

CHAPTER 15

Multiple Regression

CONTENTS

STATISTICS IN PRACTICE:
dunnhumby

- 15.1** MULTIPLE REGRESSION MODEL
Regression Model and Regression Equation
Estimated Multiple Regression Equation
- 15.2** LEAST SQUARES METHOD
An Example: Butler Trucking Company
Note on Interpretation of Coefficients
- 15.3** MULTIPLE COEFFICIENT OF DETERMINATION
- 15.4** MODEL ASSUMPTIONS
- 15.5** TESTING FOR SIGNIFICANCE
F Test
t Test
Multicollinearity
- 15.6** USING THE ESTIMATED REGRESSION EQUATION FOR ESTIMATION AND PREDICTION
- 15.7** CATEGORICAL INDEPENDENT VARIABLES
An Example: Johnson Filtration, Inc.
Interpreting the Parameters
More Complex Categorical Variables
- 15.8** RESIDUAL ANALYSIS
Detecting Outliers
Studentized Deleted Residuals and Outliers
Influential Observations
Using Cook's Distance Measure to Identify Influential Observations
- 15.9** LOGISTIC REGRESSION
Logistic Regression Equation
Estimating the Logistic Regression Equation
Testing for Significance
Managerial Use
Interpreting the Logistic Regression Equation
Logit Transformation

STATISTICS *in* PRACTICE

dunnhumby*

LONDON, ENGLAND

Founded in 1989 by the husband-and-wife team of Clive Humby (a mathematician) and Edwina Dunn (a marketer), dunnhumby combines proven natural abilities with big ideas to find clues and patterns as to what customers are buying and why. The company turns these insights into actionable strategies that create dramatic growth and sustainable loyalty, ultimately improving brand value and the customer experience.

Employing more than 950 people in Europe, Asia, and the Americas, dunnhumby serves a prestigious list of companies, including Kroger, Tesco, Coca-Cola, General Mills, Kimberly-Clark, PepsiCo, Procter & Gamble, and Home Depot. dunnhumbyUSA is a joint venture between the Kroger Company and dunnhumby and has offices in New York, Chicago, Atlanta, Minneapolis, Cincinnati, and Portland.

The company's research begins with data collected about a client's customers. Data come from customer reward or discount card purchase records, electronic point-of-sale transactions, and traditional market research. Analysis of the data often translates billions of data points into detailed insights about the behavior, preferences, and lifestyles of the customers. Such insights allow for more effective merchandising programs to be activated, including strategy recommendations on pricing, promotion, advertising, and product assortment decisions.

Researchers have used a multiple regression technique referred to as logistic regression to help in their analysis of customer-based data. Using logistic regression, an estimated multiple regression equation of the following form is developed.

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_px_p$$

The dependent variable \hat{y} is a prediction of the probability that a customer belongs to a particular customer group. The independent variables $x_1, x_2, x_3, \dots, x_p$ are measures of the customer's actual shopping behavior and may include the specific items purchased, number of items purchased, amount purchased, day of the week, hour of the day, and so on. The analysis helps identify the independent variables that are most relevant in predict-

*The authors are indebted to Paul Hunter, Senior Vice President of Solutions for dunnhumby for providing this Statistics in Practice.



dunnhumby uses logistic regression to predict customer shopping behavior. © Micro 10x/Shutterstock.com.

ing the customer's group and provides a better understanding of the customer population, enabling further analysis with far greater confidence. The focus of the analysis is on understanding the customer to the point of developing merchandising, marketing, and direct marketing programs that will maximize the relevancy and service to the customer group.

In this chapter, we will introduce multiple regression and show how the concepts of simple linear regression introduced in Chapter 14 can be extended to the multiple regression case. In addition, we will show how computer software packages are used for multiple regression. In the final section of the chapter we introduce logistic regression using an example that illustrates how the technique is used in a marketing research application.

In Chapter 14 we presented simple linear regression and demonstrated its use in developing an estimated regression equation that describes the relationship between two variables. Recall that the variable being predicted or explained is called the dependent variable and the variable being used to predict or explain the dependent variable is called the independent variable. In this chapter we continue our study of regression analysis by considering situations involving two or more independent variables. This subject area, called **multiple regression analysis**, enables us to consider more factors and thus obtain better predictions than are possible with simple linear regression.

