

14 Introduction to Multiple Regression

USING STATISTICS @ OmniFoods

14.1 Developing a Multiple Regression Model

Visualizing Multiple Regression Data
Interpreting the Regression Coefficients
Predicting the Dependent Variable Y

14.2 r^2 , Adjusted r^2 , and the Overall F Test

Coefficient of Multiple Determination
Adjusted r^2
Test for the Significance of the Overall Multiple Regression Model

14.3 Residual Analysis for the Multiple Regression Model

14.4 Inferences Concerning the Population Regression Coefficients

Tests of Hypothesis
Confidence Interval Estimation

14.5 Testing Portions of the Multiple Regression Model

Coefficients of Partial Determination

14.6 Using Dummy Variables and Interaction Terms in Regression Models

Dummy Variables
Interactions

14.7 Logistic Regression

USING STATISTICS @ OmniFoods Revisited

CHAPTER 14 EXCEL GUIDE

CHAPTER 14 MINITAB GUIDE

Learning Objectives

In this chapter, you learn:

- How to develop a multiple regression model
- How to interpret the regression coefficients
- How to determine which independent variables to include in the regression model
- How to determine which independent variables are most important in predicting a dependent variable
- How to use categorical independent variables in a regression model
- How to predict a categorical dependent variable using logistic regression



USING STATISTICS

@ OmniFoods

Y

ou are the marketing manager for OmniFoods, a large food products company. The company is planning a nationwide introduction of OmniPower, a new high-energy bar. Originally marketed to runners, mountain climbers, and other athletes, high-energy bars are now popular with the general public. OmniFoods is anxious to capture a share of this thriving market.

Because the marketplace already contains several successful energy bars, you need to develop an effective marketing strategy. In particular, you need to determine the effect that price and in-store promotions will have on sales of OmniPower. Before marketing the bar nationwide, you plan to conduct a test-market study of OmniPower sales, using a sample of 34 stores in a supermarket chain. How can you extend the linear regression methods discussed in Chapter 13 to incorporate the effects of price *and* promotion into the same model? How can you use this model to improve the success of the nationwide introduction of OmniPower?



Chapter 13 focused on simple linear regression models that use *one* numerical independent variable, X , to predict the value of a numerical dependent variable, Y . Often you can make better predictions by using *more than one* independent variable. This chapter introduces you to **multiple regression models** that use two or more independent variables to predict the value of a dependent variable.

14.1 Developing a Multiple Regression Model

The business objective facing the marketing manager at OmniFoods is to develop a model to predict monthly sales volume per store of OmniPower bars and to determine what variables influence sales. Two independent variables are considered here: the price of an OmniPower bar, as measured in cents (X_1), and the monthly budget for in-store promotional expenditures, measured in dollars (X_2). In-store promotional expenditures typically include signs and displays, in-store coupons, and free samples. The dependent variable Y is the number of OmniPower bars sold in a month. Data are collected from a sample of 34 stores in a supermarket chain selected for a test-market study of OmniPower. All the stores selected have approximately the same monthly sales volume. The data are organized and stored in **OmniPower** and presented in Table 14.1.

TABLE 14.1
Monthly OmniPower
Sales, Price, and
Promotional
Expenditures

Store	Sales	Price	Promotion	Store	Sales	Price	Promotion
1	4,141	59	200	18	2,730	79	400
2	3,842	59	200	19	2,618	79	400
3	3,056	59	200	20	4,421	79	400
4	3,519	59	200	21	4,113	79	600
5	4,226	59	400	22	3,746	79	600
6	4,630	59	400	23	3,532	79	600
7	3,507	59	400	24	3,825	79	600
8	3,754	59	400	25	1,096	99	200
9	5,000	59	600	26	761	99	200
10	5,120	59	600	27	2,088	99	200
11	4,011	59	600	28	820	99	200
12	5,015	59	600	29	2,114	99	400
13	1,916	79	200	30	1,882	99	400
14	675	79	200	31	2,159	99	400
15	3,636	79	200	32	1,602	99	400
16	3,224	79	200	33	3,354	99	600
17	2,295	79	400	34	2,927	99	600

Visualizing Multiple Regression Data

With the special case of two independent variables and one dependent variable, you can visualize your data with a three-dimensional scatter plot. Figure 14.1 on page 579 presents a three-dimensional Minitab plot of the OmniPower data. This figure shows the points plotted at a height equal to their sales with drop lines down to their promotion expense and price values.

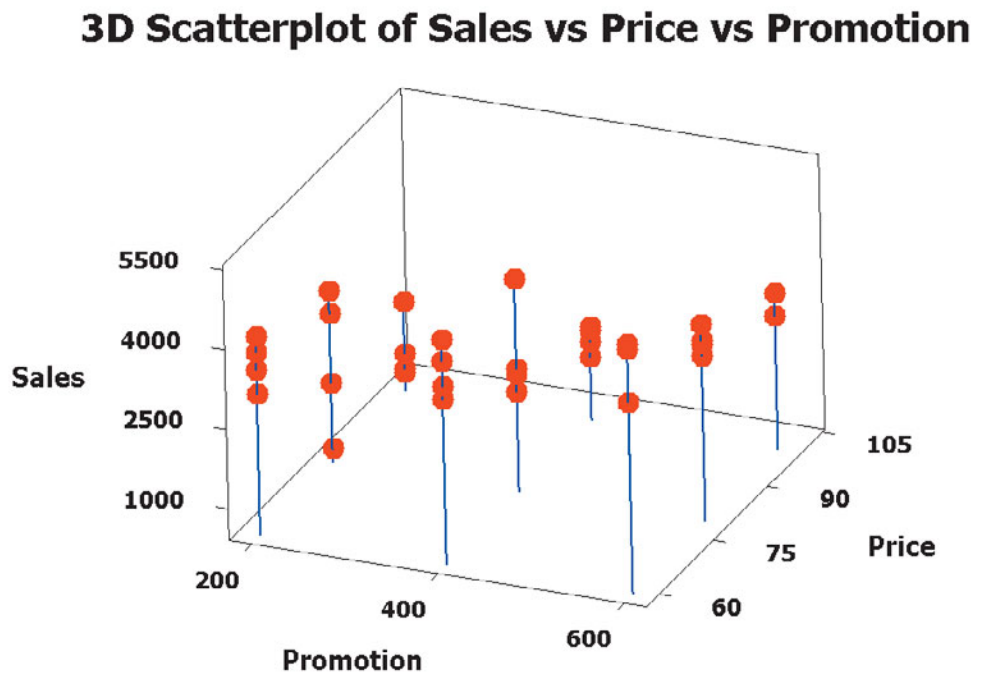
Interpreting the Regression Coefficients

When there are several independent variables, you can extend the simple linear regression model of Equation (13.1) on page 522 by assuming a linear relationship between each independent variable and the dependent variable. For example, with k independent variables, the multiple regression model is expressed in Equation (14.1).

FIGURE 14.1

Minitab three-dimensional plot of monthly OmniPower sales, price, and promotional expenditures

Excel does not include the capability to create three-dimensional scatter plots.



MULTIPLE REGRESSION MODEL WITH k INDEPENDENT VARIABLES

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + \varepsilon_i \quad (14.1)$$

where

β_0 = Y intercept

β_1 = slope of Y with variable X_1 , holding variables X_2, X_3, \dots, X_k constant

β_2 = slope of Y with variable X_2 , holding variables X_1, X_3, \dots, X_k constant

β_3 = slope of Y with variable X_3 , holding variables $X_1, X_2, X_4, \dots, X_k$ constant

\vdots

β_k = slope of Y with variable X_k , holding variables $X_1, X_2, X_3, \dots, X_{k-1}$ constant

ε_i = random error in Y for observation i

Equation (14.2) defines the multiple regression model with two independent variables.

MULTIPLE REGRESSION MODEL WITH TWO INDEPENDENT VARIABLES

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \quad (14.2)$$

where

β_0 = Y intercept

β_1 = slope of Y with variable X_1 , holding variable X_2 constant

β_2 = slope of Y with variable X_2 , holding variable X_1 constant

ε_i = random error in Y for observation i

Compare the multiple regression model to the simple linear regression model [Equation (13.1) on page 522]:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

In the simple linear regression model, the slope, β_1 , represents the change in the mean of Y per unit change in X and does not take into account any other variables. In the multiple regression model with two independent variables [Equation (14.2)], the slope, β_1 , represents the change in the mean of Y per unit change in X_1 , taking into account the effect of X_2 .

As in the case of simple linear regression, you use the least-squares method to compute sample regression coefficients (b_0 , b_1 , and b_2) as estimates of the population parameters (β_0 , β_1 , and β_2). Equation (14.3) defines the regression equation for a multiple regression model with two independent variables.

MULTIPLE REGRESSION EQUATION WITH TWO INDEPENDENT VARIABLES

$$\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i} \quad (14.3)$$

Figure 14.2 shows Excel and Minitab results for the OmniPower sales data multiple regression model. From Figure 14.2, the computed values of the three regression coefficients are

$$b_0 = 5,837.5208 \quad b_1 = -53.2173 \quad b_2 = 3.6131$$

Therefore, the multiple regression equation is

$$\hat{Y}_i = 5,837.5208 - 53.2173 X_{1i} + 3.6131 X_{2i}$$

where

\hat{Y}_i = predicted monthly sales of OmniPower bars for store i

X_{1i} = price of OmniPower bar (in cents) for store i

X_{2i} = monthly in-store promotional expenditures (in dollars) for store i

FIGURE 14.2

Excel and Minitab results for the OmniPower sales data multiple regression model

	A	B	C	D	E	F	G
1	Multiple Regression						
2							
3	Regression Statistics						
4	Multiple R	0.8705					
5	R Square	0.7577					
6	Adjusted R Square	0.7421					
7	Standard Error	638.0653					
8	Observations	34					
9							
10	ANOVA						
11		df	SS	MS	F	Significance F	
12	Regression	2	39472730.7730	19736365.3865	48.4771	0.0000	
13	Residual	31	12620946.6682	407127.3119			
14	Total	33	52093677.4412				
15							
16		Coefficients	Standard error	t Stat	P value	Lower 95%	Upper 95%
17	Intercept	5837.5208	628.1502	9.2932	0.0000	4556.3999	7118.6416
18	Price	-53.2173	6.8522	-7.7664	0.0000	-67.1925	-39.2421
19	Promotion	3.6131	0.6852	5.2728	0.0000	2.2155	5.0106

Regression Analysis: Sales versus Price, Promotion

The regression equation is
Sales = 5838 - 53.2 Price + 3.61 Promotion

Predictor	Coef	SE Coef	T	P
Constant	5837.5	628.2	9.29	0.000
Price	-53.217	6.852	-7.77	0.000
Promotion	3.6131	0.6852	5.27	0.000

S = 638.065 R-Sq = 75.8% R-Sq(adj) = 74.2%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	39472731	19736365	48.48	0.000
Residual Error	31	12620947	407127		
Total	33	52093677			

Predicted Values for New Observations

New Obs	Fit	SE Fit	95% CI	95% PI
1	3079	110	(2854, 3303)	(1758, 4399)

Values of Predictors for New Observations

New Obs	Price	Promotion
1	79.0	400

The sample Y intercept ($b_0 = 5,837.5208$) estimates the number of OmniPower bars sold in a month if the price is \$0.00 and the total amount spent on promotional expenditures is also \$0.00. Because these values of price and promotion are outside the range of price and promotion used in the test-market study, and because they make no sense in the context of the problem, the value of b_0 has little or no practical interpretation.

The slope of price with OmniPower sales ($b_1 = -53.2173$) indicates that, for a given amount of monthly promotional expenditures, the predicted sales of OmniPower are estimated to decrease by 53.2173 bars per month for each 1-cent increase in the price. The slope of monthly promotional expenditures with OmniPower sales ($b_2 = 3.6131$) indicates that, for a given price, the estimated sales of OmniPower are predicted to increase by 3.6131 bars for each additional \$1 spent on promotions. These estimates allow you to better understand the likely effect that price and promotion decisions will have in the marketplace. For example, a 10-cent decrease in price is predicted to increase sales by 532.173 bars, with a fixed amount of monthly promotional expenditures. A \$100 increase in promotional expenditures is predicted to increase sales by 361.31 bars, for a given price.

Regression coefficients in multiple regression are called **net regression coefficients**; they estimate the predicted change in Y per unit change in a particular X , *holding constant the effect of the other X variables*. For example, in the study of OmniPower bar sales, for a store with a given amount of promotional expenditures, the estimated sales are predicted to decrease by 53.2173 bars per month for each 1-cent increase in the price of an OmniPower bar. Another way to interpret this “net effect” is to think of two stores with an equal amount of promotional expenditures. If the first store charges 1 cent more than the other store, the net effect of this difference is that the first store is predicted to sell 53.2173 fewer bars per month than the second store. To interpret the net effect of promotional expenditures, you can consider two stores that are charging the same price. If the first store spends \$1 more on promotional expenditures, the net effect of this difference is that the first store is predicted to sell 3.6131 more bars per month than the second store.

Predicting the Dependent Variable Y

You can use the multiple regression equation to predict values of the dependent variable. For example, what are the predicted sales for a store charging 79 cents during a month in which promotional expenditures are \$400? Using the multiple regression equation,

$$\hat{Y}_i = 5,837.5208 - 53.2173X_{1i} + 3.6131X_{2i}$$

with $X_{1i} = 79$ and $X_{2i} = 400$,

$$\begin{aligned}\hat{Y}_i &= 5,837.5208 - 53.2173(79) + 3.6131(400) \\ &= 3,078.57\end{aligned}$$

Thus, you predict that stores charging 79 cents and spending \$400 in promotional expenditures will sell 3,078.57 OmniPower bars per month.

After you have developed the regression equation, done a residual analysis (see Section 14.3), and determined the significance of the overall fitted model (see Section 14.2), you can construct a confidence interval estimate of the mean value and a prediction interval for an individual value. You should rely on software to do these computations for you, given the complex nature of the computations. Figure 14.3 presents an Excel worksheet that computes a confidence interval estimate and a prediction interval for the OmniPower sales data. (The Minitab results in Figure 14.2 include these computations.)

FIGURE 14.3

Excel confidence interval estimate and prediction interval worksheet for the OmniPower sales data

	A	B	C	D
1	Confidence Interval Estimate and Prediction Interval			
2				
3	Data			
4	Confidence Level	95%		
5		1		
6	Price given value	79		
7	Promotion given value	400		
8				
9	X'X	34	2646	13200
10		2646	214674	1018800
11		13200	1018800	6000000
12				
13	Inverse of X'X	0.9692	-0.0094	-0.0005
14		-0.0094	0.0001	0.0000
15		-0.0005	0.0000	0.0000
16				
17	X'G times Inverse of X'X	0.0121	0.0001	0.0000
18				
19	[X'G times Inverse of X'X] times XG	0.0298	=MMULT(B17:D17, B5:B7)	
20	t Statistic	2.0395	=TINV(1 - B4, COMPUTE!B13)	
21	Predicted Y (YHat)	3078.57	{=MMULT(TRANSPOSE(B5:B7), COMPUTE!B17:B19)}	
22				
23	For Average Predicted Y (YHat)			
24	Interval Half Width	224.50	=B20 * SQRT(B19) * COMPUTE!B7	
25	Confidence Interval Lower Limit	2854.07	=B21 - B24	
26	Confidence Interval Upper Limit	3303.08	=B21 + B24	
27				
28	For Individual Response Y			
29	Interval Half Width	1320.57	=B20 * SQRT(1 + B19) * COMPUTE!B7	
30	Prediction Interval Lower Limit	1758.01	=B21 - B29	
31	Prediction Interval Upper Limit	4399.14	=B21 + B29	

Also:

Cell range B9:D11 =MMULT(TRANSPOSE(MRArray!A2:C35), MRArray!A2:C35)

Cell range B13:B15 =MINVERSE(B9:D11)

Cell range B17:D17 =MMULT(TRANSPOSE(B5:B7), B13:D15)

The 95% confidence interval estimate of the mean OmniPower sales for all stores charging 79 cents and spending \$400 in promotional expenditures is 2,854.07 to 3,303.08 bars. The prediction interval for an individual store is 1,758.01 to 4,399.14 bars.

Problems for Section 14.1

LEARNING THE BASICS

14.1 For this problem, use the following multiple regression equation:

$$\hat{Y}_i = 10 + 5X_{1i} + 3X_{2i}$$

- Interpret the meaning of the slopes.
- Interpret the meaning of the Y intercept.

14.2 For this problem, use the following multiple regression equation:

$$\hat{Y}_i = 50 - 2X_{1i} + 7X_{2i}$$

- Interpret the meaning of the slopes.
- Interpret the meaning of the Y intercept.

APPLYING THE CONCEPTS

14.3 A shoe manufacturer is considering developing a new brand of running shoes. The business problem facing the marketing analyst is to determine which variables should be used to predict durability (i.e., the effect of long-term impact). Two independent variables under consideration are X_1 (FOREIMP), a measurement of the forefoot shock-absorbing capability, and X_2 (MIDSOLE), a measurement

of the change in impact properties over time. The dependent variable Y is LTIMP, a measure of the shoe's durability after a repeated impact test. Data are collected from a random sample of 15 types of currently manufactured running shoes, with the following results:

Variable	Coefficients	Standard Error	t Statistic	p -Value
Intercept	-0.02686	0.06905	-0.39	0.7034
Foreimp	0.79116	0.06295	12.57	0.0000
Midsole	0.60484	0.07174	8.43	0.0000

- State the multiple regression equation.
- Interpret the meaning of the slopes, b_1 and b_2 , in this problem.

SELF Test **14.4** A mail-order catalog business selling personal computer supplies, software, and hardware maintains a centralized warehouse. Management is currently examining the process of distribution from the warehouse. The business problem facing management relates to the factors that affect warehouse distribution costs. Currently, a small handling fee is added to each order, regardless of the amount of the order. Data collected over the past 24 months (stored in **WareCost**) indicate the warehouse distribution costs (in thousands of dollars), the sales (in thousands of dollars), and the number of orders received.

- State the multiple regression equation.
- Interpret the meaning of the slopes, b_1 and b_2 , in this problem.
- Explain why the regression coefficient, b_0 , has no practical meaning in the context of this problem.
- Predict the monthly warehouse distribution cost when sales are \$400,000 and the number of orders is 4,500.
- Construct a 95% confidence interval estimate for the mean monthly warehouse distribution cost when sales are \$400,000 and the number of orders is 4,500.
- Construct a 95% prediction interval for the monthly warehouse distribution cost for a particular month when sales are \$400,000 and the number of orders is 4,500.
- Explain why the interval in (e) is narrower than the interval in (f).

14.5 How does horsepower and weight affect the mileage of family sedans? Data from a sample of twenty 2010 family sedans were collected and organized and stored in **Auto2010**. (Data extracted from "Top 2010 Cars," *Consumer Reports*, April 2010, pp. 38–70.) Develop a regression model to predict mileage (as measured by miles per gallon) based on the horsepower of the car's engine and the weight of the car (in pounds).

- State the multiple regression equation.
- Interpret the meaning of the slopes, b_1 and b_2 , in this problem.
- Explain why the regression coefficient, b_0 , has no practical meaning in the context of this problem.

- Predict the miles per gallon for cars that have 190 horsepower and weigh 3,500 pounds.
- Construct a 95% confidence interval estimate for the mean miles per gallon for cars that have 190 horsepower and weigh 3,500 pounds.
- Construct a 95% prediction interval for the miles per gallon for an individual car that has 190 horsepower and weighs 3,500 pounds.

14.6 The business problem facing a consumer products company is to measure the effectiveness of different types of advertising media in the promotion of its products. Specifically, the company is interested in the effectiveness of radio advertising and newspaper advertising (including the cost of discount coupons). During a one month test period, data were collected from a sample of 22 cities with approximately equal populations. Each city is allocated a specific expenditure level for radio advertising and for newspaper advertising. The sales of the product (in thousands of dollars) and also the levels of media expenditure (in thousands of dollars) during the test month are recorded, with the following results shown below and stored in **Advertise**:

City	Sales (\$Thousands)	Radio Advertising (\$Thousands)	Newspaper Advertising (\$Thousands)
1	973	0	40
2	1,119	0	40
3	875	25	25
4	625	25	25
5	910	30	30
6	971	30	30
7	931	35	35
8	1,177	35	35
9	882	40	25
10	982	40	25
11	1,628	45	45
12	1,577	45	45
13	1,044	50	0
14	914	50	0
15	1,329	55	25
16	1,330	55	25
17	1,405	60	30
18	1,436	60	30
19	1,521	65	35
20	1,741	65	35
21	1,866	70	40
22	1,717	70	40

- State the multiple regression equation.
- Interpret the meaning of the slopes, b_1 and b_2 , in this problem.
- Interpret the meaning of the regression coefficient, b_0 .
- Which type of advertising is more effective? Explain.

