sqrt{1 + \text{distance value}}]$$
$$= [956.6 \pm 2.365(36.6856)\sqrt{1.1671826}]$$
$$= [956.6 \pm 93.73]$$
$$= [862.9, 1050.3]$$

Here $t_{\alpha/2} = t_{.025} = 2.365$ is based on $n - (k + 1) = 10 - 3 = 7$ degrees of freedom.

Exercises for Section 15.6

CONCEPTS

connect

15.20 What is the difference between a confidence interval and a prediction interval?

15.21 What does the distance value measure? How does the distance value affect a confidence or prediction interval? (Note: You must read the optional technical note to answer this question.)

METHODS AND APPLICATIONS

15.22 THE NATURAL GAS CONSUMPTION CASE GasCon2

The following partial MINITAB regression output for the natural gas consumption data relates to predicting the city's natural gas consumption (in MMcF) in a week that has an average hourly temperature of 40°F and a chill index of 10.

New Obs	Fit	SE Fit	95% CI	95% PI
1	10.333	0.170	(9.895, 10.771)	(9.293, 11.374)

a Report (as shown on the computer output) a point estimate of and a 95 percent confidence interval for the mean natural gas consumption for all weeks having an average hourly temperature of 40°F and a chill index of 10.

b Report (as shown on the computer output) a point prediction of and a 95 percent prediction interval for the natural gas consumption in a single week that has an average hourly temperature of 40°F and a chill index of 10.

c Suppose that next week the city's average hourly temperature will be 40°F and the city's chill index will be 10. Also, suppose the city's natural gas company will use the point prediction $\hat{y} = 10.333$ and order 10.333 MMcF of natural gas to be shipped to the city by a pipeline transmission system. The gas company will have to pay a fine to the transmission system if the city's actual gas usage y differs from the order of 10.333 MMcF by more than 10.5 percent—that is, is outside of the range $[10.333 \pm .105(10.333)] = [9.248, 11.418]$. Discuss why the 95 percent prediction interval for y, [9.293, 11.374], says that y is likely to be inside the allowable range and thus makes the gas company 95 percent confident that it will avoid paying a fine.

d Find 99 percent confidence and prediction intervals for the mean and actual natural gas consumption referred to in parts *a* and *b*. Hint: $n = 8$ and $s = .367078$. Optional technical note needed.

15.23 THE REAL ESTATE SALES PRICE CASE RealEst2

The following MINITAB output relates to a house having 2,000 square feet and a rating of 8.

New Obs	Fit	SE Fit	95% CI	95% PI
1	172.28	1.57	(168.56, 175.99)	(163.76, 180.80)

a Report (as shown on the output) a point estimate of and a 95 percent confidence interval for the mean sales price of all houses having 2,000 square feet and a rating of 8.

b Report (as shown on the output) a point prediction of and a 95 percent prediction interval for the actual sales price of an individual house having 2,000 square feet and a rating of 8.

c Find 99 percent confidence and prediction intervals for the mean and actual sales prices referred to in parts *a* and *b*. Hint: $n = 10$ and $s = 3.24164$. Optional technical note needed.

15.24 THE FRESH DETERGENT CASE Fresh2

Consider the demand for Fresh Detergent in a future sales period when Enterprise Industries' price for Fresh will be $x_1 = 3.70$, the average price of competitors' similar detergents will be $x_2 = 3.90$, and Enterprise Industries' advertising expenditure for Fresh will be $x_3 = 6.50$. A 95 percent prediction interval for this demand is given on the following Excel add-in (MegaStat) output:

	95% Confidence Interval		95% Prediction Interval		
Predicted	lower	upper	lower	upper	Leverage
8.4107	8.3143	8.5070	7.9188	8.9025	0.040

a Find and report the 95 percent prediction interval on the output. If Enterprise Industries plans to have in inventory the number of bottles implied by the upper limit of this interval, it can be very confident that it will have enough bottles to meet demand for Fresh in the future sales period. How many bottles is this? If we multiply the number of bottles implied by the lower limit of the prediction interval by the price of Fresh ($3.70), we can be very confident that the resulting dollar amount will be the minimum revenue from Fresh in the future sales period. What is this dollar amount?

b Calculate a 99 percent prediction interval for the demand for Fresh in the future sales period. Hint: $n = 30$ and $s = .235$. Optional technical note needed. Note that the distance value equals **Leverage.**

15.25 THE HOSPITAL LABOR NEEDS CASE HospLab

Consider a questionable hospital for which XRay = 56,194, BedDays = 14,077.88, and LengthStay = 6.89. A 95 percent prediction interval for the labor hours corresponding to this combination of values of the independent variables is given in the following Excel add-in (MegaStat) output:

	95% Confidence Interval		95% Prediction Interval		
Predicted	lower	upper	lower	upper	Leverage
15,896.2473	15,378.0313	16,414.4632	14,906.2361	16,886.2584	0.3774

Find and report the prediction interval on the output. Then, use this interval to determine if the actual number of labor hours used by the questionable hospital ($y = 17{,}207.31$) is unusually low or high.

15.7 The Sales Representative Case: Evaluating Employee Performance

Suppose the sales manager of a company wishes to evaluate the performance of the company's sales representatives. Each sales representative is solely responsible for one sales territory, and the manager decides that it is reasonable to measure the performance, *y*, of a sales representative by using the yearly sales of the company's product in the representative's sales territory. The manager feels that sales performance *y* substantially depends on five independent variables:

x_1 = number of months the representative has been employed by the company

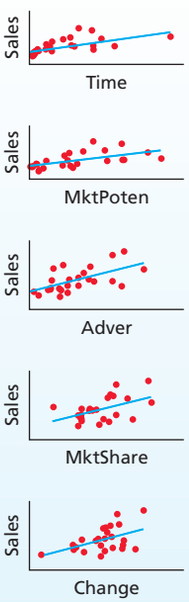
x_2 = sales of the company's product and competing products in the sales territory (a measure of sales potential)

x_3 = dollar advertising expenditure in the territory

x_4 = weighted average of the company's market share in the territory for the previous four years

x_5 = change in the company's market share in the territory over the previous four years

FIGURE 15.10 Sales Representative Performance Study Data SalePerf



Sales, y	Time with Company, x_1	Market Potential, x_2	Advertising, x_3	Market Share, x_4	Market Share Change, x_5
3,669.88	43.10	74,065.11	4,582.88	2.51	0.34
3,473.95	108.13	58,117.30	5,539.78	5.51	0.15
2,295.10	13.82	21,118.49	2,950.38	10.91	−0.72
4,675.56	186.18	68,521.27	2,243.07	8.27	0.17
6,125.96	161.79	57,805.11	7,747.08	9.15	0.50
2,134.94	8.94	37,806.94	402.44	5.51	0.15
5,031.66	365.04	50,935.26	3,140.62	8.54	0.55
3,367.45	220.32	35,602.08	2,086.16	7.07	−0.49
6,519.45	127.64	46,176.77	8,846.25	12.54	1.24
4,876.37	105.69	42,053.24	5,673.11	8.85	0.31
2,468.27	57.72	36,829.71	2,761.76	5.38	0.37
2,533.31	23.58	33,612.67	1,991.85	5.43	−0.65
2,408.11	13.82	21,412.79	1,971.52	8.48	0.64
2,337.38	13.82	20,416.87	1,737.38	7.80	1.01
4,586.95	86.99	36,272.00	10,694.20	10.34	0.11
2,729.24	165.85	23,093.26	8,618.61	5.15	0.04
3,289.40	116.26	26,878.59	7,747.89	6.64	0.68
2,800.78	42.28	39,571.96	4,565.81	5.45	0.66
3,264.20	52.84	51,866.15	6,022.70	6.31	−0.10
3,453.62	165.04	58,749.82	3,721.10	6.35	−0.03
1,741.45	10.57	23,990.82	860.97	7.37	−1.63
2,035.75	13.82	25,694.86	3,571.51	8.39	−0.43
1,578.00	8.13	23,736.35	2,845.50	5.15	0.04
4,167.44	58.54	34,314.29	5,060.11	12.88	0.22
2,799.97	21.14	22,809.53	3,552.00	9.14	−0.74

Source: This data set is from a research study published in "An Analytical Approach for Evaluation of Sales Territory Performance," *Journal of Marketing,* January 1972, 31–37 (authors are David W. Cravens, Robert B. Woodruff, and Joseph C. Stamper). We have updated the situation in our case study to be more modern.

In Figure 15.10 we present values of y and x_1 through x_5 for 25 randomly selected sales representatives. To understand the values of y and x_2 in the table, note that sales of the company's product or any competing product are measured in hundreds of units of the product sold. Therefore, for example, the first sales figure of 3,669.88 in Figure 15.10 means that the first randomly selected sales representative sold 366,988 units of the company's product during the year.

Plots of y versus x_1 through x_5 are given in Figure 15.10. Because each plot has an approximate straight-line appearance, it is reasonable to relate y to x_1 through x_5 by using the regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon$$

The main objective of the regression analysis is to help the sales manager evaluate sales performance by comparing actual performance to predicted performance. The manager has randomly selected the 25 representatives from all the representatives the company considers to be effective and wishes to use a regression model based on effective representatives to evaluate questionable representatives.

Figure 15.11 on the next page gives the Excel output of a regression analysis of the sales representative performance data using the five independent variable model. This output tells us that the least squares point estimates of the model parameters are $b_0 = -1{,}113.7879$, $b_1 = 3.6121$, $b_2 = .0421$, $b_3 = .1289$, $b_4 = 256.9555$, and $b_5 = 324.5334$. In addition, because the output tells us that the p-values associated with Time, MktPoten, Adver, and MktShare are all less than .01, we have very strong evidence that these variables are significantly related to y and, thus, are important in this model. Because the p-value associated with Change is .0530, we have close to strong evidence that this variable is also important.

FIGURE 15.11 Excel Output of a Regression Analysis of the Sales Representative Performance Data Using the Model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon$

(a) The Excel output

Regression Statistics

Multiple R	0.9566
R Square	0.9150
Adjusted R Square	0.8926
Standard Error	430.2319
Observations	25

ANOVA	df	SS	MS	F	Significance F
Regression	5	37862658.9002	7572531.7800	40.9106	0.0000
Residual	19	3516890.0266	185099.4751		
Total	24	41379548.9269			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-1113.7879	419.8869	-2.6526	0.0157	-1992.6213	-234.9545
Time	3.6121	1.1817	3.0567	0.0065	1.1388	6.0854
MktPoten	0.0421	0.0067	6.2527	0.0000	0.0280	0.0562
Adver	0.1289	0.0370	3.4792	0.0025	0.0513	0.2064
MktShare	256.9555	39.1361	6.5657	0.0000	175.0428	338.8683
Change	324.5334	157.2831	2.0634	0.0530	-4.6638	653.7307

(b) Prediction using an Excel add-in (MegaStat)

Predicted values for: Sales

	95% Confidence Interval		95% Prediction Interval		
Predicted	lower	upper	lower	upper	Leverage
4,181.74333	3,884.90651	4,478.58015	3,233.59431	5,129.89235	0.109

Consider a questionable sales representative for whom Time = 85.42, MktPoten = 35,182.73, Adver = 7,281.65, MktShare = 9.64, and Change = .28. The point prediction of the sales, y, corresponding to this combination of values of the independent variables is

$$\begin{aligned}\hat{y} &= -1{,}113.7879 + 3.6121(85.42) + .0421(35{,}182.73) \\ &\quad + .1289(7{,}281.65) + 256.9555(9.64) + 324.5334(.28) \\ &= 4{,}181.74 \text{ (that is, 418,174 units)}\end{aligned}$$

In addition to giving this point prediction, the Excel output tells us that a 95 percent prediction interval for y is [3233.59, 5129.89]. Furthermore, suppose that the actual sales y for the questionable representative were 3,087.52. This actual sales figure is less than the point prediction $\hat{y} = 4{,}181.74$ and is less than the lower bound of the 95 percent prediction interval for y, [3233.59, 5129.89]. Therefore, we conclude that there is strong evidence that the actual performance of the questionable representative is less than predicted performance. We should investigate the reason for this. Perhaps the questionable representative needs special training.

BI

LO15-7 Use dummy variables to model qualitative independent variables.

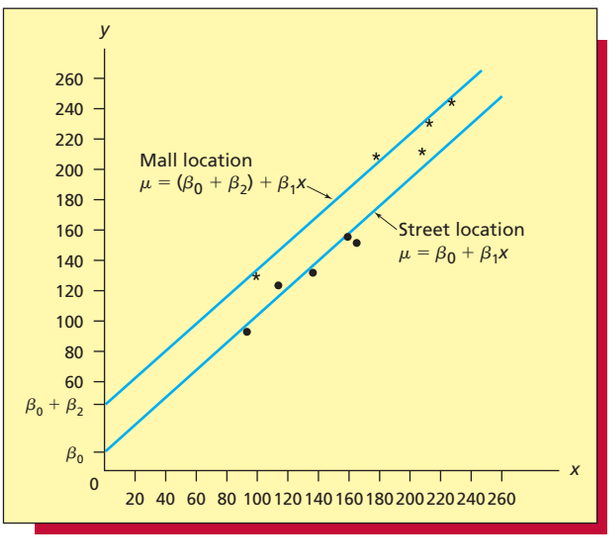
15.8 Using Dummy Variables to Model Qualitative Independent Variables

While the levels (or values) of a quantitative independent variable are numerical, the levels of a **qualitative** independent variable are defined by describing them. For instance, the type of sales technique used by a door-to-door salesperson is a qualitative independent variable. Here we might define three different levels—high pressure, medium pressure, and low pressure.

TABLE 15.8 The Electronics World Sales Volume Data
Electronics1

Store	Number of Households, x	Location	Sales Volume, y
1	161	Street	157.27
2	99	Street	93.28
3	135	Street	136.81
4	120	Street	123.79
5	164	Street	153.51
6	221	Mall	241.74
7	179	Mall	201.54
8	204	Mall	206.71
9	214	Mall	229.78
10	101	Mall	135.22

FIGURE 15.12 Plot of the Sales Volume Data and a Geometrical Interpretation of the Model $y = \beta_0 + \beta_1 x + \beta_2 D_M + \varepsilon$



We can model the effects of the different levels of a qualitative independent variable by using what we call **dummy variables** (also called **indicator variables**). Such variables are usually defined so that they take on two values—either 0 or 1. To see how we use dummy variables, we begin with an example.

EXAMPLE 15.6 The Electronics World Case: Comparing Three Kinds of Store Locations

Part 1: The data and data plots Suppose that Electronics World, a chain of stores that sells audio and video equipment, has gathered the data in Table 15.8. These data concern store sales volume in July of last year (y, measured in thousands of dollars), the number of households in the store's area (x, measured in thousands), and the location of the store (on a suburban street or in a suburban shopping mall—a qualitative independent variable). Figure 15.12 gives a data plot of y versus x. Stores having a street location are plotted as solid dots, while stores having a mall location are plotted as asterisks. Notice that the line relating y to x for mall locations has a higher y-intercept than does the line relating y to x for street locations.

Part 2: A dummy variable model In order to model the effects of the street and shopping mall locations, we define a dummy variable denoted D_M as follows:

$$D_M = \begin{cases} 1 & \text{if a store is in a mall location} \\ 0 & \text{otherwise} \end{cases}$$

Using this dummy variable, we consider the regression model

$$y = \beta_0 + \beta_1 x + \beta_2 D_M + \varepsilon$$

This model and the definition of D_M imply that

1 For a street location, mean sales volume equals

$$\begin{aligned} \beta_0 + \beta_1 x + \beta_2 D_M &= \beta_0 + \beta_1 x + \beta_2(0) \\ &= \beta_0 + \beta_1 x \end{aligned}$$

2 For a mall location, mean sales volume equals

$$\begin{aligned} \beta_0 + \beta_1 x + \beta_2 D_M &= \beta_0 + \beta_1 x + \beta_2(1) \\ &= (\beta_0 + \beta_2) + \beta_1 x \end{aligned}$$

FIGURE 15.13 Excel Output of a Regression Analysis of the Sales Volume Data Using the Model $y = \beta_0 + \beta_1 x + \beta_2 D_M + \varepsilon$

Regression Statistics	
Multiple R	0.9913
R Square	0.9827
Adjusted R Square	0.9778
Standard Error	7.3288
Observations	10

ANOVA	df	SS	MS	F	Significance F
Regression	2	21411.7977	10705.8989	199.3216	6.75E-07
Residual	7	375.9817	53.7117		
Total	9	21787.7795			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	17.3598	9.4470	1.8376	0.1087	-4.9788	39.6985
Households (x)	0.8510	0.0652	13.0439	3.63E-06	0.6968	1.0053
DummyMall	29.2157	5.5940	5.2227	0.0012	15.9881	42.4434

Thus, the dummy variable allows us to model the situation illustrated in Figure 15.12. Here, the lines relating mean sales volume to x for street and mall locations have different y intercepts, β_0 and $(\beta_0 + \beta_2)$, and the same slope β_1. Note that β_2 is the difference between the mean monthly sales volume for stores in mall locations and the mean monthly sales volume for stores in street locations, when all these stores have the same number of households in their areas. That is, we can say that β_2 represents the effect on mean sales of a mall location compared to a street location. The Excel output in Figure 15.13 tells us that the least squares point estimate of β_2 is $b_2 = 29.2157$. This says that for any given number of households in a store's area, we estimate that the mean monthly sales volume in a mall location is \$29,215.70 greater than the mean monthly sales volume in a street location.

Part 3: A dummy variable model for comparing three locations In addition to the data concerning street and mall locations in Table 15.8, Electronics World has also collected data concerning downtown locations. The complete data set is given in Table 15.9 and plotted in Figure 15.14. Here stores having a downtown location are plotted as open circles. A model describing these data is

$$y = \beta_0 + \beta_1 x + \beta_2 D_M + \beta_3 D_D + \varepsilon$$

Here the dummy variable D_M is as previously defined and the dummy variable D_D is defined as follows

$$D_D = \begin{cases} 1 & \text{if a store is in a downtown location} \\ 0 & \text{otherwise} \end{cases}$$

It follows that

1 For a street location, mean sales volume equals

$$\begin{aligned} \beta_0 + \beta_1 x + \beta_2 D_M + \beta_3 D_D &= \beta_0 + \beta_1 x + \beta_2(0) + \beta_3(0) \\ &= \beta_0 + \beta_1 x \end{aligned}$$

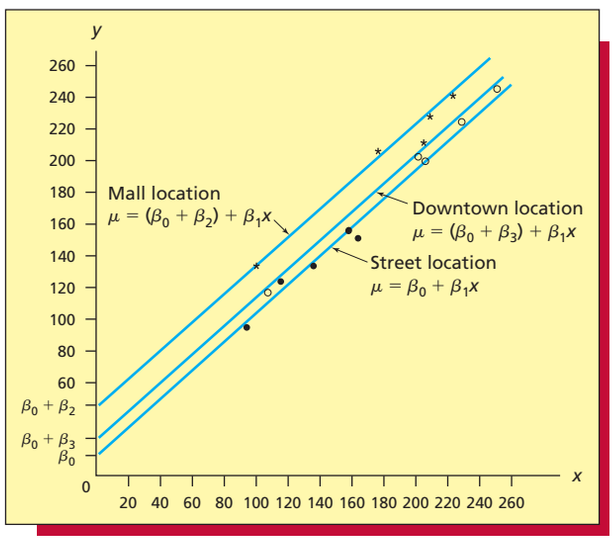
2 For a mall location, mean sales volume equals

$$\begin{aligned} \beta_0 + \beta_1 x + \beta_2 D_M + \beta_3 D_D &= \beta_0 + \beta_1 x + \beta_2(1) + \beta_3(0) \\ &= (\beta_0 + \beta_2) + \beta_1 x \end{aligned}$$

TABLE 15.9 The Complete Electronics World Sales Volume Data
Electronics2

Store	Number of Households, x	Location	Sales Volume, y
1	161	Street	157.27
2	99	Street	93.28
3	135	Street	136.81
4	120	Street	123.79
5	164	Street	153.51
6	221	Mall	241.74
7	179	Mall	201.54
8	204	Mall	206.71
9	214	Mall	229.78
10	101	Mall	135.22
11	231	Downtown	224.71
12	206	Downtown	195.29
13	248	Downtown	242.16
14	107	Downtown	115.21
15	205	Downtown	197.82

FIGURE 15.14 Plot of the Complete Electronics World Sales Volume Data and a Geometrical Interpretation of the Model $y = \beta_0 + \beta_1 x + \beta_2 D_M + \beta_3 D_D + \varepsilon$



3 For a downtown location, mean sales volume equals

$$\begin{aligned}\beta_0 + \beta_1 x + \beta_2 D_M + \beta_3 D_D &= \beta_0 + \beta_1 x + \beta_2(0) + \beta_3(1) \\ &= (\beta_0 + \beta_3) + \beta_1 x\end{aligned}$$

Thus, the dummy variables allow us to model the situation illustrated in Figure 15.14. Here the lines relating mean sales volume to x for street, mall, and downtown locations have different y-intercepts, β_0, $(\beta_0 + \beta_2)$, and $(\beta_0 + \beta_3)$, and the same slope β_1. Note that β_2 represents the effect on mean sales of a mall location compared to a street location, and β_3 represents the effect on mean sales of a downtown location compared to a street location. Furthermore, the difference between β_2 and β_3, $\beta_2 - \beta_3$, represents the effect on mean sales of a mall location compared to a downtown location.

Part 4: Comparing the three locations Figure 15.15 on the next page gives the MINITAB and Excel outputs of a regression analysis of the sales volume data using the dummy variable model. These outputs tell us that the least squares point estimate of β_2 is $b_2 = 28.374$. This says that for any given number of households in a store's area, we estimate that the mean monthly sales volume in a mall location is \$28,374 greater than the mean monthly sales volume in a street location. Furthermore, because the Excel output tells us that a 95 percent confidence interval for β_2 is [18.5545, 38.193], we are 95 percent confident that for any given number of households in a store's area, the mean monthly sales volume in a mall location is between \$18,554.50 and \$38,193 greater than the mean monthly sales volume in a street location. The MINITAB and Excel outputs also show that the t statistic for testing H_0: $\beta_2 = 0$ versus H_a: $\beta_2 \neq 0$ equals 6.36 and that the related p-value is less than .001. Therefore, we have very strong evidence that there is a difference between the mean monthly sales volumes in mall and street locations.

We next note that the outputs in Figure 15.15 show that the least squares point estimate of β_3 is $b_3 = 6.864$. Therefore, we estimate that for any given number of households in a store's area, the mean monthly sales volume in a downtown location is \$6,864 greater than the mean monthly sales volume in a street location. Furthermore, the Excel output shows that a 95 percent confidence interval for β_3 is [−3.636, 17.3635]. This says we are 95 percent confident that for any given number of households in a store's area, the mean monthly sales volume in a downtown

FIGURE 15.15 MINITAB and Excel Outputs of a Regression Analysis of the Complete Sales Volume Data Using the Model $y = \beta_0 + \beta_1 x + \beta_2 D_M + \beta_3 D_D + \varepsilon$

(a) The MINITAB output

```
The regression equation is
Sales = 15.0 + 0.869 Households + 28.4 DMall + 6.86 DDowntown

Predictor       Coef  SE Coef      T      P
Constant      14.978    6.188   2.42  0.034
Households   0.86859  0.04049  21.45  0.000
DMall         28.374    4.461   6.36  0.000
DDowntown      6.864    4.770   1.44  0.178

S = 6.34941   R-Sq = 98.7%   R-Sq(adj) = 98.3%

Analysis of Variance
Source           DF     SS     MS       F      P
Regression        3  33269  11090  275.07  0.000
Residual Error   11    443     40
Total            14  33712

Values of Predictors for New Obs       Predicted Values for New Observations
New Obs  Households  DMall DDowntown   New Obs      Fit  SE Fit       95% CI            95% PI
      1         200      1         0         1   217.07    2.91  (210.65, 223.48)  (201.69, 232.45)
```

(b) The Excel output

Regression Statistics	
Multiple R	0.9934
R Square	0.9868
Adjusted R Square	0.9833
Standard Error	6.3494
Observations	15

ANOVA	df	SS	MS	F	Significance F
Regression	3	33268.6953	11089.5651	275.0729	1.27E-10
Residual	11	443.4650	40.3150		
Total	14	33712.1603			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	14.9777	6.1884	2.4203	0.0340	1.3570	28.5984
Households (x)	0.8686	0.0405	21.4520	2.52E-10	0.7795	0.9577
DummyMall	28.3738	4.4613	6.3600	5.37E-05	18.5545	38.1930
DummyDtown	6.8638	4.7705	1.4388	0.1780	-3.6360	17.3635

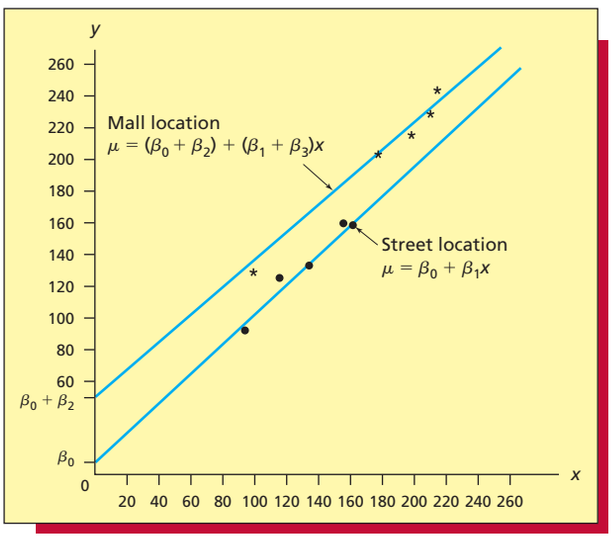
location is between \$3,636 less than and \$17,363.50 greater than the mean monthly sales volume in a street location. The MINITAB and Excel outputs also show that the t statistic and p-value for testing H_0: $\beta_3 = 0$ versus H_a: $\beta_3 \neq 0$ are $t = 1.44$ and p-value $= .178$. Therefore, we do not have strong evidence that there is a difference between the mean monthly sales volumes in downtown and street locations.

Finally, note that, because $b_2 = 28.374$ and $b_3 = 6.864$, the point estimate of $\beta_2 - \beta_3$ is $b_2 - b_3 = 28.374 - 6.864 = 21.51$. Therefore, we estimate that mean monthly sales volume in a mall location is \$21,510 higher than mean monthly sales volume in a downtown location. Near the end of this section we show how to compare the mall and downtown locations by using a confidence interval and a hypothesis test. We will find that there is very strong evidence that the mean monthly sales volume in a mall location is higher than the mean monthly sales volume in a downtown location. In summary, the mall location seems to give a higher mean monthly sales volume than either the street or downtown location.

BI

Part 5: Predicting a future sales volume Suppose that Electronics World wishes to predict the sales volume in a future month for an individual store that has 200,000 households in its area and is located in a shopping mall. The point prediction of this sales volume is (note $D_M = 1$

FIGURE 15.16 Geometrical Interpretation of the Sales Volume Model
$y = \beta_0 + \beta_1 x + \beta_2 D_M + \beta_3 x D_M + \varepsilon$



and $D_D = 0$ when a store is in a shopping mall)

$$\begin{aligned}\hat{y} &= b_0 + b_1(200) + b_2(1) + b_3(0)\\ &= 14.978 + .8686(200) + 28.374(1)\\ &= 217.07\end{aligned}$$

This point prediction is given at the bottom of the MINITAB output in Figure 15.15(a). The corresponding 95 percent prediction interval, which is [201.69, 232.45], says we are 95 percent confident that the sales volume in a future sales period for an individual mall store that has 200,000 households in its area will be between \$201,690 and \$232,450.