15.1 Multiple Regression Model

Multiple regression analysis is the study of how a dependent variable y is related to two or more independent variables. In the general case, we will use p to denote the number of independent variables.

Regression Model and Regression Equation

The concepts of a regression model and a regression equation introduced in the preceding chapter are applicable in the multiple regression case. The equation that describes how the dependent variable y is related to the independent variables x_1, x_2, \dots, x_p and an error term is called the **multiple regression model**. We begin with the assumption that the multiple regression model takes the following form.

MULTIPLE REGRESSION MODEL

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_px_p + \epsilon \quad (15.1)$$

In the multiple regression model, $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ are the parameters and the error term ϵ (the Greek letter epsilon) is a random variable. A close examination of this model reveals that y is a linear function of x_1, x_2, \dots, x_p (the $\beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_px_p$ part) plus the error term ϵ . The error term accounts for the variability in y that cannot be explained by the linear effect of the p independent variables.

In Section 15.4 we will discuss the assumptions for the multiple regression model and ϵ . One of the assumptions is that the mean or expected value of ϵ is zero. A consequence of this assumption is that the mean or expected value of y , denoted $E(y)$, is equal to $\beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_px_p$. The equation that describes how the mean value of y is related to x_1, x_2, \dots, x_p is called the **multiple regression equation**.

MULTIPLE REGRESSION EQUATION

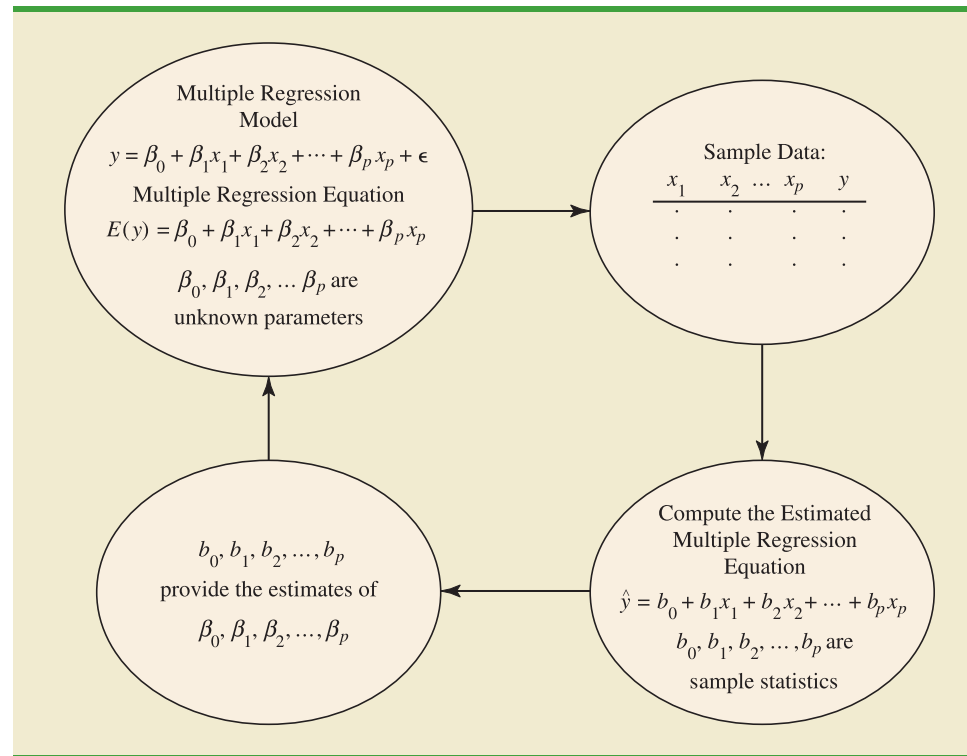
$$E(y) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_px_p \quad (15.2)$$

Estimated Multiple Regression Equation

If the values of $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ were known, equation (15.2) could be used to compute the mean value of y at given values of x_1, x_2, \dots, x_p . Unfortunately, these parameter values will not, in general, be known and must be estimated from sample data. A simple random sample is used to compute sample statistics $b_0, b_1, b_2, \dots, b_p$ that are used as the point

FIGURE 15.1 THE ESTIMATION PROCESS FOR MULTIPLE REGRESSION

In simple linear regression, b_0 and b_1 were the sample statistics used to estimate the parameters β_0 and β_1 . Multiple regression parallels this statistical inference process, with $b_0, b_1, b_2, \dots, b_p$ denoting the sample statistics used to estimate the parameters $\beta_0, \beta_1, \beta_2, \dots, \beta_p$.



estimators of the parameters $\beta_0, \beta_1, \beta_2, \dots, \beta_p$. These sample statistics provide the following **estimated multiple regression equation**.