14.7 The business problem facing the director of broadcasting operations for a television station was the issue of standby hours (i.e., hours in which unionized graphic artists at the station are paid but are not actually involved in any activity) and what factors were related to standby hours. The study included the following variables:

Standby hours (Y)—Total number of standby hours in a week

Total staff present (X_1)—Weekly total of people-days

Remote hours (X_2)—Total number of hours worked by employees at locations away from the central plant

Data were collected for 26 weeks; these data are organized and stored in **Standby**.

- State the multiple regression equation.
- Interpret the meaning of the slopes, b_1 and b_2 , in this problem.
- Explain why the regression coefficient, b_0 , has no practical meaning in the context of this problem.
- Predict the standby hours for a week in which the total staff present have 310 people-days and the remote hours are 400.
- Construct a 95% confidence interval estimate for the mean standby hours for weeks in which the total staff present have 310 people-days and the remote hours are 400.

- Construct a 95% prediction interval for the standby hours for a single week in which the total staff present have 310 people-days and the remote hours are 400.

14.8 Nassau County is located approximately 25 miles east of New York City. The data organized and stored in **GlenCove** include the appraised value, land area of the property in acres, and age, in years, for a sample of 30 single-family homes located in Glen Cove, a small city in Nassau County. Develop a multiple linear regression model to predict appraised value based on land area of the property and age, in years.

- State the multiple regression equation.
- Interpret the meaning of the slopes, b_1 and b_2 , in this problem.
- Explain why the regression coefficient, b_0 , has no practical meaning in the context of this problem.
- Predict the appraised value for a house that has a land area of 0.25 acres and is 45 years old.
- Construct a 95% confidence interval estimate for the mean appraised value for houses that have a land area of 0.25 acres and are 45 years old.
- Construct a 95% prediction interval estimate for the appraised value for an individual house that has a land area of 0.25 acres and is 45 years old.

14.2 r^2 , Adjusted r^2 , and the Overall F Test

This section discusses three methods you can use to evaluate the overall multiple regression model: the coefficient of multiple determination, r^2 , the adjusted r^2 , and the overall F test.

Coefficient of Multiple Determination

Recall from Section 13.3 that the coefficient of determination, r^2 , measures the proportion of the variation in Y that is explained by the independent variable X in the simple linear regression model. In multiple regression, the **coefficient of multiple determination** represents the proportion of the variation in Y that is explained by the set of independent variables. Equation (14.4) defines the coefficient of multiple determination for a multiple regression model with two or more independent variables.

COEFFICIENT OF MULTIPLE DETERMINATION

The coefficient of multiple determination is equal to the regression sum of squares (SSR) divided by the total sum of squares (SST).

$$r^2 = \frac{\text{Regression sum of squares}}{\text{Total sum of squares}} = \frac{SSR}{SST} \quad (14.4)$$

where

SSR = regression sum of squares

SST = total sum of squares

In the OmniPower example, from Figure 14.2 on page 580, $SSR = 39,472,730.77$ and $SST = 52,093,677.44$. Thus,

$$r^2 = \frac{SSR}{SST} = \frac{39,472,730.77}{52,093,677.44} = 0.7577$$

The coefficient of multiple determination ($r^2 = 0.7577$) indicates that 75.77% of the variation in sales is explained by the variation in the price and in the promotional expenditures. The coefficient of multiple determination also appears in the Figure 14.2 results on page 580, and is labeled R Square in the Excel results and R-Sq in the Minitab results.

Adjusted r^2

When considering multiple regression models, some statisticians suggest that you should use the **adjusted r^2** to take into account both the number of independent variables in the model and the sample size. Reporting the adjusted r^2 is extremely important when you are comparing two or more regression models that predict the same dependent variable but have a different number of independent variables. Equation (14.5) defines the adjusted r^2 .

ADJUSTED r^2

$$r_{\text{adj}}^2 = 1 - \left[(1 - r^2) \frac{n - 1}{n - k - 1} \right] \quad (14.5)$$

where k is the number of independent variables in the regression equation.

Thus, for the OmniPower data, because $r^2 = 0.7577$, $n = 34$, and $k = 2$,

$$\begin{aligned} r_{\text{adj}}^2 &= 1 - \left[(1 - 0.7577) \frac{34 - 1}{34 - 2 - 1} \right] \\ &= 1 - \left[(0.2423) \frac{33}{31} \right] \\ &= 1 - 0.2579 \\ &= 0.7421 \end{aligned}$$

Therefore, 74.21% of the variation in sales is explained by the multiple regression model—adjusted for the number of independent variables and sample size. The adjusted r^2 also appears in the Figure 14.2 results on page 580, and is labeled Adjusted R Square in the Excel results and R-Sq(adj) in the Minitab results.

Test for the Significance of the Overall Multiple Regression Model

You use the **overall F test** to determine whether there is a significant relationship between the dependent variable and the entire set of independent variables (the overall multiple regression model). Because there is more than one independent variable, you use the following null and alternative hypotheses:

$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$ (There is no linear relationship between the dependent variable and the independent variables.)

H_1 : At least one $\beta_j \neq 0$, $j = 1, 2, \dots, k$ (There is a linear relationship between the dependent variable and at least one of the independent variables.)

Equation (14.6) defines the overall F test statistic. Table 14.2 presents the ANOVA summary table.

OVERALL F TEST

The F_{STAT} test statistic is equal to the regression mean square (MSR) divided by the mean square error (MSE).

$$F_{STAT} = \frac{MSR}{MSE} \tag{14.6}$$

where

F_{STAT} = test statistic from an F distribution with k and $n - k - 1$ degrees of freedom

k = number of independent variables in the regression model

TABLE 14.2
ANOVA Summary Table
for the Overall F Test

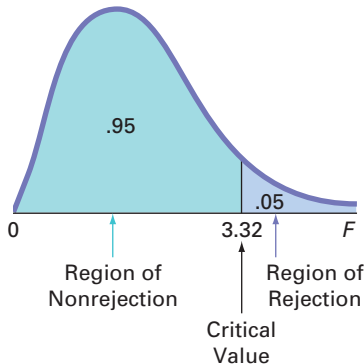
Source	Degrees of Freedom	Sum of Squares	Mean Squares (Variance)	F
Regression	k	SSR	$MSR = \frac{SSR}{k}$	$F_{STAT} = \frac{MSR}{MSE}$
Error	$n - k - 1$	SSE	$MSE = \frac{SSE}{n - k - 1}$	
Total	$n - 1$	SST		

The decision rule is

Reject H_0 at the α level of significance if $F_{STAT} > F_\alpha$;
otherwise, do not reject H_0 .

Using a 0.05 level of significance, the critical value of the F distribution with 2 and 31 degrees of freedom found from Table E.5 is approximately 3.32 (see Figure 14.4 below). From Figure 14.2 on page 580, the F_{STAT} test statistic given in the ANOVA summary table is 48.4771. Because $48.4771 > 3.32$, or because the p -value = 0.000 < 0.05, you reject H_0 and conclude that at least one of the independent variables (price and/or promotional expenditures) is related to sales.

FIGURE 14.4
Testing for the
significance of a set of
regression coefficients at
the 0.05 level of
significance, with 2 and
31 degrees of freedom



Problems for Section 14.2

LEARNING THE BASICS

14.9 The following ANOVA summary table is for a multiple regression model with two independent variables:

Source	Degrees of Freedom	Sum of Squares	Mean Squares	F
Regression	2	60		
Error	18	120		
Total	20	180		

- Determine the regression mean square (MSR) and the mean square error (MSE).
- Compute the overall F_{STAT} test statistic.
- Determine whether there is a significant relationship between Y and the two independent variables at the 0.05 level of significance.
- Compute the coefficient of multiple determination, r^2 , and interpret its meaning.
- Compute the adjusted r^2 .

14.10 The following ANOVA summary table is for a multiple regression model with two independent variables:

Source	Degrees of Freedom	Sum of Squares	Mean Squares	F
Regression	2	30		
Error	10	120		
Total	12	150		

- Determine the regression mean square (MSR) and the mean square error (MSE).
- Compute the overall F_{STAT} test statistic.
- Determine whether there is a significant relationship between Y and the two independent variables at the 0.05 level of significance.
- Compute the coefficient of multiple determination, r^2 , and interpret its meaning.
- Compute the adjusted r^2 .

APPLYING THE CONCEPTS

14.11 Eileen M. Van Aken and Brian M. Kleiner, professors at Virginia Polytechnic Institute and State University, investigated the factors that contribute to the effectiveness of teams. (Data extracted from “Determinants of Effectiveness for Cross-Functional Organizational Design Teams,” *Quality Management Journal*, 4 (1997), 51–79.) The researchers studied 34 independent variables, such as team skills, diversity, meeting frequency, and clarity in expectations. For each of the teams studied, each of the variables was given a value of 1 through 100, based on the results of interviews and survey data, where

100 represents the highest rating. The dependent variable, team performance, was also given a value of 1 through 100, with 100 representing the highest rating. Many different regression models were explored, including the following:

Model 1

$$\text{Team performance} = \beta_0 + \beta_1 (\text{Team skills}) + \varepsilon$$

$$r^2_{\text{adj}} = 0.68$$

Model 2

$$\text{Team performance} = \beta_0 + \beta_1 (\text{Clarity in expectations}) + \varepsilon$$

$$r^2_{\text{adj}} = 0.78$$

Model 3

$$\text{Team performance} = \beta_0 + \beta_1 (\text{Team skills})$$

$$+ \beta_2 (\text{Clarity in expectations}) + \varepsilon$$

$$r^2_{\text{adj}} = 0.97$$

- Interpret the adjusted r^2 for each of the three models.
- Which of these three models do you think is the best predictor of team performance?

14.12 In Problem 14.3 on page 582, you predicted the durability of a brand of running shoe, based on the forefoot shock-absorbing capability and the change in impact properties over time. The regression analysis resulted in the following ANOVA summary table:

Source	Degrees of Freedom	Sum of Squares	Mean Squares	F	p -Value
Regression	2	12.61020	6.30510	97.69	0.0001
Error	12	0.77453	0.06454		
Total	14	13.38473			

- Determine whether there is a significant relationship between durability and the two independent variables at the 0.05 level of significance.
- Interpret the meaning of the p -value.
- Compute the coefficient of multiple determination, r^2 , and interpret its meaning.
- Compute the adjusted r^2 .

14.13 In Problem 14.5 on page 583, you used horsepower and weight to predict mileage (stored in **Auto2010**). Using the results from that problem,

- determine whether there is a significant relationship between mileage and the two independent variables (horsepower and weight) at the 0.05 level of significance.
- interpret the meaning of the p -value.
- compute the coefficient of multiple determination, r^2 , and interpret its meaning.
- compute the adjusted r^2 .

SELF Test **14.14** In Problem 14.4 on page 583, you used sales and number of orders to predict distribution costs at a mail-order catalog business (stored in **WareCost**). Use the results from that problem.

- Determine whether there is a significant relationship between distribution costs and the two independent variables (sales and number of orders) at the 0.05 level of significance.
- Interpret the meaning of the p -value.
- Compute the coefficient of multiple determination, r^2 , and interpret its meaning.
- Compute the adjusted r^2 .

14.15 In Problem 14.7 on page 584, you used the total staff present and remote hours to predict standby hours (stored in **Standby**). Use the results from that problem.

- Determine whether there is a significant relationship between standby hours and the two independent variables (total staff present and remote hours) at the 0.05 level of significance.
- Interpret the meaning of the p -value.
- Compute the coefficient of multiple determination, r^2 , and interpret its meaning.
- Compute the adjusted r^2 .

14.16 In Problem 14.6 on page 583, you used radio advertising and newspaper advertising to predict sales (stored in **Advertise**). Use the results from that problem.

- Determine whether there is a significant relationship between sales and the two independent variables (radio advertising and newspaper advertising) at the 0.05 level of significance.
- Interpret the meaning of the p -value.
- Compute the coefficient of multiple determination, r^2 , and interpret its meaning.
- Compute the adjusted r^2 .

14.17 In Problem 14.8 on page 584, you used the land area of a property and the age of a house to predict appraised value (stored in **GlenCove**). Use the results from that problem.

- Determine whether there is a significant relationship between appraised value and the two independent variables (land area of a property and age of a house) at the 0.05 level of significance.
- Interpret the meaning of the p -value.
- Compute the coefficient of multiple determination, r^2 , and interpret its meaning.
- Compute the adjusted r^2 .

14.3 Residual Analysis for the Multiple Regression Model

In Section 13.5, you used residual analysis to evaluate the fit of the simple linear regression model. For the multiple regression model with two independent variables, you need to construct and analyze the following residual plots:

- Residuals versus \hat{Y}_i
- Residuals versus X_{1i}
- Residuals versus X_{2i}
- Residuals versus time

The first residual plot examines the pattern of residuals versus the predicted values of Y . If the residuals show a pattern for the predicted values of Y , there is evidence of a possible curvilinear effect (see Section 15.1) in at least one independent variable, a possible violation of the assumption of equal variance (see Figure 13.13 on page 542), and/or the need to transform the Y variable.

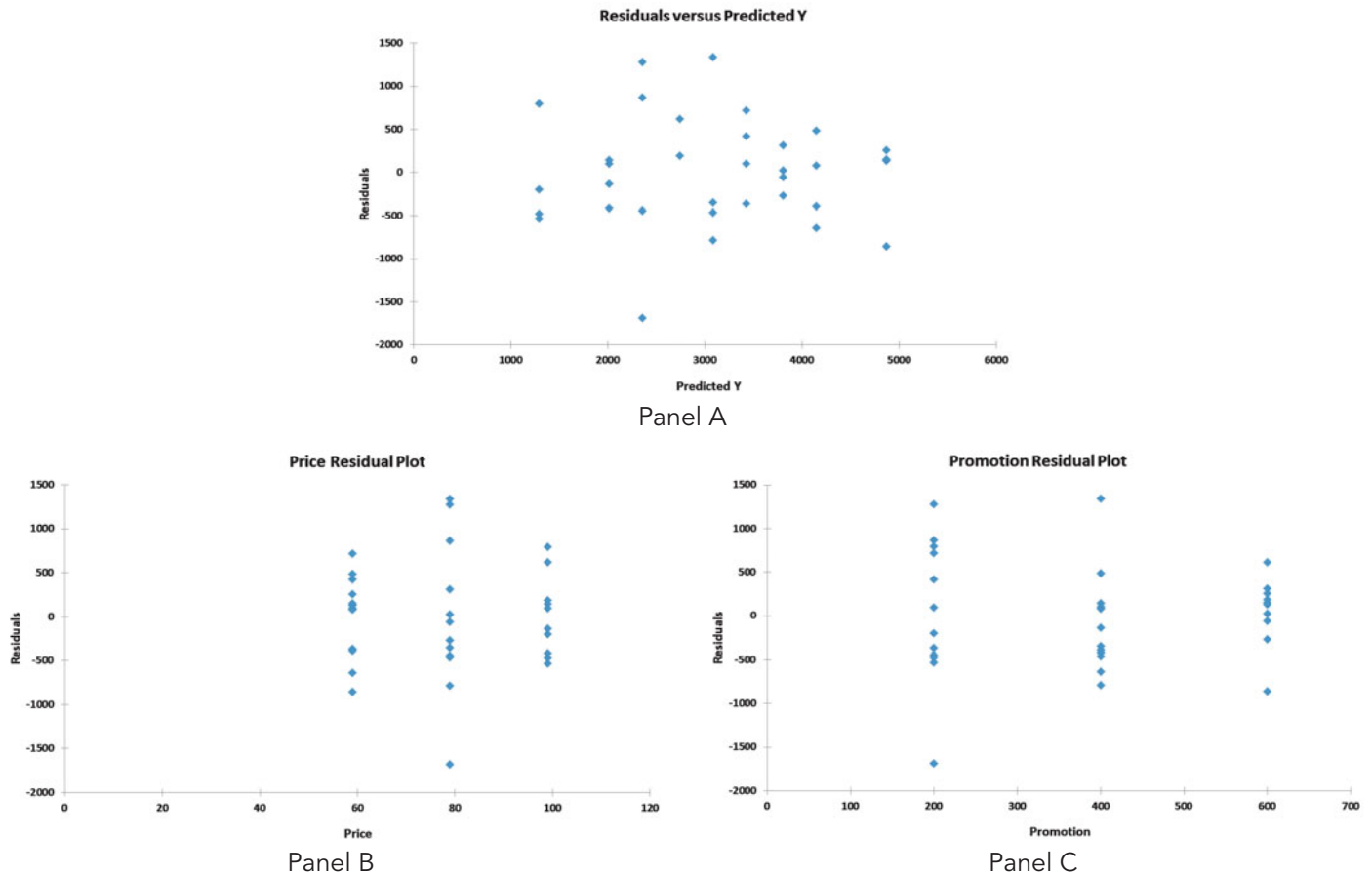
The second and third residual plots involve the independent variables. Patterns in the plot of the residuals versus an independent variable may indicate the existence of a curvilinear effect and, therefore, the need to add a curvilinear independent variable to the multiple regression model (see Section 15.1).

The fourth plot is used to investigate patterns in the residuals in order to validate the independence assumption when the data are collected in time order. Associated with this residual plot, as in Section 13.6, you can compute the Durbin-Watson statistic to determine the existence of positive autocorrelation among the residuals.

Figure 14.5 presents the residual plots for the OmniPower sales example. There is very little or no pattern in the relationship between the residuals and the predicted value of Y , the value of X_1 (price), or the value of X_2 (promotional expenditures). Thus, you can conclude that the multiple regression model is appropriate for predicting sales. There is no need to plot the residuals versus time because the data were not collected in time order.

FIGURE 14.5

Residual plots for the OmniPower sales data: Panel A, residuals versus predicted \hat{Y} ; Panel B, residuals versus price; Panel C, residuals versus promotional expenditures



Problems for Section 14.3

APPLYING THE CONCEPTS

14.18 In Problem 14.4 on page 583, you used sales and number of orders to predict distribution costs at a mail-order catalog business (stored in **WareCost**).

- Plot the residuals versus \hat{Y}_i .
- Plot the residuals versus X_{1i} .
- Plot the residuals versus X_{2i} .
- Plot the residuals versus time.
- In the residual plots created in (a) through (d), is there any evidence of a violation of the regression assumptions? Explain.
- Determine the Durbin-Watson statistic.
- At the 0.05 level of significance, is there evidence of positive autocorrelation in the residuals?

14.19 In Problem 14.5 on page 583, you used horsepower and weight to predict mileage (stored in **Auto2010**).

- Plot the residuals versus \hat{Y}_i .
- Plot the residuals versus X_{1i} .
- Plot the residuals versus X_{2i} .
- In the residual plots created in (a) through (c), is there any evidence of a violation of the regression assumptions? Explain.
- Should you compute the Durbin-Watson statistic for these data? Explain.

14.20 In Problem 14.6 on page 583, you used radio advertising and newspaper advertising to predict sales (stored in **Advertise**).

- Perform a residual analysis on your results.
- If appropriate, perform the Durbin-Watson test, using $\alpha = 0.05$.
- Are the regression assumptions valid for these data?

14.21 In Problem 14.7 on page 584, you used the total staff present and remote hours to predict standby hours (stored in **Standby**).

- Perform a residual analysis on your results.
- If appropriate, perform the Durbin-Watson test, using $\alpha = 0.05$.
- Are the regression assumptions valid for these data?

14.22 In Problem 14.8 on page 584, you used the land area of a property and the age of a house to predict appraised value (stored in **GlenCove**).

- Perform a residual analysis on your results.
- If appropriate, perform the Durbin-Watson test, using $\alpha = 0.05$.
- Are the regression assumptions valid for these data?

14.4 Inferences Concerning the Population Regression Coefficients

In Section 13.7, you tested the slope in a simple linear regression model to determine the significance of the relationship between X and Y . In addition, you constructed a confidence interval estimate of the population slope. This section extends those procedures to multiple regression.

Tests of Hypothesis

In a simple linear regression model, to test a hypothesis concerning the population slope, β_1 , you used Equation (13.16) on page 548:

$$t_{STAT} = \frac{b_1 - \beta_1}{S_{b_1}}$$

Equation (14.7) generalizes this equation for multiple regression.