Part 6: Interaction models Consider the Electronics World data for street and mall locations given in Table 15.8 (page 581) and the model

$$y = \beta_0 + \beta_1 x + \beta_2 D_M + \beta_3 x D_M + \varepsilon$$

This model uses the *cross-product,* or *interaction, term* xD_M and implies that

1 For a street location, mean sales volume equals (because $D_M = 0$)

$$\beta_0 + \beta_1 x + \beta_2(0) + \beta_3 x(0) = \beta_0 + \beta_1 x$$

2 For a mall location, mean sales volume equals (because $D_M = 1$)

$$\beta_0 + \beta_1 x + \beta_2(1) + \beta_3 x(1) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)x$$

As illustrated in Figure 15.16, if we use this model, then the straight lines relating mean sales volume to x for street and mall locations have *different y-intercepts* and *different slopes.* The *different slopes* imply that this model assumes *interaction* between x and store location. Such a model is appropriate if the relationship between mean sales volume and x depends on (that is, is different for) the street and mall store locations. In general, **interaction** exists between two independent variables if the relationship between (for example, the slope of the line relating) the mean value of the dependent variable and one of the independent variables depends upon the value (or level) of the other independent variable. Figure 15.17 gives the Excel output of a regression analysis of the sales volume data using the interaction model. Here D_M and xD_M are labeled as DM and XDM, respectively, on the output. The Excel output tells us that the p-value related to the significance of xD_M is .5886. This large p-value tells us that the interaction term is not significant. It follows that the no-interaction model on page 581 seems best.

FIGURE 15.17 Excel Output Using the Interaction Model $y = \beta_0 + \beta_1 x + \beta_2 D_M + \beta_3 x D_M + \varepsilon$

Regression Statistics	
Multiple R	0.9918
R Square	0.9836
Adjusted R Square	0.9755
Standard Error	7.7092
Observations	10

ANOVA	df	SS	MS	F	Significance F
Regression	3	21431.1861	7143.7287	120.1995988	9.531E-06
Residual	6	356.5933	59.4322		
Total	9	21787.7795			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	7.9004	19.3142	0.4090	0.6967	-39.3598	55.1606
Households	0.9207	0.1399	6.5792	0.0006	0.5783	1.2631
DM	42.7297	24.3812	1.7526	0.1302	-16.9289	102.3884
XDM	-0.0917	0.1606	-0.5712	0.5886	-0.4846	0.3012

Next, consider the Electronics World data for street, mall, and downtown locations given in Table 15.9 (page 583). In modeling these data, if we believe that interaction exists between the number of households in a store's area and store location, we might consider using the model

$$y = \beta_0 + \beta_1 x + \beta_2 D_M + \beta_3 D_D + \beta_4 x D_M + \beta_5 x D_D + \varepsilon$$

Similar to Figure 15.16, this model implies that the straight lines relating mean sales volume to x for the street, mall, and downtown locations have *different y-intercepts* and *different slopes.* If we perform a regression analysis of the sales volume data using this interaction model, we find that the p-values related to the significance of xD_M and xD_D are large (.5334 and .8132, respectively). Because these interaction terms are not significant, it seems best to employ the no-interaction model on page 582.

In general, if we wish to model the effect of a qualitative independent variable having a levels, we use $a - 1$ dummy variables. Consider the k^{th} such dummy variable D_k (k = one of the values $1, 2, \ldots, a - 1$). The parameter β_k multiplying D_k represents the mean difference between the level of y when the qualitative variable assumes level k and when it assumes the level a. For example, if we wish to compare the effects on sales, y, of four different types of advertising campaigns—television (T), radio (R), magazine (M), and mailed coupons (C)—we might employ the model

$$y = \beta_0 + \beta_1 D_T + \beta_2 D_R + \beta_3 D_M + \varepsilon$$

Because this model does not use a dummy variable to represent the mailed coupon advertising campaign, the parameter β_1 is the difference between mean sales when a television advertising campaign is used and mean sales when a mailed coupon advertising campaign is used. The interpretations of β_2 and β_3 follow similarly. As another example, if we wish to employ a confidence interval and a hypothesis test to compare the mall and downtown locations in the Electronics World example, we can use the model

$$y = \beta_0 + \beta_1 x + \beta_2 D_S + \beta_3 D_M + \varepsilon$$

Here the dummy variable D_M is as previously defined, and

$$D_S = \begin{cases} 1 & \text{if a store is in a street location} \\ 0 & \text{otherwise} \end{cases}$$

Because this model does not use a dummy variable to represent the downtown location, the parameter β_2 expresses the effect on mean sales of a street location compared to a downtown

location, and the parameter β_3 expresses the effect on mean sales of a mall location compared to a downtown location.

The Excel output of the least squares point estimates of the parameters of this model is as follows:

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	21.8415	8.5585	2.5520	0.0269	3.0044	40.6785
Households (x)	0.8686	0.0405	21.4520	2.52E-10	0.7795	0.9577
DummyStreet	-6.8638	4.7705	-1.4388	0.1780	-17.3635	3.6360
DummyMall	21.5100	4.0651	5.2914	0.0003	12.5628	30.4572

Because the least squares point estimate of β_3 is $b_3 = 21.51$, we estimate that for any given number of households in a store's area, the mean monthly sales volume in a mall location is \$21,510 higher than the mean monthly sales volume in a downtown location. The Excel output tells us that a 95 percent confidence interval for β_3 is [12.5628, 30.4572]. Therefore, we are 95 percent confident that for any given number of households in a store's area, the mean monthly sales volume in a mall location is between \$12,562.80 and \$30,457.20 greater than the mean monthly sales volume in a downtown location. The Excel output also shows that the t statistic and p-value for testing H_0: $\beta_3 = 0$ versus H_a: $\beta_3 \neq 0$ in this model are, respectively, 5.2914 and .0003. Therefore, we have very strong evidence that there is a difference between the mean monthly sales volumes in mall and downtown locations.

In some situations dummy variables represent the effects of unusual events or occurrences that may have an important impact on the dependent variable. For instance, suppose we wish to build a regression model relating quarterly sales of automobiles (y) to automobile prices (x_1), fuel prices (x_2), and personal income (x_3). If an autoworkers' strike occurred in a particular quarter that had a major impact on automobile sales, then we might define a dummy variable D_S to be equal to 1 if an autoworkers' strike occurs and to be equal to 0 otherwise. The least squares point estimate of the regression parameter multiplied by D_S would estimate the effect of the strike on mean auto sales. Finally, dummy variables can be used to model the impact of regularly occuring *seasonal* influences on time series data—for example, the impact of the hot summer months on soft drink sales. This is discussed in Supplementary Exercise 15.58 at the end of this chapter.

Exercises for Section 15.8

CONCEPTS

connect

15.26 What is a qualitative independent variable?

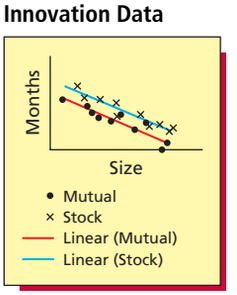
15.27 How do we use dummy variables to model the effects of a qualitative independent variable?

15.28 What does the parameter multiplied by a dummy variable express?

METHODS AND APPLICATIONS

15.29 Neter, Kutner, Nachtsheim, and Wasserman (1996) relate the speed, y, with which a particular insurance innovation is adopted to the size of the insurance firm, x, and the type of firm. The dependent variable y is measured by the number of months elapsed between the time the first firm adopted the innovation and the time the firm being considered adopted the innovation. The size of the firm, x, is measured by the total assets of the firm, and the type of firm—a qualitative independent variable—is either a mutual company or a stock company. The data in Table 15.10 on the next page are observed. DS InsInnov

Plot of the Insurance Innovation Data



a Discuss why the data plot in the page margin indicates that the model

$$y = \beta_0 + \beta_1 x + \beta_2 D_S + \varepsilon$$

might appropriately describe the observed data. Here D_S equals 1 if the firm is a stock company and 0 if the firm is a mutual company.

b The model of part *a* implies that the mean adoption time of an insurance innovation by mutual companies having an asset size x equals

$$\beta_0 + \beta_1 x + \beta_2(0) = \beta_0 + \beta_1 x$$

TABLE 15.10 The Insurance Innovation Data InsInnov

Firm	Number of Months Elapsed, y	Size of Firm (Millions of Dollars), x	Type of Firm	Firm	Number of Months Elapsed, y	Size of Firm (Millions of Dollars), x	Type of Firm
1	17	151	Mutual	11	28	164	Stock
2	26	92	Mutual	12	15	272	Stock
3	21	175	Mutual	13	11	295	Stock
4	30	31	Mutual	14	38	68	Stock
5	22	104	Mutual	15	31	85	Stock
6	0	277	Mutual	16	21	224	Stock
7	12	210	Mutual	17	20	166	Stock
8	19	120	Mutual	18	13	305	Stock
9	4	290	Mutual	19	30	124	Stock
10	16	238	Mutual	20	14	246	Stock

FIGURE 15.18 Excel Output of a Regression Analysis of the Insurance Innovation Data Using the Model $y = \beta_0 + \beta_1 x + \beta_2 D_S + \varepsilon$

Regression Statistics	
Multiple R	0.9461
R Square	0.8951
Adjusted R Square	0.8827
Standard Error	3.2211
Observations	20

ANOVA	df	SS	MS	F	Significance F
Regression	2	1,504.4133	752.2067	72.4971	4.77E-09
Residual	17	176.3867	10.3757		
Total	19	1,680.8			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	33.8741	1.8139	18.6751	9.15E-13	30.0472	37.7010
Size of Firm (x)	-0.1017	0.0089	-11.4430	2.07E-09	-0.1205	-0.0830
DummyStock	8.0555	1.4591	5.5208	3.74E-05	4.9770	11.1339

and that the mean adoption time by stock companies having an asset size x equals

$$\beta_0 + \beta_1 x + \beta_2(1) = \beta_0 + \beta_1 x + \beta_2$$

The difference between these two means equals the model parameter β_2. In your own words, interpret the practical meaning of β_2.

c Figure 15.18 presents the Excel output of a regression analysis of the insurance innovation data using the model of part *a*. **(1)** Using the output, test H_0: $\beta_2 = 0$ versus H_a: $\beta_2 \neq 0$ by setting $\alpha = .05$ and .01. **(2)** Interpret the practical meaning of the result of this test. **(3)** Also, use the computer output to find, report, and interpret a 95 percent confidence interval for β_2.

d If we add the interaction term xD_S to the model of part *a*, we find that the *p*-value related to this term is .9821. What does this imply?

15.30 THE FLORIDA POOL HOME CASE PoolHome

Table 3.11 (page 143) gives the selling price (Price, expressed in thousands of dollars), the square footage (SqrFt), the number of bathrooms (Bathrms), and the niceness rating (Niceness, expressed as an integer from 1 to 7) of 80 homes randomly selected from all homes sold in a Florida city during the last six months. (The random selections were made from homes having between 2,000 and 3,500 square feet.) Table 3.11 also gives values of the dummy variable Pool?, which equals 1 if a home has a pool and 0 otherwise. Figure 15.19 presents the MINITAB output of a regression analysis of these data using the model

$$\text{Price} = \beta_0 + \beta_1 \times \text{SqrFt} + \beta_2 \times \text{Bathrms} + \beta_3 \times \text{Niceness} + \beta_4 \times \text{Pool?} + \varepsilon$$

a Noting that β_4 is the effect on mean sales price of a home having a pool, find (on the output) a point estimate of this effect. If the average current purchase price of the pools in the sample is

FIGURE 15.19 MINITAB Output of a Regression Analysis of the Florida Pool Home Data Using the Model Price $= \beta_0 + \beta_1 \times$ SqrFt $+ \beta_2 \times$ Bathrms $+ \beta_3 \times$ Niceness $+ \beta_4 \times$ Pool? $+ \varepsilon$

```
The regression equation is
Price = 25.0 + 0.0526 SqrFt + 10.0 Bathrms + 10.0 Niceness + 25.9 Pool?

Predictor     Coef  SE Coef       T      P
Constant     24.98    16.63    1.50  0.137
SqrFt      0.05264  0.00659    7.98  0.000
Bathrms     10.043    3.729    2.69  0.009
Niceness    10.042   0.7915   12.69  0.000
Pool?       25.862    3.575    7.23  0.000

S = 13.532    R-Sq = 87.40%  R-Sq(adj) = 86.80%

Analysis of Variance
Source           DF       SS      MS       F      P
Regression        4    95665   23916  130.61  0.000
Residual Error   75    13734     183
Total            79   109399
```

$32,500, find a point estimate of the percentage of a pool's cost that a customer buying a pool can expect to recoup when selling his (or her) home.

b If we add various combinations of the interaction terms SqrFt × Pool?, Bathrooms × Pool?, and Niceness × Pool? to the above model, we find that the p-values related to these terms are greater than .05. What does this imply?

15.31 THE SUPERMARKET CASE BakeSale

The Tastee Bakery Company supplies a bakery product to many supermarkets in a metropolitan area. The company wishes to study the effect of the height of the shelf display employed by the supermarkets on monthly sales, y (measured in cases of 10 units each), for this product. Shelf display height has three levels—bottom (B), middle (M), and top (T). For each shelf display height, six supermarkets of equal sales potential will be randomly selected, and each supermarket will display the product using its assigned shelf height for a month. At the end of the month, sales of the bakery product at the 18 participating stores will be recorded. When the experiment is carried out, the data in Table 15.11 are obtained. Here we assume that the set of sales amounts for each display height is a sample that has been randomly selected from the population of all sales amounts that could be obtained (at supermarkets of the given sales potential) when using that display height. To compare the population mean sales amounts μ_B, μ_M, and μ_T that would be obtained by using the bottom, middle, and top display heights, we use the following dummy variable regression model:

$$y = \beta_B + \beta_M D_M + \beta_T D_T + \varepsilon$$

Here D_M equals 1 if a middle display height is used and 0 otherwise; D_T equals 1 if a top display height is used and 0 otherwise. Figure 15.20 on the next page presents the MINITAB output of a regression analysis of the bakery sales study data using this model.[1]

TABLE 15.11 Bakery Sales Study Data (Sales in Cases) BakeSale

Shelf Display Height		
Bottom (*B*)	Middle (*M*)	Top (*T*)
58.2	73.0	52.4
53.7	78.1	49.7
55.8	75.4	50.9
55.7	76.2	54.0
52.5	78.4	52.1
58.9	82.1	49.9

a By using the definitions of the dummy variables, show that

$$\mu_B = \beta_B \qquad \mu_M = \beta_B + \beta_M \qquad \mu_T = \beta_B + \beta_T$$

b Use the overall F statistic to test H_0: $\beta_M = \beta_T = 0$, or, equivalently, H_0: $\mu_B = \mu_M = \mu_T$. Interpret the practical meaning of the result of this test.

c Use your results in part a to do the following: **(1)** Show that

$$\mu_M - \mu_B = \beta_M \qquad \mu_T - \mu_B = \beta_T \qquad \mu_M - \mu_T = \beta_M - \beta_T$$

[1] In general, the regression approach of this exercise produces the same comparisons of several population means that are produced by **one-way analysis of variance** (see Section 12.2).

FIGURE 15.20 MINITAB Output of a Dummy Variable Regression Analysis of the Bakery Sales Data in Table 15.11

```
The regression equation is
Bakery Sales = 55.8 + 21.4 DMiddle - 4.30 DTop

Predictor     Coef  SE Coef      T      P
Constant    55.800    1.013  55.07  0.000
DMiddle     21.400    1.433  14.93  0.000
DTop        -4.300    1.433  -3.00  0.009

S = 2.48193   R-Sq = 96.1%   R-Sq(adj) = 95.6%

Analysis of Variance
Source           DF      SS      MS       F      P
Regression        2  2273.9  1136.9  184.57  0.000
Residual Error   15    92.4     6.2
Total            17  2366.3

Values of Predictors for New Obs    Predicted Values for New Observations
New Obs  DMiddle  DTop              New Obs     Fit  SE Fit       95% CI            95% PI
      1        1     0                    1  77.200   1.013  (75.040, 79.360)  (71.486, 82.914)
```

TABLE 15.12 Advertising Campaigns Used by Enterprise Industries

Fresh3

Sales Period	Advertising Campaign
1	*B*
2	*B*
3	*B*
4	*A*
5	*C*
6	*A*
7	*C*
8	*C*
9	*B*
10	*C*
11	*A*
12	*C*
13	*C*
14	*A*
15	*B*
16	*B*
17	*B*
18	*A*
19	*B*
20	*B*
21	*C*
22	*A*
23	*A*
24	*A*
25	*A*
26	*B*
27	*C*
28	*B*
29	*C*
30	*C*

(2) Use the least squares point estimates of the model parameters to find a point estimate of each of the three differences in means. **(3)** Find a 95 percent confidence interval for and test the significance of each of the first two differences in means. Interpret your results.

d Find **(1)** a point estimate of mean sales when using a middle display height, **(2)** a 95 percent confidence interval for mean sales when using a middle display height, and **(3)** a 95 percent prediction interval for sales at an individual supermarket that employs a middle display height. Hint: See the bottom of the MINITAB output in Figure 15.20.

e Consider the following alternative model

$$y = \beta_T + \beta_B D_B + \beta_M D_M + \varepsilon$$

Here D_B equals 1 if a bottom display height is used and 0 otherwise. The MINITAB output of the least squares point estimates of the parameters of this model is as follows:

Predictor	Coef	SE Coef	T	P
Constant	51.500	1.013	50.83	0.000
DBottom	4.300	1.433	3.00	0.009
DMiddle	25.700	1.433	17.94	0.000

Because β_M expresses the effect of the middle display height with respect to the effect of the top display height, β_M equals $\mu_M - \mu_T$. Use the MINITAB output to calculate a 95 percent confidence interval for and test the significance of $\mu_M - \mu_T$. Interpret your results.

15.32 THE FRESH DETERGENT CASE

Fresh3

Recall from Exercise 15.5 that Enterprise Industries has observed the historical data in Table 15.5 (page 563) concerning y (demand for Fresh liquid laundry detergent), x_1 (the price of Fresh), x_2 (the average industry price of competitors' similar detergents), and x_3 (Enterprise Industries' advertising expenditure for Fresh). To ultimately increase the demand for Fresh, Enterprise Industries' marketing department is comparing the effectiveness of three different advertising campaigns. These campaigns are denoted as campaigns *A*, *B*, and *C*. Campaign *A* consists entirely of television commercials, campaign *B* consists of a balanced mixture of television and radio commercials, and campaign *C* consists of a balanced mixture of television, radio, newspaper, and magazine ads. To conduct the study, Enterprise Industries has randomly selected one advertising campaign to be used in each of the 30 sales periods in Table 15.5. Although logic would indicate that each of campaigns *A*, *B*, and *C* should be used in 10 of the 30 sales periods, Enterprise Industries has made previous commitments to the advertising media involved in the study. As a result, campaigns *A*, *B*, and *C* were randomly assigned to, respectively, 9, 11, and 10 sales periods. Furthermore, advertising was done in only the first three weeks of each sales period, so that the carryover effect of the campaign used in a sales period to the next sales period would be minimized. Table 15.12 lists the campaigns used in the sales periods.

To compare the effectiveness of advertising campaigns *A*, *B*, and *C*, we define two dummy variables. Specifically, we define the dummy variable D_B to equal 1 if campaign *B* is used in a

FIGURE 15.21 Excel Output of a Dummy Variable Regression Model Analysis of the Fresh Demand Data

Regression Statistics	
Multiple R	0.9797
R Square	0.9597
Adjusted R Square	0.9513
Standard Error	0.1503
Observations	30

ANOVA	df	SS	MS	F	Significance F
Regression	5	12.9166	2.5833	114.3862	6.237E-16
Residual	24	0.5420	0.0226		
Total	29	13.4586			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	8.7154	1.5849	5.4989	1.1821E-05	5.4443	11.9866
Price (X1)	-2.7680	0.4144	-6.6790	6.5789E-07	-3.6234	-1.9127
Ind Price (X2)	1.6667	0.1913	8.7110	6.7695E-09	1.2718	2.0616
AdvExp (X3)	0.4927	0.0806	6.1100	2.6016E-06	0.3263	0.6592
DB	0.2695	0.0695	3.8804	0.0007	0.1262	0.4128
DC	0.4396	0.0703	6.2496	1.8506E-06	0.2944	0.5847

Predicted values for: Demand using an Excel add-in (MegaStat)

	95% Confidence Interval		95% Prediction Interval		
Predicted	lower	upper	lower	upper	Leverage
8.61621	8.51380	8.71862	8.28958	8.94285	0.109

sales period and 0 otherwise. Furthermore, we define the dummy variable D_C to equal 1 if campaign C is used in a sales period and 0 otherwise. Figure 15.21 presents the Excel add-in (MegaStat) output of a regression analysis of the Fresh demand data by using the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 D_B + \beta_5 D_C + \varepsilon$$

a In this model the parameter β_4 represents the effect on mean demand of advertising campaign B compared to advertising campaign A, and the parameter β_5 represents the effect on mean demand of advertising campaign C compared to advertising campaign A. **(1)** Use the regression output to find and report a point estimate of each of the above effects and to test the significance of each of the above effects. **(2)** Find and report a 95 percent confidence interval for each of the above effects. **(3)** Interpret your results.

b The prediction results at the bottom of the output correspond to a future period when Fresh's price will be $x_1 = 3.70$, the average price of similar detergents will be $x_2 = 3.90$, Fresh's advertising expenditure will be $x_3 = 6.50$, and advertising campaign C will be used. **(1)** Show (within rounding) how $\hat{y} = 8.61621$ is calculated. **(2)** Find, report, and interpret a 95 percent confidence interval for mean demand and a 95 percent prediction interval for an individual demand when $x_1 = 3.70$, $x_2 = 3.90$, $x_3 = 6.50$, and campaign C is used.

c Consider the alternative model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 D_A + \beta_5 D_C + \varepsilon$$

Here D_A equals 1 if advertising campaign A is used and equals 0 otherwise. Describe the effect represented by the regression parameter β_5.

d The Excel output of the least squares point estimates of the parameters of the model of part c is as follows.

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	8.9849	1.5971	5.6259	8.61E-06	5.6888	12.2811
Price (X1)	-2.7680	0.4144	-6.6790	6.58E-07	-3.6234	-1.9127
Ind Price (X2)	1.6667	0.1913	8.7110	6.77E-09	1.2718	2.0616
AdvExp (X3)	0.4927	0.0806	6.1100	2.60E-06	0.3263	0.6592
DA	-0.2695	0.0695	-3.8804	0.0007	-0.4128	-0.1262
DC	0.1701	0.0669	2.5429	0.0179	0.0320	0.3081

Use the Excel output to **(1)** test the significance of the effect represented by β_5 and **(2)** find a 95 percent confidence interval for β_5. **(3)** Interpret your results.

15.33 THE FRESH DETERGENT CASE DS Fresh3

If we add the independent variables x_3D_B and x_3D_C to the first model of Exercise 15.32, we find that the p-values related to these independent variables are .2797 and .2486. What does this imply?

LO15-8 Use squared and interaction variables.

15.9 Using Squared and Interaction Variables

Using squared variables One useful form of the multiple regression model is what we call the **quadratic regression model.** Assuming that we have obtained n observations—each consisting of an observed value of y and a corresponding value of x—the model is as follows:

The Quadratic Regression Model

The **quadratic regression model** relating y to x is

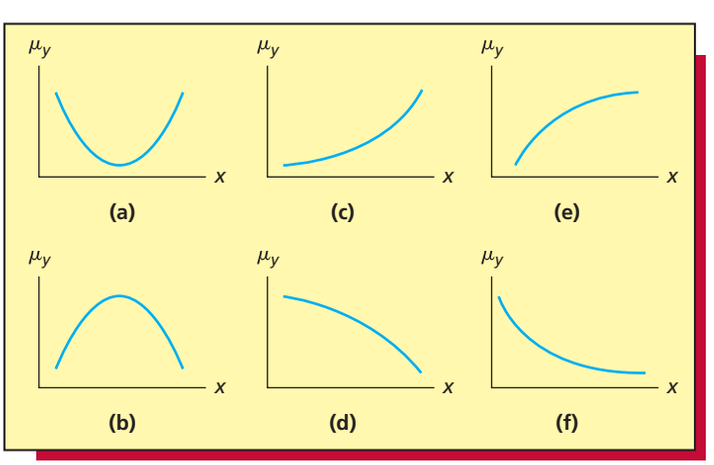
$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

where

1. $\beta_0 + \beta_1 x + \beta_2 x^2$ is μ_y, the mean value of the dependent variable y when the value of the independent variable is x.
2. β_0, β_1, and β_2 are (unknown) **regression parameters** relating the mean value of y to x.
3. ε is an error term that describes the effects on y of all factors other than x and x^2.

The quadratic equation $\mu_y = \beta_0 + \beta_1 x + \beta_2 x^2$ that relates μ_y to x is the equation of a **parabola.** Two parabolas are shown in Figure 15.22(a) and (b) and help to explain the meanings of the parameters β_0, β_1, and β_2. Here β_0 is the ***y*-intercept** of the parabola (the value of μ_y when $x = 0$). Furthermore, β_1 is the **shift parameter** of the parabola: the value of β_1 shifts the parabola to the left or right. Specifically, increasing the value of β_1 shifts the parabola to the left. Lastly, β_2 is the **rate of curvature** of the parabola. If β_2 is greater than 0, the parabola opens upward [see Figure 15.22(a)]. If β_2 is less than 0, the parabola opens downward [see Figure 15.22(b)]. If a scatter plot of y versus x shows points scattered around a parabola, or a part of a parabola [some typical parts are shown in Figure 15.22(c), (d), (e), and (f)], then the quadratic regression model might appropriately relate y to x.

FIGURE 15.22 **The Mean Value of the Dependent Variable Changing in a Quadratic Fashion as x Increases ($\mu_y = \beta_0 + \beta_1 x + \beta_2 x^2$)**



EXAMPLE 15.7 The Gasoline Additive Case: Maximizing Mileage

C

An oil company wishes to improve the gasoline mileage obtained by cars that use its regular unleaded gasoline. Company chemists suggest that an additive, ST-3000, be blended with the gasoline. In order to study the effects of this additive, mileage tests are carried out in a laboratory using test equipment that simulates driving under prescribed conditions. The amount of additive ST-3000 blended with the gasoline is varied, and the gasoline mileage for each test run is recorded. Table 15.13 gives the results of the test runs. Here the dependent variable y is gasoline mileage (in miles per gallon) and the independent variable x is the amount of additive ST-3000 used (measured as the number of units of additive added to each gallon of gasoline). One of the study's goals is to determine the number of units of additive that should be blended with the gasoline to maximize gasoline mileage. The company would also like to predict the maximum mileage that can be achieved using additive ST-3000.