ESTIMATED MULTIPLE REGRESSION EQUATION

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_p x_p \quad (15.3)$$

where

$$b_0, b_1, b_2, \dots, b_p \text{ are the estimates of } \beta_0, \beta_1, \beta_2, \dots, \beta_p$$

$$\hat{y} = \text{predicted value of the dependent variable}$$

The estimation process for multiple regression is shown in Figure 15.1.

15.2 Least Squares Method

In Chapter 14, we used the **least squares method** to develop the estimated regression equation that best approximated the straight-line relationship between the dependent and independent variables. This same approach is used to develop the estimated multiple regression equation. The least squares criterion is restated as follows.

LEAST SQUARES CRITERION

$$\min \sum (y_i - \hat{y}_i)^2 \quad (15.4)$$

where

y_i = observed value of the dependent variable for the i th observation

\hat{y}_i = predicted value of the dependent variable for the i th observation

The predicted values of the dependent variable are computed by using the estimated multiple regression equation,

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

As expression (15.4) shows, the least squares method uses sample data to provide the values of $b_0, b_1, b_2, \dots, b_p$ that make the sum of squared residuals [the deviations between the observed values of the dependent variable (y_i) and the predicted values of the dependent variable (\hat{y}_i)] a minimum.

In Chapter 14 we presented formulas for computing the least squares estimators b_0 and b_1 for the estimated simple linear regression equation $\hat{y} = b_0 + b_1x$. With relatively small data sets, we were able to use those formulas to compute b_0 and b_1 by manual calculations. In multiple regression, however, the presentation of the formulas for the regression coefficients $b_0, b_1, b_2, \dots, b_p$ involves the use of matrix algebra and is beyond the scope of this text. Therefore, in presenting multiple regression, we focus on how computer software packages can be used to obtain the estimated regression equation and other information. The emphasis will be on how to interpret the computer output rather than on how to make the multiple regression computations.

An Example: Butler Trucking Company

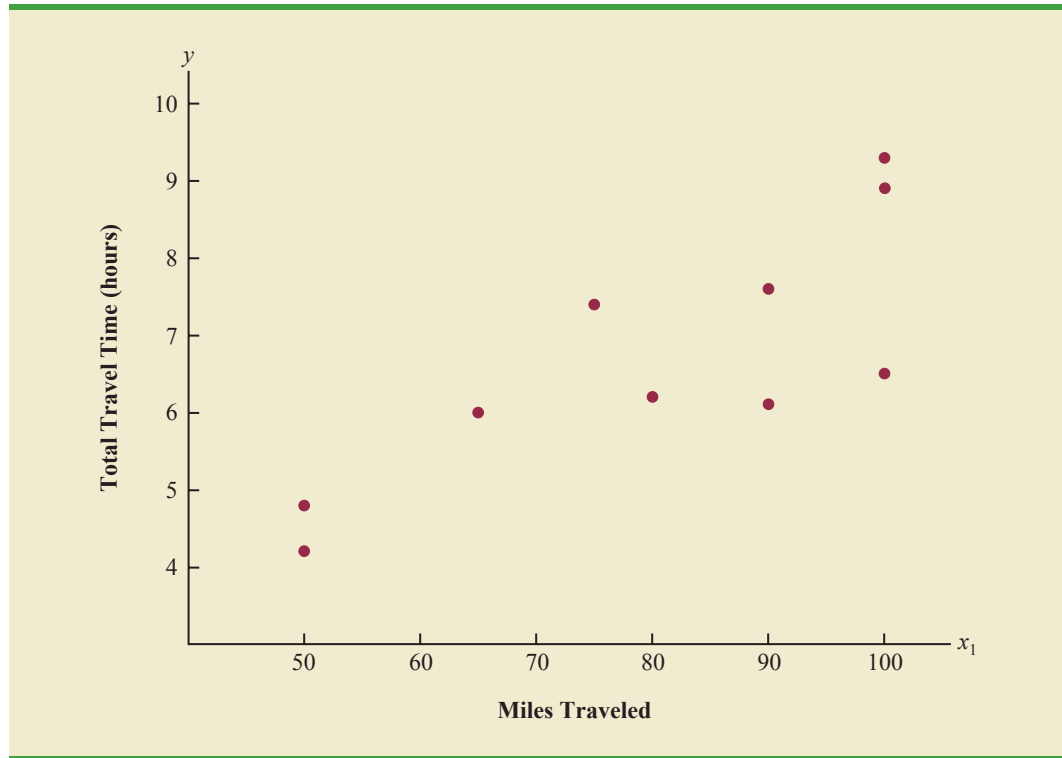
As an illustration of multiple regression analysis, we will consider a problem faced by the Butler Trucking Company, an independent trucking company in southern California. A major portion of Butler's business involves deliveries throughout its local area. To develop better work schedules, the managers want to predict the total daily travel time for their drivers.