TESTING FOR THE SLOPE IN MULTIPLE REGRESSION

$$t_{STAT} = \frac{b_j - \beta_j}{S_{b_j}} \quad (14.7)$$

where

b_j = slope of variable j with Y , holding constant the effects of all other independent variables

S_{b_j} = standard error of the regression coefficient b_j

t_{STAT} = test statistic for a t distribution with $n - k - 1$ degrees of freedom

k = number of independent variables in the regression equation

β_j = hypothesized value of the population slope for variable j , holding constant the effects of all other independent variables

To determine whether variable X_2 (amount of promotional expenditures) has a significant effect on sales, taking into account the price of OmniPower bars, the null and alternative hypotheses are

$$H_0: \beta_2 = 0$$

$$H_1: \beta_2 \neq 0$$

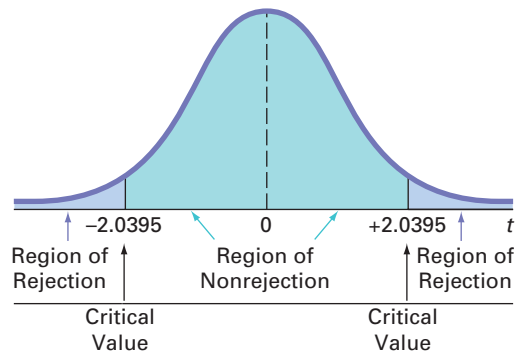
From Equation (14.7) and Figure 14.2 on page 580,

$$\begin{aligned} t_{STAT} &= \frac{b_2 - \beta_2}{S_{b_2}} \\ &= \frac{3.6131 - 0}{0.6852} = 5.2728 \end{aligned}$$

If you select a level of significance of 0.05, the critical values of t for 31 degrees of freedom from Table E.3 are -2.0395 and $+2.0395$ (see Figure 14.6).

FIGURE 14.6

Testing for significance of a regression coefficient at the 0.05 level of significance, with 31 degrees of freedom



From Figure 14.2 on page 580, observe that the computed t_{STAT} test statistic is 5.2728. Because $t_{STAT} = 5.2728 > 2.0395$ or because the p -value is approximately zero, you reject H_0 and conclude that there is a significant relationship between the variable X_2 (promotional expenditures) and sales, taking into account the price, X_1 . The extremely small p -value allows you to strongly reject the null hypothesis that there is no linear relationship between sales and promotional expenditures. Example 14.1 presents the test for the significance of β_1 , the slope of sales with price.

EXAMPLE 14.1

Testing for the Significance of the Slope of Sales with Price

At the 0.05 level of significance, is there evidence that the slope of sales with price is different from zero?

SOLUTION From Figure 14.2 on page 580, $t_{STAT} = -7.7664 < -2.0395$ (the critical value for $\alpha = 0.05$) or the p -value $= 0.0000 < 0.05$. Thus, there is a significant relationship between price, X_1 , and sales, taking into account the promotional expenditures, X_2 .

As shown with these two independent variables, the test of significance for a specific regression coefficient in multiple regression is a test for the significance of adding that variable into a regression model, given that the other variable is included. In other words, the t test for the regression coefficient is actually a test for the contribution of each independent variable.

Confidence Interval Estimation

Instead of testing the significance of a population slope, you may want to estimate the value of a population slope. Equation (14.8) defines the confidence interval estimate for a population slope in multiple regression.

CONFIDENCE INTERVAL ESTIMATE FOR THE SLOPE

$$b_j \pm t_{\alpha/2} S_{b_j} \quad (14.8)$$

where $t_{\alpha/2}$ is the critical value corresponding to an upper-tail probability of $\alpha/2$ from the t distribution with $n-k-1$ degrees of freedom (i.e., a cumulative area of $1 - \alpha/2$), and k is the number of independent variables.

To construct a 95% confidence interval estimate of the population slope, β_1 (the effect of price, X_1 , on sales, Y , holding constant the effect of promotional expenditures, X_2), the critical

value of t at the 95% confidence level with 31 degrees of freedom is 2.0395 (see Table E.3). Then, using Equation (14.8) and Figure 14.2 on page 580,

$$\begin{aligned} b_1 \pm t_{\alpha/2} S_{b_1} \\ -53.2173 \pm (2.0395)(6.8522) \\ -53.2173 \pm 13.9752 \\ -67.1925 \leq \beta_1 \leq -39.2421 \end{aligned}$$

Taking into account the effect of promotional expenditures, the estimated effect of a 1-cent increase in price is to reduce mean sales by approximately 39.2 to 67.2 bars. You have 95% confidence that this interval correctly estimates the relationship between these variables. From a hypothesis-testing viewpoint, because this confidence interval does not include 0, you conclude that the regression coefficient, β_1 , has a significant effect.

Example 14.2 constructs and interprets a confidence interval estimate for the slope of sales with promotional expenditures.

EXAMPLE 14.2

Constructing a Confidence Interval Estimate for the Slope of Sales with Promotional Expenditures

Construct a 95% confidence interval estimate of the population slope of sales with promotional expenditures.

SOLUTION The critical value of t at the 95% confidence level, with 31 degrees of freedom, is 2.0395 (see Table E.3). Using Equation (14.8) and Figure 14.2 on page 580,

$$\begin{aligned} b_2 \pm t_{\alpha/2} S_{b_2} \\ 3.6131 \pm (2.0395)(0.6852) \\ 3.6131 \pm 1.3975 \\ 2.2156 \leq \beta_2 \leq 5.0106 \end{aligned}$$

Thus, taking into account the effect of price, the estimated effect of each additional dollar of promotional expenditures is to increase mean sales by approximately 2.22 to 5.01 bars. You have 95% confidence that this interval correctly estimates the relationship between these variables. From a hypothesis-testing viewpoint, because this confidence interval does not include 0, you can conclude that the regression coefficient, β_2 , has a significant effect.

Problems for Section 14.4

LEARNING THE BASICS

14.23 Use the following information from a multiple regression analysis:

$$n = 25 \quad b_1 = 5 \quad b_2 = 10 \quad S_{b_1} = 2 \quad S_{b_2} = 8$$

- Which variable has the largest slope, in units of a t statistic?
- Construct a 95% confidence interval estimate of the population slope, β_1 .
- At the 0.05 level of significance, determine whether each independent variable makes a significant contribution to the regression model. On the basis of these results, indicate the independent variables to include in this model.

14.24 Use the following information from a multiple regression analysis:

$$n = 20 \quad b_1 = 4 \quad b_2 = 3 \quad S_{b_1} = 1.2 \quad S_{b_2} = 0.8$$


- Which variable has the largest slope, in units of a t statistic?
- Construct a 95% confidence interval estimate of the population slope, β_1 .
- At the 0.05 level of significance, determine whether each independent variable makes a significant contribution to the regression model. On the basis of these results, indicate the independent variables to include in this model.

APPLYING THE CONCEPTS

14.25 In Problem 14.3 on page 582, you predicted the durability of a brand of running shoe, based on the forefoot shock-absorbing capability (FOREIMP) and the change in impact properties over time (MIDSOLE) for a sample of 15 pairs of shoes. Use the following results:

Variable	Coefficient	Standard Error	<i>t</i> Statistic	<i>p</i> -value
Intercept	−0.02686	0.06905	−0.39	0.7034
Foreimp	0.79116	0.06295	12.57	0.0000
Midsole	0.60484	0.07174	8.43	0.0000

- Construct a 95% confidence interval estimate of the population slope between durability and forefoot shock-absorbing capability.
- At the 0.05 level of significance, determine whether each independent variable makes a significant contribution to the regression model. On the basis of these results, indicate the independent variables to include in this model.

 **14.26** In Problem 14.4 on page 583, you used sales and number of orders to predict distribution costs at a mail-order catalog business (stored in **WareCost**). Use the results from that problem.

- Construct a 95% confidence interval estimate of the population slope between distribution cost and sales.
- At the 0.05 level of significance, determine whether each independent variable makes a significant contribution to the regression model. On the basis of these results, indicate the independent variables to include in this model.

14.27 In Problem 14.5 on page 583, you used horsepower and weight to predict mileage (stored in **Auto2010**). Use the results from that problem.

- Construct a 95% confidence interval estimate of the population slope between mileage and horsepower.
- At the 0.05 level of significance, determine whether each independent variable makes a significant contribution to the regression model. On the basis of these results, indicate the independent variables to include in this model.

14.28 In Problem 14.6 on page 583, you used radio advertising and newspaper advertising to predict sales (stored in **Advertise**). Use the results from that problem.

- Construct a 95% confidence interval estimate of the population slope between sales and radio advertising.
- At the 0.05 level of significance, determine whether each independent variable makes a significant contribution to the regression model. On the basis of these results, indicate the independent variables to include in this model.

14.29 In Problem 14.7 on page 584, you used the total number of staff present and remote hours to predict standby hours (stored in **Standby**). Use the results from that problem.

- Construct a 95% confidence interval estimate of the population slope between standby hours and total number of staff present.
- At the 0.05 level of significance, determine whether each independent variable makes a significant contribution to the regression model. On the basis of these results, indicate the independent variables to include in this model.

14.30 In Problem 14.8 on page 584, you used land area of a property and age of a house to predict appraised value (stored in **GlenCove**). Use the results from that problem.

- Construct a 95% confidence interval estimate of the population slope between appraised value and land area of a property.
- At the 0.05 level of significance, determine whether each independent variable makes a significant contribution to the regression model. On the basis of these results, indicate the independent variables to include in this model.

14.5 Testing Portions of the Multiple Regression Model

In developing a multiple regression model, you want to use only those independent variables that significantly reduce the error in predicting the value of a dependent variable. If an independent variable does not improve the prediction, you can delete it from the multiple regression model and use a model with fewer independent variables.

The **partial *F* test** is an alternative method to the *t* test discussed in Section 14.4 for determining the contribution of an independent variable. Using this method, you determine the contribution to the regression sum of squares made by each independent variable after all the other independent variables have been included in the model. The new independent variable is included only if it significantly improves the model.

To conduct partial *F* tests for the OmniPower sales example, you need to evaluate the contribution of promotional expenditures (X_2) after price (X_1) has been included in the model, and also evaluate the contribution of price (X_1) after promotional expenditures (X_2) have been included in the model.

In general, if there are several independent variables, you determine the contribution of each independent variable by taking into account the regression sum of squares of a model that includes all independent variables except the one of interest, j . This regression sum of squares is denoted SSR (all X s except j). Equation (14.9) determines the contribution of variable j , assuming that all other variables are already included.

DETERMINING THE CONTRIBUTION OF AN INDEPENDENT VARIABLE TO THE REGRESSION MODEL

$$SSR(X_j | \text{All } X\text{s except } j) = SSR(\text{All } X\text{s}) - SSR(\text{All } X\text{s except } j) \quad (14.9)$$

If there are two independent variables, you use Equations (14.10a) and (14.10b) to determine the contribution of each.

CONTRIBUTION OF VARIABLE X_1 , GIVEN THAT X_2 HAS BEEN INCLUDED

$$SSR(X_1 | X_2) = SSR(X_1 \text{ and } X_2) - SSR(X_2) \quad (14.10a)$$

CONTRIBUTION OF VARIABLE X_2 , GIVEN THAT X_1 HAS BEEN INCLUDED

$$SSR(X_2 | X_1) = SSR(X_1 \text{ and } X_2) - SSR(X_1) \quad (14.10b)$$

The term $SSR(X_2)$ represents the sum of squares due to regression for a model that includes only the independent variable X_2 (promotional expenditures). Similarly, $SSR(X_1)$ represents the sum of squares due to regression for a model that includes only the independent variable X_1 (price). Figures 14.7 and 14.8 present results for these two models.

From Figure 14.7, $SSR(X_2) = 14,915,814.10$ and from Figure 14.2 on page 580, $SSR(X_1 \text{ and } X_2) = 39,472,730.77$. Then, using Equation (14.10a),

$$\begin{aligned} SSR(X_1 | X_2) &= SSR(X_1 \text{ and } X_2) - SSR(X_2) \\ &= 39,472,730.77 - 14,915,814.10 \\ &= 24,556,916.67 \end{aligned}$$

FIGURE 14.7

Excel and Minitab regression results for a simple linear regression model of sales with promotional expenditures, $SSR(X_2)$

	A	B	C	D	E	F	G
1	Sales and Promotional Expenses Analysis						
2							
3	Regression Statistics						
4	Multiple R	0.5351					
5	R Square	0.2863					
6	Adjusted R Square	0.2640					
7	Standard Error	1077.8721					
8	Observations	34					
9							
10	ANOVA						
11		df	SS	MS	F	Significance F	
12	Regression	1	14915814.1025	14915814.1025	12.8384	0.0011	
13	Residual	32	37177863.3387	1161808.2293			
14	Total	33	52093677.4412				
15							
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
17	Intercept	1496.0161	483.9789	3.0911	0.0041	510.1843	2481.8480
18	Promotion	4.1281	1.1521	3.5831	0.0011	1.7813	6.4748

Regression Analysis: Sales versus Promotion

The regression equation is
Sales = 1496 + 4.13 Promotion

Predictor	Coef	SE Coef	T	P
Constant	1496.0	484.0	3.09	0.004
Promotion	4.128	1.152	3.58	0.001

S = 1077.87 R-Sq = 28.6% R-Sq(adj) = 26.4%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	14915814	14915814	12.84	0.001
Residual Error	32	37177863	1161808		
Total	33	52093677			

FIGURE 14.8

Excel and Minitab regression results for a simple linear regression model of sales with price, $SSR(X_1)$

	A	B	C	D	E	F	G
1	Sales and Price Analysis						
2							
3	Regression Statistics						
4	Multiple R	0.7351					
5	R Square	0.5404					
6	Adjusted R Square	0.5261					
7	Standard Error	864.9457					
8	Observations	34					
9							
10	ANOVA						
11		df	SS	MS	F	Significance F	
12	Regression	1	28153486.1482	28153486.1482	37.6318	0.0000	
13	Residual	32	23940191.2930	748130.9779			
14	Total	33	52093677.4412				
15							
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
17	Intercept	7512.3480	734.6189	10.2262	0.0000	6015.9796	9008.7164
18	Price	-56.7138	9.2451	-6.1345	0.0000	-75.5455	-37.8822

Regression Analysis: Sales versus Price

The regression equation is
Sales = 7512 - 56.7 Price

Predictor	Coef	SE Coef	T	P
Constant	7512.3	734.6	10.23	0.000
Price	-56.714	9.245	-6.13	0.000

S = 864.946 R-Sq = 54.0% R-Sq(adj) = 52.6%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	28153486	28153486	37.63	0.000
Residual Error	32	23940191	748131		
Total	33	52093677			

To determine whether X_1 significantly improves the model after X_2 has been included, you divide the regression sum of squares into two component parts, as shown in Table 14.3.

TABLE 14.3

ANOVA Table Dividing the Regression Sum of Squares into Components to Determine the Contribution of Variable X_1

Source	Degrees of Freedom	Sum of Squares	Mean Square (Variance)	F
Regression	2	39,472,730.77	19,736,365.39	
$\left\{ \begin{array}{l} X_2 \\ X_1 X_2 \end{array} \right\}$	$\left\{ \begin{array}{l} 1 \\ 1 \end{array} \right\}$	$\left\{ \begin{array}{l} 14,915,814.10 \\ 24,556,916.67 \end{array} \right\}$	24,556,916.67	60.32
Error	31	12,620,946.67	407,127.31	
Total	33	52,093,677.44		

The null and alternative hypotheses to test for the contribution of X_1 to the model are

H_0 : Variable X_1 does not significantly improve the model after variable X_2 has been included.

H_1 : Variable X_1 significantly improves the model after variable X_2 has been included.

Equation (14.11) defines the partial F test statistic for testing the contribution of an independent variable.

PARTIAL F TEST STATISTIC

$$F_{STAT} = \frac{SSR(X_j | \text{All } X\text{s except } j)}{MSE} \quad (14.11)$$

The partial F test statistic follows an F distribution with 1 and $n-k-1$ degrees of freedom.

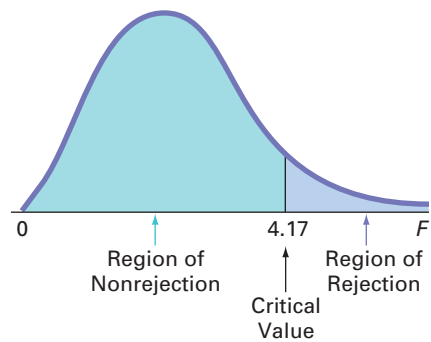
From Table 14.3,

$$F_{STAT} = \frac{24,556,916.67}{407,127.31} = 60.32$$

The partial F_{STAT} test statistic has 1 and $n-k-1 = 34-2-1 = 31$ degrees of freedom. Using a level of significance of 0.05, the critical value from Table E.5 is approximately 4.17 (see Figure 14.9).

FIGURE 14.9

Testing for the contribution of a regression coefficient to a multiple regression model at the 0.05 level of significance, with 1 and 31 degrees of freedom



Because the computed partial F_{STAT} test statistic (60.32) is greater than this critical F value (4.17), you reject H_0 . You can conclude that the addition of variable X_1 (price) significantly improves a regression model that already contains variable X_2 (promotional expenditures).

To evaluate the contribution of variable X_2 (promotional expenditures) to a model in which variable X_1 (price) has been included, you need to use Equation (14.10b). First, from Figure 14.8 on page 595, observe that $SSR(X_1) = 28,153,486.15$. Second, from Table 14.3, observe that $SSR(X_1 \text{ and } X_2) = 39,472,730.77$. Then, using Equation (14.10b) on page 594,

$$SSR(X_2 | X_1) = 39,472,730.77 - 28,153,486.15 = 11,319,244.62$$

To determine whether X_2 significantly improves a model after X_1 has been included, you can divide the regression sum of squares into two component parts, as shown in Table 14.4.

TABLE 14.4

ANOVA Table Dividing the Regression Sum of Squares into Components to Determine the Contribution of Variable X_2

Source	Degrees of Freedom	Sum of Squares	Mean Square (Variance)	F
Regression	2	39,472,730.77	19,736,365.39	
$\left\{ \begin{array}{l} X_1 \\ X_2 X_1 \end{array} \right\}$	$\left\{ \begin{array}{l} 1 \\ 1 \end{array} \right\}$	$\left\{ \begin{array}{l} 28,153,486.15 \\ 11,319,244.62 \end{array} \right\}$	11,319,244.62	27.80
Error	31	12,620,946.67	407,127.31	
Total	33	52,093,677.44		

The null and alternative hypotheses to test for the contribution of X_2 to the model are

H_0 : Variable X_2 does not significantly improve the model after variable X_1 has been included.

H_1 : Variable X_2 significantly improves the model after variable X_1 has been included.

Using Equation (14.11) and Table 14.4,

$$F_{STAT} = \frac{11,319,244.62}{407,127.31} = 27.80$$

In Figure 14.9, you can see that, using a 0.05 level of significance, the critical value of F , with 1 and 31 degrees of freedom, is approximately 4.17. Because the computed partial F_{STAT} test statistic (27.80) is greater than this critical value (4.17), you reject H_0 . You can conclude that the addition of variable X_2 (promotional expenditures) significantly improves the multiple regression model already containing X_1 (price).

Thus, by testing for the contribution of each independent variable after the other has been included in the model, you determine that each of the two independent variables significantly improves the model. Therefore, the multiple regression model should include both price, X_1 , and promotional expenditures, X_2 .

The partial F -test statistic developed in this section and the t -test statistic of Equation (14.7) on page 590 are both used to determine the contribution of an independent variable to a

¹This relationship holds only when the F_{STAT} statistic has 1 degree of freedom in the numerator.

multiple regression model. The hypothesis tests associated with these two statistics always result in the same decision (i.e., the p -values are identical). The t_{STAT} test statistics for the OmniPower regression model are -7.7664 and $+5.2728$, and the corresponding F_{STAT} test statistics are 60.32 and 27.80 . Equation (14.12) states this relationship between t and F .¹

RELATIONSHIP BETWEEN A t STATISTIC AND AN F STATISTIC

$$t_{STAT}^2 = F_{STAT} \quad (14.12)$$

Coefficients of Partial Determination

Recall from Section 14.2 that the coefficient of multiple determination, r^2 , measures the proportion of the variation in Y that is explained by variation in the independent variables. The **coefficients of partial determination** ($r_{Y1.2}^2$ and $r_{Y2.1}^2$) measure the proportion of the variation in the dependent variable that is explained by each independent variable while controlling for, or holding constant, the other independent variable. Equation (14.13) defines the coefficients of partial determination for a multiple regression model with two independent variables.