TABLE 15.13 The Gasoline Additive Data

DS GasAdd

Additive Units (x)	Gasoline Mileage (y)
0	25.8
0	26.1
0	25.4
1	29.6
1	29.2
1	29.8
2	32.0
2	31.4
2	31.7
3	31.7
3	31.5
3	31.2
4	29.4
4	29.0
4	29.5

Figure 15.23(a) gives a scatter plot of y versus x. Because the scatter plot has the appearance of a quadratic curve (that is, part of a parabola), it seems reasonable to relate y to x by using the quadratic model

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

Figure 15.23(b) and (c) gives the MINITAB output of a regression analysis of the data using this quadratic model. Here the squared term x^2 is denoted as UnitsSq on the output. The MINITAB output tells us that the least squares point estimates of the model parameters are $b_0 = 25.7152$, $b_1 = 4.9762$, and $b_2 = -1.01905$. These estimates give us the least squares prediction equation

$$\hat{y} = 25.7152 + 4.9762x - 1.01905x^2$$

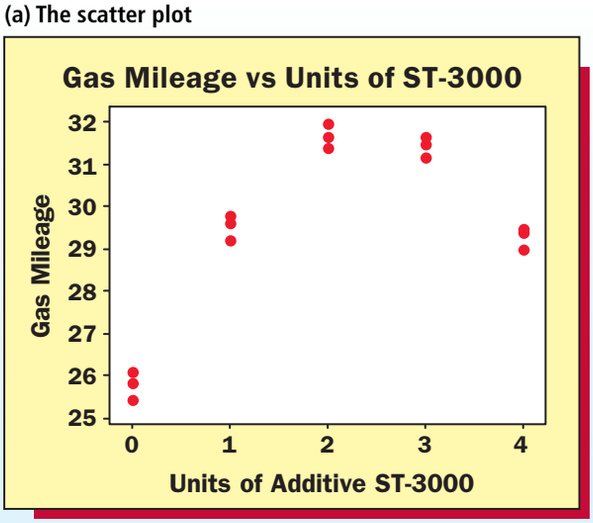
Intuitively, this is the equation of the best quadratic curve that can be fitted to the data plotted in Figure 15.23(a). The MINITAB output also tells us that the p-values related to x and x^2 are less than .001. This implies that we have very strong evidence that each of these model components is significant. The fact that x^2 seems significant confirms the graphical evidence that there is a quadratic relationship between y and x. Once we have such confirmation, we usually retain the linear term x in the model no matter what the size of its p-value. The reason is that geometrical considerations indicate that it is best to use both x and x^2 to model a quadratic relationship.

The oil company wishes to find the value of x that results in the highest predicted mileage. Using calculus, it can be shown that the value $x = 2.44$ maximizes predicted gas mileage. Therefore, the oil company can maximize predicted mileage by blending 2.44 units of additive

BI

FIGURE 15.23 MINITAB Scatter Plot and Quadratic Model Regression Analysis of the Gasoline Additive Data

(a) The scatter plot



(b) The least squares point estimates

```
The regression equation is
Mileage = 25.7 + 4.98 Units - 1.02 UnitsSq

Predictor      Coef  SE Coef       T      P
Constant    25.7152   0.1554  165.43  0.000
Units        4.9762   0.1841   27.02  0.000
UnitsSq     -1.0191   0.0441  -23.09  0.000

S = 0.286079   R-Sq = 98.6%   R-Sq(adj) = 98.3%

Analysis of Variance
Source           DF      SS      MS       F      P
Regression        2  67.915  33.958  414.92  0.000
Residual Error   12   0.982   0.082
Total            14  68.897
```

(c) Prediction

```
Values of Predictors for New Obs   Predicted Values for New Observations
New Obs  Unit   UnitsSq            New Obs      Fit  SE Fit        95% CI            95% PI
      1  2.44   5.9536                   1  31.7901  0.1111 (31.5481, 32.0322) (31.1215, 32.4588)
```

ST-3000 with each gallon of gasoline. This will result in a predicted gas mileage equal to

$$\hat{y} = 25.7152 + 4.9762(2.44) - 1.01905(2.44)^2$$
$$= 31.7901 \text{ miles per gallon}$$

This predicted mileage is the point estimate of the mean mileage that would be obtained by all gallons of the gasoline (when blended as just described) and is the point prediction of the mileage that would be obtained by an individual gallon of the gasoline. Note that $\hat{y} = 31.7901$ is given at the bottom of the MINITAB output in Figure 15.23(c). In addition, the MINITAB output tells us that a 95 percent confidence interval for the mean mileage that would be obtained by all gallons of the gasoline is [31.5481, 32.0322]. If the test equipment simulates driving conditions in a particular automobile, this confidence interval implies that an owner of the automobile can be 95 percent confident that he or she will average between 31.5481 mpg and 32.0322 mpg when using a very large number of gallons of the gasoline. The MINITAB output also tells us that a 95 percent prediction interval for the mileage that would be obtained by an individual gallon of the gasoline is [31.1215, 32.4588].

Using interaction variables Multiple regression models often contain **interaction variables.** We form an interaction variable by multiplying two independent variables together. For instance, if a regression model includes the independent variables x_1 and x_2, then we can form the interaction variable x_1x_2. It is appropriate to employ an interaction variable if the relationship between the dependent variable y and one of the independent variables depends upon the value of the other independent variable. In the following example we consider a multiple regression model that uses a linear variable, a squared variable, and an interaction variable.

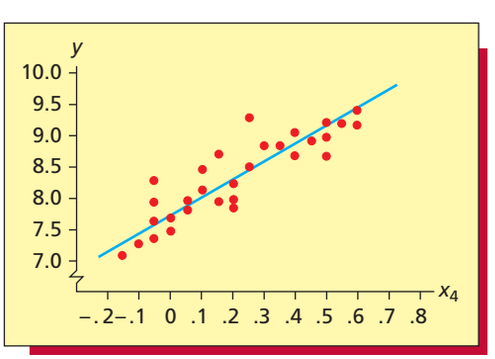
EXAMPLE 15.8 The Fresh Detergent Case: Predicting Demand

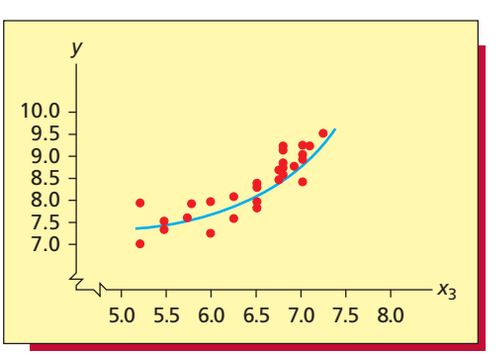
Enterprise Industries produces Fresh, a brand of liquid laundry detergent. In order to manage its inventory more effectively and make revenue projections, the company would like to better predict demand for Fresh. To develop a prediction model, the company has gathered data concerning demand for Fresh over the last 30 sales periods (each sales period is defined to be a four-week period). For each sales period, these data consist of recorded values of the demands (in hundreds of thousands of bottles) for the large size bottle of Fresh (y), the price (in dollars) of Fresh as offered by Enterprise Industries (x_1), the average industry price (in dollars) of competitors' similar detergents (x_2), and Enterprise Industries' advertising expenditure (in hundreds of thousands of dollars) to promote Fresh (x_3). The data have been given in Table 15.5 on page 563. To begin our analysis, suppose that Enterprise Industries believes on theoretical grounds that the difference between x_1 and x_2 (the *price difference* $x_4 = x_2 - x_1$) adequately describes the effects of x_1 and x_2 on y. That is, perhaps demand for Fresh depends more on how the price for Fresh compares to competitors' prices than it does on the absolute levels of the prices for Fresh and other competing detergents. This makes sense because most consumers must buy a certain amount of detergent no matter what the price might be. The values of the price difference $x_4 = x_2 - x_1$, calculated from the data in Table 15.5, are given in the page margin. Now, consider the following plots:

Price Difference, $x_4 = x_2 - x_1$ (Dollars)

−.05
.25
.60
0
.25
.20
.15
.05
−.15
.15
.20
.10
.40
.45
.35
.30
.50
.50
.40
−.05
−.05
−.10
.20
.10
.50
.60
−.05
0
.05
.55

Fresh2





Because the plot on the left shows a linear relationship between y and x_4, we should use x_4 to predict y. Because the plot on the right shows a quadratic relationship between y and x_3, we

FIGURE 15.24 Excel and Excel add-in (MegaStat) Output of a Regression Analysis of the Fresh Demand Data by Using the Interaction Model $y = \beta_0 + \beta_1 x_4 + \beta_2 x_3 + \beta_3 x_3^2 + \beta_4 x_4 x_3 + \varepsilon$

(a) The Excel output

Regression Statistics

Multiple R	0.9596	Adjusted R Square	0.9083	Observations	30
R Square	0.9209	Standard Error	0.2063		

ANOVA	df	SS	MS	F	Significance F
Regression	4	12.3942	3.0985	72.7771	2.11E-13
Residual	25	1.0644	0.0426		
Total	29	13.4586			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	29.1133	7.4832	3.8905	0.0007	13.7013	44.5252
PriceDif (x4)	11.1342	4.4459	2.5044	0.0192	1.9778	20.2906
AdvExp (x3)	-7.6080	2.4691	-3.0813	0.0050	-12.6932	-2.5228
x3sq	0.6712	0.2027	3.3115	0.0028	0.2538	1.0887
x4x3	-1.4777	0.6672	-2.2149	0.0361	-2.8518	-0.1037

(b) Prediction using an Excel add-in (MegaStat)

Predicted values for: Y

	95% Confidence Interval		95% Prediction Interval		
Predicted	lower	upper	lower	upper	Leverage
8.32725	8.21121	8.44329	7.88673	8.76777	0.075

should use x_3 and x_3^2 to predict y. Moreover, if x_4 and x_3 interact, then we should use the interaction variable x_4x_3 to predict y. This gives the model

$$y = \beta_0 + \beta_1 x_4 + \beta_2 x_3 + \beta_3 x_3^2 + \beta_4 x_4 x_3 + \varepsilon.$$

Figure 15.24 presents the Excel output obtained when this model is fit to the Fresh demand data. The p-values for testing the significance of the intercept and the independent variables are all below .05. Therefore, we have strong evidence that each of these terms should be included in the model. In particular, because the p-value related to x_4x_3 is .0361, we have strong evidence that x_4 and x_3 interact. We will examine the nature of this interaction in the discussion to come.

Suppose that Enterprise Industries wishes to predict demand for Fresh in a future sales period when the price difference will be \$.20 (that is, 20 cents) and when advertising expenditure will be \$650,000. Using the least squares point estimates in Figure 15.24(a), the point prediction is

$$\begin{aligned}\hat{y} &= 29.1133 + 11.1342(.20) - 7.6080(6.50) + 0.6712(6.50)^2 - 1.4777(.20)(6.50) \\ &= 8.32725 \text{ (832,725 bottles).}\end{aligned}$$

Figure 15.24(b) gives this point prediction along with the 95 percent confidence interval for mean demand and the 95 percent prediction interval for the actual demand when x_4 equals 0.20 and x_3 equals 6.50.

To investigate the nature of the interaction between x_3 and x_4, consider the prediction equation

$$\hat{y} = 29.1133 + 11.1342x_4 - 7.6080x_3 + 0.6712x_3^2 - 1.4777x_4x_3$$

obtained from the least squares point estimates in Figure 15.24(a). Also, consider the six combinations of price difference x_4 and advertising expenditure x_3 obtained by combining the x_4 values .10 and .30 with the x_3 values 6.0, 6.4, and 6.8. When we use the prediction equation to predict the demands for Fresh corresponding to these six combinations, we obtain the predicted demands ($\hat{y}$ values) shown in Figure 15.25(a) on the next page. (Note that we consider *two* x_4 values because there is a *linear* relationship between y and x_4, and we consider *three* x_3 values because there is a *quadratic* relationship between y and x_3.) Now:

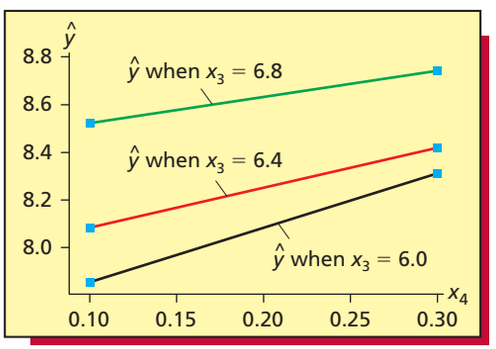
1 If we fix x_3 at 6.0 in Figure 15.25(a) and plot the corresponding $\hat{y}$ values 7.86 and 8.31 versus the x_4 values .10 and .30, we obtain the two squares connected by the lowest line in

FIGURE 15.25 Interaction between x_4 and x_3

(a) Predicted demands ($\hat{y}$ values)

x_3 \ x_4	.10	.30
6.0	7.86	8.31
6.4	8.08	8.42
6.8	8.52	8.74

(b) Plots of $\hat{y}$ versus x_4 for different x_3 values



(c) Plots of $\hat{y}$ versus x_3 for different x_4 values

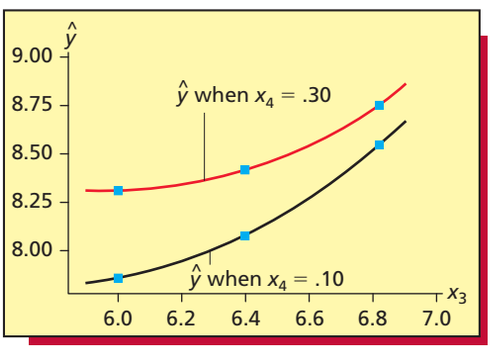


Figure 15.25(b). Similarly, if we fix x_3 at 6.4 and plot the corresponding $\hat{y}$ values 8.08 and 8.42 versus the x_4 values .10 and .30, we obtain the two squares connected by the middle line in Figure 15.25(b). Also, if we fix x_3 at 6.8 and plot the corresponding $\hat{y}$ values 8.52 and 8.74 versus the x_4 values .10 and .30, we obtain the two squares connected by the highest line in Figure 15.25(b). Examining the three lines relating $\hat{y}$ to x_4, we see that the slopes of these lines decrease as x_3 increases from 6.0 to 6.4 to 6.8. This says that as the price difference x_4 increases from .10 to .30 (that is, as Fresh becomes less expensive compared to its competitors), the *rate of increase* of predicted demand $\hat{y}$ is slower when advertising expenditure x_3 is higher than when advertising expenditure x_3 is lower. Moreover, this might be logical because it says that when a higher advertising expenditure makes more customers aware of Fresh's cleaning abilities and thus causes customer demand for Fresh to be higher, there is less opportunity for an increased price difference to increase demand for Fresh.

2 If we fix x_4 at .10 in Figure 15.25(a) and plot the corresponding $\hat{y}$ values 7.86, 8.08, and 8.52 versus the x_3 values 6.0, 6.4, and 6.8, we obtain the three squares connected by the lower quadratic curve in Figure 15.25(c). Similarly, if we fix x_4 at .30 and plot the corresponding $\hat{y}$ values 8.31, 8.42, and 8.74 versus the x_3 values 6.0, 6.4, and 6.8, we obtain the three squares connected by the higher quadratic curve in Figure 15.25(c). The non-parallel quadratic curves in Figure 15.25(c) say that as advertising expenditure x_3 increases from 6.0 to 6.8, the rate of increase of predicted demand $\hat{y}$ is slower when the price difference x_4 is larger (that is, $x_4 = .30$) than when the price difference x_4 is smaller (that is, $x_4 = .10$). Moreover, this might be logical because it says that when a larger price difference causes customer demand for Fresh to be higher, there is less opportunity for an increased advertising expenditure to increase demand for Fresh.

To summarize the nature of the interaction between x_4 and x_3, we might say that a higher value of each of these independent variables somewhat weakens the impact of the other independent variable on predicted demand. In Exercise 15.37 we will consider a situation where a higher value of each of two independent variables somewhat strengthens the impact of the other independent variable on the predicted value of the dependent variable. Moreover, if the p-value related to x_4x_3 in the Fresh detergent situation had been large and thus we had removed x_4x_3 from the model (that is, *no interaction*), then the plotted lines in Figure 15.25(b) would have been *parallel* and the plotted quadratic curves in Figure 15.25(c) would have been *parallel*. This would say that as each independent variable increases, predicted demand increases at the same rate whether the other independent variable is larger or smaller.

Exercises for Section 15.9

CONCEPTS

15.34 When does a scatter plot suggest the use of the quadratic regression model?

15.35 How do we model the interaction between two independent variables?

METHODS AND APPLICATIONS

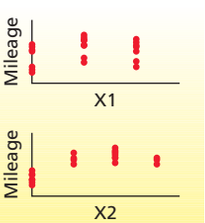
15.36 United Oil Company is attempting to develop a reasonably priced unleaded gasoline that will deliver higher gasoline mileages than can be achieved by its current unleaded gasolines. As part of its development process, United Oil wishes to study the effect of two independent variables—x_1, amount of gasoline additive RST (0, 1, or 2 units), and x_2, amount of gasoline additive XST (0, 1, 2, or 3 units), on gasoline mileage, y. Mileage tests are carried out using equipment that simulates driving under prescribed conditions. The combinations of x_1 and x_2 used in the experiment, along with the corresponding values of y, are given on the page margin.

a Discuss why the data plots given on the page margin indicate that the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_2^2 + \varepsilon$ might appropriately relate y to x_1 and x_2. If we fit this model to the data on the page margin, we find that the least squares point estimates of the model parameters and their associated p-values (given in parentheses) are $b_0 = 28.1589$ (<.001), $b_1 = 3.3133$ (<.001), $b_2 = -1.4111$ (<.001), $b_3 = 5.2752$ (<.001), and $b_4 = -1.3964$ (<.001). Moreover, consider the mean mileage obtained by all gallons of the gasoline when it is made with one unit of RST and two units of XST (a combination that the data on the page margin indicates would maximize mean mileage). A point estimate of and a 95 percent confidence interval for this mean mileage are 35.0261 and [34.4997, 35.5525]. Using the above model, show how the point estimate is calculated.

b If we add the independent variable x_1x_2 to the model in part(a), we find that the p-value related to x_1x_2 is .9777. What does this imply?

RST Units (x_1)	XST Units (x_2)	Gas Mileage (y, mpg)
0	0	27.4
0	0	28.0
0	0	28.6
1	0	29.6
1	0	30.6
2	0	28.6
2	0	29.8
0	1	32.0
0	1	33.0
1	1	33.3
1	1	34.5
0	2	32.3
0	2	33.5
1	2	34.4
1	2	35.0
1	2	35.6
2	2	33.3
2	2	34.0
2	2	34.7
1	3	33.4
2	3	32.0
2	3	33.0

DS UnitedOil



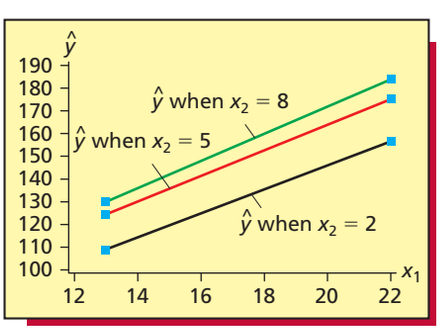
15.37 If we fit the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2 + \beta_4 x_1 x_2 + \varepsilon$ to the real estate sales price data in Table 15.4 (page 561), we find that the least squares point estimates of the model parameters and their associated p-values (given in parentheses) are $b_0 = 27.438$ (<.001), $b_1 = 5.0813$ (<.001), $b_2 = 7.2899$ (<.001), $b_3 = -.5311$ (.001), and $b_4 = .11473$ (.014).

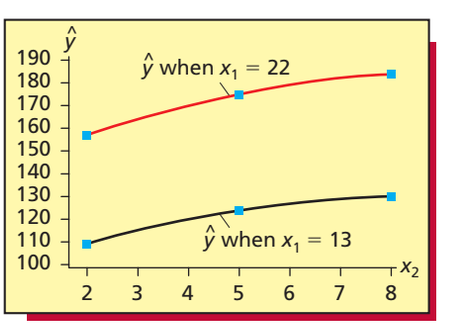
a A point prediction of and a 95 percent prediction interval for the sales price of a house having 2000 square feet ($x_1 = 20$) and a niceness rating of 8 ($x_2 = 8$) are 171.751 ($171,751) and [168.836, 174.665]. Using the above model, show how the point prediction is calculated.

b Below we give model predictions of sales prices of houses for six combinations of x_1 and x_2, along with plots of the predictions needed to interpret the interaction between x_1 and x_2. Carefully interpret this interaction. DS RealEst2

Predicted Sales Prices

x_2 \ x_1	13	22
2	108.933	156.730
5	124.124	175.019
8	129.756	183.748





15.10 Model Building and the Effects of Multicollinearity

LO15-9 Describe multicollinearity and build a multiple regression model.

Multicollinearity Recall the sales representative performance data in Figure 15.10 (page 579). These data consist of values of the dependent variable y (SALES) and of the independent variables x_1 (TIME), x_2 (MKTPOTEN), x_3 (ADVER), x_4 (MKTSHARE), and x_5 (CHANGE). The complete sales representative performance data analyzed by Cravens, Woodruff, and Stomper (1972) consist of the data presented in Figure 15.10 and data concerning three additional independent variables. These three additional variables are x_6 = number of accounts handled by the representative (ACCTS); x_7 = average workload per account, measured by using a weighting

based on the sizes of the orders by the accounts and other workload-related criteria (WKLOAD); and x_8 = an aggregate rating on eight dimensions of the representative's performance, made by a sales manager and expressed on a 1–7 scale (RATING).

TABLE 15.14 Values of ACCTS, WKLOAD, and RATING

SalePerf2

Accounts, x_6	Workload, x_7	Rating, x_8
74.86	15.05	4.9
107.32	19.97	5.1
96.75	17.34	2.9
195.12	13.40	3.4
180.44	17.64	4.6
104.88	16.22	4.5
256.10	18.80	4.6
126.83	19.86	2.3
203.25	17.42	4.9
119.51	21.41	2.8
116.26	16.32	3.1
142.28	14.51	4.2
89.43	19.35	4.3
84.55	20.02	4.2
119.51	15.26	5.5
80.49	15.87	3.6
136.58	7.81	3.4
78.86	16.00	4.2
136.58	17.44	3.6
138.21	17.98	3.1
75.61	20.99	1.6
102.44	21.66	3.4
76.42	21.46	2.7
136.58	24.78	2.8
88.62	24.96	3.9

Table 15.14 gives the observed values of x_6, x_7, and x_8, and Figure 15.26(a) presents the MINITAB output of a **correlation matrix** for the sales representative performance data. Examining the first column of this matrix, we see that the simple correlation coefficient between SALES and WKLOAD is −.117 and that the p-value for testing the significance of the relationship between SALES and WKLOAD is .577. This indicates that there is little or no relationship between SALES and WKLOAD. However, the simple correlation coefficients between SALES and the other seven independent variables range from .402 to .754, with associated p-values ranging from .046 to .000. This indicates the existence of potentially useful relationships between SALES and these seven independent variables.

While simple correlation coefficients (and scatter plots) give us a preliminary understanding of the data, they cannot be relied upon alone to tell us which independent variables are significantly related to the dependent variable. One reason for this is a condition called *multicollinearity.* **Multicollinearity** is said to exist among the independent variables in a regression situation if these independent variables are related to or dependent upon each other. One way to investigate multicollinearity is to examine the correlation matrix. To understand this, note that all of the simple correlation coefficients not located in the first column of this matrix measure the **simple correlations between the independent variables.** For example, the simple correlation coefficient between ACCTS and TIME is .758, which says that the ACCTS values increase as the TIME values increase. Such a relationship makes sense because it is logical that the longer a sales representative has been with the company, the more accounts he or she handles. Statisticians often regard multicollinearity in a data set to be severe if at least one simple correlation coefficient between the independent variables is at least .9. Because the largest such simple correlation coefficient in Figure 15.26(a) is .758, this is not true for the sales representative performance data. Note, however, that even moderate multicollinearity can be a potential problem. This will be demonstrated later using the sales representative performance data.

Another way to measure multicollinearity is to use **variance inflation factors.** Consider a regression model relating a dependent variable y to a set of independent variables $x_1, \ldots, x_{j-1}, x_j, x_{j+1}, \ldots, x_k$. The **variance inflation factor for the independent variable x_j** in this set is denoted $\boldsymbol{VIF_j}$ and is defined by the equation

$$VIF_j = \frac{1}{1 - R_j^2}$$

where R_j^2 is the multiple coefficient of determination for the regression model that relates x_j to all the other independent variables $x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k$ in the set. For example, Figure 15.26(b) gives the MINITAB output of the t statistics, p-values, and variance inflation factors

FIGURE 15.26 MINITAB Output of the Correlation Matrix and a Regression Analysis for the Sales Representative Performance Situation

(a) The correlation matrix

	Sales	Time	MktPoten	Adver	MktShare	Change	Accts	WkLoad
Time	0.623							
	0.001							
MktPoten	0.598	0.454						
	0.002	0.023						
Adver	0.596	0.249	0.174					
	0.002	0.230	0.405					
MktShare	0.484	0.106	-0.211	0.264				
	0.014	0.613	0.312	0.201				
Change	0.489	0.251	0.268	0.377	0.085			
	0.013	0.225	0.195	0.064	0.685			
Accts	0.754	0.758	0.479	0.200	0.403	0.327		
	0.000	0.000	0.016	0.338	0.046	0.110		
WkLoad	-0.117	-0.179	-0.259	-0.272	0.349	-0.288	-0.199	
	0.577	0.391	0.212	0.188	0.087	0.163	0.341	
Rating	0.402	0.101	0.359	0.411	-0.024	0.549	0.229	-0.277
	0.046	0.631	0.078	0.041	0.911	0.004	0.272	0.180

Cell contents: Pearson correlation
P-Value

(b) The *t* statistics, *p*-values, and variance inflation factors for the eight-independent-variables model.

Predictor	Coef	SE Coef	T	P	VIF
Constant	-1507.8	778.6	-1.94	0.071	
Time	2.010	1.931	1.04	0.313	3.343
MktPoten	0.037205	0.008202	4.54	0.000	1.978
Adver	0.15099	0.04711	3.21	0.006	1.910
MktShare	199.02	67.03	2.97	0.009	3.236
Change	290.9	186.8	1.56	0.139	1.602
Accts	5.551	4.776	1.16	0.262	5.639
WkLoad	19.79	33.68	0.59	0.565	1.818
Rating	8.2	128.5	0.06	0.950	1.809

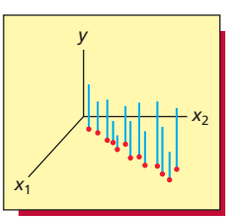
for the sales representative performance model that relates y to all eight independent variables. The largest variance inflation factor is $VIF_6 = 5.639$. To calculate VIF_6, MINITAB first calculates the multiple coefficient of determination for the regression model that relates x_6 to x_1, x_2, x_3, x_4, x_5, x_7, and x_8 to be $R_6^2 = .822673$. It then follows that

$$VIF_6 = \frac{1}{1 - R_6^2} = \frac{1}{1 - .822673} = 5.639$$

In general, if $R_j^2 = 0$, which says that x_j is not related to the other independent variables, then the variance inflation factor VIF_j equals 1. On the other hand, if $R_j^2 > 0$, which says that x_j is related to the other independent variables, then $(1 - R_j^2)$ is less than 1, making VIF_j greater than 1. Generally, the multicollinearity between independent variables is considered **(1)** severe if the largest variance inflation factor is greater than 10 and **(2)** moderately strong if the largest variance inflation factor is greater than 5. Moreover, if the mean of the variance inflation factors is substantially greater than 1 (sometimes a difficult criterion to assess), multicollinearity might be problematic. In the sales representative performance model, the largest variance inflation factor, $VIF_6 = 5.639$, is greater than 5. Therefore, we might classify the multicollinearity as being moderately strong.