Initially the managers believed that the total daily travel time would be closely related to the number of miles traveled in making the daily deliveries. A simple random sample of 10 driving assignments provided the data shown in Table 15.1 and the scatter diagram shown in Figure 15.2. After reviewing this scatter diagram, the managers hypothesized that the simple linear regression model $y = \beta_0 + \beta_1x_1 + \epsilon$ could be used to describe the relationship between the total travel time (y) and the number of miles traveled (x_1). To estimate

TABLE 15.1 PRELIMINARY DATA FOR BUTLER TRUCKING

Driving Assignment	x_1 = Miles Traveled	y = Travel Time (hours)
1	100	9.3
2	50	4.8
3	100	8.9
4	100	6.5
5	50	4.2
6	80	6.2
7	75	7.4
8	65	6.0
9	90	7.6
10	90	6.1



FIGURE 15.2 SCATTER DIAGRAM OF PRELIMINARY DATA FOR BUTLER TRUCKING

the parameters β_0 and β_1 , the least squares method was used to develop the estimated regression equation.

$$\hat{y} = b_0 + b_1x_1 \quad (15.5)$$

In Figure 15.3, we show the Minitab computer output from applying simple linear regression to the data in Table 15.1. The estimated regression equation is

$$\hat{y} = 1.27 + .0678x_1$$

At the .05 level of significance, the F value of 15.81 and its corresponding p -value of .004 indicate that the relationship is significant; that is, we can reject $H_0: \beta_1 = 0$ because the p -value is less than $\alpha = .05$. Note that the same conclusion is obtained from the t value of 3.98 and its associated p -value of .004. Thus, we can conclude that the relationship between the total travel time and the number of miles traveled is significant; longer travel times are associated with more miles traveled. With a coefficient of determination (expressed as a percentage) of $R\text{-Sq} = 66.4\%$, we see that 66.4% of the variability in travel time can be explained by the linear effect of the number of miles traveled. This finding is fairly good, but the managers might want to consider adding a second independent variable to explain some of the remaining variability in the dependent variable.

In attempting to identify another independent variable, the managers felt that the number of deliveries could also contribute to the total travel time. The Butler Trucking data, with the number of deliveries added, are shown in Table 15.2. The Minitab computer solution with both miles traveled (x_1) and number of deliveries (x_2) as independent variables is shown in Figure 15.4. The estimated regression equation is

$$\hat{y} = -.869 + .0611x_1 + .923x_2 \quad (15.6)$$

The Minitab steps necessary to generate the output shown in Figure 15.4 are given in Appendix 15.1.

FIGURE 15.3 MINITAB OUTPUT FOR BUTLER TRUCKING WITH ONE INDEPENDENT VARIABLE

The regression equation is
Time = 1.27 + 0.0678 Miles

Predictor	Coef	SE Coef	T	p
Constant	1.274	1.401	0.91	0.390
Miles	0.06783	0.01706	3.98	0.004

S = 1.00179 R-Sq = 66.4% R-Sq(adj) = 62.2%

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	1	15.871	15.871	15.81	0.004
Residual Error	8	8.029	1.004		
Total	9	23.900			

In the Minitab output the variable names Miles and Time were entered as the column headings on the worksheet; thus, x_1 = Miles and y = Time.