COEFFICIENTS OF PARTIAL DETERMINATION FOR A MULTIPLE REGRESSION MODEL CONTAINING TWO INDEPENDENT VARIABLES

$$r_{Y1.2}^2 = \frac{SSR(X_1 | X_2)}{SST - SSR(X_1 \text{ and } X_2) + SSR(X_1 | X_2)} \quad (14.13a)$$

and

$$r_{Y2.1}^2 = \frac{SSR(X_2 | X_1)}{SST - SSR(X_1 \text{ and } X_2) + SSR(X_2 | X_1)} \quad (14.13b)$$

where

$SSR(X_1 | X_2)$ = sum of squares of the contribution of variable X_1 to the regression model, given that variable X_2 has been included in the model

SST = total sum of squares for Y

$SSR(X_1 \text{ and } X_2)$ = regression sum of squares when variables X_1 and X_2 are both included in the multiple regression model

$SSR(X_2 | X_1)$ = sum of squares of the contribution of variable X_2 to the regression model, given that variable X_1 has been included in the model

For the OmniPower sales example,

$$\begin{aligned} r_{Y1.2}^2 &= \frac{24,556,916.67}{52,093,677.44 - 39,472,730.77 + 24,556,916.67} \\ &= 0.6605 \end{aligned}$$

$$\begin{aligned} r_{Y2.1}^2 &= \frac{11,319,244.62}{52,093,677.44 - 39,472,730.77 + 11,319,244.62} \\ &= 0.4728 \end{aligned}$$

The coefficient of partial determination, $r_{Y1.2}^2$, of variable Y with X_1 while holding X_2 constant is 0.6605. Thus, for a given (constant) amount of promotional expenditures, 66.05% of the variation in OmniPower sales is explained by the variation in the price. The coefficient of partial determination, $r_{Y2.1}^2$, of variable Y with X_2 while holding X_1 constant is 0.4728. Thus, for a given (constant) price, 47.28% of the variation in sales of OmniPower bars is explained by variation in the amount of promotional expenditures.

Equation (14.14) defines the coefficient of partial determination for the j th variable in a multiple regression model containing several (k) independent variables.

COEFFICIENT OF PARTIAL DETERMINATION FOR A MULTIPLE REGRESSION MODEL CONTAINING k INDEPENDENT VARIABLES

$$r_{Yj(\text{All variables except } j)}^2 = \frac{SSR(X_j | \text{All } X\text{s except } j)}{SST - SSR(\text{All } X\text{s}) + SSR(X_j | \text{All } X\text{s except } j)} \quad (14.14)$$

Problems for Section 14.5

LEARNING THE BASICS

14.31 The following is the ANOVA summary table for a multiple regression model with two independent variables:

Source	Degrees of Freedom	Sum of Squares	Mean Squares	F
Regression	2	60		
Error	18	120		
Total	20	180		

If $SSR(X_1) = 45$ and $SSR(X_2) = 25$,

- determine whether there is a significant relationship between Y and each independent variable at the 0.05 level of significance.
- compute the coefficients of partial determination, $r_{Y1.2}^2$ and $r_{Y2.1}^2$, and interpret their meaning.

14.32 The following is the ANOVA summary table for a multiple regression model with two independent variables:

Source	Degrees of Freedom	Sum of Squares	Mean Squares	F
Regression	2	30		
Error	10	120		
Total	12	150		

If $SSR(X_1) = 20$ and $SSR(X_2) = 15$,

- determine whether there is a significant relationship between Y and each independent variable at the 0.05 level of significance.

- compute the coefficients of partial determination, $r_{Y1.2}^2$ and $r_{Y2.1}^2$, and interpret their meaning.

APPLYING THE CONCEPTS

14.33 In Problem 14.5 on page 583, you used horsepower and weight to predict mileage (stored in **Auto2010**). Use the results from that problem.

- At the 0.05 level of significance, determine whether each independent variable makes a significant contribution to the regression model. On the basis of these results, indicate the most appropriate regression model for this set of data.
- Compute the coefficients of partial determination, $r_{Y1.2}^2$ and $r_{Y2.1}^2$, and interpret their meaning.



14.34 In Problem 14.4 on page 583, you used sales and number of orders to predict distribution costs at a mail-order catalog business (stored in **WareCost**). Use the results from that problem.

- At the 0.05 level of significance, determine whether each independent variable makes a significant contribution to the regression model. On the basis of these results, indicate the most appropriate regression model for this set of data.
- Compute the coefficients of partial determination, $r_{Y1.2}^2$ and $r_{Y2.1}^2$, and interpret their meaning.

14.35 In Problem 14.7 on page 584, you used the total staff present and remote hours to predict standby hours (stored in **Standby**). Use the results from that problem.

- At the 0.05 level of significance, determine whether each independent variable makes a significant contribution to the regression model. On the basis of these results, indicate the most appropriate regression model for this set of data.

- b. Compute the coefficients of partial determination, $r_{Y1.2}^2$ and $r_{Y2.1}^2$, and interpret their meaning.

14.36 In Problem 14.6 on page 583, you used radio advertising and newspaper advertising to predict sales (stored in **Advertise**). Use the results from that problem.

- a. At the 0.05 level of significance, determine whether each independent variable makes a significant contribution to the regression model. On the basis of these results, indicate the most appropriate regression model for this set of data.
- b. Compute the coefficients of partial determination, $r_{Y1.2}^2$ and $r_{Y2.1}^2$, and interpret their meaning.

14.37 In Problem 14.8 on page 584, you used land area of a property and age of a house to predict appraised value (stored in **GlenCove**). Use the results from that problem.

- a. At the 0.05 level of significance, determine whether each independent variable makes a significant contribution to the regression model. On the basis of these results, indicate the most appropriate regression model for this set of data.
- b. Compute the coefficients of partial determination, $r_{Y1.2}^2$ and $r_{Y2.1}^2$, and interpret their meaning.

14.6 Using Dummy Variables and Interaction Terms in Regression Models

The multiple regression models discussed in Sections 14.1 through 14.5 assumed that each independent variable is a numerical variable. For example, in Section 14.1, you used price and promotional expenditures, two numerical independent variables, to predict the monthly sales of OmniPower energy bars. However, for some models, you might want to include the effect of a categorical independent variable. For example, to predict the monthly sales of the OmniPower bars, you might want to include the categorical variable shelf location (not end-aisle or end-aisle) in the model.

Dummy Variables

To include a categorical independent variable in a regression model, you use a **dummy variable**. A dummy variable recodes the categories of a categorical variable using the numeric values 0 and 1. Where appropriate, the value of 0 is assigned to the absence of a characteristic and the value 1 is assigned to the presence of the characteristic. If a given categorical independent variable has only two categories, such as shelf location in the previous example, then you can define one dummy variable, X_d , to represent the two categories as

$X_d = 0$ if the observation is in category 1 (not end-aisle in the example)

$X_d = 1$ if the observation is in category 2 (end-aisle in the example)

To illustrate using dummy variables in regression, consider a business problem that involves developing a model for predicting the assessed value of houses (\$000), based on the size of the house (in thousands of square feet) and whether the house has a fireplace. To include the categorical variable for the presence of a fireplace, the dummy variable X_2 is defined as

$X_2 = 0$ if the house does not have a fireplace

$X_2 = 1$ if the house has a fireplace

Data collected from a sample of 15 houses are organized and stored in **House3**. Table 14.5 presents the data. In the last column of Table 14.5, you can see how the categorical values are converted to numerical values.

Assuming that the slope of assessed value with the size of the house is the same for houses that have and do not have a fireplace, the multiple regression model is

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

TABLE 14.5

Predicting Assessed Value, Based on Size of House and Presence of a Fireplace

Assessed Value	Size	Fireplace	Fireplace Coded
234.4	2.00	Yes	1
227.4	1.71	No	0
225.7	1.45	No	0
235.9	1.76	Yes	1
229.1	1.93	No	0
220.4	1.20	Yes	1
225.8	1.55	Yes	1
235.9	1.93	Yes	1
228.5	1.59	Yes	1
229.2	1.50	Yes	1
236.7	1.90	Yes	1
229.3	1.39	Yes	1
224.5	1.54	No	0
233.8	1.89	Yes	1
226.8	1.59	No	0

where

Y_i = assessed value, in thousands of dollars, for house i

β_0 = Y intercept

X_{1i} = size of the house, in thousands of square feet, for house i

β_1 = slope of assessed value with size of the house, holding constant the presence or absence of a fireplace

X_{2i} = dummy variable representing the absence or presence of a fireplace for house i

β_2 = net effect of the presence of a fireplace on assessed value, holding constant the size of the house

ε_i = random error in Y for house i

Figure 14.10 presents the regression results for this model.

FIGURE 14.10

Excel and Minitab regression results for the model that includes size of house and presence of fireplace

	A	B	C	D	E	F	G
1	Assessed Value Analysis						
2							
3	Regression Statistics						
4	Multiple R	0.9006					
5	R Square	0.8111					
6	Adjusted R Square	0.7796					
7	Standard Error	2.2626					
8	Observations	15					
9							
10	ANOVA						
11		df	SS	MS	F	Significance F	
12	Regression	2	263.7039	131.8520	25.7557	0.0000	
13	Residual	12	61.4321	5.1193			
14	Total	14	325.1360				
15							
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
17	Intercept	200.0905	4.3517	45.9803	0.0000	190.6090	209.5719
18	Size	16.1858	2.5744	6.2871	0.0000	10.5766	21.7951
19	FireplaceCoded	3.8530	1.2412	3.1042	0.0091	1.1486	6.5574

Regression Analysis: Value versus Size, FireplaceCoded

The regression equation is

Value = 200 + 16.2 Size + 3.85 FireplaceCoded

Predictor	Coef	SE Coef	T	P
Constant	200.090	4.352	45.98	0.000
Size	16.186	2.574	6.29	0.000
FireplaceCoded	3.853	1.241	3.10	0.009

S = 2.26260 R-Sq = 81.1% R-Sq(adj) = 78.0%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	263.70	131.85	25.76	0.000
Residual Error	12	61.43	5.12		
Total	14	325.14			

From Figure 14.10, the regression equation is

$$\hat{Y}_i = 200.0905 + 16.1858X_{1i} + 3.8530X_{2i}$$

For houses without a fireplace, you substitute $X_2 = 0$ into the regression equation:

$$\begin{aligned}\hat{Y}_i &= 200.0905 + 16.1858X_{1i} + 3.8530X_{2i} \\ &= 200.0905 + 16.1858X_{1i} + 3.8530(0) \\ &= 200.0905 + 16.1858X_{1i}\end{aligned}$$

For houses with a fireplace, you substitute $X_2 = 1$ into the regression equation:

$$\begin{aligned}\hat{Y}_i &= 200.0905 + 16.1858X_{1i} + 3.8530X_{2i} \\ &= 200.0905 + 16.1858X_{1i} + 3.8530(1) \\ &= 203.9435 + 16.1858X_{1i}\end{aligned}$$

In this model, the regression coefficients are interpreted as follows:

- Holding constant whether a house has a fireplace, for each increase of 1.0 thousand square feet in the size of the house, the predicted assessed value is estimated to increase by 16.1858 thousand dollars (i.e., \$16,185.80).
- Holding constant the size of the house, the presence of a fireplace is estimated to increase the predicted assessed value of the house by 3.8530 thousand dollars (i.e., \$3,853).

In Figure 14.10, the t_{STAT} test statistic for the slope of the size of the house with assessed value is 6.2871, and the p -value is approximately 0.000; the t_{STAT} test statistic for presence of a fireplace is 3.1042, and the p -value is 0.0091. Thus, each of the two variables makes a significant contribution to the model at the 0.01 level of significance. In addition, the coefficient of multiple determination indicates that 81.11% of the variation in assessed value is explained by variation in the size of the house and whether the house has a fireplace.

EXAMPLE 14.3

Modeling a Three-Level Categorical Variable

Define a multiple regression model using sales as the dependent variable and package design and price as independent variables. Package design is a three-level categorical variable with designs A , B , or C .

SOLUTION To model the three-level categorical variable package design, two dummy variables, X_1 and X_2 , are needed:

$$X_{1i} = 1 \text{ if package design } A \text{ is used in observation } i; 0 \text{ otherwise}$$

$$X_{2i} = 1 \text{ if package design } B \text{ is used in observation } i; 0 \text{ otherwise}$$

Thus, if observation i uses package design A , then $X_{1i} = 1$ and $X_{2i} = 0$; if observation i uses package design B , then $X_{1i} = 0$ and $X_{2i} = 1$; and if observation i uses package design C , then $X_{1i} = X_{2i} = 0$. A third independent variable is used for price:

$$X_{3i} = \text{price for observation } i$$

Thus, the regression model for this example is

$$Y_i = \beta_0 + \beta_1X_{1i} + \beta_2X_{2i} + \beta_3X_{3i} + \varepsilon_i$$

where

$$Y_i = \text{sales for observation } i$$

$$\beta_0 = Y \text{ intercept}$$

$$\beta_1 = \text{difference between the predicted sales of design } A \text{ and the predicted sales of design } C, \text{ holding price constant}$$

$$\beta_2 = \text{difference between the predicted sales of design } B \text{ and the predicted sales of design } C, \text{ holding price constant}$$

$$\beta_3 = \text{slope of sales with price, holding the package design constant}$$

$$\varepsilon_i = \text{random error in } Y \text{ for observation } i$$

Interactions

In all the regression models discussed so far, the effect an independent variable has on the dependent variable has been assumed to be independent of the other independent variables in the model. An **interaction** occurs if the effect of an independent variable on the dependent variable changes according to the *value* of a second independent variable. For example, it is possible for advertising to have a large effect on the sales of a product when the price of a product is low. However, if the price of the product is too high, increases in advertising will not dramatically change sales. In this case, price and advertising are said to interact. In other words, you cannot make general statements about the effect of advertising on sales. The effect that advertising has on sales is *dependent* on the price. You use an **interaction term** (sometimes referred to as a **cross-product term**) to model an interaction effect in a regression model.

To illustrate the concept of interaction and use of an interaction term, return to the example concerning the assessed values of homes discussed on pages 599–601. In the regression model, you assumed that the effect the size of the home has on the assessed value is independent of whether the house has a fireplace. In other words, you assumed that the slope of assessed value with size is the same for houses with fireplaces as it is for houses without fireplaces. If these two slopes are different, an interaction exists between the size of the home and the fireplace.

To evaluate whether an interaction exists, you first define an interaction term that is the product of the independent variable X_1 (size of house) and the dummy variable X_2 (FireplaceCoded). You then test whether this interaction variable makes a significant contribution to the regression model. If the interaction is significant, you cannot use the original model for prediction. For the data of Table 14.5 on page 600, you define the following:

$$X_3 = X_1 \times X_2$$

Figure 14.11 presents the results for this regression model, which includes the size of the house, X_1 , the presence of a fireplace, X_2 , and the interaction of X_1 and X_2 (defined as X_3).

FIGURE 14.11

Excel and Minitab regression results for a model that includes size, presence of fireplace, and interaction of size and fireplace

	A	B	C	D	E	F	G
1	Assessed Value Analysis						
2							
3	Regression Statistics						
4	Multiple R	0.9179					
5	R Square	0.8426					
6	Adjusted R Square	0.7996					
7	Standard Error	2.1573					
8	Observations	15					
9							
10	ANOVA						
11		df	SS	MS	F	Significance F	
12	Regression	3	273.9441	91.3147	19.6215	0.0001	
13	Residual	11	51.1919	4.6538			
14	Total	14	325.1360				
15							
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
17	Intercept	212.9522	9.6122	22.1544	0.0000	191.7959	234.1084
18	Size	8.3624	5.8173	1.4375	0.1784	-4.4414	21.1662
19	FireplaceCoded	-11.8404	10.6455	-1.1122	0.2898	-35.2710	11.5902
20	Size * FireplaceCoded	9.5180	6.4165	1.4834	0.1661	-4.6046	23.6406

Regression Analysis: Value versus Size, FireplaceCoded, Size*FireplaceCo

The regression equation is

$$\text{Value} = 213 + 8.36 \text{ Size} - 11.8 \text{ FireplaceCoded} + 9.52 \text{ Size*FireplaceCoded}$$

Predictor	Coef	SE Coef	T	P
Constant	212.952	9.612	22.15	0.000
Size	8.362	5.817	1.44	0.178
FireplaceCoded	-11.84	10.65	-1.11	0.290
Size*FireplaceCoded	9.518	6.416	1.48	0.166

$$S = 2.15727 \quad R\text{-Sq} = 84.3\% \quad R\text{-Sq(adj)} = 80.0\%$$

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	273.944	91.315	19.62	0.000
Residual Error	11	51.192	4.654		
Total	14	325.136			

To test for the existence of an interaction, you use the null hypothesis:

$$H_0: \beta_3 = 0$$

versus the alternative hypothesis:

$$H_1: \beta_3 \neq 0.$$

In Figure 14.11, the t_{STAT} test statistic for the interaction of size and fireplace is 1.4834. Because $t_{STAT} = 1.4834 < 2.201$ or the $p\text{-value} = 0.1661 > 0.05$, you do not reject the null hypothesis. Therefore, the interaction does not make a significant contribution to the model,

given that size and presence of a fireplace are already included. You can conclude that the slope of assessed value with size is the same for houses with fireplaces and without fireplaces.

Regression models can have several numerical independent variables. Example 14.4 illustrates a regression model in which there are two numerical independent variables and a categorical independent variable.

EXAMPLE 14.4

Studying a Regression Model That Contains a Dummy Variable

The business problem facing a real estate developer involves predicting heating oil consumption in single-family houses. The independent variables considered are atmospheric temperature, X_1 , and the amount of attic insulation, X_2 . Data are collected from a sample of 15 single-family houses. Of the 15 houses selected, houses 1, 4, 6, 7, 8, 10, and 12 are ranch-style houses. The data are organized and stored in [HeatingOil](#). Develop and analyze an appropriate regression model, using these three independent variables X_1 , X_2 , and X_3 (where X_3 is the dummy variable for ranch-style houses).

SOLUTION Define X_3 , a dummy variable for ranch-style house, as follows:

$$X_3 = 0 \text{ if the style is not ranch}$$

$$X_3 = 1 \text{ if the style is ranch}$$

Assuming that the slope between heating oil consumption and atmospheric temperature, X_1 , and between heating oil consumption and the amount of attic insulation, X_2 , is the same for both styles of houses, the regression model is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$$

where

Y_i = monthly heating oil consumption, in gallons, for house i

β_0 = Y intercept

β_1 = slope of heating oil consumption with atmospheric temperature, holding constant the effect of attic insulation and the style of the house

β_2 = slope of heating oil consumption with attic insulation, holding constant the effect of atmospheric temperature and the style of the house

β_3 = incremental effect of the presence of a ranch-style house, holding constant the effect of atmospheric temperature and attic insulation

ε_i = random error in Y for house i

Figure 14.12 presents results for this regression model.

FIGURE 14.12

Excel and Minitab results for a regression model that includes temperature, insulation, and ranch-style for the heating oil data

	A	B	C	D	E	F	G
1	Heating Oil Consumption Analysis						
2							
3	Regression Statistics						
4	Multiple R	0.9942					
5	R Square	0.9884					
6	Adjusted R Square	0.9853					
7	Standard Error	15.7489					
8	Observations	15					
9							
10	ANOVA						
11		df	SS	MS	F	Significance F	
12	Regression	3	233406.9094	77802.3031	313.6822	0.0000	
13	Residual	11	2728.3200	248.0291			
14	Total	14	236135.2293				
15							
16		Coefficients	Standard Error	t Stat	P value	Lower 95%	Upper 95%
17	Intercept	592.5401	14.3370	41.3295	0.0000	560.9846	624.0956
18	Temperature	-5.5251	0.2044	-27.0267	0.0000	-5.9751	-5.0752
19	Insulation	-21.3761	1.4480	-14.7623	0.0000	-24.5632	-18.1891
20	Ranch-style	-38.9727	8.3584	-4.6627	0.0007	-57.3695	-20.5759

Regression Analysis: Gallons versus Temperature, Insulation, Ranch-style

The regression equation is

Gallons = 593 - 5.53 Temperature - 21.4 Insulation - 39.0 Ranch-style

Predictor	Coef	SE Coef	T	P
Constant	592.54	14.34	41.33	0.000
Temperature	-5.5251	0.2044	-27.03	0.000
Insulation	-21.376	1.448	-14.76	0.000
Ranch-style	-38.973	8.358	-4.66	0.001

S = 15.7489 R-Sq = 98.8% R-Sq(adj) = 98.5%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	233407	77802	313.68	0.000
Residual Error	11	2728	248		
Total	14	236135			

From the results in Figure 14.12, the regression equation is

$$\hat{Y}_i = 592.5401 - 5.5251X_{1i} - 21.3761X_{2i} - 38.9727X_{3i}$$

For houses that are not ranch style, because $X_3 = 0$, the regression equation reduces to

$$\hat{Y}_i = 592.5401 - 5.5251X_{1i} - 21.3761X_{2i}$$

For houses that are ranch style, because $X_3 = 1$, the regression equation reduces to

$$\hat{Y}_i = 553.5674 - 5.5251X_{1i} - 21.3761X_{2i}$$

In this model, the regression coefficients are interpreted as follows:

- Holding constant the attic insulation and the house style, for each additional 1°F increase in atmospheric temperature, you estimate that the predicted heating oil consumption decreases by 5.5251 gallons.
- Holding constant the atmospheric temperature and the house style, for each additional 1-inch increase in attic insulation, you estimate that the predicted heating oil consumption decreases by 21.3761 gallons.
- b_3 measures the effect on oil consumption of having a ranch-style house ($X_3 = 1$) compared with having a house that is not ranch style ($X_3 = 0$). Thus, with atmospheric temperature and attic insulation held constant, you estimate that the predicted heating oil consumption is 38.9727 gallons less for a ranch-style house than for a house that is not ranch style.