The reason that VIF_j is called the variance inflation factor is that it can be shown that, when VIF_j is greater than 1, then the standard deviation σ_{b_j} of the population of all possible values of the least squares point estimate b_j is likely to be inflated beyond its value when $R_j^2 = 0$. If σ_{b_j} is greatly inflated, two slightly different samples of values of the dependent variable can yield two substantially different values of b_j. To intuitively understand why strong multicollinearity can significantly affect the least squares point estimates, consider the so-called "picket fence" display in the page margin. This figure depicts two independent variables (x_1 and x_2) exhibiting strong multicollinearity (note that as x_1 increases, x_2 increases). The heights of the pickets on the fence represent the y observations. If we assume that the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ adequately describes this data, then calculating the least squares point estimates amounts to fitting a plane to the points on the top of the picket fence. Clearly, this plane would be quite unstable. That is, a slightly different height of one of the pickets (a slightly different y value) could cause the slant of the fitted plane (and the least squares point estimates that determine this slant) to change radically. It follows that, when strong multicollinearity exists, sampling variation can result in least squares point estimates that differ substantially from the true values of the regression parameters. In fact, some of the least squares point estimates may have a sign (positive or negative) that differs from the sign of the true value of the parameter (we will see an example of this in the exercises). Therefore, when strong multicollinearity exists, it is dangerous to interpret the individual least squares point estimates.

The Picket Fence Display



The most important problem caused by multicollinearity is that, even when multicollinearity is not severe, it can hinder our ability to use the t statistics and related p-values to assess the importance of the independent variables. Recall that we can reject H_0: $\beta_j = 0$ in favor of H_a: $\beta_j \neq 0$ at level of significance α if and only if the absolute value of the corresponding t statistic is greater than $t_{\alpha/2}$ based on $n - (k + 1)$ degrees of freedom, or, equivalently, if and only if the related p-value is less than α. Thus the larger (in absolute value) the t statistic is and the smaller the p-value is, the stronger is the evidence that we should reject H_0: $\beta_j = 0$ and the stronger is the evidence that the independent variable x_j is significant. When multicollinearity exists, the sizes of the t statistic and of the related p-value **measure the additional importance of the independent variable x_j over the combined importance of the other independent variables in the regression model.** Because two or more correlated independent variables contribute redundant information, multicollinearity often causes the t statistics obtained by relating a dependent variable to a set of correlated independent variables to be smaller (in absolute value) than the t statistics that would be obtained if separate regression analyses were run, where each separate regression analysis relates the dependent variable to a smaller set (for example, only one) of the correlated independent variables. Thus multicollinearity can cause some of the correlated independent variables to appear less important—in terms of having small absolute t statistics and large p-values—than they really are. Another way to understand this is to note that because multicollinearity inflates σ_{b_j}, it inflates the point estimate s_{b_j} of σ_{b_j}. Because $t = b_j / s_{b_j}$, an inflated value of s_{b_j} can (depending on the

size of b_j) cause t to be small (and the related p-value to be large). This would suggest that x_j is not significant even though x_j may really be important.

For example, Figure 15.26(b) on page 598 tells us that when we perform a regression analysis of the sales representative performance data using a model that relates y to all eight independent variables, the p-values related to TIME, MKTPOTEN, ADVER, MKTSHARE, CHANGE, ACCTS, WKLOAD, and RATING are, respectively, .313, .000, .006, .009, .139, .262, .565, and .950. By contrast, recall from Figure 15.11 (page 580) that when we perform a regression analysis of the sales representative performance data using a model that relates y to the first five independent variables, the p-values related to TIME, MKTPOTEN, ADVER, MKTSHARE, and CHANGE are, respectively, .0065, .0001, .0025, .0001, and .0530. Note that TIME (p-value = .0065) is significant at the .01 level and CHANGE (p-value = .0530) is significant at the .06 level in the five independent variable model. However, when we consider the model that uses all eight independent variables, neither TIME (p-value = .313) nor CHANGE (p-value = .139) is significant at the .10 level. The reason that TIME and CHANGE seem more significant in the five independent variable model is that, because this model uses fewer variables, TIME and CHANGE contribute less overlapping information and thus have more additional importance in this model.

Comparing regression models on the basis of R^2, s, adjusted R^2, prediction interval length, and the C statistic We have seen that when multicollinearity exists in a model, the p-value associated with an independent variable in the model measures the additional importance of the variable over the combined importance of the other variables in the model. Therefore, it can be difficult to use the p-values to determine which variables to retain and which variables to remove from a model. The implication of this is that we need to evaluate more than the *additional importance* of each independent variable in a regression model. We also need to evaluate how well the independent variables *work together* to accurately describe, predict, and control the dependent variable. One way to do this is to determine if the *overall* model gives a high R^2 and $\overline{R}^2$, a small s, and short prediction intervals.

It can be proved that **adding any independent variable to a regression model, even an unimportant independent variable, will decrease the unexplained variation and will increase the explained variation.** Therefore, because the total variation $\Sigma(y_i - \bar{y})^2$ depends only on the observed y values and thus remains unchanged when we add an independent variable to a regression model, it follows that **adding any independent variable to a regression model will increase R^2,** which equals the explained variation divided by the total variation. This implies that R^2 cannot tell us (by decreasing) that adding an independent variable is undesirable. That is, although we wish to obtain a model with a large R^2, there are better criteria than R^2 that can be used to *compare* regression models.

One better criterion is the standard error $s = \sqrt{SSE/[n - (k + 1)]}$. When we add an independent variable to a regression model, the number of model parameters $(k + 1)$ increases by one, and thus the number of degrees of freedom $n - (k + 1)$ decreases by one. If the decrease in $n - (k + 1)$, which is used in the denominator to calculate s, is proportionally more than the decrease in SSE (the unexplained variation) that is caused by adding the independent variable to the model, then s will increase. **If s increases, this tells us that we should not add the independent variable to the model,** because the new model would give longer prediction intervals and thus less accurate predictions. If s decreases, the new model is likely to give shorter prediction intervals but (as we will see) is not guaranteed to do so. Thus, it can be useful to compare the lengths of prediction intervals for different models. Also, it can be shown that the standard error s decreases if and only if $\overline{R}^2$ (adjusted R^2) increases. It follows that, if we are comparing regression models, the model that gives the smallest s gives the largest $\overline{R}^2$.

EXAMPLE 15.9 The Sales Representative Case: Model Comparisons

Figure 15.27 gives the MINITAB output resulting from calculating R^2, $\overline{R}^2$, and s for **all possible regression models** based on all possible combinations of the eight independent variables in the sales representative performance situation (the values of C-p on the output will be explained after we complete this example). The MINITAB output gives the two best models of each size in terms

FIGURE 15.27 MINITAB Output of the Best Two Sales Representative Performance Regression Models of Each Size

Vars	R-Sq	R-Sq(adj)	Mallows C-p	S	Time	MktPoten	Adver	MktShare	Change	Accts	Wkload	Rating
1	56.8	55.0	67.6	881.09						X		
1	38.8	36.1	104.6	1049.3	X							
2	77.5	75.5	27.2	650.39			X			X		
2	74.6	72.3	33.1	691.10		X		X				
3	84.9	82.7	14.0	545.51		X	X	X				
3	82.8	80.3	18.4	582.64		X	X			X		
4	90.0	88.1	5.4	453.84		X	X	X		X		
4	89.6	87.5	6.4	463.95	X	X	X	X				
5	91.5	89.3	4.4	430.23	X	X	X	X	X			
5	91.2	88.9	5.0	436.75		X	X	X	X	X		
6	92.0	89.4	5.4	428.00	X	X	X	X	X	X		
6	91.6	88.9	6.1	438.20		X	X	X	X	X	X	
7	92.2	89.0	7.0	435.67	X	X	X	X	X	X	X	
7	92.0	88.8	7.3	440.30	X	X	X	X	X	X		X
8	92.2	88.3	9.0	449.03	X	X	X	X	X	X	X	X

of s and $\overline{R}^2$—the two best one-variable models, the two best two-variable models, the two best three-variable models, and so on. Examining the output, we see that the three models having the smallest values of s and largest values of $\overline{R}^2$ are

1 The six-variable model that contains

TIME, MKTPOTEN, ADVER, MKTSHARE, CHANGE, ACCTS

and has $s = 428.00$ and $\overline{R}^2 = 89.4$; we refer to this model as Model 1.

2 The five-variable model that contains

TIME, MKTPOTEN, ADVER, MKTSHARE, CHANGE

and has $s = 430.23$ and $\overline{R}^2 = 89.3$; we refer to this model as Model 2.

3 The seven-variable model that contains

TIME, MKTPOTEN, ADVER, MKTSHARE, CHANGE, ACCTS, WKLOAD

and has $s = 435.67$ and $\overline{R}^2 = 89.0$; we refer to this model as Model 3.

To see that s can increase when we add an independent variable to a regression model, note that s increases from 428.00 to 435.67 when we add WKLOAD to Model 1 to form Model 3. In this case, although it can be verified that adding WKLOAD decreases the unexplained variation from 3,297,279.3342 to 3,226,756.2751, this decrease has not been enough to offset the change in the denominator of $s^2 = SSE/[n - (k + 1)]$, which decreases from $25 - 7 = 18$ to $25 - 8 = 17$. To see that prediction interval lengths might increase even though s decreases, consider adding ACCTS to Model 2 to form Model 1. This decreases s from 430.23 to 428.00. However, consider a questionable sales representative for whom TIME = 85.42, MKTPOTEN = 35,182.73, ADVER = 7,281.65, MKTSHARE = 9.64, CHANGE = .28, and ACCTS = 120.61. The 95 percent prediction interval given by Model 2 for sales corresponding to this combination of values of the independent variables is [3,233.59, 5,129.89] and has length $5{,}129.89 - 3{,}233.59 = 1896.3$. The 95 percent prediction interval given by Model 1 for such sales is [3,193.86, 5,093.14] and has length $5{,}093.14 - 3{,}193.86 = 1{,}899.28$. In other words, even though Model 2 has a slightly larger s, it gives a slightly shorter prediction interval. (For those who have studied the formula for

a prediction interval, $[\hat{y} \pm t_{\alpha/2} s\sqrt{1 + \text{distance value}}]$, Model 2 gives a slightly shorter 95 percent prediction interval because it uses one less variable and thus can be verified to give slightly smaller values of $t_{.025}$ and the distance value.) In addition, the extra independent variable ACCTS in Model 1 can be verified to have a p-value of .2881. Therefore, we conclude that Model 2 is better than Model 1 and is, in fact, the "best" sales representative performance model (using only linear independent variables).

Another quantity that can be used for comparing regression models is called the ***C* statistic** (also often called the ***C_p* statistic**). To show how to calculate the C statistic, suppose that we wish to choose an appropriate set of independent variables from p potential independent variables. We first calculate the mean square error, which we denote as s_p^2, for the model using all p potential independent variables. Then, if SSE denotes the unexplained variation for another particular model that has k independent variables, it follows that the C statistic for this model is

$$C = \frac{SSE}{s_p^2} - [n - 2(k + 1)]$$

For example, consider the sales representative case. It can be verified that the mean square error for the model using all $p = 8$ independent variables is 201,621.21 and that the SSE for the model using the first $k = 5$ independent variables (Model 2 in the previous example) is 3,516,812.7933. It follows that the C statistic for this latter model is

$$C = \frac{3{,}516{,}812.7933}{201{,}621.21} - [25 - 2(5 + 1)] = 4.4$$

Because the C statistic for a given model is a function of the model's SSE, and because we want SSE to be small, **we want *C* to be small.** Although adding an unimportant independent variable to a regression model will decrease SSE, adding such a variable can increase C. This can happen when the decrease in SSE caused by the addition of the extra independent variable is not enough to offset the decrease in $n - 2(k + 1)$ caused by the addition of the extra independent variable (which increases k by 1). It should be noted that although adding an unimportant independent variable to a regression model can increase both s^2 and C, there is no exact relationship between s^2 and C.

While we want C to be small, it can be shown from the theory behind the C statistic that **we also wish to find a model for which the *C* statistic roughly equals $k + 1$,** the number of parameters in the model. **If a model has a *C* statistic substantially greater than $k + 1$, it can be shown that this model has substantial *bias* and is undesirable.** Thus, although we want to find a model for which C is as small as possible, if C for such a model is substantially greater than $k + 1$, we may prefer to choose a different model for which C is slightly larger and more nearly equal to the number of parameters in that (different) model. **If a particular model has a small value of *C* and *C* for this model is less than $k + 1$, then the model should be considered desirable.** Finally, it should be noted that for the model that includes all p potential independent variables (and thus utilizes $p + 1$ parameters), it can be shown that $C = p + 1$.

If we examine Figure 15.27, we see that Model 2 of the previous example has the smallest C statistic. The C statistic for this model (denoted C-p on the MINITAB output) equals 4.4. Because $C = 4.4$ is less than $k + 1 = 6$, the model is not biased. Therefore, this model should be considered best with respect to the C statistic.

Stepwise regression In some situations it is useful to employ an **iterative model selection procedure,** where at each step a single independent variable is added to or deleted from a regression model, and a new regression model is evaluated. We now discuss one such procedure—**stepwise regression.**

Stepwise regression begins by considering all of the one-independent-variable models and choosing the model for which the p-value related to the independent variable in the model is the smallest. If this p-value is less than α_{entry}, an α value for "entering" a variable, the independent variable is the first variable entered into the stepwise regression model and stepwise regression continues. Stepwise regression then considers the remaining independent variables not in the stepwise model and chooses the independent variable which, when paired with the first independent variable entered, has the smallest p-value. If this p-value is less than α_{entry}, the new variable is entered into the stepwise model. Moreover, the stepwise procedure checks to see if the p-value

FIGURE 15.28 The MINITAB Output of the Sales Representative Case Stepwise Regression

($\alpha_{entry} = \alpha_{stay} = .10$)

Step	1	2	3	4
Constant	709.32	50.30	-327.23	-1441.94
Accts	21.7	19.0	15.6	9.2
T-Value	5.50	6.41	5.19	3.22
P-Value	0.000	0.000	0.000	0.004
Adver		0.227	0.216	0.175
T-Value		4.50	4.77	4.74
P-Value		0.000	0.000	0.000
MktPoten			0.0219	0.0382
T-Value			2.53	4.79
P-Value			0.019	0.000
MktShare				190
T-Value				3.82
P-Value				0.001
S	881	650	583	454
R-Sq	56.85	77.51	82.77	90.04
R-Sq(adj)	54.97	75.47	80.31	88.05
Mallows C-p	67.6	27.2	18.4	5.4

TABLE 15.15 Squared Variables and Interaction Variables in the Sales Representative Case

SQT	=	TIME*TIME
SQMP	=	MKTPOTEN*MKTPOTEN
SQA	=	ADVER*ADVER
SQMS	=	MKTSHARE*MKTSHARE
SQC	=	CHANGE*CHANGE
TMP	=	TIME*MKTPOTEN
TA	=	TIME*ADVER
TMS	=	TIME*MKTSHARE
TC	=	TIME*CHANGE
MPA	=	MKTPOTEN*ADVER
MPMS	=	MKTPOTEN*MKTSHARE
MPC	=	MKTPOTEN*CHANGE
AMS	=	ADVER*MKTSHARE
AC	=	ADVER*CHANGE
MSC	=	MKTSHARE*CHANGE

related to the first variable entered into the stepwise model is less than α_{stay}, an α value for allowing a variable to stay in the stepwise model. This is done because multicollinearity could have changed the p-value of the first variable entered into the stepwise model. The stepwise procedure continues this process and concludes when no new independent variable can be entered into the stepwise model. It is common practice to set both α_{entry} and α_{stay} equal to .05 or .10.

For example, again consider the sales representative performance data. We let x_1, x_2, x_3, x_4, x_5, x_6, x_7, and x_8 be the eight potential independent variables employed in the stepwise procedure. Figure 15.28 gives the MINITAB output of the stepwise regression employing these independent variables where both α_{entry} and α_{stay} have been set equal to .10. The stepwise procedure (1) adds ACCTS (x_6) on the first step; (2) adds ADVER (x_3) and retains ACCTS on the second step; (3) adds MKTPOTEN (x_2) and retains ACCTS and ADVER on the third step; and (4) adds MKTSHARE (x_4) and retains ACCTS, ADVER, and MKTPOTEN on the fourth step. The procedure terminates after step 4 when no more independent variables can be added. Therefore, the stepwise procedure arrives at the model that utilizes x_2, x_3, x_4, and x_6. Note that this model is not the model using x_1, x_2, x_3, x_4, and x_5 that was obtained by evaluating all possible regression models and that has the smallest C statistic of 4.4. In general, stepwise regression can miss finding the best regression model but is useful in **data mining,** where a massive number of independent variables exist and all possible regression models cannot be evaluated.

Some advanced model-building methods: Using squared and interaction variables and the partial *F*-test We have concluded that perhaps the best sales representative performance model using only linear independent variables is the model using TIME, MKTPOTEN, ADVER, MKTSHARE, and CHANGE. We have also seen in Section 15.9 that using squared variables (which model quadratic curvature) and interaction variables can improve a regression model. In Table 15.15 we present the five squared variables and the ten (pairwise) interaction variables that can be formed using TIME, MKTPOTEN, ADVER, MKTSHARE, and CHANGE. Consider having MINITAB evaluate all possible models involving these squared and interaction variables, where the five linear variables are included in each possible model. If we have MINITAB do this and find the best model of each size in terms of s, we obtain the output in Figure 15.29 on the next page. (Note that we do not include values of the C statistic on the output because it can be shown that this statistic can give misleading results when using squared and interaction variables.) Examining the output, we see that the model that uses 12 squared and interaction variables (or a total of

FIGURE 15.29 MINITAB Comparisons Using Additional Squared and Interaction Variables

Note: all models include the 5 linear variables

Total Vars	Squared and Interaction Vars	R-Sq	R-Sq(adj)	S	S Q T	S Q P	S Q A	S Q S	S Q C	S T P	S T A	T S C	T M A	M P C	M P S	M P C	A M S	A C	M S C
6	1	94.2	92.2	365.87								X							
7	2	95.8	94.1	318.19	X							X							
8	3	96.5	94.7	301.61	X							X		X					
9	4	97.0	95.3	285.53	X						X	X			X				
10	5	97.5	95.7	272.05	X						X	X			X		X		
11	6	98.1	96.5	244.00	X		X				X	X			X			X	
12	7	98.7	97.4	210.70	X	X					X	X			X		X	X	
13	8	99.0	97.8	193.95	X	X			X		X	X			X		X	X	
14	9	99.2	98.0	185.44	X	X		X			X	X			X		X	X	X
15	10	99.3	98.2	175.70	X	X		X			X	X			X	X	X	X	X
16	11	99.4	98.2	177.09	X	X		X		X	X	X			X	X	X	X	X
17	12	99.5	98.2	174.60	X	X		X	X	X	X	X			X	X	X	X	X
18	13	99.5	98.1	183.22	X	X	X		X	X	X	X	X		X	X	X	X	X
19	14	99.6	97.9	189.77	X	X		X	X	X	X	X	X	X	X	X	X	X	X
20	15	99.6	97.4	210.78	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

17 variables, including the 5 linear variables) has the smallest s (= 174.6) of any model. Unfortunately, although we would like to test the significance of the independent variables in this model, extreme multicollinearity (relationships between the independent variables) exists when using squared and interaction variables. Thus, the usual t tests for assessing the significance of individual independent variables might not be reliable. As an alternative, we will use a *partial F-test.* Specifically, considering the model with the smallest s of 174.6 and a total of 17 variables to be a *complete model,* we will use this test to assess whether at least one variable in the *subset* of 12 squared and interaction variables in this model is significant.

The Partial *F*-Test: An *F*-Test for a Portion of a Regression Model

Suppose that the regression assumptions hold, and consider a **complete model** that uses k independent variables. To assess whether at least one of the independent variables in a subset of k^* independent variables in this model is significant, we test the null hypothesis

H_0: All of the β_j coefficients corresponding to the independent variables in the subset are zero

which says that none of the independent variables in the subset are significant. We test H_0 versus

H_a: At least one of the β_j coefficients corresponding to the independent variables in the subset is not equal to zero

which says that at least one of the independent variables in the subset is significant. Let SSE_C denote the unexplained variation for the complete model, and let SSE_R denote the unexplained variation for the **reduced model** that uses all k independent variables **except** for the k^* independent variables in the subset. Also, define

$$F(\text{partial}) = \frac{(SSE_R - SSE_C)/k^*}{SSE_C/[n - (k + 1)]}$$

and define the p-value related to F(partial) to be the area under the curve of the F distribution (having k^* and $[n - (k + 1)]$ degrees of freedom) to the right of F(partial). Then, we can reject H_0 in favor of H_a at level of significance α if either of the following equivalent conditions holds:

1. $F(\text{partial}) > F_\alpha$
2. $p\text{-value} < \alpha$

Here the point F_α is based on k^* numerator and $n - (k + 1)$ denominator degrees of freedom.

Using Excel or MINITAB, we find that the unexplained variation for the *complete model* that uses all $k = 17$ variables is $SSE_C = 213{,}396.12$ and the unexplained variation for the *reduced model* that does not use the $k^* = 12$ squared and interaction variables (and thus uses only the 5 linear variables) is $SSE_R = 3{,}516{,}859.2$. It follows that

$$F(\text{partial}) = \frac{(SSE_R - SSE_C)/k^*}{SSE_C/[n - (k + 1)]}$$

$$= \frac{(3{,}516{,}859.2 - 213{,}396.12)/12}{213{,}396.12/[25 - (17 + 1)]}$$

$$= \frac{3{,}303{,}463.1/12}{213{,}396.12/7} = 9.03$$

Because $F_{.05}$ based on $k^* = 12$ numerator and $n-(k+1) = 7$ denominator degrees of freedom is 3.57 (see Table A.7 on page 796), and because F(partial) = 9.03 is greater than $F_{.05} = 3.57$, we conclude (at an α of .05) that at least one of the 12 squared and interaction variables in the 17-variable model is significant. In the exercises, the reader will do further analysis and use another partial F-test to pick a model that is perhaps better than the 17-variable model.

Exercises for Section 15.10

CONCEPTS

15.38 What is multicollinearity? What problems can be caused by multicollinearity?

15.39 Discuss how we compare regression models.

METHODS AND APPLICATIONS

15.40 THE HOSPITAL LABOR NEEDS CASE HospLab2

Recall that Table 15.6 (page 564) presents data concerning the need for labor in 16 U.S. Navy hospitals. This table gives values of the dependent variable Hours (monthly labor hours) and of the independent variables Xray (monthly X-ray exposures), BedDays (monthly occupied bed days—a hospital has one occupied bed day if one bed is occupied for an entire day), and Length (average length of patients' stay, in days). The data in Table 15.6 are part of a larger data set analyzed by the Navy. The complete data set consists of two additional independent variables—Load (average daily patient load) and Pop (eligible population in the area, in thousands)—values of which are given on the page margin. Figures 15.30 and 15.31 (on the next page) give Excel and MINITAB outputs of multicollinearity analysis and model building for the complete hospital labor needs data set.

Load	Pop
15.57	18.0
44.02	9.5
20.42	12.8
18.74	36.7
49.20	35.7
44.92	24.0
55.48	43.3
59.28	46.7
94.39	78.7
128.02	180.5
96.00	60.9
131.42	103.7
127.21	126.8
409.20	169.4
463.70	331.4
510.22	371.6

a **(1)** Find the three largest simple correlation coefficients among the independent variables and the three largest variance inflation factors in Figure 15.30. **(2)** Discuss why these statistics imply that the independent variables BedDays, Load, and Pop are most strongly involved in multicollinearity and thus contribute possibly redundant information for predicting Hours. Note that, although we have reasoned in Exercise 15.6(a) on page 565 that a negative coefficient (that is, least squares point estimate) for Length might be intuitively reasonable, the negative coefficients for Load and Pop [see Figure 15.30(b)] are not intuitively reasonable and are a

FIGURE 15.30 Multicollinearity and a Regression Analysis in the Hospital Labor Needs Situation

(a) Excel output of the correlation matrix

	Hours(y)	Xray(x1)	BedDays(x2)	Length(x3)	Load(x4)	Pop(x5)
Hours(y)	1					
Xray(x1)	0.9425	1				
BedDays(x2)	0.9889	0.9048	1			
Length(x3)	0.5603	0.4243	0.6609	1		
Load(x4)	0.9886	0.9051	0.9999	0.6610	1	
Pop(x5)	0.9465	0.9124	0.9328	0.4515	0.9353	1

(b) A MINITAB regression analysis

Predictor	Coef	SE Coef	T	P	VIF
Constant	2270.4	670.8	3.38	0.007	
Xray(x1)	0.04112	0.01368	3.01	0.013	8.1
BedDays(x2)	1.413	1.925	0.73	0.48	8684.2
Length(x3)	-467.9	131.6	-3.55	0.005	4.2
Load(x4)	-9.30	60.81	-0.15	0.882	9334.5
Pop(x5)	-3.223	4.474	-0.72	0.488	23.0

FIGURE 15.31 The MINITAB Output of Model Building for the Hospital Labor Needs Data

(a) The two best models of each size

Vars	R-Sq	R-Sq(adj)	Mallows C-p	S	Xray 1	DayS 2	Load 3	Load 4	Pop 5
1	97.8	97.6	52.3	856.71		X			
1	97.7	97.6	54	867.67				X	
2	99.3	99.2	9.5	489.13		X	X		
2	99.3	99.2	11.1	509.82			X	X	
3	99.6	99.5	3.3	387.16	X	X	X		
3	99.6	99.4	5	415.47	X		X	X	
4	99.7	99.5	4	381.56	X	X	X		X
4	99.6	99.5	4.5	390.88	X	X	X	X	
5	99.7	99.5	6	399.71	X	X	X	X	X

(b) Stepwise regression ($\alpha_{entry} = \alpha_{stay} = .10$)

Step	1	2	3
Constant	-70.23	2741.24	1946.80
BedDays	1.101	1.223	1.039
T-Value	24.87	36.30	15.39
P-Value	0.000	0.000	0.000
Length		-572	-414
T-Value		-5.47	-4.20
P-Value		0.000	0.001
XRay			0.039
T-Value			2.96
P-Value			0.012
S	857	489	387
R-Sq	97.79	99.33	99.61

further indication of strong multicollinearity. We conclude that a final regression model for predicting Hours may not need all three of the potentially redundant independent variables BedDays, Load, and Pop.

b Figure 15.31(a) indicates that the two best hospital labor needs models are the model using Xray, BedDays, Pop, and Length, which we will call Model 1, and the model using Xray, BedDays, and Length, which we will call Model 2. **(1)** Which model gives the smallest value of s and the largest value of $\overline{R}^2$? **(2)** Which model gives the smallest value of C? **(3)** Consider a questionable hospital for which Xray = 56,194, BedDays = 14,077.88, Pop = 329.7, and Length = 6.89. The 95 percent prediction intervals given by Models 1 and 2 for labor hours corresponding to this combination of values of the independent variables are, respectively, [14,888.43, 16,861.30] and [14,906.24, 16,886.26]. Which model gives the shortest prediction interval?

c **(1)** Which model is chosen by stepwise regression in Figure 15.31(b)? **(2)** Overall, which model seems best? **(3)** Which of BedDays, Load, and Pop does this best model use?

15.41 THE SALES REPRESENTATIVE CASE

Consider Figure 15.29 on page 604. The model using 12 squared and interaction variables has the smallest s. However, if we desire a somewhat simpler model, note that s does not increase substantially until we move from a model having seven squared and interaction variables to a model having six such variables. Moreover, we might subjectively conclude that the s of 210.70 for the model using 7 squared and interaction variables is not that much larger than the s of 174.6 for the model using 12 squared and interaction variables. Using the fact that the unexplained variations for these respective models are 532,733.88 and 213,396.12, perform a partial F-test to assess whether at least one of the extra five squared and interaction variables is significant. If none of the five extra variables are significant, we might consider the simpler model to be best. SalePerf

LO15-10 Use residual analysis to check the assumptions of multiple regression.