In the next section we will discuss the use of the coefficient of multiple determination in measuring how good a fit is provided by this estimated regression equation. Before doing so, let us examine more carefully the values of $b_1 = .0611$ and $b_2 = .923$ in equation (15.6).

Note on Interpretation of Coefficients

One observation can be made at this point about the relationship between the estimated regression equation with only the miles traveled as an independent variable and the equation that includes the number of deliveries as a second independent variable. The value of b_1 is not the same in both cases. In simple linear regression, we interpret b_1 as an estimate of the change in y for a one-unit change in the independent variable. In multiple regression analysis, this interpretation must be modified somewhat. That is, in multiple regression analysis, we interpret each regression coefficient as follows: b_i represents an estimate of the change in y corresponding to a one-unit change in x_i when all other independent variables are held constant. In the Butler Trucking example involving two independent variables, $b_1 = .0611$. Thus,

TABLE 15.2 DATA FOR BUTLER TRUCKING WITH MILES TRAVELED (x_1) AND NUMBER OF DELIVERIES (x_2) AS THE INDEPENDENT VARIABLES

Driving Assignment	x_1 = Miles Traveled	x_2 = Number of Deliveries	y = Travel Time (hours)
1	100	4	9.3
2	50	3	4.8
3	100	4	8.9
4	100	2	6.5
5	50	2	4.2
6	80	2	6.2
7	75	3	7.4
8	65	4	6.0
9	90	3	7.6
10	90	2	6.1



FIGURE 15.4 MINITAB OUTPUT FOR BUTLER TRUCKING WITH TWO INDEPENDENT VARIABLES

In the Minitab output the variable names Miles, Deliveries, and Time were entered as the column headings on the worksheet; thus, x_1 = Miles, x_2 = Deliveries, and y = Time.

The regression equation is
 Time = - 0.869 + 0.0611 Miles + 0.923 Deliveries

Predictor	Coef	SE Coef	T	p
Constant	-0.8687	0.9515	-0.91	0.392
Miles	0.061135	0.009888	6.18	0.000
Deliveries	0.9234	0.2211	4.18	0.004

S = 0.573142 R-Sq = 90.4% R-Sq(adj) = 87.6%

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	2	21.601	10.800	32.88	0.000
Residual Error	7	2.299	0.328		
Total	9	23.900			

.0611 hours is an estimate of the expected increase in travel time corresponding to an increase of one mile in the distance traveled when the number of deliveries is held constant. Similarly, because $b_2 = .923$, an estimate of the expected increase in travel time corresponding to an increase of one delivery when the number of miles traveled is held constant is .923 hours.

Exercises

Note to student: The exercises involving data in this and subsequent sections were designed to be solved using a computer software package.

Methods

- The estimated regression equation for a model involving two independent variables and 10 observations follows.

$$\hat{y} = 29.1270 + .5906x_1 + .4980x_2$$

- Interpret b_1 and b_2 in this estimated regression equation.
 - Predict y when $x_1 = 180$ and $x_2 = 310$.
- Consider the following data for a dependent variable y and two independent variables, x_1 and x_2 .

SELF test

WEB file

Exer2

x_1	x_2	y
30	12	94
47	10	108
25	17	112
51	16	178
40	5	94
51	19	175
74	7	170
36	12	117
59	13	142
76	16	211

- Develop an estimated regression equation relating y to x_1 . Predict y if $x_1 = 45$.
 - Develop an estimated regression equation relating y to x_2 . Predict y if $x_2 = 15$.
 - Develop an estimated regression equation relating y to x_1 and x_2 . Predict y if $x_1 = 45$ and $x_2 = 15$.
3. In a regression analysis involving 30 observations, the following estimated regression equation was obtained.

$$\hat{y} = 17.6 + 3.8x_1 - 2.3x_2 + 7.6x_3 + 2.7x_4$$

- Interpret b_1 , b_2 , b_3 , and b_4 in this estimated regression equation.
- Predict y when $x_1 = 10$, $x_2 = 5$, $x_3 = 1$, and $x_4 = 2$.