The three t_{STAT} test statistics representing the slopes for temperature, insulation, and ranch style are -27.0267 , -14.7623 , and -4.6627 . Each of the corresponding p -values is extremely small (less than 0.001). Thus, each of the three variables makes a significant contribution to the model. In addition, the coefficient of multiple determination indicates that 98.84% of the variation in oil usage is explained by variation in temperature, insulation, and whether the house is ranch style.

Before you can use the model in Example 14.4, you need to determine whether the independent variables interact with each other. In Example 14.5, three interaction terms are added to the model.

EXAMPLE 14.5

Evaluating a Regression Model with Several Interactions

For the data of Example 14.4, determine whether adding the interaction terms make a significant contribution to the regression model.

SOLUTION To evaluate possible interactions between the independent variables, three interaction terms are constructed as follows: $X_4 = X_1 \times X_2$, $X_5 = X_1 \times X_3$, and $X_6 = X_2 \times X_3$. The regression model is now

$$Y_i = \beta_0 + \beta_1X_{1i} + \beta_2X_{2i} + \beta_3X_{3i} + \beta_4X_{4i} + \beta_5X_{5i} + \beta_6X_{6i} + \varepsilon_i$$

where X_1 is temperature, X_2 is insulation, X_3 is the dummy variable ranch style, X_4 is the interaction between temperature and insulation, X_5 is the interaction between temperature and ranch style, and X_6 is the interaction between insulation and ranch style. Figure 14.13 presents the results for this regression model.

FIGURE 14.13

Excel and Minitab regression results for a model that includes temperature, X_1 ; insulation, X_2 ; the dummy variable ranch-style, X_3 ; the interaction of temperature and insulation, X_4 ; the interaction of temperature and ranch-style, X_5 ; and the interaction of insulation and ranch-style, X_6

	A	B	C	D	E	F	G
1	Heating Oil Consumption Analysis						
2							
3	Regression Statistics						
4	Multiple R	0.9966					
5	R Square	0.9931					
6	Adjusted R Square	0.9880					
7	Standard Error	14.2506					
8	Observations	15					
9							
10	ANOVA						
11		df	SS	MS	F	Significance F	
12	Regression	6	234510.5818	39085.0970	192.4607	0.0000	
13	Residual	8	1624.6475	203.0809			
14	Total	14	236135.2293				
15							
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
17	Intercept	642.8867	26.7059	24.0728	0.0000	581.3027	704.4707
18	Temperature	-6.9263	0.7531	-9.1969	0.0000	-8.6629	-5.1896
19	Insulation	-27.8825	3.5801	-7.7882	0.0001	-36.1383	-19.6268
20	Style	-84.6088	29.9956	-2.8207	0.0225	-153.7788	-15.4389
21	Temperature * Insulation	0.1702	0.0886	1.9204	0.0911	-0.0342	0.3746
22	Temperature * Ranch-style	0.6596	0.4617	1.4286	0.1910	-0.4051	1.7242
23	Insulation * Ranch-style	4.9870	3.5137	1.4193	0.1936	-3.1156	13.0895

Regression Analysis: Gallons versus Temperature, Insulation, —

The regression equation is

$$\text{Gallons} = 643 - 6.93 \text{ Temperature} - 27.9 \text{ Insulation} - 84.6 \text{ Ranch-style} + 0.170 \text{ Temperature*Insulation} + 0.660 \text{ Temperature*Ranch-style} + 4.99 \text{ Insulation*Ranch-style}$$

Predictor	Coef	SE Coef	T	P
Constant	642.89	26.71	24.07	0.000
Temperature	-6.9263	0.7531	-9.20	0.000
Insulation	-27.883	3.580	-7.79	0.000
Ranch-style	-84.61	30.00	-2.82	0.022
Temperature*Insulation	0.17021	0.08863	1.92	0.091
Temperature*Ranch-style	0.6596	0.4617	1.43	0.191
Insulation*Ranch-style	4.987	3.514	1.42	0.194

S = 14.2506 R-Sq = 99.3% R-Sq(adj) = 98.8%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	6	234511	39085	192.46	0.000
Residual Error	8	1625	203		
Total	14	236135			

To test whether the three interactions significantly improve the regression model, you use the partial F test. The null and alternative hypotheses are

$$H_0: \beta_4 = \beta_5 = \beta_6 = 0 \text{ (There are no interaction among } X_1, X_2, \text{ and } X_3.)$$

$$H_1: \beta_4 \neq 0 \text{ and/or } \beta_5 \neq 0 \text{ and/or } \beta_6 \neq 0 \text{ (} X_1 \text{ interacts with } X_2, \text{ and/or } X_1 \text{ interacts with } X_3, \text{ and/or } X_2 \text{ interacts with } X_3.)$$

From Figure 14.13,

$$SSR(X_1, X_2, X_3, X_4, X_5, X_6) = 234,510.5818 \text{ with 6 degrees of freedom}$$

and from Figure 14.12 on page 603, $SSR(X_1, X_2, X_3) = 233,406.9094$ with 3 degrees of freedom. Thus,

$$SSR(X_1, X_2, X_3, X_4, X_5, X_6) - SSR(X_1, X_2, X_3) = 234,510.5818 - 233,406.9094 = 1,103.6724.$$

The difference in degrees of freedom is $6 - 3 = 3$.

To use the partial F test for the simultaneous contribution of three variables to a model, you use an extension of Equation (14.11) on page 616.² The partial F_{STAT} test statistic is

$$F_{STAT} = \frac{[SSR(X_1, X_2, X_3, X_4, X_5, X_6) - SSR(X_1, X_2, X_3)]/3}{MSE(X_1, X_2, X_3, X_4, X_5, X_6)} = \frac{1,103.6724/3}{203.0809} = 1.8115$$

You compare the computed F_{STAT} test statistic to the critical F value for 3 and 8 degrees of freedom. Using a level of significance of 0.05, the critical F value from Table E.5 is 4.07. Because $F_{STAT} = 1.8115 < 4.07$, you conclude that the interactions do not make a significant contribution to the model, given that the model already includes temperature, X_1 ; insulation, X_2 ; and whether the house is ranch style, X_3 . Therefore, the multiple regression model using X_1, X_2 , and X_3 but no interaction terms is the better model. If you rejected this null hypothesis, you would then test the contribution of each interaction separately in order to determine which interaction terms to include in the model.

²In general, if a model has several independent variables and you want to test whether additional independent variables contribute to the model, the numerator of the F test is SSR (for all independent variables) minus SSR (for the initial set of variables) divided by the number of independent variables whose contribution is being tested.

Problems for Section 14.6

LEARNING THE BASICS

14.38 Suppose X_1 is a numerical variable and X_2 is a dummy variable and the regression equation for a sample of $n = 20$ is

$$\hat{Y}_i = 6 + 4X_{1i} + 2X_{2i}$$

- Interpret the regression coefficient associated with variable X_1 .
- Interpret the regression coefficient associated with variable X_2 .
- Suppose that the t_{STAT} test statistic for testing the contribution of variable X_2 is 3.27. At the 0.05 level of significance, is there evidence that variable X_2 makes a significant contribution to the model?

APPLYING THE CONCEPTS

14.39 The chair of the accounting department plans to develop a regression model to predict the grade point average in accounting for those students who are graduating and have completed the accounting major, based on the student's SAT score and whether the student received a grade of B or higher in the introductory statistics course (0 = no and 1 = yes).

- Explain the steps involved in developing a regression model for these data. Be sure to indicate the particular models you need to evaluate and compare.
- Suppose the regression coefficient for the variable whether the student received a grade of B or higher in the introductory statistics course is +0.30. How do you interpret this result?

14.40 A real estate association in a suburban community would like to study the relationship between the size of a single-family house (as measured by the number of rooms) and the selling price of the house (in thousands of dollars). Two different neighborhoods are included in the study, one on the east side of the community (=0) and the other on the west side (=1). A random sample of 20 houses was selected, with the results stored in **Neighbor**. For (a) through (k), do not include an interaction term.

- State the multiple regression equation that predicts the selling price, based on the number of rooms and the neighborhood.
- Interpret the regression coefficients in (a).
- Predict the selling price for a house with nine rooms that is located in an east-side neighborhood. Construct a 95% confidence interval estimate and a 95% prediction interval.
- Perform a residual analysis on the results and determine whether the regression assumptions are valid.
- Is there a significant relationship between selling price and the two independent variables (rooms and neighborhood) at the 0.05 level of significance?

- At the 0.05 level of significance, determine whether each independent variable makes a contribution to the regression model. Indicate the most appropriate regression model for this set of data.
- Construct and interpret a 95% confidence interval estimate of the population slope for the relationship between selling price and number of rooms.
- Construct and interpret a 95% confidence interval estimate of the population slope for the relationship between selling price and neighborhood.
- Compute and interpret the adjusted r^2 .
- Compute the coefficients of partial determination and interpret their meaning.
- What assumption do you need to make about the slope of selling price with number of rooms?
- Add an interaction term to the model and, at the 0.05 level of significance, determine whether it makes a significant contribution to the model.
- On the basis of the results of (f) and (l), which model is most appropriate? Explain.

14.41 The marketing manager of a large supermarket chain faced the business problem of determining the effect on the sales of pet food of shelf space and whether the product was placed at the front (=1) or back (=0) of the aisle. Data are collected from a random sample of 12 equal-sized stores. The results are shown in the following table (and organized and stored in **Petfood**):

Store	Shelf Space (Feet)	Location	Weekly Sales (\$)
1	5	Back	160
2	5	Front	220
3	5	Back	140
4	10	Back	190
5	10	Back	240
6	10	Front	260
7	15	Back	230
8	15	Back	270
9	15	Front	280
10	20	Back	260
11	20	Back	290
12	20	Front	310

For (a) through (m), do not include an interaction term.

- State the multiple regression equation that predicts weekly sales based on shelf space and location.
- Interpret the regression coefficients in (a).
- Predict the weekly sales of pet food for a store with 8 feet of shelf space situated at the back of the aisle. Construct a 95% confidence interval estimate and a 95% prediction interval.

- d. Perform a residual analysis on the results and determine whether the regression assumptions are valid.
- e. Is there a significant relationship between sales and the two independent variables (shelf space and aisle position) at the 0.05 level of significance?
- f. At the 0.05 level of significance, determine whether each independent variable makes a contribution to the regression model. Indicate the most appropriate regression model for this set of data.
- g. Construct and interpret 95% confidence interval estimates of the population slope for the relationship between sales and shelf space and between sales and aisle location.
- h. Compare the slope in (b) with the slope for the simple linear regression model of Problem 13.4 on page 531. Explain the difference in the results.
- i. Compute and interpret the meaning of the coefficient of multiple determination, r^2 .
- j. Compute and interpret the adjusted r^2 .
- k. Compare r^2 with the r^2 value computed in Problem 13.16 (a) on page 537.
- l. Compute the coefficients of partial determination and interpret their meaning.
- m. What assumption about the slope of shelf space with sales do you need to make in this problem?
- n. Add an interaction term to the model and, at the 0.05 level of significance, determine whether it makes a significant contribution to the model.
- o. On the basis of the results of (f) and (n), which model is most appropriate? Explain.

14.42 In mining engineering, holes are often drilled through rock, using drill bits. As a drill hole gets deeper, additional rods are added to the drill bit to enable additional drilling to take place. It is expected that drilling time increases with depth. This increased drilling time could be caused by several factors, including the mass of the drill rods that are strung together. The business problem relates to whether drilling is faster using dry drilling holes or wet drilling holes. Using dry drilling holes involves forcing compressed air down the drill rods to flush the cuttings and drive the hammer. Using wet drilling holes involves forcing water rather than air down the hole. Data have been collected from a sample of 50 drill holes that contains measurements of the time to drill each additional 5 feet (in minutes), the depth (in feet), and whether the hole was a dry drilling hole or a wet drilling hole. The data are organized and stored in. Develop a model to predict additional drilling time, based on depth and type of drilling hole (dry or wet). For (a) through (k) do not include an interaction term **Drill**.


Source: Data extracted from R. Penner and D. G. Watts, "Mining Information," *The American Statistician*, 45, 1991, pp. 4–9.

- a. State the multiple regression equation.
- b. Interpret the regression coefficients in (a).
- c. Predict the additional drilling time for a dry drilling hole at a depth of 100 feet. Construct a 95% confidence interval estimate and a 95% prediction interval.
- d. Perform a residual analysis on the results and determine whether the regression assumptions are valid.
- e. Is there a significant relationship between sales and the two independent variables (shelf space and aisle position) at the 0.05 level of significance?
- f. At the 0.05 level of significance, determine whether each independent variable makes a contribution to the regression model. Indicate the most appropriate regression model for this set of data.
- g. Construct a 95% confidence interval estimate of the population slope for the relationship between sales and shelf space and between sales and aisle location.
- h. Compare the slope in (b) with the slope for the simple linear regression model of Problem 13.4 on page 531. Explain the difference in the results.
- i. Compute and interpret the meaning of the coefficient of multiple determination, r^2 .
- j. Compute and interpret the adjusted r^2 .
- k. Compare r^2 with the r^2 value computed in Problem 13.16 (a) on page 537.
- l. Compute the coefficients of partial determination and interpret their meaning.
- m. What assumption about the slope of shelf space with sales do you need to make in this problem?
- n. Add an interaction term to the model and, at the 0.05 level of significance, determine whether it makes a significant contribution to the model.
- o. On the basis of the results of (f) and (n), which model is most appropriate? Explain.

14.43 The owner of a moving company typically has his most experienced manager predict the total number of labor hours that will be required to complete an upcoming move. This approach has proved useful in the past, but the owner has the business objective of developing a more accurate method of predicting labor hours. In a preliminary effort to provide a more accurate method, the owner has decided to use the number of cubic feet moved and whether there is an elevator in the apartment building as the independent variables and has collected data for 36 moves in which the origin and destination were within the borough of Manhattan in New York City and the travel time was an insignificant portion of the hours worked. The data are organized and stored in. For (a) through (k), do not include an interaction term **Moving**.

- a. State the multiple regression equation for predicting labor hours, using the number of cubic feet moved and whether there is an elevator.
- b. Interpret the regression coefficients in (a).
- c. Predict the labor hours for moving 500 cubic feet in an apartment building that has an elevator and construct a 95% confidence interval estimate and a 95% prediction interval.
- d. Perform a residual analysis on the results and determine whether the regression assumptions are valid.
- e. Is there a significant relationship between labor hours and the two independent variables (cubic feet moved and whether there is an elevator in the apartment building) at the 0.05 level of significance?

- f. At the 0.05 level of significance, determine whether each independent variable makes a contribution to the regression model. Indicate the most appropriate regression model for this set of data.
- g. Construct a 95% confidence interval estimate of the population slope for the relationship between labor hours and cubic feet moved.
- h. Construct a 95% confidence interval estimate for the relationship between labor hours and the presence of an elevator.
- i. Compute and interpret the adjusted r^2 .
- j. Compute the coefficients of partial determination and interpret their meaning.
- k. What assumption do you need to make about the slope of labor hours with cubic feet moved?
- l. Add an interaction term to the model and, at the 0.05 level of significance, determine whether it makes a significant contribution to the model.
- m. On the basis of the results of (f) and (l), which model is most appropriate? Explain.

 **14.44** In Problem 14.4 on page 583, you used sales and orders to predict distribution cost (stored in **WareCost**). Develop a regression model to predict distribution cost that includes sales, orders, and the interaction of sales and orders.

- a. At the 0.05 level of significance, is there evidence that the interaction term makes a significant contribution to the model?
- b. Which regression model is more appropriate, the one used in (a) or the one used in Problem 14.4? Explain.

14.45 Zagat's publishes restaurant ratings for various locations in the United States. The file **Restaurants** contains the Zagat rating for food, décor, service, and cost per person for a sample of 50 restaurants located in a city and 50 restaurants located in a suburb. Develop a regression model to predict the cost per person, based on a variable that represents the sum of the ratings for food, décor, and service and a dummy variable concerning location (city vs. suburban). For (a) through (m), do not include an interaction term.

Sources: Extracted from *Zagat Survey 2010, New York City Restaurants*; and *Zagat Survey 2009–2010, Long Island Restaurants*.

- a. State the multiple regression equation.
- b. Interpret the regression coefficients in (a).
- c. Predict the cost for a restaurant with a summated rating of 60 that is located in a city and construct a 95% confidence interval estimate and a 95% prediction interval.
- d. Perform a residual analysis on the results and determine whether the regression assumptions are satisfied.
- e. Is there a significant relationship between price and the two independent variables (summated rating and location) at the 0.05 level of significance?
- f. At the 0.05 level of significance, determine whether each independent variable makes a contribution to the regression model. Indicate the most appropriate regression model for this set of data.

- g. Construct a 95% confidence interval estimate of the population slope for the relationship between cost and summated rating.
- h. Compare the slope in (b) with the slope for the simple linear regression model of Problem 13.5 on page 531. Explain the difference in the results.
- i. Compute and interpret the meaning of the coefficient of multiple determination.
- j. Compute and interpret the adjusted r^2 .
- k. Compare r^2 with the r^2 value computed in Problem 13.17 (b) on page 537.
- l. Compute the coefficients of partial determination and interpret their meaning.
- m. What assumption about the slope of cost with summated rating do you need to make in this problem?
- n. Add an interaction term to the model and, at the 0.05 level of significance, determine whether it makes a significant contribution to the model.
- o. On the basis of the results of (f) and (n), which model is most appropriate? Explain.

14.46 In Problem 14.6 on page 583, you used radio advertising and newspaper advertising to predict sales (stored in **Advertise**). Develop a regression model to predict sales that includes radio advertising, newspaper advertising, and the interaction of radio advertising and newspaper advertising.

- a. At the 0.05 level of significance, is there evidence that the interaction term makes a significant contribution to the model?
- b. Which regression model is more appropriate, the one used in this problem or the one used in Problem 14.6? Explain.

14.47 In Problem 14.5 on page 583, horsepower and weight were used to predict miles per gallon (stored in **Auto2010**). Develop a regression model that includes horsepower, weight, and the interaction of horsepower and weight to predict miles per gallon.

- a. At the 0.05 level of significance, is there evidence that the interaction term makes a significant contribution to the model?
- b. Which regression model is more appropriate, the one used in this problem or the one used in Problem 14.5? Explain.

14.48 In Problem 14.7 on page 584, you used total staff present and remote hours to predict standby hours (stored in **Standby**). Develop a regression model to predict standby hours that includes total staff present, remote hours, and the interaction of total staff present and remote hours.

- a. At the 0.05 level of significance, is there evidence that the interaction term makes a significant contribution to the model?
- b. Which regression model is more appropriate, the one used in this problem or the one used in Problem 14.7? Explain.

14.49 The director of a training program for a large insurance company has the business objective of determining which training method is best for training underwriters. The three methods to be evaluated are traditional, CD-ROM based, and Web based. The 30 trainees are divided into three randomly assigned groups of 10. Before the start of the training, each trainee is given a proficiency exam that measures mathematics and computer skills. At the end of the training, all students take the same end-of-training exam. The results are organized and stored in **Underwriting**.

Develop a multiple regression model to predict the score on the end-of-training exam, based on the score on the proficiency exam and the method of training used. For (a) through (k), do not include an interaction term.