15.11 Residual Analysis in Multiple Regression

Basic residual analysis For a multiple regression model we plot the residuals given by the model against (1) values of each independent variable, (2) values of the predicted value of the dependent variable, and (3) the time order in which the data have been observed (if the regression data are time series data). A fanning-out pattern on a residual plot indicates an increasing error variance; a funneling-in pattern indicates a decreasing error variance. Both violate the constant variance assumption. A curved pattern on a residual plot indicates that the functional form of the regression model is incorrect. If the regression data are time series data, a cyclical pattern on the residual plot versus time suggests positive autocorrelation, while an alternating pattern suggests negative autocorrelation. Both violate the independence assumption. On the other hand, if all residual plots have (at least approximately) a horizontal band appearance, then it is reasonable to believe that the constant variance, correct functional form, and independence assumptions approximately hold. To check the normality assumption, we can construct a histogram, stem-and-leaf display, and normal plot of the residuals. The histogram and stem-and-leaf display should look bell-shaped and symmetric about 0; the normal plot should have a straight-line appearance.

FIGURE 15.32 Residual Plots for the Sales Representative Performance Model

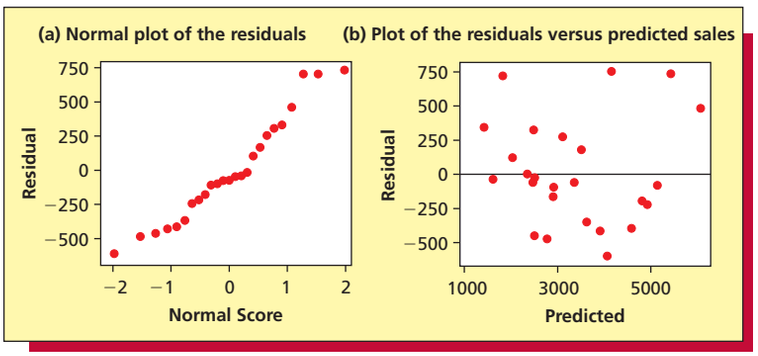
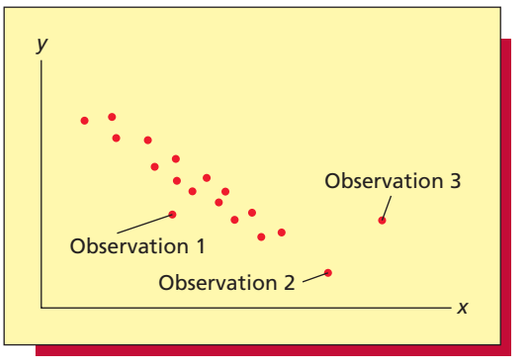


FIGURE 15.33 Outlying Observations



To illustrate these ideas, consider the sales representative performance data in Figure 15.10 (page 579). Figure 15.11 (page 580) gives the Excel output of a regression analysis of these data using the model that relates y to x_1, x_2, x_3, x_4, and x_5. The least squares point estimates on the output give the prediction equation

$$\hat{y} = -1{,}113.7879 + 3.6121x_1 + .0421x_2 + .1289x_3 + 256.9555x_4 + 324.5334x_5$$

Obs	Predicted	Residual
1	3,504.990	164.890
2	3,901.180	−427.230
3	2,774.866	−479.766
4	4,911.872	−236.312
5	5,415.196	710.764
6	2,026.090	108.850
7	5,126.127	−94.467
8	3,106.925	260.525
9	6,055.297	464.153
10	4,143.597	732.773
11	2,503.165	−34.895
12	1,827.065	706.245
13	2,478.083	−69.973
14	2,351.344	−13.964
15	4,797.688	−210.738
16	2,904.099	−174.859
17	3,362.660	−73.260
18	2,907.376	−106.596
19	3,625.026	−360.826
20	4,056.443	−602.823
21	1,409.835	331.615
22	2,494.101	−458.351
23	1,617.561	−39.561
24	4,574.903	−407.463
25	2,488.700	311.270

Using this prediction equation, we can calculate the predicted sales values and residuals given on the page margin. For example, observation 10 corresponds to a sales representative for whom $x_1 = 105.69$, $x_2 = 42{,}053.24$, $x_3 = 5{,}673.11$, $x_4 = 8.85$, and $x_5 = .31$. If we insert these values into the prediction equation, we obtain a predicted sales value of $\hat{y}_{10} = 4{,}143.597$. Because the actual sales for the sales representative are $y_{10} = 4{,}876.370$, the residual e_{10} equals the difference between $y_{10} = 4{,}876.370$ and $\hat{y}_{10} = 4{,}143.597$, which is 732.773. The normal plot of the residuals in Figure 15.32(a) has an approximate straight-line appearance. The plot of the residuals versus predicted sales in Figure 15.32(b) has a horizontal band appearance, as do the plots of the residuals versus the independent variables (these plots are not given here). We conclude that the regression assumptions approximately hold for the sales representative performance model.

Outliers An observation that is well separated from the rest of the data is called an **outlier,** and an observation may be an outlier with respect to its y value and/or its x values. We illustrate these ideas by considering Figure 15.33, which is a hypothetical plot of the values of a dependent variable y against an independent variable x. Observation 1 in this figure is outlying with respect to its y value, but not with respect to its x value. Observation 2 is outlying with respect to its x value, but because its y value is consistent with the regression relationship displayed by the nonoutlying observations, it is not outlying with respect to its y value. Observation 3 is an outlier with respect to its x value and its y value.

It is important to identify outliers because (as we will see) outliers can have adverse effects on a regression analysis and thus are candidates for removal from a data set. Moreover, in addition to using data plots, we can use more sophisticated procedures to detect outliers. For example, suppose that the U.S. Navy wishes to develop a regression model based on efficiently run Navy hospitals to evaluate the labor needs of questionably run Navy hospitals. Figure 15.34(a) on the next page gives labor needs data for 17 Navy hospitals. Specifically, this table gives values of the dependent variable Hours (y, monthly labor hours required) and of the independent variables Xray (x_1, monthly X-ray exposures), BedDays (x_2, monthly occupied bed days—a hospital has one occupied bed day if one bed is occupied for an entire day), and Length (x_3, average length of patients' stay, in days). When we perform a regression analysis of these data using the model relating y to x_1, x_2, and x_3, we obtain the Excel add-in (MegaStat) output of residuals and outlier diagnostics shown in Figure 15.34(b), as well as the residual plot shown in Figure 15.34(c). (MINITAB gives the same diagnostics, and at the end of this section we will give formulas for most of these diagnostics). We now explain the meanings of the diagnostics.

FIGURE 15.34 Hospital Labor Needs Data, Outlier Diagnostics, and Residual Plots

(a) The data DS HospLab3

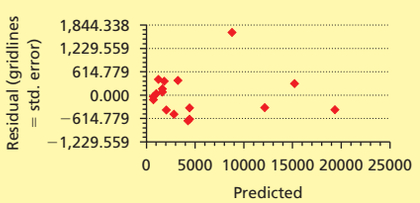
Hospital	Hours y	Xray x_1	BedDays x_2	Length x_3
1	566.52	2463	472.92	4.45
2	696.82	2048	1339.75	6.92
3	1033.15	3940	620.25	4.28
4	1603.62	6505	568.33	3.90
5	1611.37	5723	1497.60	5.50
6	1613.27	11520	1365.83	4.60
7	1854.17	5779	1687.00	5.62
8	2160.55	5969	1639.92	5.15
9	2305.58	8461	2872.33	6.18
10	3503.93	20106	3655.08	6.15
11	3571.89	13313	2912.00	5.88
12	3741.40	10771	3921.00	4.88
13	4026.52	15543	3865.67	5.50
14	10343.81	36194	7684.10	7.00
15	11732.17	34703	12446.33	10.78
16	15414.94	39204	14098.40	7.05
17	18854.45	86533	15524.00	6.35

(b) Excel add-in (MegaStat) outlier diagnostics for the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$

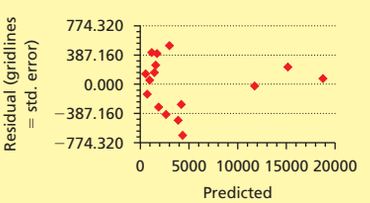
Observation	Residual	Leverage	Studentized Residual	Studentized Deleted Residual
1	-121.889	0.121	-0.211	-0.203
2	-25.028	0.226	-0.046	-0.044
3	67.757	0.130	0.118	0.114
4	431.156	0.159	0.765	0.752
5	84.590	0.085	0.144	0.138
6	-380.599	0.112	-0.657	-0.642
7	177.612	0.084	0.302	0.291
8	369.145	0.083	0.627	0.612
9	-493.181	0.085	-0.838	-0.828
10	-687.403	0.120	-1.192	-1.214
11	380.933	0.077	0.645	0.630
12	-623.102	0.177	-1.117	-1.129
13	-337.709	0.064	-0.568	-0.553
14	1,630.503	0.146	2.871	4.558
15	-348.694	0.682	-1.005	-1.006
16	281.914	0.785	0.990	0.989
17	-406.003	0.863	-1.786	-1.975

Source: Hospital Labor Needs Data from Procedures and Analysis for Staffing Standards Development: Regression Analysis Handbook, © 1979.

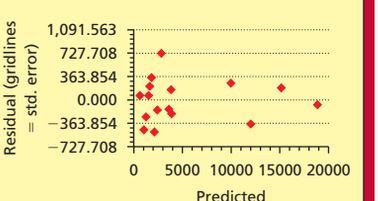
(c) Plot of residuals in Figure 15.34(b)



(d) Plot of residuals for Option 1



(e) Plot of residuals for Option 2



Leverage values The **leverage value** for an observation is the **distance value** that has been discussed in the optional technical note at the end of Section 15.6 (page 576). This value is a measure of the distance between the observation's x values and the center of all of the observed x values. **If the leverage value for an observation is large, the observation is outlying with respect to its x values and thus would have substantial leverage in determining the least squares prediction equation.** For example, each of observations 2 and 3 in Figure 15.33 on the previous page is an outlier with respect to its x value and thus would have substantial leverage in determining the position of the least squares line. Moreover, because observations 2 and 3 have inconsistent y values, they would pull the least squares line in opposite directions. A **leverage value is considered to be large if it is greater than twice the average of all of the leverage values,** which can be shown to be equal to $2(k + 1)/n$. [The Excel add-in (MegaStat) shades such a leverage value in dark blue.] For example, because there are $n = 17$ observations in Figure 15.34(a) and because the model relating y to x_1, x_2, and x_3 utilizes $k = 3$ independent variables, twice the average leverage value is $2(k + 1)/n = 2(3 + 1)/17 = .4706$. Looking at Figure 15.34(b), we see that the leverage values for hospitals 15, 16, and 17 are, respectively, .682, .785, and .863. Because these leverage values are greater than .4706, we conclude that **hospitals 15, 16, and 17 are outliers with respect to their x values.** Intuitively, this is because Figure15.34(a) indicates that x_2 (monthly occupied bed days) is substantially larger for hospitals 15, 16, and 17 than for hospitals 1 through 14. Also note that both x_1 (monthly X-ray exposures) and x_2 (monthly occupied bed days) are substantially larger for hospital 14 than for hospitals 1 through 13. To summarize, we might classify, hospitals 1 through 13 as small to medium sized hospitals and hospitals 14, 15, 16, and 17 as larger hospitals.

Residuals and studentized residuals To identify outliers with respect to their y values, we can use residuals. Any residual that is substantially different from the others is suspect. For example, note from Figure 15.34(b) that the residual for hospital 14, $e_{14} = 1630.503$, seems much larger than the other residuals. Assuming that the labor hours of 10,343.81 for hospital 14 has not been misrecorded, the residual of 1630.503 says that the labor hours are 1630.503 hours more than predicted by the regression model. If we divide an observation's residual by the residual's standard error, we obtain a **studentized residual.** For example, Figure 15.34(b) tells us that the studentized residual for hospital 14 is 2.871. If the studentized residual for an observation is greater than 2 in absolute value, we have some evidence that the observation is an outlier with respect to its y value.

Deleted residuals and studentized deleted residuals Consider again Figure 15.33 on page 607, and suppose that we use observation 3 to help determine the least squares line. Doing this might draw the least squares line toward observation 3, causing the point prediction $\hat{y}_3$ given by the line to be near y_3 and thus the usual residual $y_3 - \hat{y}_3$ to be small. This would falsely imply that observation 3 is not an outlier with respect to its y value. Moreover, this sort of situation shows the need for computing a **deleted residual.** For a particular observation, observation i, the deleted residual is found by subtracting from y_i the point prediction $\hat{y}_{(i)}$ computed using least squares point estimates based on all n observations except for observation i. Standard statistical software packages calculate the deleted residual for each observation and divide this residual by its standard error to form the **studentized deleted residual.** The experience of the authors leads us to suggest that one should conclude that an observation is an outlier with respect to its y value if (and only if) the studentized deleted residual is greater in absolute value than $t_{.005}$, which is based on $n - k - 2$ degrees of freedom. [The Excel add-in (MegaStat) shades such a studentized deleted residual in dark blue.] For the hospital labor needs model, $n - k - 2 = 17 - 3 - 2 = 12$, and therefore $t_{.005} = 3.055$. The studentized deleted residual for hospital 14, which equals 4.558 [see Figure 15.34(b)], is greater in absolute value than $t_{.005} = 3.055$. Therefore, we conclude that **hospital 14 is an outlier with respect to its y value.**

An example of dealing with outliers One option for dealing with the fact that hospital 14 is an outlier with respect to its y value is to assume that hospital 14 has been run inefficiently. Because we need to develop a regression model using efficiently run hospitals, based on this assumption we would remove hospital 14 from the data set. If we perform a regression analysis using a model relating y to x_1, x_2, and x_3 with hospital 14 removed from the data set (we call this **Option 1**), we obtain a standard error of $s = 387.16$. This s is considerably smaller than the large standard error of 614.779 caused by hospital 14's large residual when we use all 17 hospitals to relate y to x_1, x_2, and x_3.

A second option is motivated by the fact that *large organizations sometimes exhibit inherent inefficiencies*. To assess whether there might be a *general large hospital inefficiency,* we define a dummy variable D_L that equals 1 for the larger hospitals 14–17 and 0 for the smaller hospitals 1–13. If we fit the resulting regression model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 D_L + \varepsilon$ to all 17 hospitals (we call this **Option 2**), we obtain a b_4 of 2871.78 and a p-value for testing H_0: $\beta_4 = 0$ of .0003. This indicates the existence of a large hospital inefficiency that is estimated to be an extra 2871.78 hours per month. In addition, the dummy variable model's s is 363.854, which is slightly smaller than the s of 387.16 obtained using Option 1. The studentized deleted residual for hospital 14 using the dummy variable model tells us what would happen if we removed hospital 14 from the data set and predicted y_{14} by using a newly fitted dummy variable model. In the exercises the reader will show that **the prediction obtained, which uses a large hospital inefficiency estimate based on the remaining large hospitals 15, 16, and 17, indicates that hospital 14's labor hours are not unusually large.** This justifies leaving hospital 14 in the data set when using the dummy variable model. In summary, both Options 1 and 2 seem reasonable. The reader will further compare these options in the exercises.

Cook's D, Dfbetas, and Dffits (Optional) If a particular observation, observation i, is an outlier with respect to its y and/or x values, it might significantly *influence* the least squares point estimates of the model parameters. To detect such influence, we compute **Cook's distance measure** (or **Cook's D**) for observation i, which we denote as D_i. To understand D_i, let $F_{.50}$

Obs	Cook's D
1	0.002
2	0.000
3	0.001
4	0.028
5	0.000
6	0.014
7	0.002
8	0.009
9	0.016
10	0.049
11	0.009
12	0.067
13	0.006
14	0.353
15	0.541
16	0.897
17	5.033

denote the 50th percentile of the F distribution based on $(k + 1)$ numerator and $n - (k + 1)$ denominator degrees of freedom. It can be shown that **if D_i is greater than $F_{.50}$, then removing observation i from the data set would significantly change (as a group) the least squares point estimates of the model parameters.** In this case we say that **observation i is influential.** For example, suppose that we relate y to x_1, x_2, and x_3 using all $n = 17$ observations in Figure 15.34(a). Noting that $k + 1 = 4$ and $n - (k + 1) = 13$, we find (using Excel) that $F_{.50} = .8845$. The Excel add-in (MegaStat) output in the page margin tells us that both $D_{16} = .897$ and $D_{17} = 5.033$ are greater than $F_{.50} = .8845$ (see the dark blue shading). It follows that removing either hospital 16 or 17 from the data set would significantly change (as a group) the least squares estimates of the model parameters.

To assess whether a particular least squares point estimate would significantly change, we can use an advanced statistical software package such as SAS, which gives the following **difference in estimate of β_j statistics (Dfbetas)** for hospitals 16 and 17:

	INTERCEP	X1	X2	X3
Obs	Dfbetas	Dfbetas	Dfbetas	Dfbetas
16	0.9880	−1.4289	1.7339	−1.1029
17	0.0294	−3.0114	1.2688	0.3155

Examining the Dfbetas statistics, we see that hospital 17's Dfbetas for the independent variable x_1 (monthly X-ray exposures) equals −3.0114, which is *negative and greater in absolute value than 2,* a sometimes used critical value for Dfbetas statistics. This implies that removing hospital 17 from the data set would *significantly decrease* the least squares point estimate of the effect, β_1, of monthly X-ray exposures on monthly labor hours. One possible consequence might then be that our model would *significantly underpredict* the monthly labor hours for a hospital which [like hospital 17—see Figure 15.34(a)] has a particularly large number of monthly X-ray exposures. In fact, consider the MINITAB output in the page margin of the **difference in fits statistic (Dffits).** Dffits for hospital 17 equals −4.96226, which is *negative and greater in absolute value than the critical value 2*. This implies that removing hospital 17 from the data set would *significantly decrease* the point prediction of the monthly labor hours for a hospital that has the same values of x_1, x_2, and x_3 as does hospital 17. Moreover, although it can be verified that using the previously discussed Option 1 or Option 2 to deal with hospital 14's large residual substantially reduces Cook's D, Dfbetas for x_1, and Dffits for hospital 17, these or similar statistics remain or become somewhat significant for the large hospitals 15, 16, and 17. The practical implication is that if we wish to predict monthly labor hours for questionably run *large* hospitals, it is very important to keep all of the efficiently run large hospitals 15, 16, and 17 in the data set. (Furthermore, it would be desirable to add information for additional efficiently run large hospitals to the data set).

Hosp	Dffits
1	−0.07541
2	−0.02404
3	0.04383
4	0.32657
5	0.04213
6	−0.22799
7	0.08818
8	0.18406
9	−0.25179
10	−0.44871
11	0.18237
12	−0.52368
13	−0.14509
14	1.88820
15	−1.47227
16	1.89295
17	−4.96226

A technical note (Optional) Let h_i and e_i denote the leverage value and residual for observation i based on a regression model using k independent variables. Then, for observation i, the studentized residual is e_i divided by $s\sqrt{1 - h_i}$, the studentized deleted residual (denoted t_i) is $e_i\sqrt{n - k - 2}$ divided by $\sqrt{SSE(1 - h_i) - e_i^2}$, Cook's distance measure is $e_i^2 h_i$ divided by $(k + 1)s^2(1 - h_i)^2$, and the difference in fits statistic is $t_i[h_i/(1 - h_i)]^{1/2}$. The formula for the difference in estimate of β_j statistics is very complicated and will not be given here.

Exercises for Section 15.11

connect

CONCEPTS

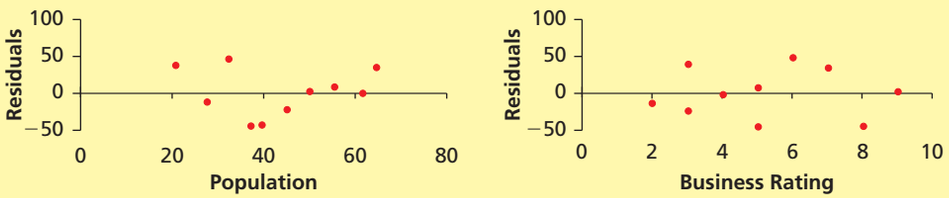
15.42 Discuss how we use residual plots to check the regression assumptions for multiple regression.

15.43 Discuss how we identify an outlier with respect to its y value and/or x values.

METHODS AND APPLICATIONS

15.44 For each of the following cases, use the indicated residual plots to check for any violations of the regression assumptions.

a **The Tasty Sub Shop Case:** For the model relating *Revenue* to *Population* and *Business Rating,* use the following plots: DS TastySub2



b **The Natural Gas Consumption Case:** For the model relating *GasCons* to *Temp* and *Chill,* use the plots shown on pages 623 (in Appendix 15.2) and 626 (in Appendix 15.3). DS GasCon2

15.45 **THE HOSPITAL LABOR NEEDS CASE** DS HospLab DS HospLab4

(1) Analyze the studentized deleted residuals on the page margin for Options 1 and 2 (see **SDR1** and **SDR2**). **(2)** Is hospital 14 an outlier with respect to its y value when using Option 2? **(3)** Consider a questionable large hospital ($D_L = 1$) for which Xray = 56,194, BedDays = 14,077.88, and Length = 6.89. Also, consider the labor needs in an efficiently run large hospital described by this combination of values of the independent variables. The 95 percent prediction intervals for these labor needs given by the models of Options 1 and 2 are, respectively, [14,906.24, 16,886.26] and [15,175.04, 17,030.01]. By comparing these prediction intervals, by analyzing the residual plots for Options 1 and 2 given in Figure 15.34(d) and (e) on page 608, and by using your conclusions regarding the studentized deleted residuals, recommend which option should be used. **(4)** What would you conclude if the questionable large hospital used 17,207.31 monthly labor hours?

15.46 Recall that Figure14.26 (a) on page 533 gives $n = 16$ weekly values of Pages Bookstore sales (y), Pages advertising expenditure (x_1), and competitor's advertising expenditure (x_2). When we fit the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ to the data, we find that the Durbin–Watson statistic is $d = 1.6\overline{3}$. Use the partial Durbin–Watson table on the page margin to test for positive autocorrelation by setting α equal to .05.

Obs	SDR1	SDR2
1	−0.333	−1.439
2	0.404	0.233
3	0.161	−0.750
4	1.234	0.202
5	0.425	0.213
6	−0.795	−1.490
7	0.677	0.617
8	1.117	1.010
9	−1.078	−0.409
10	−1.359	−0.400
11	1.461	2.571
12	−2.224	−0.624
13	−0.685	0.464
14		1.406
15	−0.137	−2.049
16	1.254	1.108
17	0.597	−0.639

	$k = 2$	
n	$d_{L,.05}$	$d_{U,.05}$
15	0.95	1.54
16	0.98	1.54
17	1.02	1.54
18	1.05	1.53

15.12 Logistic Regression

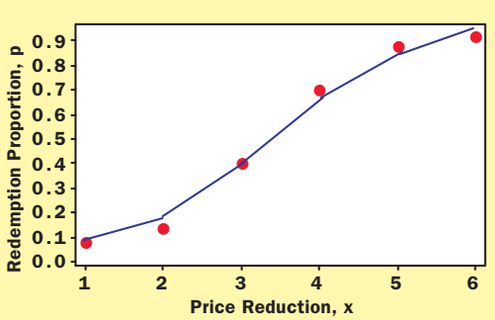
LO15-11 Use a logistic model to estimate probabilities and odds ratios.

Suppose that in a study of the effectiveness of offering a price reduction on a given product, 300 households having similar incomes were selected. A coupon offering a price reduction, x, on the product, as well as advertising material for the product, was sent to each household. The coupons offered different price reductions (10, 20, 30, 40, 50, and 60 dollars), and 50 homes were assigned at random to each price reduction. The table in the page margin summarizes the number, y, and proportion, $\hat{p}$, of households redeeming coupons for each price reduction, x (expressed in units of \$10). On the left side of Figure 15.35 we plot the $\hat{p}$ values versus the

x	y	$\hat{p}$
1	4	.08
2	7	.14
3	20	.40
4	35	.70
5	44	.88
6	46	.92

DS PrcRed

FIGURE 15.35 MINITAB Output of a Logistic Regression of the Price Reduction Data



Logistic Regression Table

Predictor	Coef	SE Coef	Z	P
Constant	-3.7456	0.434355	-8.62	0.000
x	1.1109	0.119364	9.31	0.000

Price Reduction, x	Probability Estimate	Price Reduction, x	Probability Estimate
1	0.066943	4	0.667791
2	0.178920	5	0.859260
3	0.398256	6	0.948831

TABLE 15.16
The Performance Data DS PerfTest

Group	Test 1	Test 2
1	96	85
1	96	88
1	91	81
1	95	78
1	92	85
1	93	87
1	98	84
1	92	82
1	97	89
1	95	96
1	99	93
1	89	90
1	94	90
1	92	94
1	94	84
1	90	92
1	91	70
1	90	81
1	86	81
1	90	76
1	91	79
1	88	83
1	87	82
0	93	74
0	90	84
0	91	81
0	91	78
0	88	78
0	86	86
0	79	81
0	83	84
0	79	77
0	88	75
0	81	85
0	85	83
0	82	72
0	82	81
0	81	77
0	86	76
0	81	84
0	85	78
0	83	77
0	81	71

Source: Performance data from T.E. Dielman, *Applied Regression Analysis for Business and Economics,* 2nd ed. © 1996. Reprinted with permission of Brooks/Cole, a division of Cengage Learning, www.cengagerights.com, Fax 800-730-2215.

x values and draw a hypothetical curve through the plotted points. A theoretical curve having the shape of the curve in Figure 15.35 is the **logistic curve**

$$p(x) = \frac{e^{(\beta_0 + \beta_1 x)}}{1 + e^{(\beta_0 + \beta_1 x)}}$$

where $p(x)$ denotes the probability that a household receiving a coupon having a price reduction of x will redeem the coupon. The MINITAB output in Figure 15.35 tells us that the point estimates of β_0 and β_1 are $b_0 = -3.7456$ and $b_1 = 1.1109$. (The point estimates in logistic regression are usually obtained by an advanced statistical procedure called *maximum likelihood estimation.*) Using these estimates, it follows that, for example,

$$\hat{p}(5) = \frac{e^{(-3.7456 + 1.1109(5))}}{1 + e^{(-3.7456 + 1.1109(5))}} = \frac{6.1037}{1 + 6.1037} = .8593$$

That is, $\hat{p}(5) = .8593$ is the point estimate of the probability that a household receiving a coupon having a price reduction of $50 will redeem the coupon. The MINITAB output in Figure 15.35 gives the values of $\hat{p}(x)$ for $x = 1, 2, 3, 4, 5$, and 6.

The **general logistic regression model** relates the probability that an event (such as redeeming a coupon) will occur to k independent variables $x_1, x_2, \ldots, x_k$. This general model is

$$p(x_1, x_2, \ldots, x_k) = \frac{e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}}{1 + e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}}$$

where $p(x_1, x_2, \ldots, x_k)$ is the probability that the event will occur when the values of the independent variables are $x_1, x_2, \ldots, x_k$. In order to estimate $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ we obtain n observations, with each observation consisting of observed values of $x_1, x_2, \ldots, x_k$ and of a dependent variable y. Here, y is a **dummy variable** that equals 1 if the event has occurred and 0 otherwise.