Applications

4. A shoe store developed the following estimated regression equation relating sales to inventory investment and advertising expenditures.

$$\hat{y} = 25 + 10x_1 + 8x_2$$

where

x_1 = inventory investment (\$1000s)

x_2 = advertising expenditures (\$1000s)

y = sales (\$1000s)

- Predict the sales resulting from a \$15,000 investment in inventory and an advertising budget of \$10,000.
 - Interpret b_1 and b_2 in this estimated regression equation.
5. The owner of Showtime Movie Theaters, Inc., would like to predict weekly gross revenue as a function of advertising expenditures. Historical data for a sample of eight weeks follow.

SELF test

WEB file

Showtime

Weekly Gross Revenue (\$1000s)	Television Advertising (\$1000s)	Newspaper Advertising (\$1000s)
96	5.0	1.5
90	2.0	2.0
95	4.0	1.5
92	2.5	2.5
95	3.0	3.3
94	3.5	2.3
94	2.5	4.2
94	3.0	2.5

- Develop an estimated regression equation with the amount of television advertising as the independent variable.
 - Develop an estimated regression equation with both television advertising and newspaper advertising as the independent variables.
 - Is the estimated regression equation coefficient for television advertising expenditures the same in part (a) and in part (b)? Interpret the coefficient in each case.
 - Predict weekly gross revenue for a week when \$3500 is spent on television advertising and \$1800 is spent on newspaper advertising?
6. The National Football League (NFL) records a variety of performance data for individuals and teams. To investigate the importance of passing on the percentage of games won by a team, the following data show the conference (Conf), average number of passing yards per

attempt (Yds/Att), the number of interceptions thrown per attempt (Int/Att), and the percentage of games won (Win%) for a random sample of 16 NFL teams for the 2011 season (NFL website, February 12, 2012).



Team	Conf	Yds/Att	Int/Att	Win%
Arizona Cardinals	NFC	6.5	.042	50.0
Atlanta Falcons	NFC	7.1	.022	62.5
Carolina Panthers	NFC	7.4	.033	37.5
Cincinnati Bengals	AFC	6.2	.026	56.3
Detroit Lions	NFC	7.2	.024	62.5
Green Bay Packers	NFC	8.9	.014	93.8
Houston Texans	AFC	7.5	.019	62.5
Indianapolis Colts	AFC	5.6	.026	12.5
Jacksonville Jaguars	AFC	4.6	.032	31.3
Minnesota Vikings	NFC	5.8	.033	18.8
New England Patriots	AFC	8.3	.020	81.3
New Orleans Saints	NFC	8.1	.021	81.3
Oakland Raiders	AFC	7.6	.044	50.0
San Francisco 49ers	NFC	6.5	.011	81.3
Tennessee Titans	AFC	6.7	.024	56.3
Washington Redskins	NFC	6.4	.041	31.3

- Develop the estimated regression equation that could be used to predict the percentage of games won given the average number of passing yards per attempt.
 - Develop the estimated regression equation that could be used to predict the percentage of games won given the number of interceptions thrown per attempt.
 - Develop the estimated regression equation that could be used to predict the percentage of games won given the average number of passing yards per attempt and the number of interceptions thrown per attempt.
 - The average number of passing yards per attempt for the Kansas City Chiefs was 6.2 and the number of interceptions thrown per attempt was .036. Use the estimated regression equation developed in part (c) to predict the percentage of games won by the Kansas City Chiefs. (*Note:* For the 2011 season the Kansas City Chiefs' record was 7 wins and 9 losses.) Compare your prediction to the actual percentage of games won by the Kansas City Chiefs.
7. *PC World* rated four component characteristics for 10 ultraportable laptop computers: features; performance; design; and price. Each characteristic was rated using a 0–100 point scale. An overall rating, referred to as the *PCW World Rating*, was then developed for each laptop. The following table shows the performance rating, features rating, and the *PCW World Rating* for the 10 laptop computers (*PC World* website, February 5, 2009).