- State the multiple regression equation.
- Interpret the regression coefficients in (a).
- Predict the end-of-training exam score for a student with a proficiency exam score of 100 who had Web-based training.
- Perform a residual analysis on your results and determine whether the regression assumptions are valid.
- Is there a significant relationship between the end-of-training exam score and the independent variables (proficiency score and training method) at the 0.05 level of significance?
- At the 0.05 level of significance, determine whether each independent variable makes a contribution to the regression model. Indicate the most appropriate regression model for this set of data.
- Construct and interpret a 95% confidence interval estimate of the population slope for the relationship between the end-of-training exam score and the proficiency exam score.
- Construct and interpret a 95% confidence interval estimate of the population slope for the relationship between the end-of-training exam score and type of training method.
- Compute and interpret the adjusted r^2 .
- Compute the coefficients of partial determination and interpret their meaning.
- What assumption about the slope of proficiency score with end-of-training exam score do you need to make in this problem?
- Add interaction terms to the model and, at the 0.05 level of significance, determine whether any interaction terms make a significant contribution to the model.
- On the basis of the results of (f) and (l), which model is most appropriate? Explain.

14.7 Logistic Regression

The discussion of the simple linear regression model in Chapter 13 and the multiple regression models in Sections 14.1 through 14.6 only considered *numerical* dependent variables. However, in many instances, the dependent variable is a *categorical* variable that takes on one of only two possible values such as a customer prefers Brand *A* or a customer prefers Brand *B*. Using a categorical dependent variable violates the normality assumption of least-squares and can also result in predicted Y values that are impossible.

An alternative approach to least-squares regression originally applied to survival data in the health sciences (see reference 1), **logistic regression**, enables you to use regression models to predict the probability of a particular categorical response for a given set of independent variables. The logistic regression model uses the **odds ratio**, which represents the probability of an event of interest compared with the probability of not having an event of interest. Equation (14.15) defines the odds ratio.

ODDS RATIO

$$\text{Odds ratio} = \frac{\text{Probability of an event of interest}}{1 - \text{Probability of an event of interest}} \quad (14.15)$$

Using Equation (14.15), if the probability of an event of interest is 0.50, the odds ratio is

$$\text{Odds ratio} = \frac{0.50}{1 - 0.50} = 1.0, \text{ or } 1 \text{ to } 1$$

If the probability of an event of interest is 0.75, the odds ratio is

$$\text{Odds ratio} = \frac{0.75}{1 - 0.75} = 3.0, \text{ or } 3 \text{ to } 1$$

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

The logistic regression model is based on the natural logarithm (ln) of this odds ratio.³ Equation (14.16) defines the logistic regression model for k independent variables.

LOGISTIC REGRESSION MODEL

$$\ln(\text{Odds ratio}) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i \quad (14.16)$$

where

k = number of independent variables in the model

ε_i = random error in observation i

In Sections 13.2 and 14.1, the method of least squares was used to develop a regression equation. In logistic regression, a mathematical method called *maximum likelihood estimation* is usually used to develop a regression equation to predict the natural logarithm of this odds ratio. Equation (14.17) defines the logistic regression equation.

LOGISTIC REGRESSION EQUATION

$$\ln(\text{Estimated odds ratio}) = b_0 + b_1 X_{1i} + b_2 X_{2i} + \cdots + b_k X_{ki} \quad (14.17)$$

Once you have determined the logistic regression equation, you use Equation (14.18) to compute the estimated odds ratio.

ESTIMATED ODDS RATIO

$$\text{Estimated odds ratio} = e^{\ln(\text{Estimated odds ratio})} \quad (14.18)$$

Once you have computed the estimated odds ratio, you use Equation (14.19) to find the estimated probability of an event of interest.

ESTIMATED PROBABILITY OF AN EVENT OF INTEREST

$$\text{Estimated probability of an event of interest} = \frac{\text{Estimated odds ratio}}{1 + \text{Estimated odds ratio}} \quad (14.19)$$

To illustrate the logistic regression model, the marketing department for a credit card company wants to organize a campaign to convince existing holders of the company's standard credit card to upgrade to the company's premium card for a nominal annual fee. The marketing department begins with the question "Which of the existing standard credit cardholders should be the target for the campaign?"

The department has access to data from a sample of 30 cardholders who were contacted during last year's campaign. That data indicates whether the cardholder upgraded to a premium

card (0 = no, 1 = yes). The department wants to predict the categorical variable (i.e., did the customer upgrade to a premium card?) using two independent variables: total amount of credit card purchases (in thousands of dollars) in the prior year (X_1), and whether the cardholder ordered additional credit cards (at extra cost) for other members of the household (X_2 ; 0 = no, 1 = yes). Figure 14.14 presents results for the logistic regression model, the data for which are stored in [Logpurch](#).

FIGURE 14.14

Minitab logistic regression results for the credit card marketing data

Excel does not contain any logistic regression functions, but logistic regression analysis can be done using the Excel Solver add-in (beyond the scope of this book).

Binary Logistic Regression: Upgraded versus Purchases, Extra Cards

Link Function: Logit

Response Information

Variable	Value	Count	
Upgraded	1	13	(Event)
	0	17	
Total		30	

Logistic Regression Table

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI Lower	95% CI Upper
Constant	-6.93984	2.94712	-2.35	0.019			
Purchases	0.139469	0.0680641	2.05	0.040	1.15	1.01	1.31
Extra Cards							
1	2.77434	1.19267	2.33	0.020	16.03	1.55	165.99

Log-Likelihood = -10.038

Test that all slopes are zero: G = 20.977, DF = 2, P-Value = 0.000

Goodness-of-Fit Tests

Method	Chi-Square	DF	P
Pearson	18.5186	27	0.887
Deviance	20.0769	27	0.828
Hosmer-Lemeshow	6.5174	8	0.589

In this model, the regression coefficients are interpreted as follows:

- The regression constant b_0 is -6.940 . This means that for a credit cardholder who did not charge any purchases last year and who does not have additional cards, the estimated natural logarithm of the odds ratio of purchasing the premium card is -6.940 .
- The regression coefficient b_1 is 0.13947 . This means that holding constant the effect of whether the credit cardholder has additional cards for members of the household, for each increase of \$1,000 in annual credit card spending using the company's card, the estimated natural logarithm of the odds ratio of purchasing the premium card increases by 0.13947 . Therefore, cardholders who charged more in the previous year are more likely to upgrade to a premium card.
- The regression coefficient b_2 is 2.774 . This means that holding constant the annual credit card spending, the estimated natural logarithm of the odds ratio of purchasing the premium card increases by 2.774 for a credit cardholder who has additional cards for members of the household compared with one who does not have additional cards. Therefore, cardholders possessing additional cards for other members of the household are much more likely to upgrade to a premium card.

The regression coefficients suggest that the credit card company should develop a marketing campaign that targets cardholders who tend to charge large amounts to their cards, and to households that possess more than one card.

As was the case with least-squares regression models, a main purpose of performing logistic regression analysis is to provide predictions of a dependent variable. For example, consider a cardholder who charged \$36,000 last year and possesses additional cards for members of the household. What is the probability the cardholder will upgrade to the

premium card during the marketing campaign? Using $X_1 = 36, X_2 = 1$, Equation (14.17) on page 610, and the results displayed in Figure 14.14 on page 611,

$$\begin{aligned}\ln(\text{estimated odds of purchasing versus not purchasing}) &= -6.94 + (0.13947)(36) + (2.774)(1) \\ &= 0.85492\end{aligned}$$

Then, using Equation (14.18) on page 610,

$$\text{Estimated odds ratio} = e^{0.85492} = 2.3512$$

Therefore, the odds are 2.3512 to 1 that a credit cardholder who spent \$36,000 last year and has additional cards will purchase the premium card during the campaign. Using Equation (14.19) on page 610, you can convert this odds ratio to a probability:

$$\begin{aligned}\text{estimated probability of purchasing premium card} &= \frac{2.3512}{1 + 2.3512} \\ &= 0.7016\end{aligned}$$

Thus, the estimated probability is 0.7016 that a credit cardholder who spent \$36,000 last year and has additional cards will purchase the premium card during the campaign. In other words, you predict 70.16% of such individuals will purchase the premium card.

Now that you have used the logistic regression model for prediction, you need to determine whether or not the model is a good-fitting model. The **deviance statistic** is frequently used to determine whether or not the current model provides a good fit to the data. This statistic measures the fit of the current model compared with a model that has as many parameters as there are data points (what is called a *saturated* model). The deviance statistic follows a chi-square distribution with $n-k-1$ degrees of freedom. The null and alternative hypotheses are

H_0 : The model is a good-fitting model.

H_1 : The model is not a good-fitting model.

When using the deviance statistic for logistic regression, the null hypothesis represents a good-fitting model, which is the opposite of the null hypothesis when using the overall F test for the multiple regression model (see Section 14.2). Using the α level of significance, the decision rule is

$$\begin{aligned}\text{Reject } H_0 \text{ if deviance} &> \chi_{\alpha}^2 \\ \text{Otherwise, do not reject } H_0.\end{aligned}$$

The critical value for a χ^2 statistic with $n-k-1 = 30-2-1 = 27$ degrees of freedom is 40.113 (see Table E.4). From Figure 14.14 on page 611, the deviance = 20.08 < 40.113, or the p -value = 0.828 > 0.05. Thus, you do not reject H_0 , and you conclude that the model is a good-fitting one.

Now that you have concluded that the model is a good-fitting one, you need to evaluate whether each of the independent variables makes a significant contribution to the model in the presence of the others. As was the case with linear regression in Sections 13.7 and 14.4, the test statistic is based on the ratio of the regression coefficient to the standard error of the regression coefficient. In logistic regression, this ratio is defined by the **Wald statistic**, which approximately follows the normal distribution. From Figure 14.14, the Wald statistic (labeled Z) is 2.05 for X_1 and 2.33 for X_2 . Each of these is greater than the critical value of +1.96 for the normal distribution at the 0.05 level of significance (the p -values are 0.04 and 0.02). You can conclude that each of the two independent variables makes a contribution to the model in the presence of the other. Therefore, you should include both these independent variables in the model.

Problems for Section 14.7

LEARNING THE BASICS

14.50 Interpret the meaning of a slope coefficient equal to 2.2 in logistic regression.

14.51 Given an estimated odds ratio of 2.5, compute the estimated probability of an event of interest.

14.52 Given an estimated odds ratio of 0.75, compute the estimated probability of an event of interest.

14.53 Consider the following logistic regression equation:

$$\ln(\text{Estimated odds ratio}) = 0.1 + 0.5X_{1i} + 0.2X_{2i}$$

- Interpret the meaning of the logistic regression coefficients.
- If $X_1 = 2$ and $X_2 = 1.5$, compute the estimated odds ratio and interpret its meaning.
- On the basis of the results of (b), compute the estimated probability of an event of interest.

APPLYING THE CONCEPTS

14.54 Refer to Figure 14.14 on page 611.

- Predict the probability that a cardholder who charged \$36,000 last year and does not have any additional credit cards for members of the household will purchase the premium card during the marketing campaign.
- Compare the results in (a) with those for a person with additional credit cards.
- Predict the probability that a cardholder who charged \$18,000 and does not have any additional credit cards for members of the household will purchase the premium card during the marketing campaign.
- Compare the results of (a) and (c) and indicate what implications these results might have for the strategy for the marketing campaign.

14.55 Undergraduate students at Miami University in Oxford, Ohio, were surveyed in order to evaluate the effect of price on the purchase of a pizza from Pizza Hut. The students were asked to suppose that they were going to have a large two-topping pizza delivered to their residence. Then they were asked to select from either Pizza Hut or another pizzeria of their choice. The price they would have to pay to get a Pizza Hut pizza differed from survey to survey. For example, some surveys used the price \$11.49. Other prices investigated were \$8.49, \$9.49, \$10.49, \$12.49, \$13.49, and \$14.49. The dependent variable for this study is whether or not a student will select Pizza Hut. Possible independent variables are the price of a Pizza Hut pizza and the gender of the student. The data set [PizzaHut](#) has 220 observations and three variables:

Gender (1=male, 0=female)

Price (8.49, 9.49, 10.49, 11.49, 12.49, 13.49, or 14.49)

Purchase (1=the student selected Pizza Hut, 0=the student selected another pizzeria)

- Develop a logistic regression model to predict the probability that a student selects Pizza Hut based on the price of the pizza. Is price an important indicator of purchase selection?
- Develop a logistic regression model to predict the probability that a student selects Pizza Hut based on the price of the pizza and the gender of the student. Is price an important indicator of purchase selection? Is gender an important indicator of purchase selection?
- Compare the results from (a) and (b). Which model would you choose? Discuss.
- Using the model selected in (c), predict the probability that a student will select Pizza Hut if the price is \$8.99.
- Using the model selected in (c), predict the probability that a student will select Pizza Hut if the price is \$11.49.
- Using the model selected in (c), predict the probability that a student will select Pizza Hut if the price is \$13.99.

14.56 The director of graduate studies at a college of business wants to predict the success of students in an MBA program using two independent variables, undergraduate grade point average (GPA) and GMAT score. A random sample of 30 students (stored in [MBA](#)) indicates that 20 successfully completed the program (coded as 1) and 10 did not (coded as 0).

Success in MBA Program	Under- graduate GPA	GMAT Score	Success in MBA Program	Under- graduate GPA	GMAT Score
0	2.93	617	1	3.17	639
0	3.05	557	1	3.24	632
0	3.11	599	1	3.41	639
0	3.24	616	1	3.37	619
0	3.36	594	1	3.46	665
0	3.41	567	1	3.57	694
0	3.45	542	1	3.62	641
0	3.60	551	1	3.66	594
0	3.64	573	1	3.69	678
0	3.57	536	1	3.70	624
1	2.75	688	1	3.78	654
1	2.81	647	1	3.84	718
1	3.03	652	1	3.77	692
1	3.10	608	1	3.79	632
1	3.06	680	1	3.97	784

- Develop a logistic regression model to predict the probability of successful completion of the MBA program based on undergraduate grade point average and GMAT score.
- Explain the meaning of the regression coefficients for the model in (a).
- Predict the probability of successful completion of the program for a student with an undergraduate grade point average of 3.25 and a GMAT score of 600.
- At the 0.05 level of significance, is there evidence that a logistic regression model that uses undergraduate grade

point average and GMAT score to predict probability of success in the MBA program is a good-fitting model?

- e. At the 0.05 level of significance, is there evidence that undergraduate grade point average and GMAT score each make a significant contribution to the logistic regression model?
- f. Develop a logistic regression model that includes only undergraduate grade point average to predict probability of success in the MBA program.
- g. Develop a logistic regression model that includes only GMAT score to predict probability of success in the MBA program.
- h. Compare the models in (a), (f), and (g). Evaluate the differences among the models.

14.57 A hotel has designed a new system for room service delivery of breakfast that allows the customer to select a specific delivery time. The difference between the actual and requested delivery times was recorded (a negative time means that the breakfast was delivered before the requested time) for 30 deliveries on a particular day along with

whether the customer had previously stayed at the hotel. The data are stored in the file **Satisfaction**.

- a. Develop a logistic regression model to predict the probability that the customer will be satisfied (0 = unfavorable, 1 = favorable), based on the delivery time difference and whether the customer had previously stayed at the hotel.
- b. Explain the meaning of the regression coefficients for the model in (a).
- c. Predict the probability that the customer will be satisfied if the delivery time difference is +3 minutes and he or she did not previously stay at the hotel.
- d. At the 0.05 level of significance, is there evidence that a logistic regression model that uses delivery time difference and whether the customer had previously stayed at the hotel is a good-fitting model?
- e. At the 0.05 level of significance, is there evidence that both independent variables (delivery time difference and whether the customer had previously stayed at the hotel) make a significant contribution to the logistic regression model?

USING STATISTICS



@ OmniFoods Revisited

In the Using Statistics scenario, you were the marketing manager for OmniFoods, a large food products company planning a nationwide introduction of a new high-energy bar, OmniPower. You needed to determine the effect that price and in-store promotions would have on sales of OmniPower in order to develop an effective marketing strategy. A sample of 34 stores in a supermarket chain was selected for a test-market study. The stores charged between 59 and 99 cents per bar and were given an in-store promotion budget between \$200 and \$600.

At the end of the one-month test-market study, you performed a multiple regression analysis on the data. Two independent variables were considered: the price of an OmniPower bar and the monthly budget for in-store promotional expenditures. The dependent variable was the number of OmniPower bars sold in a month. The coefficient of determination indicated that 75.8% of the variation in sales was explained by knowing the price charged and the amount spent on in-store promotions. The model indicated that the predicted sales of OmniPower are estimated to decrease by 532 bars per month for each 10-cent increase in the price, and the predicted sales are estimated to increase by 361 bars for each additional \$100 spent on promotions.

After studying the relative effects of price and promotion, OmniFoods needs to set price and promotion standards for a nationwide introduction (obviously, lower prices and higher promotion budgets lead to more sales, but they do so at a lower profit margin). You determined that if stores spend \$400 a month for in-store promotions and charge 79 cents, the 95% confidence interval estimate of the mean monthly sales is 2,854 to 3,303 bars. OmniFoods can multiply the lower and upper bounds of this confidence interval by the number of stores included in the nationwide introduction to estimate total monthly sales. For example, if 1,000 stores are in the nationwide introduction, then total monthly sales should be between 2.854 million and 3.308 million bars.

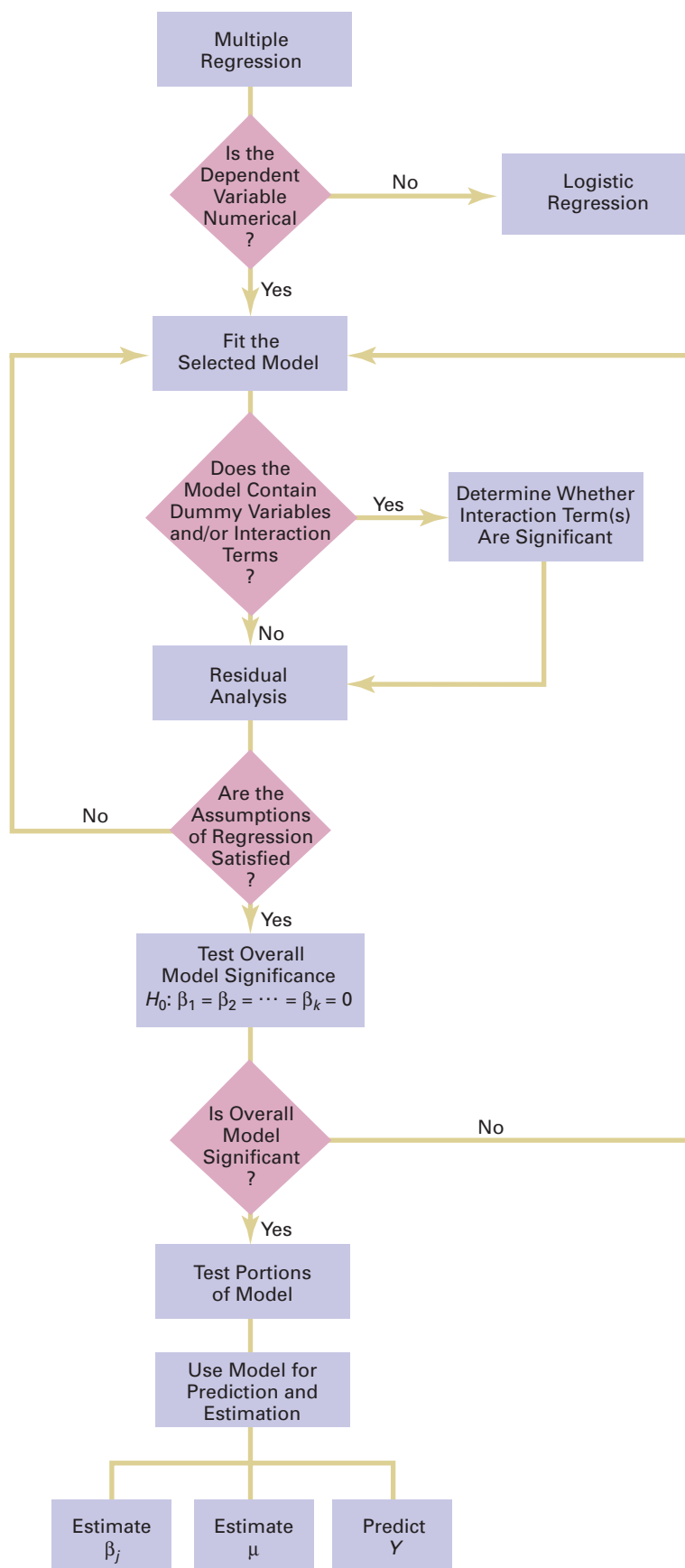
SUMMARY

In this chapter, you learned how multiple regression models allow you to use two or more independent variables to predict the value of a dependent variable. You also learned how to include categorical independent variables and interaction

terms in regression models. In addition, you used the logistic regression model to predict a categorical dependent variable. Figure 14.15 presents a roadmap of the chapter.