For example, suppose that the personnel director of a firm has developed two tests to help determine whether potential employees would perform successfully in a particular position. To help estimate the usefulness of the tests, the director gives both tests to 43 employees that currently hold the position. Table 15.16 gives the scores of each employee on both tests and indicates whether the employee is currently performing successfully or unsuccessfully in the position. If the employee is performing successfully, we set the dummy variable *Group* equal to 1; if the employee is performing unsuccessfully, we set *Group* equal to 0. Let x_1 and x_2 denote the scores of a potential employee on tests 1 and 2, and let $p(x_1, x_2)$ denote the probability that a potential employee having the scores x_1 and x_2 will perform successfully in the position. We can estimate the relationship between $p(x_1, x_2)$ and x_1 and x_2 by using the logistic regression model

$$p(x_1, x_2) = \frac{e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}}{1 + e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}}$$

The MINITAB output in Figure 15.36 tells us that the point estimates of β_0, β_1, and β_2 are $b_0 = -56.17$, $b_1 = .4833$, and $b_2 = .1652$. Consider, therefore, a potential employee who scores a 93 on test 1 and an 84 on test 2. It follows that a point estimate of the probability that the potential employee will perform successfully in the position is

$$\hat{p}(93, 84) = \frac{e^{(-56.17 + .4833(93) + .1652(84))}}{1 + e^{(-56.17 + .4833(93) + .1652(84))}} = \frac{14.206506}{15.206506} = .9342$$

If we *classify* a potential employee into *group 1* ("will perform successfully"), as opposed to *group 2* ("will not perform successfully"), if and only if $\hat{p}(x_1, x_2)$ is greater than .5, this potential employee is classified into group 1.

To further analyze the logistic regression output, we consider several hypothesis tests that are based on the chi-square distribution. We first consider testing H_0: $\beta_1 = \beta_2 = 0$ versus H_a: At least one of β_1 or β_2 does not equal 0. The p-value for this test is the area under the chi-square curve having $k = 2$ degrees of freedom to the right of the test statistic value $G = 31.483$. Although the

FIGURE 15.36 MINITAB Output of a Logistic Regression of the Performance Data

```
Response Information

Variable  Value  Count
Group     1         23  (Event)
          0         20
          Total     43

Logistic Regression Table

                                                 Odds      95% CI
Predictor       Coef   SE Coef      Z      P  Ratio  Lower  Upper
Constant     -56.17   17.4516  -3.22  0.001
Test 1       0.4833    0.1578   3.06  0.002   1.62   1.19   2.21
Test 2       0.1652    0.1021   1.62  0.106   1.18   0.97   1.44

Log-Likelihood = -13.959
Test that all slopes are zero: G = 31.483, DF = 2, P-Value = 0.000
```

calculation of G is too complicated to demonstrate in this book, the MINITAB output gives the value of G and the related p-value, which is less than .001. This p-value implies that we have extremely strong evidence that at least one of β_1 or β_2 does not equal zero. The p-value for testing H_0: $\beta_1 = 0$ versus H_a: $\beta_1 \neq 0$ is the area under the chi-square curve having one degree of freedom to the right of the square of $z = (b_1/s_{b_1}) = (.4833/.1578) = 3.06$. The MINITAB output tells us that this p-value is .002, which implies that we have very strong evidence that the score on test 1 is related to the probability of a potential employee's success. The p-value for testing H_0: $\beta_2 = 0$ versus H_a: $\beta_2 \neq 0$ is the area under the chi-square curve having one degree of freedom to the right of the square of $z = (b_2/s_{b_2}) = (.1652/.1021) = 1.62$. The MINITAB output tells us that this p-value is .106, which implies that we do not have strong evidence that the score on test 2 is related to the probability of a potential employee's success. In Exercise 15.49 we will consider a logistic regression model that uses only the score on test 1 to estimate the probability of a potential employee's success.

The **odds** of success for a potential employee is defined to be the probability of success divided by the probability of failure for the employee. That is,

$$\text{odds} = \frac{p(x_1, x_2)}{1 - p(x_1, x_2)}$$

For the potential employee who scores a 93 on test 1 and an 84 on test 2, we estimate that the odds of success are $.9342/(1 - .9342) = 14.22$. That is, we estimate that the odds of success for the potential employee are about 14 to 1. It can be shown that $e^{b_1} = e^{.4833} = 1.62$ is a point estimate of the **odds ratio for x_1,** which is the proportional change in the odds (for any potential employee) that is associated with an increase of one in x_1 when x_2 stays constant. This point estimate of the odds ratio for x_1 is shown on the MINITAB output and says that, for every one point increase in the score on test 1 when the score on test 2 stays constant, we estimate that a potential employee's odds of success increase by 62 percent. Furthermore, the 95 percent confidence interval for the odds ratio for x_1, [1.19, 2.21], does not contain 1. Therefore, as with the (equivalent) chi-square test of H_0: $\beta_1 = 0$, we conclude that there is strong evidence that the score on test 1 is related to the probability of success for a potential employee. Similarly, it can be shown that $e^{b_2} = e^{.1652} = 1.18$ is a point estimate of the **odds ratio for x_2,** which is the proportional change in the odds (for any potential employee) that is associated with an increase of one in x_2 when x_1 stays constant. This point estimate of the odds ratio for x_2 is shown on the MINITAB output and says that, for every one point increase in the score on test 2 when the score on test 1 stays constant, we estimate that a potential employee's odds of success increases by 18 percent. However, the 95 percent confidence interval for the odds ratio for x_2, [.97, 1.44], contains 1. Therefore, as with the equivalent chi-square test of H_0: $\beta_2 = 0$, we cannot conclude that there is strong evidence that the score on test 2 is related to the probability of success for a potential employee.

To conclude this section, consider the **general logistic regression model**

$$p(x_1, x_2, \ldots, x_k) = \frac{e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}}{1 + e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}}$$

where $p(x_1, x_2, \ldots, x_k)$ is the probability that the event under consideration will occur when the values of the independent variables are $x_1, x_2, \ldots, x_k$. The **odds** of the event occurring is defined to be $p(x_1, x_2, \ldots, x_k)/(1 - p(x_1, x_2, \ldots, x_k))$, which is the probability that the event will occur divided by the probability that the event will not occur. It can be shown that the odds equals $e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}$. The natural logarithm of the odds is $(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)$, which is called the **logit.** If $b_0, b_1, b_2, \ldots, b_k$ are the point estimates of $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$, the point estimate of the logit, denoted $\widehat{\ell g}$, is $(b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k)$. It follows that the point estimate of the probability that the event will occur is

$$\hat{p}(x_1, x_2, \ldots, x_k) = \frac{e^{\widehat{\ell g}}}{1 + e^{\widehat{\ell g}}} = \frac{e^{(b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k)}}{1 + e^{(b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k)}}$$

Finally, consider an arbitrary independent variable x_j. It can be shown that e^{b_j} is the point estimate of the **odds ratio for x_j,** which is the proportional change in the odds that is associated with a one unit increase in x_j when the other independent variables stay constant.

Exercises for Section 15.12

connect

CONCEPTS

15.47 What two values does the dependent variable equal in logistic regression? What do these values represent?

15.48 Define the odds of an event, and the odds ratio for x_j.

METHODS AND APPLICATIONS

15.49 If we use the logistic regression model

$$p(x_1) = \frac{e^{(\beta_0 + \beta_1 x_1)}}{1 + e^{(\beta_0 + \beta_1 x_1)}}$$

to analyze the performance data in Table 15.16 on page 612, we find that the point estimates of the model parameters and their associated p-values (given in parentheses) are $b_0 = -43.37(.001)$ and $b_1 = .4897(.001)$. **(1)** Find a point estimate of the probability of success for a potential employee who scores a 93 on test 1. **(2)** Using $b_1 = .4897$, find a point estimate of the odds ratio for x_1. **(3)** Interpret this point estimate.

15.50 Mendenhall and Sincich (1993) present data that can be used to investigate allegations of gender discrimination in the hiring practices of a particular firm. These data are given in the page margin. In this table, y is a dummy variable that equals 1 if a potential employee was hired and 0 otherwise; x_1 is the number of years of education of the potential employee; x_2 is the number of years of experience of the potential employee; and x_3 is a dummy variable that equals 1 if the potential employee was a male and 0 if the potential employee was a female. If we use the logistic regression model

$$p(x_1, x_2, x_3) = \frac{e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)}}{1 + e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)}}$$

to analyze these data, we find that the point estimates of the model parameters and their associated p-values (given in parentheses) are $b_0 = -14.2483(.0191)$, $b_1 = 1.1549(.0552)$, $b_2 = .9098$ (.0341), and $b_3 = 5.6037(.0313)$.

a Consider a potential employee having 4 years of education and 5 years of experience. Find **(1)** a point estimate of the probability that the potential employee will be hired if the potential employee is a male, and **(2)** a point estimate of the probability that the potential employee will be hired if the potential employee is a female.

b **(1)** Using $b_3 = 5.6037$, find a point estimate of the odds ratio for x_3. **(2)** Interpret this odds ratio. **(3)** Using the p-value describing the importance of x_3, can we conclude that there is strong evidence that gender is related to the probability that a potential employee will be hired?

y	x_1	x_2	x_3
0	6	2	0
0	4	0	1
1	6	6	1
1	6	3	1
0	4	1	0
1	8	3	0
0	4	2	1
0	4	4	0
0	6	1	0
1	8	10	0
0	4	2	1
0	8	5	0
0	4	2	0
0	6	7	0
1	4	5	1
0	6	4	0
0	8	0	1
1	6	1	1
0	4	7	0
0	4	1	1
0	4	5	0
0	6	0	1
1	8	5	1
0	4	9	0
0	8	1	0
0	6	1	1
1	4	10	1
1	6	12	0

DS Gender

Source: William Mendenhall and Terry Sincich, *A Second Course in Business Statistics: Regression Analysis,* Fourth edition, © 1993. Reprinted with permission of Prentice Hall.

Chapter Summary

This chapter has discussed **multiple regression analysis.** We began by considering the **multiple regression model** and the assumptions behind this model. We next discussed the **least squares point estimates** of the model parameters and some ways to judge **overall model utility**—the **standard error,** the **multiple coefficient of determination,** the **adjusted multiple coefficient of determination,** and the **overall *F*-test.** Then we considered testing the significance of a single independent variable in a multiple regression model, calculating a **confidence interval** for the mean value of the dependent variable, and calculating a **prediction interval** for an individual value of the dependent variable. We continued this chapter by explaining the use of **dummy variables** to model **qualitative** independent variables and the use of **squared and interaction variables.** We then considered **multicollinearity,** which can adversely affect the ability of the *t* statistics and associated *p*-values to assess the importance of the independent variables in a regression model. For this reason, we need to determine if the overall model gives a **high R^2,** a **small *s*,** a **high adjusted R^2, short prediction intervals,** and a **small *C*.** We explained how to compare regression models on the basis of these criteria, and we discussed **stepwise regression** and the **partial *F*-test.** We then considered using **residual analysis** (including the detection of **outliers**) to check the assumptions for multiple regression models. We concluded this chapter by discussing how to use **logistic regression** to estimate the probability of an event.

Glossary of Terms

dummy variable: A variable that takes on the values 0 or 1 and is used to describe the effects of the different levels of a qualitative independent variable in a regression model. (page 581)
interaction: The situation in which the relationship between the mean value of the dependent variable and an independent variable is dependent on the value of another independent variable. (pages 585 and 594)
multicollinearity: The situation in which the independent variables used in a regression analysis are related to each other. (page 598)
multiple regression model: An equation that describes the relationship between a dependent variable and more than one independent variable. (page 560)
stepwise regression: An iterative model building technique for selecting important predictor variables (page 602)

Important Formulas and Tests

The least squares point estimates: page 557
The multiple regression model: page 560
Point estimate of a mean value of *y*: page 560
Point prediction of an individual value of *y*: page 560
Mean square error: page 566
Standard error: page 566
Total variation: page 567
Explained variation: page 567
Unexplained variation: page 567
Multiple coefficient of determination: page 567
Multiple correlation coefficient: page 567
Adjusted multiple coefficient of determination: page 568
An *F* test for the multiple regression model: page 569
Testing the significance of an independent variable: page 572
Confidence interval for β_j: page 574
Confidence interval for a mean value of *y*: pages 575 and 576
Prediction interval for an individual value of *y*: pages 575 and 576
Distance value (in multiple regression): page 576
The quadratic regression model: page 592
Variance inflation factor: page 598
C statistic: page 602
Partial *F* test: page 604
Studentized deleted residual: page 609
Cook's D, Dfbetas, and Dffits: pages 609–610
Logistic curve: page 612
Logistic regression model: page 612
Odds and odds ratio: page 613

Supplementary Exercises

connect™

15.51 The trend in home building in recent years has been to emphasize open spaces and great rooms, rather than smaller living rooms and family rooms. A builder of speculative homes in the college community of Oxford, Ohio, had been building such homes, but his homes had been taking many months to sell and selling for substantially less than the asking price. In order to determine what types of homes would attract residents of the community, the builder contacted a statistician at a local college. The statistician went to a local real estate agency and obtained the data in Table 15.17. This table presents the sales price *y*, square footage x_1, number of rooms x_2, number of bedrooms x_3, and age x_4 for each of 63 single-family residences recently sold in the community. When we perform a regression analysis of these data using the model

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

we find that the least squares point estimates of the model parameters and their associated *p*-values (given in parentheses) are as shown in Table 15.18. Discuss why the estimates $b_2 = 6.3218$ and $b_3 = -11.1032$ suggest that it might be more profitable when building a house of a

TABLE 15.17 Measurements Taken on 63 Single-Family Residences DS OxHome

Residence	Sales Price, y (× $1,000)	Square Feet, x_1	Rooms, x_2	Bedrooms, x_3	Age, x_4	Residence	Sales Price, y (× $1,000)	Square Feet, x_1	Rooms, x_2	Bedrooms, x_3	Age, x_4
1	53.5	1,008	5	2	35	33	63.0	1,053	5	2	24
2	49.0	1,290	6	3	36	34	60.0	1,728	6	3	26
3	50.5	860	8	2	36	35	34.0	416	3	1	42
4	49.9	912	5	3	41	36	52.0	1,040	5	2	9
5	52.0	1,204	6	3	40	37	75.0	1,496	6	3	30
6	55.0	1,204	5	3	10	38	93.0	1,936	8	4	39
7	80.5	1,764	8	4	64	39	60.0	1,904	7	4	32
8	86.0	1,600	7	3	19	40	73.0	1,080	5	2	24
9	69.0	1,255	5	3	16	41	71.0	1,768	8	4	74
10	149.0	3,600	10	5	17	42	83.0	1,503	6	3	14
11	46.0	864	5	3	37	43	90.0	1,736	7	3	16
12	38.0	720	4	2	41	44	83.0	1,695	6	3	12
13	49.5	1,008	6	3	35	45	115.0	2,186	8	4	12
14	105.0	1,950	8	3	52	46	50.0	888	5	2	34
15	152.5	2,086	7	3	12	47	55.2	1,120	6	3	29
16	85.0	2,011	9	4	76	48	61.0	1,400	5	3	33
17	60.0	1,465	6	3	102	49	147.0	2,165	7	3	2
18	58.5	1,232	5	2	69	50	210.0	2,353	8	4	15
19	101.0	1,736	7	3	67	51	60.0	1,536	6	3	36
20	79.4	1,296	6	3	11	52	100.0	1,972	8	3	37
21	125.0	1,996	7	3	9	53	44.5	1,120	5	3	27
22	87.9	1,874	5	2	14	54	55.0	1,664	7	3	79
23	80.0	1,580	5	3	11	55	53.4	925	5	3	20
24	94.0	1,920	5	3	14	56	65.0	1,288	5	3	2
25	74.0	1,430	9	3	16	57	73.0	1,400	5	3	2
26	69.0	1,486	6	3	27	58	40.0	1,376	6	3	103
27	63.0	1,008	5	2	35	59	141.0	2,038	12	4	62
28	67.5	1,282	5	3	20	60	68.0	1,572	6	3	29
29	35.0	1,134	5	2	74	61	139.0	1,545	6	3	9
30	142.5	2,400	9	4	15	62	140.0	1,993	6	3	4
31	92.2	1,701	5	3	15	63	55.0	1,130	5	2	21
32	56.0	1,020	6	3	16						

TABLE 15.18 The Least Squares Point Estimates for Exercise 15.51

$b_0 = 10.3676$ (.3710)
$b_1 = .0500$ (<.001)
$b_2 = 6.3218$ (.0152)
$b_3 = -11.1032$ (.0635)
$b_4 = -.4319$ (.0002)

specified square footage (1) to include both a (smaller) living room and family room rather than a (larger) great room and (2) to not increase the number of bedrooms (at the cost of another type of room) that would normally be included in a house of the specified square footage. Note: Based on the statistical results, the builder realized that there are many families with children in a college town and that the parents in such families would rather have one living area for the children (the family room) and a separate living area for themselves (the living room). The builder started modifying his open-space homes accordingly and greatly increased his profits.

15.52 THE FRESH DETERGENT CASE DS Fresh3

Recall from Exercise 15.32 (page 590) that Enterprise Industries has advertised Fresh liquid laundry detergent by using three different advertising campaigns—advertising campaign A (television commercials), advertising campaign B (a balanced mixture of television and radio commercials) and advertising campaign C (a balanced mixture of television, radio, newspaper, and magazine ads). To compare the effectiveness of these advertising campaigns, consider using two models, Model 1 and Model 2, that are shown with corresponding partial Excel outputs in Figure 15.37. In these models y is demand for Fresh; x_4 is the price difference; x_3 is Enterprise Industries' advertising expenditure for Fresh; D_A equals 1 if advertising campaign A is used in a sales period and 0 otherwise; D_B equals 1 if advertising campaign B is used in a sales period and 0 otherwise; and D_C equals 1 if advertising campaign C is used in a sales period and 0 otherwise. Moreover, in Model 1 the parameter β_5 represents the effect on mean demand of advertising campaign B compared to advertising campaign A, and the parameter β_6 represents the effect on mean demand of advertising campaign C compared to advertising campaign A. In Model 2 the

FIGURE 15.37 Excel Output for the Fresh Detergent Case

(a) Partial Excel output for the model $y = \beta_0 + \beta_1 x_4 + \beta_2 x_3 + \beta_3 x_3^2 + \beta_4 x_4 x_3 + \beta_5 D_B + \beta_6 D_C + \varepsilon$

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	25.612696	4.7938	5.3429	2.00E-05	15.6960	35.5294
X4	9.0587	3.0317	2.9880	0.0066	2.7871	15.3302
X3	−6.5377	1.5814	−4.1342	0.0004	−9.8090	−3.2664
X3SQ	0.5844	0.1299	4.5001	0.0002	0.3158	0.8531
X4X3	−1.1565	0.4557	−2.5376	0.0184	−2.0992	−0.2137
DB	0.2137	0.0622	3.4380	0.0022	0.0851	0.3423
DC	0.3818	0.0613	6.2328	2.33E-06	0.2551	0.5085

(b) Partial Excel output for the model $y = \beta_0 + \beta_1 x_4 + \beta_2 x_3 + \beta_3 x_3^2 + \beta_4 x_4 x_3 + \beta_5 D_A + \beta_6 D_C + \varepsilon$

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	25.8264	4.7946	5.3866	1.80E-05	15.9081	35.7447
X4	9.05868	3.0317	2.9880	0.0066	2.7871	15.3302
X3	−6.5377	1.5814	−4.1342	0.0004	−9.8090	−3.2664
X3SQ	0.58444	0.1299	4.5001	0.0002	0.3158	0.8531
X4X3	−1.1565	0.4557	−2.5376	0.0184	−2.0992	−0.2137
DA	−0.2137	0.0622	−3.4380	0.0022	−0.3423	−0.0851
DC	0.16809	0.0637	2.6385	0.0147	0.0363	0.2999

parameter β_6 represents the effect on mean demand of advertising campaign C compared to advertising campaign B.

a Compare advertising compaigns A, B, and C by finding 95 percent confidence intervals for (1) β_5 and β_6 in Model 1 and (2) β_6 in Model 2. Interpret the intervals.

b Using Model 1 or Model 2, a point prediction of Fresh demand when $x_4 = .20$, $x_3 = 6.50$, and campaign C will be used is 8.50068 (that is, 850,068 bottles). Show (by hand calculation) that Model 1 and Model 2 give the same point prediction.

c Consider the alternative model

$$y = \beta_0 + \beta_1 x_4 + \beta_2 x_3 + \beta_3 x_3^2 + \beta_4 x_4 x_3 + \beta_5 D_B + \beta_6 D_C + \beta_7 x_3 D_B + \beta_8 x_3 D_C + \varepsilon$$

which we will call Model 3. The Excel output of the least squares point estimates of the parameters of this model is as follows:

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	28.6873	5.1285	5.5937	1.5E-05	18.0221	39.3526
X4	10.8253	3.2988	3.2816	0.0036	3.9651	17.6855
X3	−7.4115	1.6617	−4.4602	0.0002	−10.8671	−3.9558
X3SQ	0.6458	0.1346	4.7984	9.66E-05	0.3659	0.9257
X4X3	−1.4156	0.4929	−2.8722	0.00912	−2.4406	−0.3907
DB	−0.4807	0.7309	−0.6577	0.517904	−2.0007	1.0393
DC	−0.9351	0.8357	−1.1189	0.2758	−2.6731	0.8029
X3DB	0.10722	0.1117	0.9600	0.3480	−0.1251	0.3395
X3DC	0.20349	0.1288	1.5797	0.1291	−0.0644	0.4714

Let $\mu_{[d,a,A]}$, $\mu_{[d,a,B]}$, and $\mu_{[d,a,C]}$ denote the mean demands for Fresh when the price difference is d, the advertising expenditure is a, and we use advertising campaigns A, B, and C, respectively. The model of this part implies that

$$\mu_{[d,a,A]} = \beta_0 + \beta_1 d + \beta_2 a + \beta_3 a^2 + \beta_4 da + \beta_5(0) + \beta_6(0) + \beta_7 a(0) + \beta_8 a(0)$$

$$\mu_{[d,a,B]} = \beta_0 + \beta_1 d + \beta_2 a + \beta_3 a^2 + \beta_4 da + \beta_5(1) + \beta_6(0) + \beta_7 a(1) + \beta_8 a(0)$$

$$\mu_{[d,a,C]} = \beta_0 + \beta_1 d + \beta_2 a + \beta_3 a^2 + \beta_4 da + \beta_5(0) + \beta_6(1) + \beta_7 a(0) + \beta_8 a(1)$$

(1) Using these equations, verify that $\mu_{[d,a,C]} - \mu_{[d,a,A]}$ equals $\beta_6 + \beta_8 a$. **(2)** Using the least squares point estimates, show that a point estimate of $\mu_{[d,a,C]} - \mu_{[d,a,A]}$ equals .3266 when $a = 6.2$ and equals .4080 when $a = 6.6$. **(3)** Verify that $\mu_{[d,a,C]} - \mu_{[d,a,B]}$ equals $\beta_6 - \beta_5 + \beta_8 a - \beta_7 a$. **(4)** Using the least squares point estimates, show that a point estimate of $\mu_{[d,a,C]} - \mu_{[d,a,B]}$ equals .14266 when $a = 6.2$ and equals .18118 when $a = 6.6$. **(5)** Discuss

why these results imply that the larger that advertising expenditure a is, then the larger is the improvement in mean sales that is obtained by using advertising campaign C rather than advertising campaign A or B.

d If we use an Excel add-in (MegaStat), we can use Models 1, 2, and 3 to predict demand for Fresh in a future sales period when the price difference will be $x_4 = .20$, the advertising expenditure will be $x_3 = 6.50$, and campaign C will be used. The prediction results using Model 1 or Model 2 are as follows:

	95% Confidence Interval		95% Prediction Interval		
Predicted	lower	upper	lower	upper	Leverage
8.50068	8.40370	8.59765	8.21322	8.78813	0.128

The prediction results using Model 3 are as follows:

	95% Confidence Interval		95% Prediction Interval		
Predicted	lower	upper	lower	upper	Leverage
8.51183	8.41229	8.61136	8.22486	8.79879	0.137

e Which model gives the shortest 95 percent prediction interval for Fresh demand?

f Using all of the results in this exercise, discuss why there might be a small amount of interaction between advertising expenditure and advertising campaign.

15.53 THE FRESH DETERGENT CASE Fresh3

The unexplained variation for Model 1 of the previous exercise

$$y = \beta_0 + \beta_1 x_4 + \beta_2 x_3 + \beta_3 x_3^2 + \beta_4 x_4 x_3 + \beta_5 D_B + \beta_6 D_C + \varepsilon$$

is .3936. If we set both β_5 and β_6 in this model equal to 0 (that is, if we eliminate the dummy variable portion of this model), the resulting reduced model has an unexplained variation of 1.0644. Using an α of .05, perform a partial F-test (see page 604) of H_0: $\beta_5 = \beta_6 = 0$. (Hint: $n = 30$, $k = 6$, and $k^* = 2$.) If we reject H_0, we conclude that at least two of advertising campaigns A, B, and C have different effects on mean demand. Many statisticians believe that rejection of H_0 by using the partial F-test makes it more legitimate to make pairwise comparisons of advertising campaigns A, B, and C, as we did in part a of the previous exercise. Here, the partial F-test is regarded as a *preliminary test of significance.*

15.54 In the article "The Effect of Promotion Timing on Major League Baseball Attendance" (*Sport Marketing Quarterly,* December 1999), T. C. Boyd and T. C. Krehbiel use data from six major league baseball teams having outdoor stadiums to study the effect of promotion timing on major league baseball attendance. One of their regression models describes game attendance in 1996 as follows (p-values less than .10 are shown in parentheses under the appropriate independent variables):

$$\begin{aligned}\text{Attendance} = {} & 2{,}521 + 106.5\ \textit{Temperature} + 12.33\ \textit{Winning \%} + .2248\ \textit{OpWin \%} \\ & \qquad (<.001) \qquad\quad (<.001) \qquad\quad (<.001) \\ & - 424.2\ \textit{DayGame} + 4{,}845\ \textit{Weekend} + 1{,}192\ \textit{Rival} + 4{,}745\ \textit{Promotion} \\ & \qquad\quad (<.001) \qquad\quad (<.10) \qquad\quad (<.001) \\ & + 5{,}059\ \textit{Promo*DayGame} - 4{,}690\ \textit{Promo*Weekend} + 696.5\ \textit{Promo*Rival} \\ & \qquad\quad (<.001) \qquad\quad\quad (<.001)\end{aligned}$$

In this model, *Temperature* is the high temperature recorded in the city on game day; *Winning %* is the home team's winning percentage at the start of the game; *OpWin %* is a dummy variable that equals 1 if the opponent's winning percentage was .500 or higher and 0 otherwise; *DayGame* is a dummy variable that equals 1 if the game was a day game and 0 otherwise; *Weekend* is a dummy variable that equals 1 if the game was on a Friday, Saturday, or Sunday and 0 otherwise; *Rival* is a dummy variable that equals 1 if the opponent was a rival and 0 otherwise; *Promotion* is a dummy variable that equals 1 if the home team ran a promotion during the game and 0 otherwise. Using the model, which is based on 475 games and has an R^2 of .6221, Boyd and Krehbiel conclude that "promotions run during day games and on weekdays are likely to result

in greater attendance increases." Carefully explain why the following part of their model justifies this conclusion: + 5,059 *Promo*DayGame* − 4,690 *Promo*Weekend.*

15.55 Table 15.19 in the page margin gives the number of bathrooms for each of the 63 homes in Table 15.17. Using the following MINITAB output of the best single model of each size in terms of $\overline{R}^2$, s, and C, determine which overall model seems best: OxHome2

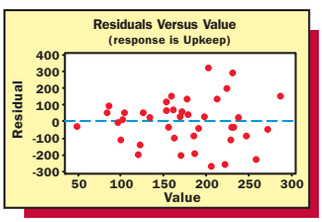
Vars	R-Sq	R-Sq(adj)	Mallows C-p	S	R SQFT	O OMS	B EDS	B AGE	A TH
1	63.3	62.7	18.9	21.382	X				
2	69.1	68.1	8.6	19.782	X			X	
3	70.9	69.4	6.9	19.372	X	X		X	
4	72.6	70.7	5.3	18.962	X	X	X	X	
5	73.2	70.8	6.0	18.917	X	X	X	X	X

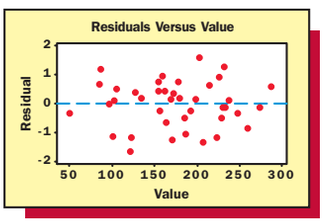
TABLE 15.19
Number of Bathrooms
OxHome2

Bath	Bath
1.0	2.0
1.0	1.5
1.0	1.0
1.0	1.5
1.0	2.0
1.5	1.5
1.5	1.0
2.0	1.5
2.0	1.5
2.5	2.0
1.0	2.5
1.0	2.0
2.0	2.5
1.5	1.0
2.0	1.0
1.5	1.0
1.0	2.0
1.5	2.5
1.0	2.0
1.5	2.0
2.5	1.0
2.0	2.0
1.0	1.0
2.5	1.5
2.0	1.5
2.0	1.0
1.0	2.0
2.0	1.0
1.0	2.0
2.5	2.0
2.0	2.0
1.0	

15.56 THE QHIC CASE

Consider the QHIC data in Figure 14.21 (page 526). When we performed a regression analysis of these data by using the simple linear regression model, plots of the model's residuals versus x (home value) and $\hat{y}$ (predicted upkeep expenditure) both fanned out and had a "dip," or slightly curved appearance (see Figure 14.22, page 528). In order to remedy the indicated violations of the constant variance and correct functional form assumptions, we transformed the dependent variable by taking the square roots of the upkeep expenditures. An alternative approach consists of two steps. First, the slightly curved appearance of the residual plots implies that it is reasonable to add the squared term x^2 to the simple linear regression model. This gives the **quadratic regression model** $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$. The upper residual plot in the MINITAB output that follows shows that a plot of the model's residuals versus x fans out, indicating a violation of the constant variance assumption.





Predictor	Coef	SE Coef	T	P
Noconstant				
1/Value	-53.50	83.20	-0.64	0.524
One	3.409	1.321	2.58	0.014
Value	0.011224	0.004627	2.43	0.020

Predicted Values for New Observations

Fit	95% CI	95% PI
5.635	(5.306, 5.964)	(3.994, 7.276)

$$\frac{\hat{y}}{220} = \frac{-53.50}{220} + 3.409 + .011224(220) = 5.635$$

To remedy this violation, we (in the second step) divide all terms in the quadratic model by x. This gives the transformed model

$$\frac{y}{x} = \beta_0\left(\frac{1}{x}\right) + \beta_1 + \beta_2 x + \frac{\varepsilon}{x}$$

The lower residual plot in the MINITAB output is the residual plot versus x for this transformed model, and the MINITAB output of a regression analysis using the transformed model is given to the right of the residual plots.

a Does the residual plot indicate the constant variance assumption holds for the transformed model?

TABLE 15.20 Quarterly Sales of the TRK-50 Mountain Bike
BikeSales

t	Sales, y_t
1 (Winter)	10
2 (Spring)	31
3 (Summer)	43
4 (Fall)	16
5	11
6	33
7	45
8	17
9	13
10	34
11	48
12	19
13	15
14	37
15	51
16	21

b Consider a home worth \$220,000. We let μ_0 represent the mean yearly upkeep expenditure for all homes worth \$220,000, and we let y_0 represent the yearly upkeep expenditure for an individual home worth \$220,000. **(1)** The bottom of the MINITAB regression output on the previous page tells us that $\hat{y}/220 = 5.635$ is a point estimate of $\mu_0/220$ and a point prediction of $y_0/220$. Multiply this result by 220 to obtain $\hat{y}$. **(2)** Multiply the ends of the confidence interval and prediction interval shown on the MINITAB output by 220. This will give a 95 percent confidence interval for μ_0 and a 95 percent prediction interval for y_0. **(3)** Suppose that QHIC has decided to send a special, more expensive advertising brochure to any home whose value makes QHIC 95 percent confident that the mean upkeep expenditure for all homes having this value is at least \$1,000. Should QHIC send a special brochure to a home worth \$220,000?

15.57 BACKWARD ELIMINATION OxHome2

Backward elimination is an iterative model selection procedure that begins by considering the model that contains all of the potential independent variables and then attempts to remove independent variables one at a time from this model. On each step, an independent variable is removed from the model if it has the largest p-value of any independent variable remaining in the model and if its p-value is greater than α_{stay}, an α value for allowing a variable to stay in the model. Backward elimination terminates when all the p-values for the independent variables remaining in the model are less than α_{stay}. For example, Figure 15.38 shows the MINITAB output of a backward elimination for the residential sales data in Tables 15.17 and 15.19 on pages 616 and 619. Here, α_{stay} has been set equal to .05. Interpret the steps in the output and write out the model that is found by the backward elimination procedure.

15.58 TREND AND SEASONAL PATTERNS BikeSales

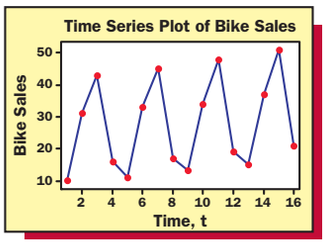


Table 15.20 in the page margin presents quarterly sales of the TRK-50 mountain bike for the previous 16 quarters at a bicycle shop in Switzerland. The time series plot under the sales data shows that the bike sales exhibit an upward *linear trend* (that is, an upward straight line movement over time) and a strong *seasonal pattern*, with bike sales being higher in the spring and summer quarters than in the winter and fall quarters. If we let y_t denote the number of TRK-50 mountain bikes sold in time period t at the Swiss bike shop, then a regression model describing y_t is

$$y_t = \beta_0 + \beta_1 t + \beta_{Q2}Q_2 + \beta_{Q3}Q_3 + \beta_{Q4}Q_4 + \varepsilon_t$$

Here the expression $(\beta_0 + \beta_1 t)$ models the linear trend evident in the time series plot, and Q_2, Q_3, and Q_4 are dummy variables defined for quarters 2, 3, and 4. Specifically, Q_2 equals 1 if quarterly bike sales were observed in quarter 2 (spring) and 0 otherwise; Q_3 equals 1 if quarterly bike sales

FIGURE 15.38 MINITAB Output of the Residential Sales Data Backward Elimination

Step	1	2	3	4
Constant	2.27061	10.36762	-0.08788	10.31302
Sqfeet	0.0448	0.0500	0.0427	0.0530
T-Value	4.82	6.17	5.87	10.66
P-Value	0.000	0.000	0.000	0.000
Rooms	5.8	6.3	4.5	
T-Value	2.26	2.50	1.89	
P-Value	0.027	0.015	0.064	
Bdrms	-9.4	-11.1		
T-Value	-1.56	-1.89		
P-Value	0.124	0.063		
Age	-0.37	-0.43	-0.43	-0.36
T-Value	-3.01	-3.94	-3.87	-3.36
P-Value	0.004	0.000	0.000	0.001
Bathrms	7.7			
T-Value	1.13			
P-Value	0.264			
S	18.9	19.0	19.4	19.8
R-Sq	73.17	72.57	70.88	69.12
R-Sq(adj)	70.82	70.68	69.40	68.09
Mallows Cp	6.0	5.3	6.9	8.6

FIGURE 15.39 MINITAB Output of the Quarterly Bike Sales Regression Analysis

```
The regression equation is
BikeSales = 8.75 + 0.500 Time + 21.0 Q2
                 + 33.5 Q3 + 4.50 Q4

Predictor     Coef  SE Coef      T      P
Constant    8.7500   0.4281  20.44  0.000
Time       0.50000  0.03769  13.27  0.000
Q2         21.0000   0.4782  43.91  0.000
Q3         33.5000   0.4827  69.41  0.000
Q4          4.5000   0.4900   9.18  0.000

S = 0.674200   R-Sq = 99.8%

               R-Sq(adj) = 99.8%

Predicted Values for New Observations
New Obs     Fit  SE Fit        95%  PI
      1  17.250   0.506  (15.395, 19.105)
      2  38.750   0.506  (36.895, 40.605)
      3  51.750   0.506  (49.895, 53.605)
      4  23.250   0.506  (21.395, 25.105)
```

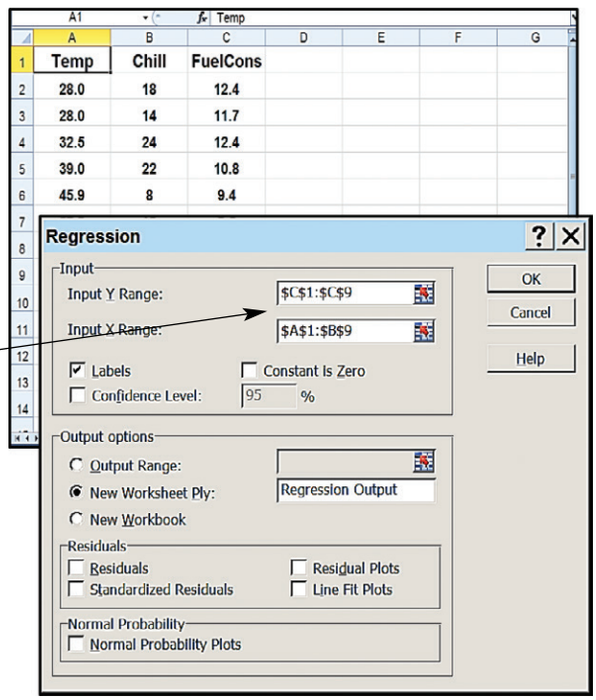
were observed in quarter 3 (summer) and 0 otherwise; Q_4 equals 1 if quarterly bike sales were observed in quarter 4 (fall) and 0 otherwise. Note that we have not defined a dummy variable for quarter 1 (winter). It follows that the regression parameters β_{Q2}, β_{Q3}, and β_{Q4} compare quarters 2, 3, and 4 with quarter 1. Intuitively, for example, β_{Q4} is the difference, excluding trend, between the level of the time series in quarter 4 (fall) and the level of the time series in quarter 1 (winter). A positive β_{Q4} would imply that, excluding trend, bike sales in the fall can be expected to be higher than bike sales in the winter. A negative β_{Q4} would imply that, excluding trend, bike sales in the fall can be expected to be lower than bike sales in the winter.

Figure 15.39 gives the MINITAB output of a regression analysis of the quarterly bike sales by using the dummy variable model. **(1)** Interpret the least squares point estimates $b_{Q2} = 21$, $b_{Q3} = 33.5$, and $b_{Q4} = 4.5$ of β_{Q2}, β_{Q3}, and β_{Q4}. **(2)** Also note that at the bottom of the MINITAB output there are point predictions of and 95 percent prediction intervals for y_{17}, y_{18}, y_{19}, and y_{20}, bike sales in periods 17, 18, 19, and 20 (quarters 1, 2, 3 and 4 of next year). Hand calculate these point predictions.

Appendix 15.1 ■ Multiple Regression Analysis Using Excel

Multiple regression in Figure 15.5(a) on page 562 (data file: GasCon2.xlsx):

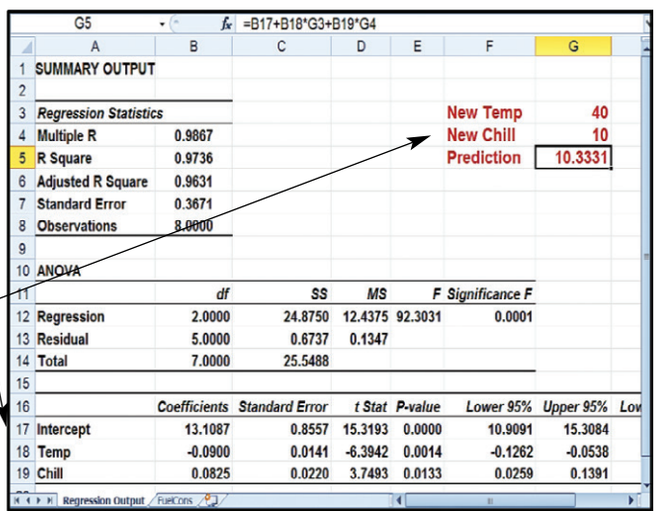
- Enter the gas consumption data from Table 15.3 (page 561)—temperatures (with label Temp) in column A, chill indexes (with label Chill) in column B, and gas consumptions (with label FuelCons) in column C.
- Select **Data : Data Analysis : Regression** and click OK in the Data Analysis dialog box.
- In the Regression dialog box:
  Enter C1 : C9 into the "Input Y Range" window.
  Enter A1 : B9 into the "Input X Range" window.
- Place a checkmark in the Labels checkbox.
- Be sure that the "Constant is Zero" checkbox is NOT checked.
- Select the "New Worksheet Ply" Output option.
- Click OK in the Regression dialog box to obtain the regression output in a new worksheet.



Note: The independent variables must be in adjacent columns because the "Input X Range" must span the range of the values for all of the independent variables.

To compute a point prediction for natural gas consumption when temperature is 40°F and the chill index is 10:

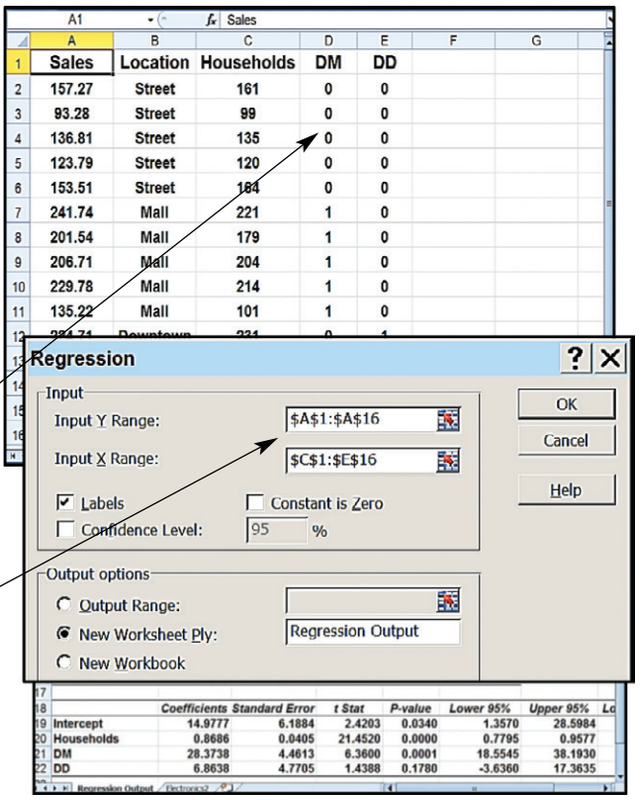
- The Excel Analysis ToolPak does not provide an option for computing point or interval predictions. A point prediction can be computed from the regression results using Excel cell formulas as follows.
- The estimated regression coefficients and their labels are in cells A17:B19 of the output worksheet and the predictor values 40 and 10 have been placed in cells G3 and G4.
- In cell G5, enter the Excel formula
  = B17+B18*G3+B19*G4
  to compute the point prediction (=10.3331).



Sales volume multiple regression with indicator (dummy) variables in Figure 15.15(b) on page 584 (data file: Electronics2.xlsx):

- Enter the sales volume data from Table 15.9 (page 583)—sales volumes (with label Sales) in column A, store locations (with label Location) in column B, and number of households (with label Households) in column C. (The order of the columns is chosen to arrange for an adjacent block of predictor variables.)
- Enter the labels DM and DD in cells D1 and E1.
- Following the definition of the dummy variables DM and DD in Example 15.6 (pages 581 and 582), enter the appropriate values of 0 and 1 for these two variables into columns D and E.
- Select **Data : Data Analysis : Regression** and click OK in the Data Analysis dialog box.
- In the Regression dialog box:

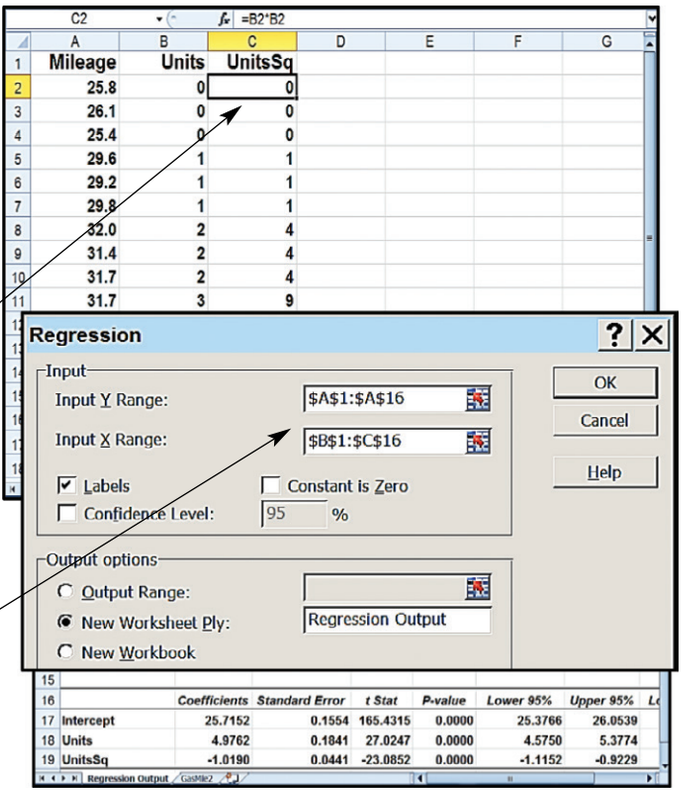
  Enter A1:A16 into the "Input Y Range" window.
  Enter C1:E16 into the "Input X Range" window.
- Place a checkmark in the Labels checkbox.
- Select the "New Worksheet Ply" Output option.
- Click OK in the Regression dialog box to obtain the regression results in a new worksheet.



Gasoline additive multiple linear regression with a quadratic term similar to Figure 15.23 on page 593 (data file: GasAdd.xlsx):

- Enter the gas mileage data from Table 15.13 (page 593)—mileages (with label Mileage) in column A and units of additive (with label Units) in column B. (Units are listed second in order to be adjacent to the squared units predictor.)
- Enter UnitsSq into cell C1.
- Click on cell C2, and enter the formula =B2*B2. Press "Enter" to compute the squared value of Units for the first observation.
- Copy the cell formula of C2 through cell C16 (by double-clicking the drag handle in the lower right corner of cell C2) to compute the squared units for the remaining observations.
- Select **Data : Data Analysis : Regression** and click OK in the Data Analysis dialog box.
- In the Regression dialog box:

  Enter A1:A16 into the "Input Y Range" window.
  Enter B1:C16 into the "Input X Range" window.
- Place a checkmark in the Labels checkbox.
- Select the "New Worksheet Ply" Output option.
- Click OK in the Regression dialog box to obtain the regression output in a new worksheet.



Appendix 15.2 ■ Multiple Regression Analysis Using MegaStat

Multiple regression similar to Figure 15.5 on page 562 (data file: GasCon2.xlsx):

- Enter the gas consumption data in Table 15.3 (page 561) as shown—temperature (with label Temp) in column A, chill index (with label Chill) in column B, and gas consumption (with label FuelCons) in column C. Note that Temp and Chill are contiguous columns (that is, they are next to each other). This is not necessary, but it makes selection of the independent variables (as described below) easiest.
- Select **Add-Ins : MegaStat : Correlation/ Regression : Regression Analysis**
- In the Regression Analysis dialog box, click in the Independent Variables window and use the AutoExpand feature to enter the range A1:B9. Note that if the independent variables are not next to each other, hold the CTRL key down while making selections and then autoexpand.
- Click in the Dependent Variable window and enter the range C1:C9.
- Check the appropriate Options and Residuals checkboxes as follows:

  1 Check "Test Intercept" to include a *y*-intercept and to test its significance.

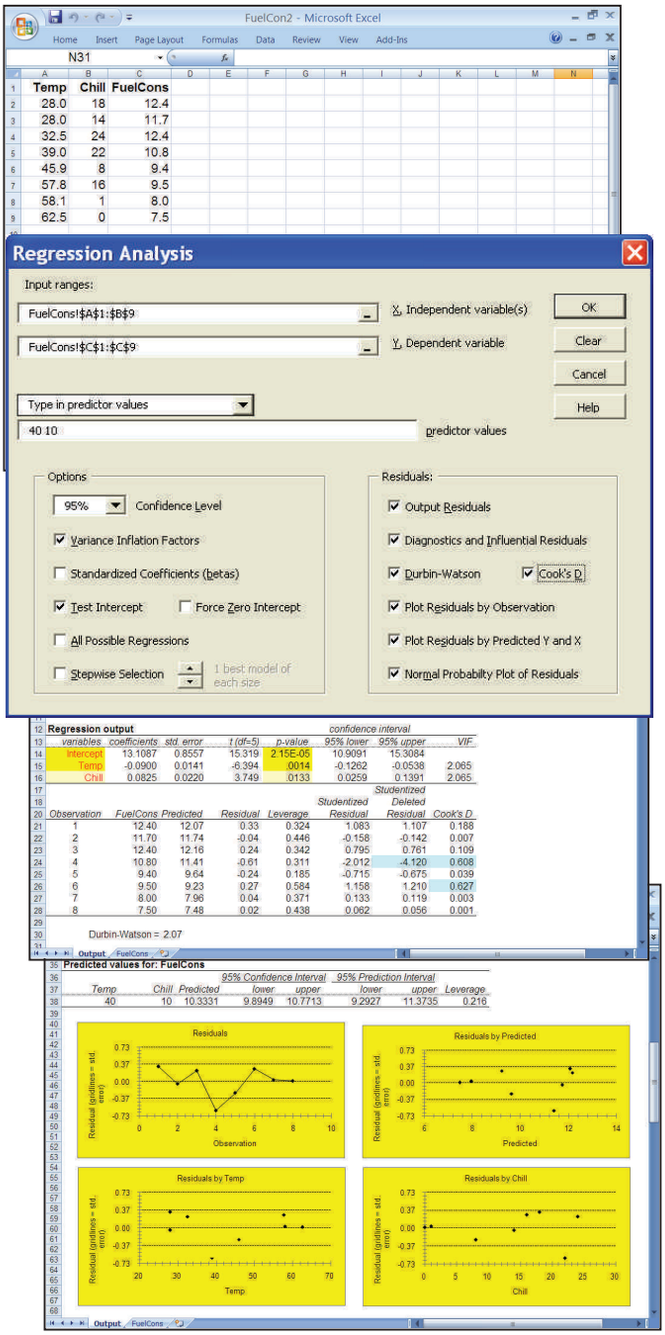
  2 Check "Output Residuals" to obtain a list of the model residuals.

  3 Check "Plot Residuals by Observation" and "Plot Residuals by Predicted Y and X" to obtain residual plots versus time, versus the predicted values of *y*, and versus the values of each independent variable (see Section 15.11).

  4 Check "Normal Probability Plot of Residuals" to obtain a normal plot (see Section 14.9).

  5 Check "Diagnostics and Influential Residuals" to obtain diagnostics (see Section 15.11).

  6 Check "Durbin-Watson" to obtain the Durbin–Watson statistic (see Section 14.9) and check "Variance Inflation Factors" (see Section 15.10).



To obtain a **point prediction** of *y* when temperature equals 40 and chill index equals 10 (as well as a confidence interval and prediction interval):

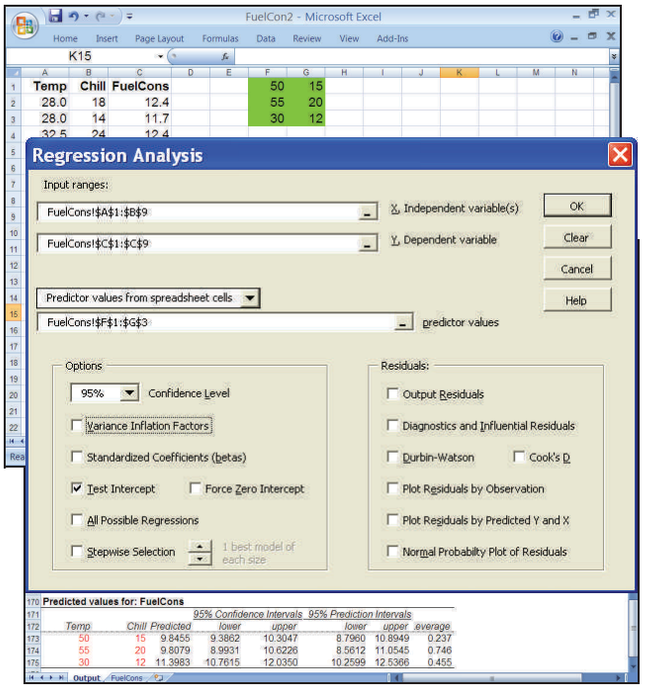
- Click on the drop-down menu above the Predictor Values window and select "Type in predictor values."
- Type 40 and 10 (separated by at least one blank space) into the Predictor Values window.

(*Continues across page*)

- Select a desired level of confidence (here 95%) from the Confidence Level drop-down menu or type in a value.
- Click OK in the Regression Analysis dialog box.

Predictions can also be obtained by placing the values of the predictor variables into spreadsheet cells. For example, suppose that we wish to compute predictions of *y* for each of the following three temperature—chill index combinations: 50 and 15; 55 and 20; 30 and 12. To do this:

- Enter the values for which predictions are desired in spreadsheet cells as illustrated in the screenshot—here temperatures are entered in column F and chill indexes are entered in column G. However, the values could be entered in any contiguous columns.
- In the drop-down menu above the Predictor Values window, select "Predictor values from spreadsheet cells."
- Select the range of cells containing the Predictor Values (here F1:G3) into the predictor values window.
- Select a desired level of confidence from the Confidence Level drop-down menu or type in a value.
- Click OK in the Regression Analysis dialog box.

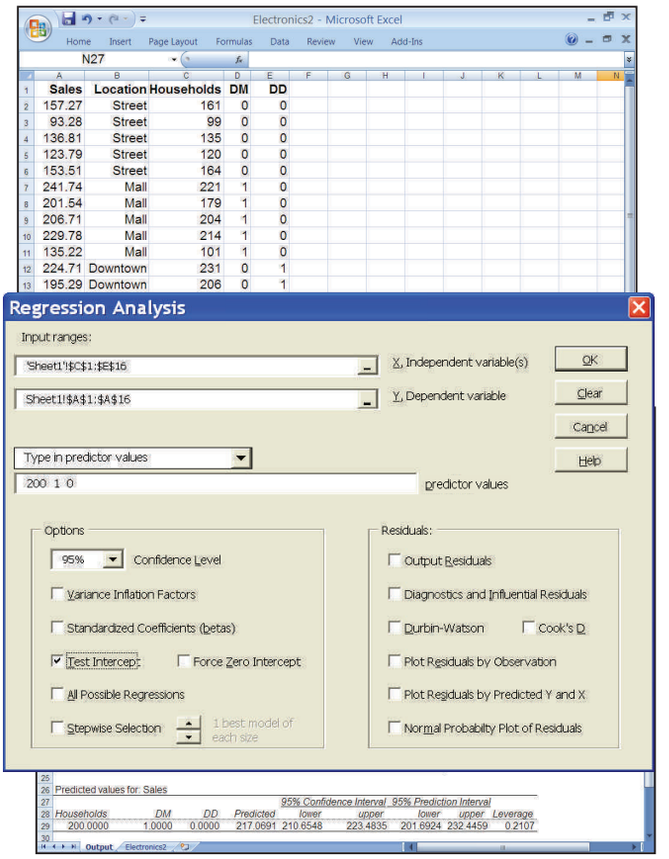


Multiple regression with indicator (dummy) variables similar to Figure 15.15 on page 584 (data file: Electronics2.xlsx):

- Enter the sales volume data from Table 15.9 (page 583)—sales volume (with label Sales) in column A, store location (with label Location) in column B, and number of households (with label Households) in column C. Again note that the order of the variables is chosen to allow for a contiguous block of predictor variables.
- Enter the labels DM and DD into cells D1 and E1.
- Following the definitions of the dummy variables DM and DD in Example 15.6 (pages 581 and 582), enter the appropriate values of 0 and 1 for these two variables into columns D and E as shown in the screen.
- Select **Add-Ins : MegaStat : Correlation/ Regression : Regression Analysis.**
- In the Regression Analysis dialog box, click in the Independent Variables window and use the autoexpand feature to enter the range C1:E16.
- Click in the Dependent Variable window and enter the range A1:A16.