FIGURE 14.15

Roadmap for multiple regression



KEY EQUATIONS

Multiple Regression Model with k Independent Variables

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + \varepsilon_i \quad (14.1)$$

Multiple Regression Model with Two Independent Variables

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \quad (14.2)$$

Multiple Regression Equation with Two Independent Variables

$$\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i} \quad (14.3)$$

Coefficient of Multiple Determination

$$r^2 = \frac{\text{Regression sum of squares}}{\text{Total sum of squares}} = \frac{SSR}{SST} \quad (14.4)$$

Adjusted r^2

$$r_{\text{adj}}^2 = 1 - \left[(1 - r^2) \frac{n - 1}{n - k - 1} \right] \quad (14.5)$$

Overall F Test

$$F_{STAT} = \frac{MSR}{MSE} \quad (14.6)$$

Testing for the Slope in Multiple Regression

$$t_{STAT} = \frac{b_j - \beta_j}{S_{b_j}} \quad (14.7)$$

Confidence Interval Estimate for the Slope

$$b_j \pm t_{\alpha/2} S_{b_j} \quad (14.8)$$

Determining the Contribution of an Independent Variable to the Regression Model

$$SSR(X_j | \text{All } X_s \text{ except } j) = SSR(\text{All } X_s) - SSR(\text{All } X_s \text{ except } j) \quad (14.9)$$

Contribution of Variable X_1 , Given That X_2 Has Been Included

$$SSR(X_1 | X_2) = SSR(X_1 \text{ and } X_2) - SSR(X_2) \quad (14.10a)$$

Contribution of Variable X_2 , Given That X_1 Has Been Included

$$SSR(X_2 | X_1) = SSR(X_1 \text{ and } X_2) - SSR(X_1) \quad (14.10b)$$

Partial F Test Statistic

$$F_{STAT} = \frac{SSR(X_j | \text{All } X_s \text{ except } j)}{MSE} \quad (14.11)$$

Relationship Between a t Statistic and an F Statistic

$$t_{STAT}^2 = F_{STAT} \quad (14.12)$$

Coefficients of Partial Determination for a Multiple Regression Model Containing Two Independent Variables

$$r_{Y1.2}^2 = \frac{SSR(X_1 | X_2)}{SST - SSR(X_1 \text{ and } X_2) + SSR(X_1 | X_2)} \quad (14.13a)$$

and

$$r_{Y2.1}^2 = \frac{SSR(X_2 | X_1)}{SST - SSR(X_1 \text{ and } X_2) + SSR(X_2 | X_1)} \quad (14.13b)$$

Coefficient of Partial Determination for a Multiple Regression Model Containing k Independent Variables

$$r_{Y_j(\text{All variables except } j)}^2 = \frac{SSR(X_j | \text{All } X_s \text{ except } j)}{SST - SSR(\text{All } X_s) + SSR(X_j | \text{All } X_s \text{ except } j)} \quad (14.14)$$

Odds Ratio

$$\text{Odds ratio} = \frac{\text{Probability of an event of interest}}{1 - \text{Probability of an event of interest}} \quad (14.15)$$

Logistic Regression Model

$$\ln(\text{Odds ratio}) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i \quad (14.16)$$

Logistic Regression Equation

$$\ln(\text{Estimated odds ratio}) = b_0 + b_1 X_{1i} + b_2 X_{2i} + \cdots + b_k X_{ki} \quad (14.17)$$

Estimated Odds Ratio

$$\text{Estimated odds ratio} = e^{\ln(\text{Estimated odds ratio})} \quad (14.18)$$

Estimated Probability of an Event of Interest

$$\text{Estimated probability of an event of interest} = \frac{\text{Estimated odds ratio}}{1 + \text{Estimated odds ratio}} \quad (14.19)$$

KEY TERMS

adjusted r^2 585
 coefficient of multiple
 determination 584
 coefficient of partial
 determination 597
 cross-product term 602

deviance statistic 612
 dummy variable 599
 interaction 602
 interaction term 602
 logistic regression 609
 multiple regression model 578

net regression coefficient 581
 odds ratio 609
 overall F test 585
 partial F test 593
 Wald statistic 612

CHAPTER REVIEW PROBLEMS

CHECKING YOUR UNDERSTANDING

14.58 What is the difference between r^2 and adjusted r^2 ?

14.59 How does the interpretation of the regression coefficients differ in multiple regression and simple linear regression?

14.60 How does testing the significance of the entire multiple regression model differ from testing the contribution of each independent variable?

14.61 How do the coefficients of partial determination differ from the coefficient of multiple determination?

14.62 Why and how do you use dummy variables?

14.63 How can you evaluate whether the slope of the dependent variable with an independent variable is the same for each level of the dummy variable?

14.64 Under what circumstances do you include an interaction term in a regression model?

14.65 When a dummy variable is included in a regression model that has one numerical independent variable, what assumption do you need to make concerning the slope between the dependent variable, Y , and the numerical independent variable, X ?

14.66 When do you use logistic regression?

highest value generating the most sales. The following regression equation is presented:

$$\hat{Y}_i = -3.888 + 1.449 X_{1i} + 1.462 X_{2i} - 0.190 X_{1i} X_{2i}$$

Suppose that X_1 is the perceived quality of the product and X_2 is the perceived value of the product. (Note: If the customer thinks the product is overpriced, he or she perceives it to be of low value and vice versa.)

- What is the predicted purchase behavior when $X_1 = 2$ and $X_2 = 2$?
- What is the predicted purchase behavior when $X_1 = 2$ and $X_2 = 7$?
- What is the predicted purchase behavior when $X_1 = 7$ and $X_2 = 2$?
- What is the predicted purchase behavior when $X_1 = 7$ and $X_2 = 7$?
- What is the regression equation when $X_2 = 2$? What is the slope for X_1 now?
- What is the regression equation when $X_2 = 7$? What is the slope for X_1 now?
- What is the regression equation when $X_1 = 2$? What is the slope for X_2 now?
- What is the regression equation when $X_1 = 7$? What is the slope for X_2 now?
- Discuss the implications of (a) through (h) within the context of increasing sales for this product with two customer satisfaction measures.

APPLYING THE CONCEPTS

14.67 Increasing customer satisfaction typically results in increased purchase behavior. For many products, there is more than one measure of customer satisfaction. In many of these instances, purchase behavior can increase dramatically with an increase in any one of the customer satisfaction measures, not necessarily all of them at the same time. Gunst and Barry ("One Way to Moderate Ceiling Effects," *Quality Progress*, October 2003, pp. 83–85) consider a product with two satisfaction measures, X_1 and X_2 , that range from the lowest level of satisfaction, 1, to the highest level of satisfaction, 7. The dependent variable, Y , is a measure of purchase behavior, with the

14.68 The owner of a moving company typically has his most experienced manager predict the total number of labor hours that will be required to complete an upcoming move. This approach has proved useful in the past, but the owner has the business objective of developing a more accurate method of predicting labor hours. In a preliminary effort to provide a more accurate method, the owner has decided to use the number of cubic feet moved and the number of pieces of large furniture as the independent variables and has collected data for 36 moves in which the origin and destination were within the borough of Manhattan in New York City and the travel time was an insignificant portion of the hours worked. The data are organized and stored in

Moving

- State the multiple regression equation.
- Interpret the meaning of the slopes in this equation.
- Predict the labor hours for moving 500 cubic feet with two large pieces of furniture.
- Perform a residual analysis on your results and determine whether the regression assumptions are valid.
- Determine whether there is a significant relationship between labor hours and the two independent variables (the number of cubic feet moved and the number of pieces of large furniture) at the 0.05 level of significance.
- Determine the p -value in (e) and interpret its meaning.
- Interpret the meaning of the coefficient of multiple determination in this problem.
- Determine the adjusted r^2 .
- At the 0.05 level of significance, determine whether each independent variable makes a significant contribution to the regression model. Indicate the most appropriate regression model for this set of data.
- Determine the p -values in (i) and interpret their meaning.
- Construct a 95% confidence interval estimate of the population slope between labor hours and the number of cubic feet moved. How does the interpretation of the slope here differ from that in Problem 13.44 on page 552?
- Compute and interpret the coefficients of partial determination.

14.69 Professional basketball has truly become a sport that generates interest among fans around the world. More and more players come from outside the United States to play in the National Basketball Association (NBA). You want to develop a regression model to predict the number of wins achieved by each NBA team, based on field goal (shots made) percentage for the team and for the opponent. The data are stored in **NBA2010**.

- State the multiple regression equation.
- Interpret the meaning of the slopes in this equation.
- Predict the number of wins for a team that has a field goal percentage of 45% and an opponent field goal percentage of 44%.
- Perform a residual analysis on your results and determine whether the regression assumptions are valid.
- Is there a significant relationship between number of wins and the two independent variables (field goal percentage for the team and for the opponent) at the 0.05 level of significance?
- Determine the p -value in (e) and interpret its meaning.
- Interpret the meaning of the coefficient of multiple determination in this problem.
- Determine the adjusted r^2 .
- At the 0.05 level of significance, determine whether each independent variable makes a significant contribution to the regression model. Indicate the most appropriate regression model for this set of data.
- Determine the p -values in (i) and interpret their meaning.
- Compute and interpret the coefficients of partial determination.

14.70 A sample of 30 recently sold single-family houses in a small city is selected. Develop a model to predict the selling price (in thousands of dollars), using the assessed value (in thousands of dollars) as well as time (in months since reassessment). The houses in the city had been reassessed at full value one year prior to the study. The results are stored in **House1**.

- State the multiple regression equation.
- Interpret the meaning of the slopes in this equation.
- Predict the selling price for a house that has an assessed value of \$170,000 and was sold 12 months after reassessment.
- Perform a residual analysis on your results and determine whether the regression assumptions are valid.
- Determine whether there is a significant relationship between selling price and the two independent variables (assessed value and time period) at the 0.05 level of significance.
- Determine the p -value in (e) and interpret its meaning.
- Interpret the meaning of the coefficient of multiple determination in this problem.
- Determine the adjusted r^2 .
- At the 0.05 level of significance, determine whether each independent variable makes a significant contribution to the regression model. Indicate the most appropriate regression model for this set of data.
- Determine the p -values in (i) and interpret their meaning.
- Construct a 95% confidence interval estimate of the population slope between selling price and assessed value. How does the interpretation of the slope here differ from that in Problem 13.76 on page 565?
- Compute and interpret the coefficients of partial determination.

14.71 Measuring the height of a California redwood tree is very difficult because these trees grow to heights over 300 feet. People familiar with these trees understand that the height of a California redwood tree is related to other characteristics of the tree, including the diameter of the tree at the breast height of a person (in inches) and the thickness of the bark of the tree (in inches). The file **Redwood** contains the height, diameter at breast height of a person, and bark thickness for a sample of 21 California redwood trees.

- State the multiple regression equation that predicts the height of a tree, based on the tree's diameter at breast height and the thickness of the bark.
- Interpret the meaning of the slopes in this equation.
- Predict the height for a tree that has a breast height diameter of 25 inches and a bark thickness of 2 inches.
- Interpret the meaning of the coefficient of multiple determination in this problem.
- Perform a residual analysis on the results and determine whether the regression assumptions are valid.
- Determine whether there is a significant relationship between the height of redwood trees and the two independent variables (breast-height diameter and bark thickness) at the 0.05 level of significance.

- g. Construct a 95% confidence interval estimate of the population slope between the height of redwood trees and breast-height diameter and between the height of redwood trees and the bark thickness.
- h. At the 0.05 level of significance, determine whether each independent variable makes a significant contribution to the regression model. Indicate the independent variables to include in this model.
- i. Construct a 95% confidence interval estimate of the mean height for trees that have a breast-height diameter of 25 inches and a bark thickness of 2 inches, along with a prediction interval for an individual tree.
- j. Compute and interpret the coefficients of partial determination.

14.72 Develop a model to predict the assessed value (in thousands of dollars), using the size of the houses (in thousands of square feet) and the age of the houses (in years) from the following table (stored in **House2**):

House	Assessed Value (\$Thousands)	Size of House (Thousands of Square Feet)	Age (Years)
1	184.4	2.00	3.42
2	177.4	1.71	11.50
3	175.7	1.45	8.33
4	185.9	1.76	0.00
5	179.1	1.93	7.42
6	170.4	1.20	32.00
7	175.8	1.55	16.00
8	185.9	1.93	2.00
9	178.5	1.59	1.75
10	179.2	1.50	2.75
11	186.7	1.90	0.00
12	179.3	1.39	0.00
13	174.5	1.54	12.58
14	183.8	1.89	2.75
15	176.8	1.59	7.17

- a. State the multiple regression equation.
- b. Interpret the meaning of the slopes in this equation.
- c. Predict the assessed value for a house that has a size of 1,750 square feet and is 10 years old.
- d. Perform a residual analysis on the results and determine whether the regression assumptions are valid.
- e. Determine whether there is a significant relationship between assessed value and the two independent variables (size and age) at the 0.05 level of significance.
- f. Determine the p -value in (e) and interpret its meaning.
- g. Interpret the meaning of the coefficient of multiple determination in this problem.
- h. Determine the adjusted r^2 .
- i. At the 0.05 level of significance, determine whether each independent variable makes a significant contribution to the regression model. Indicate the most appropriate regression model for this set of data.

- j. Determine the p -values in (i) and interpret their meaning.
- k. Construct a 95% confidence interval estimate of the population slope between assessed value and size. How does the interpretation of the slope here differ from that of Problem 13.77 on page 566?
- l. Compute and interpret the coefficients of partial determination.
- m. The real estate assessor's office has been publicly quoted as saying that the age of a house has no bearing on its assessed value. Based on your answers to (a) through (l), do you agree with this statement? Explain.

14.73 Crazy Dave, a well-known baseball analyst, wants to determine which variables are important in predicting a team's wins in a given season. He has collected data related to wins, earned run average (ERA), and runs scored for the 2009 season (stored in **BB2009**). Develop a model to predict the number of wins based on ERA and runs scored.

- a. State the multiple regression equation.
- b. Interpret the meaning of the slopes in this equation.
- c. Predict the number of wins for a team that has an ERA of 4.50 and has scored 750 runs.
- d. Perform a residual analysis on the results and determine whether the regression assumptions are valid.
- e. Is there a significant relationship between number of wins and the two independent variables (ERA and runs scored) at the 0.05 level of significance?
- f. Determine the p -value in (e) and interpret its meaning.
- g. Interpret the meaning of the coefficient of multiple determination in this problem.
- h. Determine the adjusted r^2 .
- i. At the 0.05 level of significance, determine whether each independent variable makes a significant contribution to the regression model. Indicate the most appropriate regression model for this set of data.
- j. Determine the p -values in (i) and interpret their meaning.
- k. Construct a 95% confidence interval estimate of the population slope between wins and ERA.
- l. Compute and interpret the coefficients of partial determination.
- m. Which is more important in predicting wins—pitching, as measured by ERA, or offense, as measured by runs scored? Explain.

14.74 Referring to Problem 14.73, suppose that in addition to using ERA to predict the number of wins, Crazy Dave wants to include the league (0 = American, 1 = National) as an independent variable. Develop a model to predict wins based on ERA and league. For (a) through (k), do not include an interaction term.

- a. State the multiple regression equation.
- b. Interpret the slopes in (a).
- c. Predict the number of wins for a team with an ERA of 4.50 in the American League. Construct a 95% confidence interval estimate for all teams and a 95% prediction interval for an individual team.
- d. Perform a residual analysis on the results and determine whether the regression assumptions are valid.

- e. Is there a significant relationship between wins and the two independent variables (ERA and league) at the 0.05 level of significance?
- f. At the 0.05 level of significance, determine whether each independent variable makes a contribution to the regression model. Indicate the most appropriate regression model for this set of data.
- g. Construct a 95% confidence interval estimate of the population slope for the relationship between wins and ERA.
- h. Construct a 95% confidence interval estimate of the population slope for the relationship between wins and league.
- i. Compute and interpret the adjusted r^2 .
- j. Compute and interpret the coefficients of partial determination.
- k. What assumption do you have to make about the slope of wins with ERA?
- l. Add an interaction term to the model and, at the 0.05 level of significance, determine whether it makes a significant contribution to the model.
- m. On the basis of the results of (f) and (l), which model is most appropriate? Explain.

14.75 You are a real estate broker who wants to compare property values in Glen Cove and Roslyn (which are located approximately 8 miles apart). In order to do so, you will

analyze the data in **GCRoslyn**, a file that includes samples of houses from Glen Cove and Roslyn. Making sure to include the dummy variable for location (Glen Cove or Roslyn), develop a regression model to predict appraised value, based on the land area of a property, the age of a house, and location. Be sure to determine whether any interaction terms need to be included in the model.

14.76 A recent article discussed a metal deposition process in which a piece of metal is placed in an acid bath and an alloy is layered on top of it. The business objective of engineers working on the process was to reduce variation in the thickness of the alloy layer. To begin, the temperature and the pressure in the tank holding the acid bath are to be studied as independent variables. Data are collected from 50 samples. The results are organized and stored in **Thickness**. (Data extracted from J. Conklin, “It’s a Marathon, Not a Sprint,” *Quality Progress*, June 2009, pp. 46–49.)

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

MANAGING ASHLAND MULTICOMM SERVICES

In its continuing study of the *3-For-All* subscription solicitation process, a marketing department team wants to test the effects of two types of structured sales presentations (personal formal and personal informal) and the number of hours spent on telemarketing on the number of new subscriptions. The staff has recorded these data in the file **AMS14** for the past 24 weeks.

Analyze these data and develop a multiple regression model to predict the number of new subscriptions for a week, based on the number of hours spent on telemarketing and the sales presentation type. Write a report, giving detailed findings concerning the regression model used.

DIGITAL CASE

Apply your knowledge of multiple regression models in this Digital Case, which extends the OmniFoods Using Statistics scenario from this chapter.

To ensure a successful test marketing of its OmniPower energy bars, the OmniFoods marketing department has contracted with In-Store Placements Group (ISPG), a merchandising consultancy. ISPG will work with the grocery store chain that is conducting the test-market study. Using the same 34-store sample used in the test-market study, ISPG claims that the choice of shelf location and the presence of in-store OmniPower coupon dispensers both increase sales of the energy bars.

Open **Omni_ISPGMemo.pdf** to review the ISPG claims and supporting data. Then answer the following questions:

1. Are the supporting data consistent with ISPG’s claims? Perform an appropriate statistical analysis to confirm (or discredit) the stated relationship between sales and the two independent variables of product shelf location and the presence of in-store OmniPower coupon dispensers.
2. If you were advising OmniFoods, would you recommend using a specific shelf location and in-store coupon dispensers to sell OmniPower bars?
3. What additional data would you advise collecting in order to determine the effectiveness of the sales promotion techniques used by ISPG?

REFERENCES

1. Hosmer, D. W., and S. Lemeshow, *Applied Logistic Regression*, 2nd ed. (New York: Wiley, 2001).
2. Kutner, M., C. Nachtsheim, J. Neter, and W. Li, *Applied Linear Statistical Models*, 5th ed. (New York: McGraw-Hill/Irwin, 2005).
3. *Microsoft Excel 2010* (Redmond, WA: Microsoft Corp., 2010).
4. *Minitab Release 16* (State College, PA: Minitab, Inc., 2010).

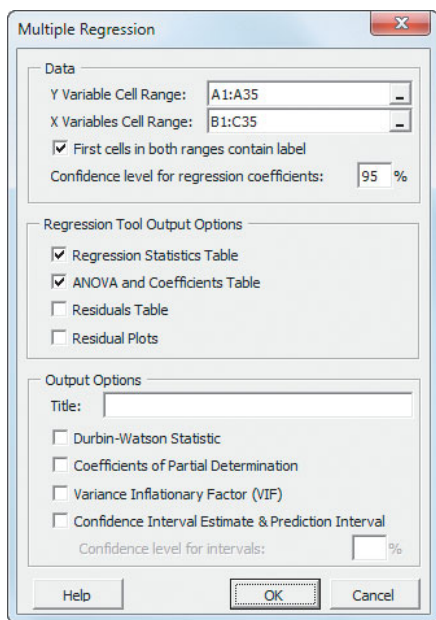
CHAPTER 14 EXCEL GUIDE

EG14.1 DEVELOPING a MULTIPLE REGRESSION MODEL

Interpreting the Regression Coefficients

PHStat2 Use **Multiple Regression** to perform a multiple regression analysis. For example, to perform the Figure 14.2 analysis of the OmniPower sales data on page 580, open to the **DATA** worksheet of the **OmniPower** workbook. Select **PHStat** → **Regression** → **Multiple Regression**, and in the procedure's dialog box (shown below):

1. Enter **A1:A35** as the **Y Variable Cell Range**.
2. Enter **B1:C35** as the **X Variables Cell Range**.
3. Check **First cells in both ranges contain label**.
4. Enter **95** as the **Confidence level for regression coefficients**.
5. Check **Regression Statistics Table** and **ANOVA and Coefficients Table**.
6. Enter a **Title** and click **OK**.