To compute a prediction of sales volume for 200,000 households and a mall location:

- Select "Type in predictor values" from the drop-down menu above the Predictor Values window.
- Type 200 1 0 into the Predictor Values window.
- Select or type a desired level of confidence (here 95%) in the Confidence Level box.
- Click the Options and Residuals checkboxes as shown (or as desired).
- Click OK in the Regression Analysis dialog box.

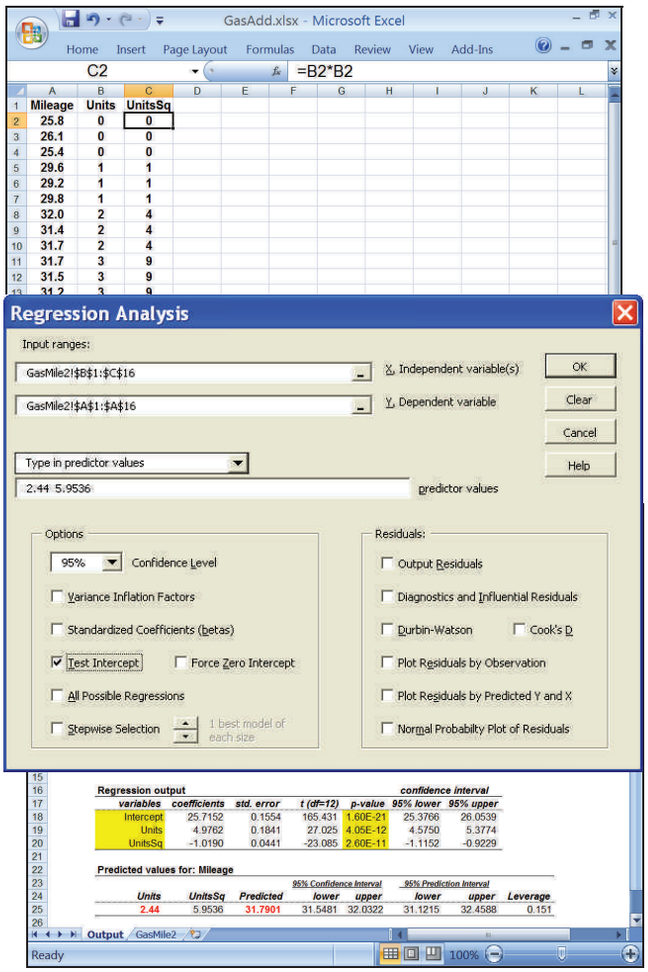


Multiple linear regression with a quadratic term similar to Figure 15.23 on page 593 (data file: GasAdd.xlsx):

- Enter the gasoline additive data from Table 15.13 (page 593)—mileages (with label Mileage) in column A and units of additive (with label Units) in column B.
- Enter the label UnitsSq in cell C1.
- Click on cell C2 and type the cell formula =B2*B2. Press enter to compute the squared value of Units for the first observation.
- Copy the cell formula of C2 through cell C16 (by double-clicking the drag handle in the lower right corner of cell C2) to compute the squared units for the remaining observations.
- Select **Add-Ins : MegaStat : Correlation/ Regression : Regression Analysis.**
- In the Regression Analysis dialog box, click in the Independent Variables window and use the AutoExpand feature to enter the range B1:C16.
- Click in the Dependent Variable window and enter the range A1:A16.

To compute a prediction for mileage when Units equals 2.44:

- Select "Type in predictor values" from the drop-down menu above the Predictor Values window.
- Type 2.44 5.9536 in the Predictor Values window. Note that (2.44)**2=5.9536 must first be hand calculated.
- Select or type the desired level of confidence (here 95%) in the Confidence Level box.
- Check the Options and Residuals checkboxes as desired.
- Click OK in the Regression Analysis dialog box.

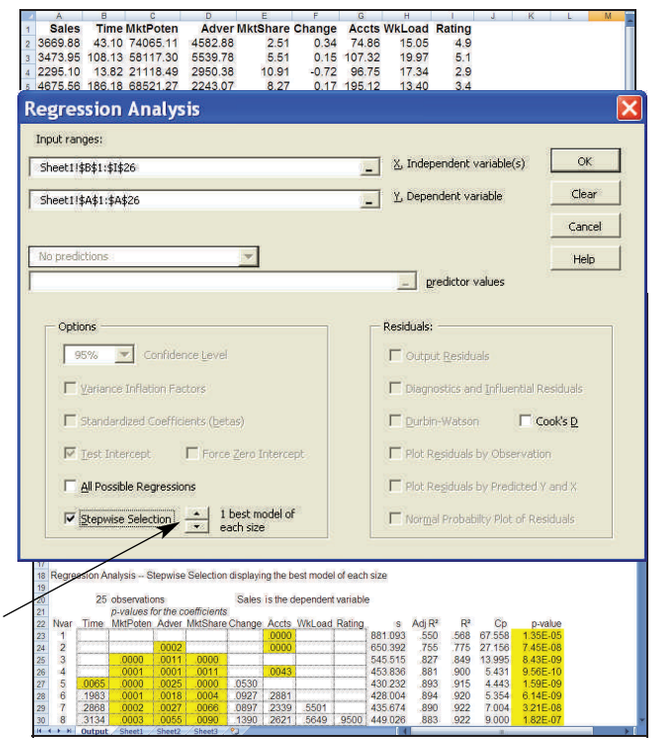


Stepwise selection similar to Figure 15.27 on page 601 (data file: SalePerf2.xlsx):

- Enter the sales performance data in Figure 15.10 (page 579) and Table 15.14 (page 598) into columns A through I with labels as shown in the screen.
- Select **Add-Ins : MegaStat : Correlation/ Regression : Regression Analysis.**
- In the Regression Analysis dialog box, click in the Independent Variables window and use the AutoExpand feature to enter the range B1:I26.
- Click in the Dependent Variable window and use the AutoExpand feature to enter the range A1:A26.
- Check the "Stepwise Selection" checkbox.
- Click OK in the Regression Analysis dialog box.

Stepwise selection will give the best model of each size (1, 2, 3, etc. independent variables). The default gives one model of each size. For more models, use the arrow buttons to request the desired number of models of each size.

- Check the "All Possible Regressions" checkbox to obtain the results for all possible regressions. This option will handle up to 12 independent variables.



Appendix 15.3 ■ Multiple Regression Analysis Using MINITAB

Multiple regression in Figure 15.5(b) on page 562 (data file: GasCon2.MTW):

- In the Data window, enter the gas consumption data from Table 15.3 (page 561)—the average hourly temperatures in column C1 with variable name Temp, the chill indexes in column C2 with variable name Chill, and the weekly gas consumptions in column C3 with variable name FuelCons.
- Select **Stat : Regression : Regression.**
- In the Regression dialog box, select FuelCons into the Response window.
- Select Temp and Chill into the Predictors window.

To compute a prediction for gas consumption when the temperature is 40° F and the chill index is 10:

- In the Regression dialog box, click on the Options . . . button.
- In the "Regression Options" dialog box, enter 40 and 10 into the "Prediction intervals for new observations" window. (The number and order of values in this window must match the Predictors list in the Regression dialog box.)
- Click OK in the Regression—Options dialog box.

To obtain residual plots:

- Click on the Graphs . . . button, select the desired plots (see Appendix 14.3), and click OK in the Regression—Graphs dialog box.

To obtain variance inflation factors (VIF):

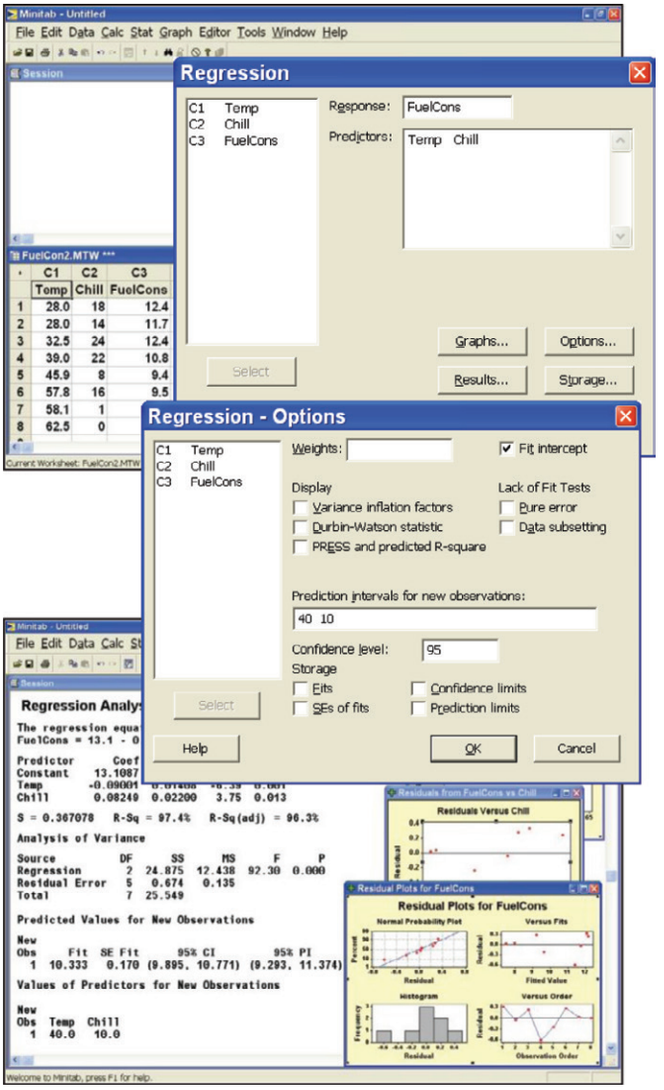
- Place a checkmark in the Variance inflation factors checkbox in the Regression—Options dialog box.

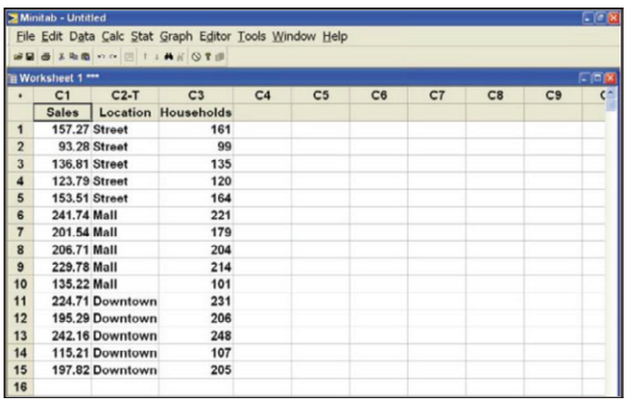
To obtain outlying and influential observation diagnostics:

- Click the Storage button, and in the Regression—Storage dialog box, place checkmarks in the following checkboxes: Fits (for predicted values), Residuals, Standardized residuals (for studentized residuals), Deleted *t* residuals (for studentized deleted residuals), Hi (for leverages), and Cook's distance. Click OK in the Regression—Storage dialog box.
- Click OK in the Regression dialog box.

Multiple regression with indicator (dummy) variables in Figure 15.15(a) on page 584 (data file: Electronics2.MTW):

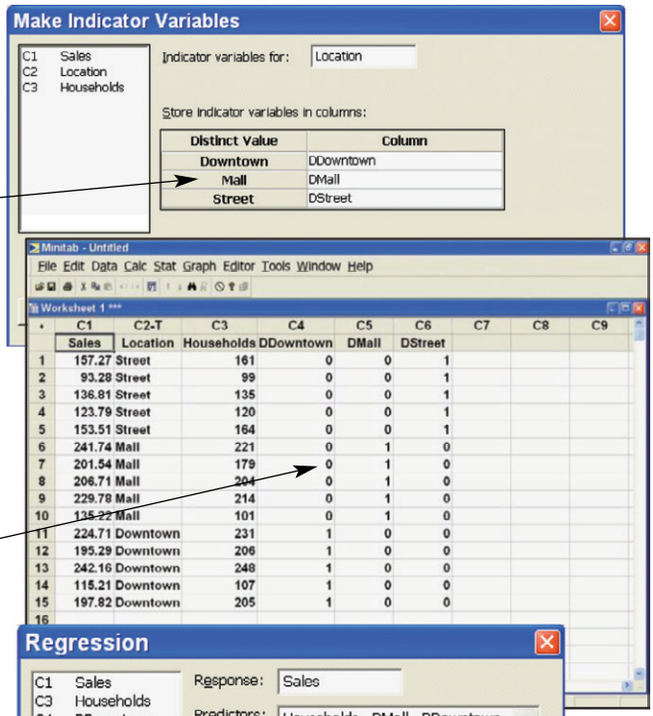
- In the Data window, enter the sales volume data from Table 15.9 on page 583 with sales volume in column C1, location in column C2, and number of households in column C3 with variable names Sales, Location, and Households.





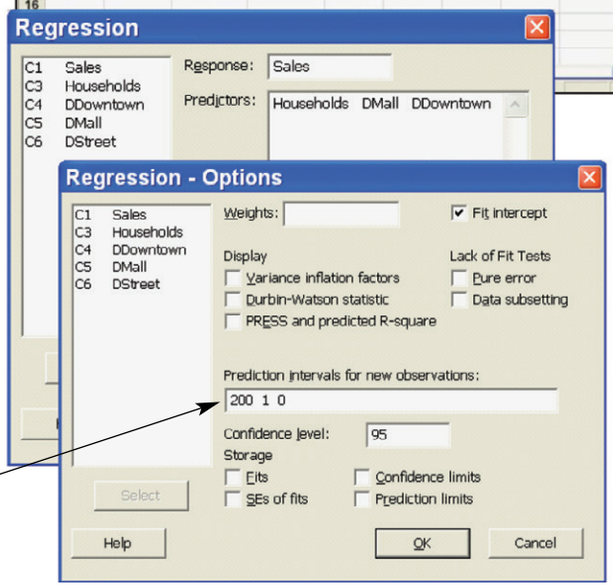
To create indicator/dummy variable predictors:

- Select **Calc : Make Indicator Variables.**
- In the "Make Indicator Variables" dialog box, enter Location into the "Indicator variables for" window.
- The "Store indicator variables in columns" window lists the distinct values of Location in alphabetical order. Corresponding to each distinct value, enter the variable name to be used for that value's indicator variable—here we have used the names DDowntown, DMall, and DStreet (or you can use default names that are supplied by MINITAB if you wish). The first indicator variable (DDowntown) will have 1's in all rows where the Location equals Downtown and 0's elsewhere. The second indicator variable (DMall) will have 1's in all rows where Location equals Mall and 0's elsewhere. The third indicator variable (DStreet) will have 1's in all rows where Location equals Street and 0's elsewhere.
- Click OK in the "Make Indicator Variables" dialog box to create the indicator variables in the Data window.



To fit the multiple regression model:

- Select **Stat : Regression : Regression.**
- In the Regression dialog box, select Sales into the Response window.
- Select Households DMall DDowntown into the Predictors window.

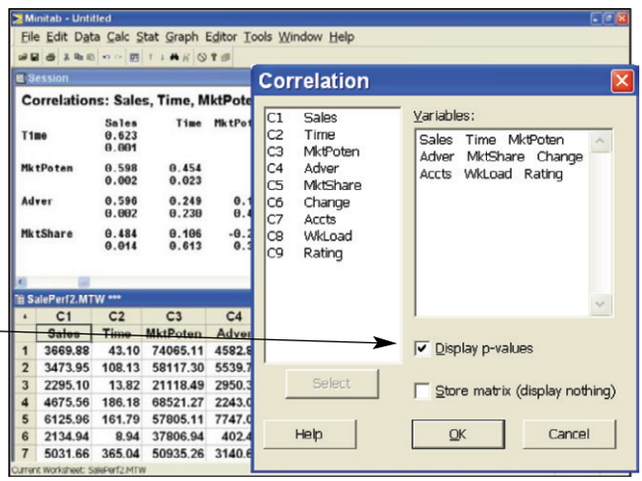


To compute a prediction of sales volume for 200,000 households and a mall location:

- Click on the Options . . . button.
- In the "Regression—Options" dialog box, type 200 1 0 into the "Prediction intervals for new observations" window.
- Click OK in the "Regression—Options" dialog box.
- Click OK in the Regression dialog box.

Correlation matrix in Figure 15.26(a) on page 598 (data file: SalePerf2.MTW):

- In the Data window, enter the sales representative performance data from Figure 15.10 (page 579) and Table 15.14 (page 598) into columns C1–C9 with variable names Sales, Time, MktPoten, Adver, MktShare, Change, Accts, WkLoad, and Rating.
- Select **Stat : Basic Statistics : Correlation.**
- In the Correlation dialog box, enter all variable names into the Variables window.
- If *p*-values are desired, make sure that the "Display *p*-values" checkbox is checked.
- Click OK in the Correlation dialog box.



Multiple linear regression with a quadratic term in Figure 15.23 on page 593 (data file: GasAdd.MTW):

- In the Data window, enter the gasoline mileage data from Table 15.13 (page 593)—mileages in column C1 with variable name Mileage and units of additive in column C2 with variable name Units.

To compute the **quadratic predictor**, UnitsSq:

- Select **Calc : Calculator.**
- In the Calculator dialog box, enter UnitsSq in the "Store result in variable" box.
- Enter Units*Units in the Expression window.
- Click OK in the Calculator dialog box to obtain the squared values in column C3 with variable name UnitsSq.

To fit the quadratic regression model:

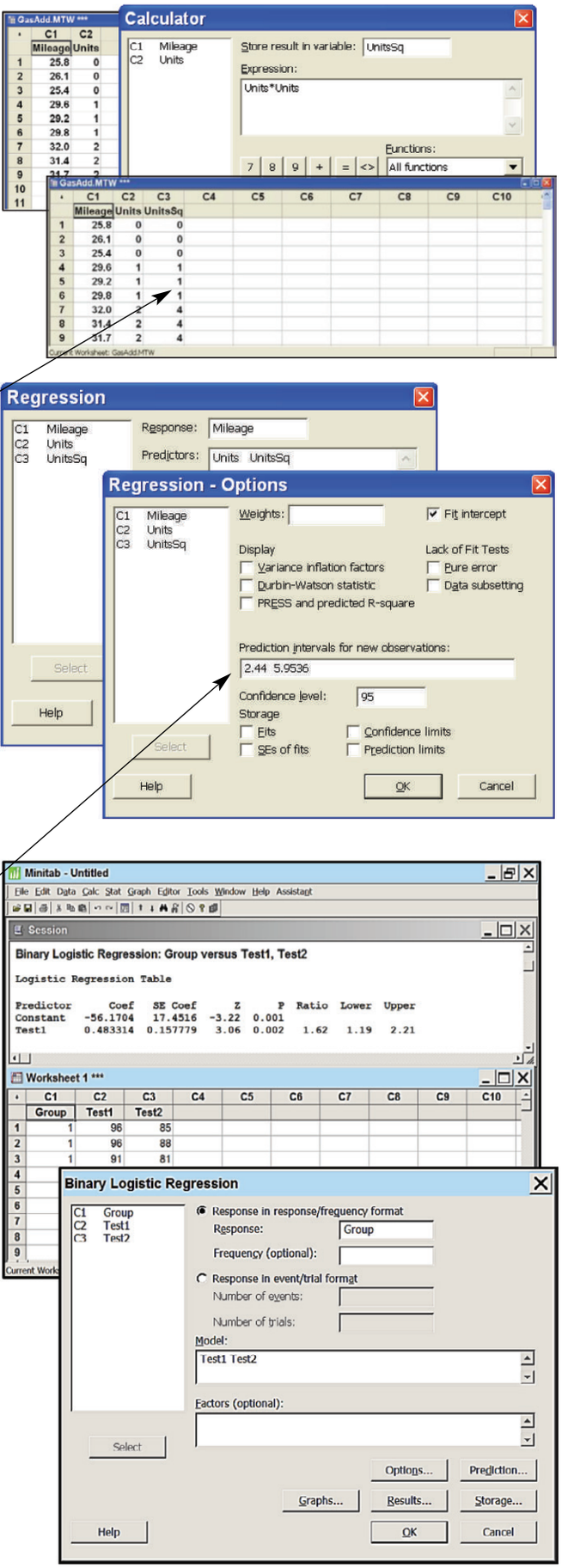
- Select **Stat : Regression : Regression.**
- In the Regression dialog box, select Mileage into the Response window.
- Select Units and UnitsSq into the Predictors window.
- Click OK in the Regression dialog box.

To compute a **prediction** for mileage when 2.44 units of additive are used:

- Click on the Options . . . button.
- In the Regression—Options dialog box, type 2.44 and 5.9536 into the "Prediction intervals for new observations" window. [$(2.44)^2 = 5.9536$ must first be calculated by hand.]
- Click OK in the Regression—Options dialog box.
- Click OK in the Regression dialog box.

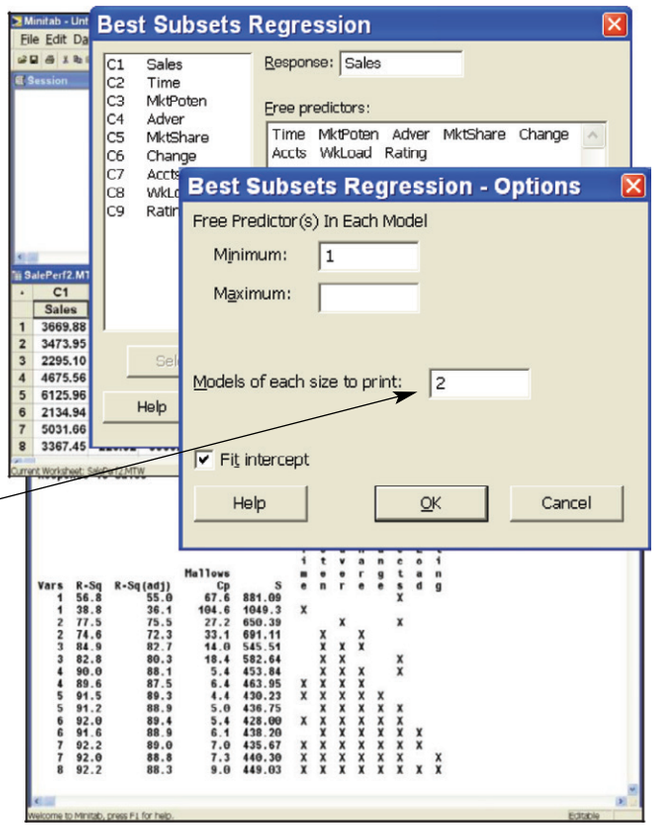
Logistic regression in Figure 15.36 on page 613 (data file: PerfTest.MTW):

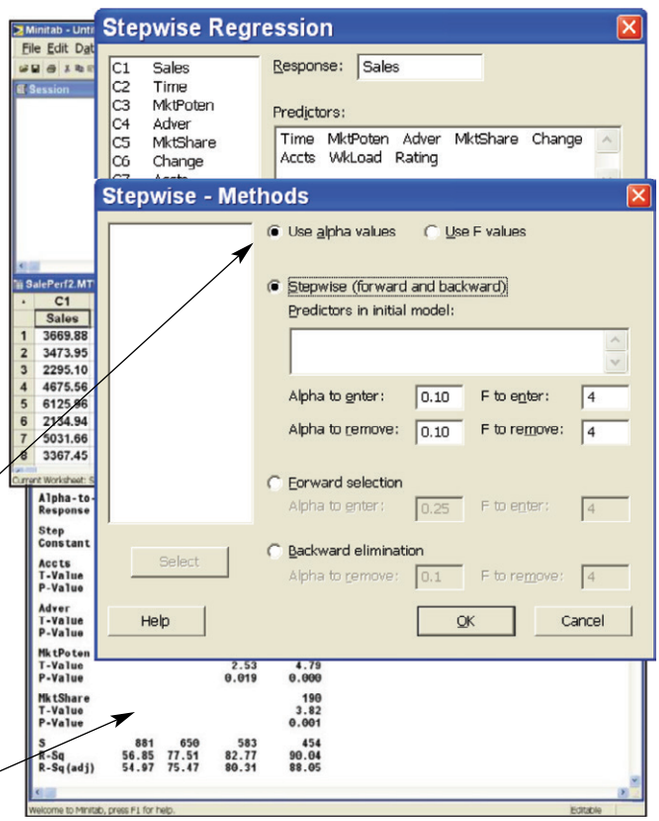
- In the data window, enter the performance data in Table 15.16 on page 612—Group (either 1 or 0) in column C1 with variable name Group, the score on test 1 in column C2 with variable name Test1, and the score on test 2 in column C3 with variable name Test2.
- Select **Stat : Regression : Binary Logistic Regression.**
- In the "Binary Logistic Regression" dialog box, enter Group into the Response window.
- Enter Test1 and Test2 into the model window.
- Click OK in the "Binary Logistic Regression" dialog box.



Best subsets regression in Figure 15.27 on page 601 (data file: SalePerf2.MTW):

- In the Data window, enter the sales territory performance data from Figure 15.10 (page 579) and Table 15.14 (page 598) into columns C1–C9 with variable names Sales, Time, MktPoten, Adver, MktShare, Change, Accts, WkLoad, and Rating.
- Select **Stat : Regression : Best Subsets.**
- In the Best Subsets Regression dialog box, enter Sales into the Response window.
- Enter the remaining variable names into the "Free predictors" window.
- Click on the Options . . . button.
- In the "Best Subsets Regression—Options" dialog box, enter 2 in the "Models of each size to print" window.
- Click OK in the "Best Subsets Regression—Options" dialog box.
- Click OK in the Best Subsets Regression dialog box.



Stepwise regression in Figure 15.28 on page 603 (data file: SalePerf2.MTW):

- In the Data window, enter the sales territory performance data from Figure 15.10 (page 579) and Table 15.14 (page 598) into columns C1–C9 with variable names Sales, Time, MktPoten, Adver, MktShare, Change, Accts, WkLoad, and Rating.
- Select **Stat : Regression : Stepwise.**
- In the Stepwise Regression dialog box, enter Sales into the Response window.
- Enter the remaining variable names into the Predictors window.
- Click on the Methods . . . button.
- In the Stepwise—Methods dialog box, select the "Use alpha values" option.
- Select the "Stepwise (Forward and Backward)" option.
- Enter 0.10 in the "Alpha to enter" and "Alpha to remove" boxes.
- Click OK in the Stepwise—Methods dialog box.
- Click OK in the Stepwise Regression dialog box.
- The results of the stepwise regression are given in the Session window.