The procedure creates a worksheet that contains a copy of your data in addition to the regression results worksheet shown in Figure 14.2. For more information about these worksheets, read the following *In-Depth Excel* section.

In-Depth Excel Use the **COMPUTE** worksheet of the **Multiple Regression** workbook, partially shown in Figure 14.2 on page 580, as a template for performing multiple

regression. Columns A through I of this worksheet duplicate the visual design of the Analysis ToolPak regression worksheet. The worksheet uses the regression data in the **MRData** worksheet to perform the regression analysis for the OmniPower sales data.

Figure 14.2 does not show the columns K through N Calculations area. This area contains a **LINEST**(*cell range of Y variable, cell range of X variable, True, True*) array formula in the cell range L2:N6 and calculations for the *t* test of the slope (see Section 13.7 on page 548). The array formula computes the b_2 , b_1 , and b_0 coefficients in cells L2, M2, and N2; the b_2 , b_1 , and b_0 standard error in cells L3, M3, and N3; r^2 and the standard error of the estimate in cells L4 and M4; the *F* test statistic and error *df* in cells L5 and M5; and *SSR* and *SSE* in cells L6 and M6. (The rest of the cell range, N4, N5, and N6, displays the #N/A message. This is not an error.)

Open to the **COMPUTE_FORMULAS** worksheet to examine all the formulas in the worksheet, some of which are discussed in the Chapter 13 Excel Guide *In-Depth Excel* sections.

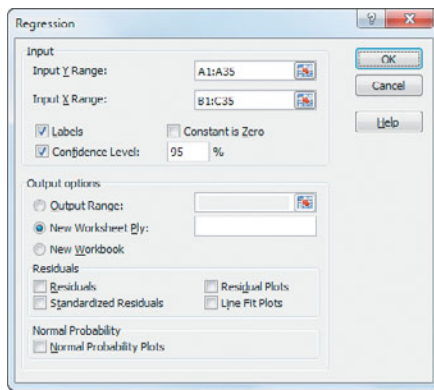
To perform multiple regression analyses for other data, paste the regression data into the MRData worksheet. Paste the values for the *Y* variable into column A. Paste the values for the *X* variables into consecutive columns, starting with column B. Open to the **COMPUTE** worksheet. First, enter the confidence level in cell L8. Then, edit the correct 5-row-by-3-column array of cells that starts with cell L2. First adjust the range of the array, adding a column for each independent variable in excess of two. Then, edit the cell ranges in the array formula, and then, while holding down the **Control** and **Shift** keys (or the **Apple** key on a Mac), press the **Enter** key.

Analysis ToolPak Use **Regression** to perform a multiple regression analysis. For example, to perform the Figure 14.2 analysis of the OmniPower sales data on page 580, open to the **DATA** worksheet of the **OmniPower** workbook and:

1. Select **Data** → **Data Analysis**.
2. In the Data Analysis dialog box, select **Regression** from the **Analysis Tools** list and then click **OK**.

In the Regression dialog box (shown on page 623):

3. Enter **A1:A35** as the **Input Y Range** and enter **B1:C35** as the **Input X Range**.
4. Check **Labels** and check **Confidence Level** and enter **95** in its box.
5. Click **New Worksheet Ply**.
6. Click **OK**.



Predicting the Dependent Variable Y

PHStat2 Use the “Interpreting the Regression Coefficients” *PHStat2* instructions but replace step 6 with the following steps 6 through 8:

6. Check **Confidence Interval Estimate & Prediction Interval** and enter **95** as the percentage for **Confidence level for intervals**.
7. Enter a **Title** and click **OK**.
8. In the new worksheet, enter **79** in cell **B6** and enter **400** in cell **B7**.

These steps create a new worksheet that is discussed in the following *In-Depth Excel* instructions.

In-Depth Excel Use the **CIEandPI worksheet** of the **Multiple Regression workbook**, shown in Figure 14.3 on page 582, as a template for computing confidence interval estimates and prediction intervals for a multiple regression model with two independent variables. The worksheet contains the data and formulas for the OmniPower sales example shown in Figure 14.3. The worksheet uses several array formulas to use functions that perform matrix operations to compute the matrix product $X'X$ (in cell range B9:D11), the inverse of the $X'X$ matrix (in cell range B13:D15), the product of $X'G$ multiplied by the inverse of $X'X$ (in cell range B17:D17), and the predicted Y (in cell B21). (Open to the **CIEandPI_FORMULAS worksheet** to examine all formulas.)

Modifying this worksheet for other models with more than two independent variables requires knowledge that is beyond the scope of this book. For other models with two independent variables, paste the data for those variables into columns B and C of the **MRArray worksheet** and adjust the number of entries in column A (all of which are 1). Then open to the **COMPUTE worksheet** and edit the array formula in cell range B9:D11 and edit the labels in cells A6 and A7.

EG14.2 r^2 , ADJUSTED r^2 , and the OVERALL F TEST

The coefficient of multiple determination, r^2 , the adjusted r^2 , and the overall F test are all computed as part of creating

the multiple regression results worksheet using the Section EG14.1 instructions. If you use either the *PHStat2* or *In-Depth Excel* instructions, formulas are used to compute these results in the **COMPUTE worksheet**. Formulas in cells B5, B7, B13, C12, C13, D12, and E12 copy values computed by an array formula in cell range L2:N6 and in cell F12, the expression **FDIST(F test statistic, 1, error degrees of freedom)** computes the p -value for the overall F test.

EG14.3 RESIDUAL ANALYSIS for the MULTIPLE REGRESSION MODEL

PHStat2 Use the Section EG14.1 “Interpreting the Regression Coefficients” *PHStat2* instructions. Modify step 5 by checking **Residuals Table** and **Residual Plots** in addition to checking **Regression Statistics Table** and **ANOVA and Coefficients Table**.

In-Depth Excel Create a worksheet that calculates residuals and then create a scatter plot of the original X variable and the residuals (plotted as the Y variable).

Use the **RESIDUALS worksheet** of the **Multiple Regression workbook** as a template for creating a residuals worksheet. The formulas in this worksheet compute the residuals for the multiple regression model for the OmniPower sales example by using the regression data in the **MRData worksheet** in the same workbook. In column D, the worksheet computes the predicted Y values by multiplying the X_1 values by the b_1 coefficient and the X_2 values by the b_2 coefficient and adding these products to the b_0 coefficient. In column F, the worksheet computes residuals by subtracting the predicted Y values from the Y values. (Open to the **RESIDUALS_FORMULAS worksheet** to examine all formulas.) For other problems, modify this worksheet as follows:

1. If the number of independent variables is greater than 2, select column D, right-click, and click **Insert** from the shortcut menu. Repeat this step as many times as necessary to create the additional columns to hold all the X variables.
2. Paste the data for the X variables into columns, starting with column B.
3. Paste Y values in column E (or in the second-to-last column if there are more than two X variables).
4. For sample sizes smaller than 34, delete the extra rows. For sample sizes greater than 34, copy the predicted Y and residuals formulas down through the row containing the last pair of X and Y values. Also, add the new observation numbers in column A.

To create residual plots, use copy-and-paste special values (see Appendix Section F.6) to paste data values on a new worksheet in the proper order before applying the Section EG2.6 scatter plot instructions.

Analysis ToolPak Use the Section EG14.1 *Analysis ToolPak* instructions. Modify step 5 by checking **Residuals** and **Residual Plots** before clicking **New Worksheet Ply** and then **OK**. (Note that the **Residuals Plots** option creates residual plots only for each independent variable.)

EG14.4 INFERENCE CONCERNING the POPULATION REGRESSION COEFFICIENTS

The regression results worksheets created by using the EG14.1 instructions include the information needed to make the inferences discussed in Section 14.4.

EG14.5 TESTING PORTIONS of the MULTIPLE REGRESSION MODEL

PHStat2 Use the Section EG14.1 “Interpreting the Regression Coefficients” *PHStat2* instructions but modify step 6 by checking **Coefficients of Partial Determination** before you click **OK**.

In-Depth Excel You compute the coefficients of partial determination by using a two-step process. You first use the Section EG14.1 *In-Depth Excel* instructions to create all possible regression results worksheets in a copy of the **Multiple Regression workbook**. For example, if you have two independent variables, you perform three regression analyses: Y with X_1 and X_2 , Y with X_1 , and Y with X_2 , to create three regression results worksheets. Then open to the **CPD worksheet** for the number of independent variables (**CPD_2**, **CPD_3**, and **CPD_4 worksheets** are included) and follow the italicized instructions to copy and paste special values from the regression results worksheets. The **CPD_2 worksheet** contains the data to compute the coefficients of partial determination for the OmniPower regression model used as an example in Section 14.5.

EG14.6 USING DUMMY VARIABLES and INTERACTION TERMS in REGRESSION MODELS

Dummy Variables

Use **Find and Replace** to create a dummy variable from a two-level categorical variable. Before using **Find and Replace**, copy and paste the categorical values to another column in order to preserve the original values.

For example, to create a dummy variable named **FireplaceCoded** from the two-level categorical variable **Fireplace** as shown in Table 14.5 on page 600, open to the **DATA worksheet** of the **House3 workbook** and:

1. Copy and paste the **Fireplace** values in column **C** to column **D** (the first empty column).
2. Select column **D**.
3. Press **Ctrl+H** (the keyboard shortcut for **Find and Replace**).

In the Find and Replace dialog box:

4. Enter **Yes** in the **Find what** box and enter **1** in the **Replace with** box.
5. Click **Replace All**. If a message box to confirm the replacement appears, click **OK** to continue.
6. Enter **No** in the **Find what** box and enter **0** in the **Replace with** box.
7. Click **Replace All**. If a message box to confirm the replacement appears, click **OK** to continue.
8. Click **Close**.

Categorical variables that have more than two levels require the use of formulas in multiple columns. For example, to create the dummy variables for Example 14.3 on page 601, two columns are needed. Assume that the three-level categorical variable mentioned in the example is in Column **D** of the opened worksheet. A first new column that contains formulas in the form **=IF(column D cell = first level, 1, 0)** and a second new column that contains formulas in the form **=IF(column D cell = second level, 1, 0)** would properly create the two dummy variables that the example requires.

Interactions

To create an interaction term, add a column of formulas that multiply one independent variable by another. For example, if the first independent variable appeared in column **B** and the second independent variable appeared in column **C**, enter the formula **= B2 * C2** in the row 2 cell of an empty new column and then copy the formula down through all rows of data to create the interaction.

EG14.7 LOGISTIC REGRESSION

There are no Excel Guide instructions for this section.

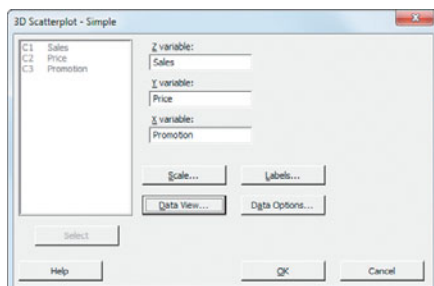
CHAPTER 14 MINITAB GUIDE

MG14.1 DEVELOPING a MULTIPLE REGRESSION MODEL

Visualizing Multiple Regression Data

Use **3D Scatterplot** to create a three-dimensional plot for the special case of a regression model that contains two independent variables. For example, to create the Figure 14.1 plot on page 579 for the OmniPower sales data, open the **OmniPower worksheet**. Select **Graph** → **3D Scatterplot**. In the 3D Scatterplots dialog box, click **Simple** and then click **OK**. In the 3D Scatterplot - Simple dialog box (shown below):

1. Double-click **C1 Sales** in the variables list to add **Sales** to the **Z variable** box.
2. Double-click **C2 Price** in the variables list to add **Price** to the **Y variable** box.
3. Double-click **C3 Promotion** in the variables list to add **Promotion** to the **X variable** box.
4. Click **Data View**.



In the 3D Scatterplot - Data View dialog box:

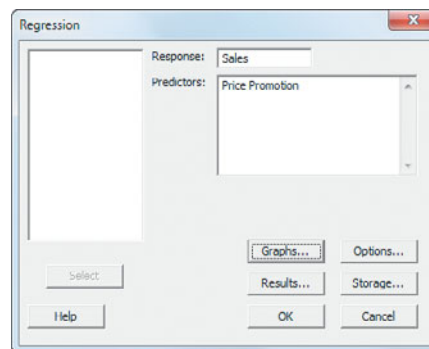
5. Check **Symbols** and **Project lines**.
6. Click **OK**.
7. Back in the 3D Scatterplot - Simple dialog box, click **OK**.

Interpreting the Regression coefficients

Use **Regression** to perform a multiple regression analysis. For example, to perform the Figure 14.2 analysis of the OmniPower sales data on page 580, open to the **OmniPower worksheet**. Select **Stat** → **Regression** → **Regression**. In the Regression dialog box (shown at the top of the next column):

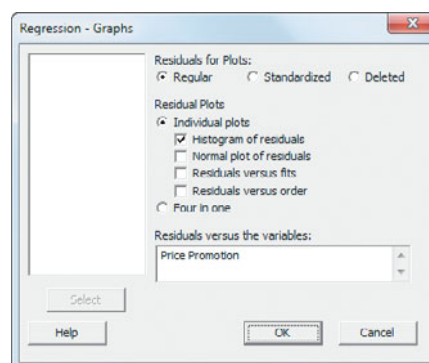
1. Double-click **C1 Sales** in the variables list to add **Sales** to the **Response** box.
2. Double-click **C2 Price** in the variables list to add **Price** to the **Predictors** box.

3. Double-click **C3 Promotion** in the variables list to add **Promotion** to the **Predictors** box.
4. Click **Graphs**.



In the Regression - Graphs dialog box (shown below):

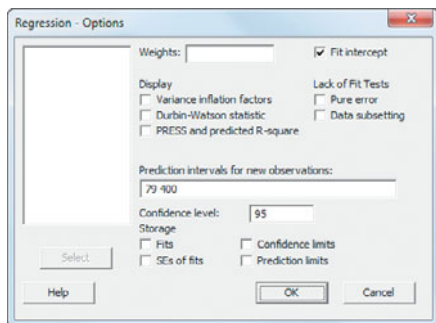
5. Click **Regular** and **Individual Plots**.
6. Check **Histogram of residuals** and clear all the other check boxes.
7. Click anywhere inside the **Residuals versus the variables** box.
8. Double-click **C2 Price** in the variables list to add **Price** in the **Residuals versus the variables** box.
9. Double-click **C3 Promotion** in the variables list to add **Promotion** in the **Residuals versus the variables** box.
10. Click **OK**.



11. Back in the Regression dialog box, click **Results**. In the Regression - Results dialog box (not shown):
12. Click **In addition, the full table of fits and residuals** and then click **OK**.
13. Back in the Regression dialog box, click **Options**.

In the Regression - Options dialog box (shown below):

14. Check **Fit Intercept**.
15. Clear all the **Display** and **Lack of Fit Test** check boxes.
16. Enter **79** and **400** in the **Prediction intervals for new observations** box.
17. Enter **95** in the **Confidence level** box.
18. Click **OK**.



19. Back in the Regression dialog box, click **OK**.

The results in the Session Window will include a table of residuals that is not shown in Figure 14.2.

MG14.2 r^2 , ADJUSTED r^2 , and the OVERALL F TEST

The coefficient of multiple determination, r^2 , the adjusted r^2 , and the overall F test are all computed as part of creating the multiple regression results using the Section MG14.1 instructions.

MG14.3 RESIDUAL ANALYSIS for the MULTIPLE REGRESSION MODEL

Residual analysis results are created using the Section MG14.1 instructions.

MG14.4 INFERENCES CONCERNING the POPULATION REGRESSION COEFFICIENTS

The regression results created by using the MG14.1 instructions include the information needed to make the inferences discussed in Section 14.4.

MG14.5 TESTING PORTIONS of the MULTIPLE REGRESSION MODEL

You compute the coefficients of partial determination by using a two-step process. You first use the Section MG14.1 instructions to create all possible regression results in the

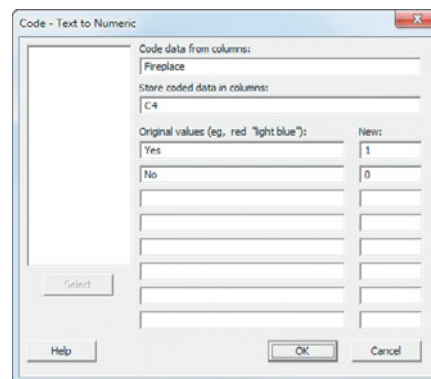
same project file. For example, if you have two independent variables, you perform three regression analyses— Y with X_1 and X_2 , Y with X_1 , and Y with X_2 —to create three sets of regression results. With those results you can then compute the partial F test and the coefficients of partial determination using the instructions in Section 14.5.

MG14.6 USING DUMMY VARIABLES and INTERACTION TERMS in REGRESSION MODELS

Dummy Variables

Use **Text to Numeric** to create a dummy variable. For example, to create a dummy variable named **FireplaceCoded** from the categorical variable **Fireplace** (see Table 14.5 on page 600), open to the **House3 worksheet** and select **Data → Code → Text to Numeric**. In the Code - Text to Numeric dialog box (shown below):

1. Double-click **C3 Fireplace** in the variables list to add **Fireplace** to the **Code data from columns** box and press **Tab**.
2. Enter **C4** in the **Store coded data in columns** box and press **Tab**. (Column C4 is the first empty column in the worksheet.)
3. In the first row, enter **Yes** in the **Original Values** (eg, red “light blue”) box and enter **1** in the **New** box.
4. In the second row, enter **No** in the **Original Values** (eg, red “light blue”) box and enter **0** in the **New** box.
5. Click **OK**.



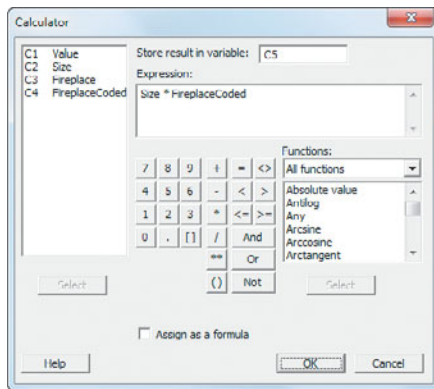
6. Enter **FireplaceCoded** as the name of column **C4**.

Interactions

Use **Calculator** to add a new column that contains the product of multiplying one independent variable by another to create an interaction term. For example, to create an interaction term of size and the dummy variable **FireplaceCoded** (see Table 14.5 on page 600), open to the **House3 worksheet**. Use the “Dummy Variables” instructions in the preceding

part to create the **FireplaceCoded** column in the worksheet. Select **Calc** → **Calculator**. In the Calculator dialog box (shown below):

1. Enter **C5** in the **Store result in variable** box and press **Tab**.
2. Enter **Size * FireplaceCoded** in the **Expression** box.
3. Click **OK**.



4. Enter **Size*FireplaceCoded** as the name for column **C5**.

MG14.7 LOGISTIC REGRESSION

Use **Binary Logistic Regression** to perform a logistic regression. For example, to perform the Figure 14.14 analysis of the credit card marketing data on page 611, open to

Logpurch worksheet. Select **Stat** → **Regression** → **Binary Logistic Regression**. In the Binary Logistic Regression dialog box (shown below):

1. Click **Response in response/frequency format** and press **Tab**.
2. Double-click **C1 Upgraded** in the variables list to add **Upgraded** in the **Response** box.
3. Click inside the **Model** box.
4. Double-click **C2 Purchases** in the variables list to add **Purchases** to the **Model** box.
5. Double-click **C3 Extra Cards** in the variables list to add 'Extra Cards' to the **Model** box and press **Tab**.
6. Double-click **C3 Extra Cards** in the variables list to add 'Extra Cards' to the **Factors** box (because **Extra Cards** is a categorical variable).
7. Click **OK**.