Model	Performance	Features	PCW Rating
Thinkpad X200	77	87	83
VGN-Z598U	97	85	82
U6V	83	80	81
Elitebook 2530P	77	75	78
X360	64	80	78
Thinkpad X300	56	76	78
Ideapad U110	55	81	77
Micro Express JFT2500	76	73	75
Toughbook W7	46	79	73
HP Voodoo Envy133	54	68	72

- a. Determine the estimated regression equation that can be used to predict the *PCW World Rating* using the performance rating as the independent variable.
 - b. Determine the estimated regression equation that can be used to predict the *PCW World Rating* using both the performance rating and the features rating.
 - c. Predict the *PCW World Rating* for a laptop computer that has a performance rating of 80 and a features rating of 70.
8. The *Condé Nast Traveler* Gold List for 2012 provided ratings for the top 20 small cruise ships (*Condé Nast Traveler* website, March 1, 2012). The data shown below are the scores each ship received based upon the results from *Condé Nast Traveler's* annual Readers' Choice Survey. Each score represents the percentage of respondents who rated a ship as excellent or very good on several criteria, including Shore Excursions and Food/Dining. An overall score was also reported and used to rank the ships. The highest ranked ship, the *Seabourn Odyssey*, has an overall score of 94.4, the highest component of which is 97.8 for Food/Dining.



Ship	Overall	Shore Excursions	Food/Dining
<i>Seabourn Odyssey</i>	94.4	90.9	97.8
<i>Seabourn Pride</i>	93.0	84.2	96.7
<i>National Geographic Endeavor</i>	92.9	100.0	88.5
<i>Seabourn Sojourn</i>	91.3	94.8	97.1
<i>Paul Gauguin</i>	90.5	87.9	91.2
<i>Seabourn Legend</i>	90.3	82.1	98.8
<i>Seabourn Spirit</i>	90.2	86.3	92.0
<i>Silver Explorer</i>	89.9	92.6	88.9
<i>Silver Spirit</i>	89.4	85.9	90.8
<i>Seven Seas Navigator</i>	89.2	83.3	90.5
<i>Silver Whisperer</i>	89.2	82.0	88.6
<i>National Geographic Explorer</i>	89.1	93.1	89.7
<i>Silver Cloud</i>	88.7	78.3	91.3
<i>Celebrity Xpedition</i>	87.2	91.7	73.6
<i>Silver Shadow</i>	87.2	75.0	89.7
<i>Silver Wind</i>	86.6	78.1	91.6
<i>SeaDream II</i>	86.2	77.4	90.9
<i>Wind Star</i>	86.1	76.5	91.5
<i>Wind Surf</i>	86.1	72.3	89.3
<i>Wind Spirit</i>	85.2	77.4	91.9

- a. Determine an estimated regression equation that can be used to predict the overall score given the score for Shore Excursions.
 - b. Consider the addition of the independent variable Food/Dining. Develop the estimated regression equation that can be used to predict the overall score given the scores for Shore Excursions and Food/Dining.
 - c. Predict the overall score for a cruise ship with a Shore Excursions score of 80 and a Food/Dining Score of 90.
9. Waterskiing and wakeboarding are two popular water-sports. Finding a model that best suits your intended needs, whether it is waterskiing, wakeboarding, or general boating, can be a difficult task. *WaterSki* magazine did extensive testing for 88 boats and provided a wide variety of information to help consumers select the best boat. A portion of the data they reported for 20 boats with a length of between 20 and 22 feet follows (*WaterSki*, January/February 2006). Beam is the maximum width of the boat in inches, HP is the horsepower of the boat's engine, and TopSpeed is the top speed in miles per hour (mph).

WEB file
Boats

Make and Model	Beam	HP	TopSpeed
Calabria Cal Air Pro V-2	100.0	330	45.3
Correct Craft Air Nautique 210	91.0	330	47.3
Correct Craft Air Nautique SV-211	93.0	375	46.9
Correct Craft Ski Nautique 206 Limited	91.0	330	46.7
Gekko GTR 22	96.0	375	50.1
Gekko GTS 20	83.0	375	52.2
Malibu Response LXi	93.5	340	47.2
Malibu Sunsetter LXi	98.0	400	46.0
Malibu Sunsetter 21 XTi	98.0	340	44.0
Malibu Sunscape 21 LSV	98.0	400	47.5
Malibu Wakesetter 21 XTi	98.0	340	44.9
Malibu Wakesetter VLX	98.0	400	47.3
Malibu vRide	93.5	340	44.5
Malibu Ride XTi	93.5	320	44.5
Mastercraft ProStar 209	96.0	350	42.5
Mastercraft X-1	90.0	310	45.8
Mastercraft X-2	94.0	310	42.8
Mastercraft X-9	96.0	350	43.2
MB Sports 190 Plus	92.0	330	45.3
Svfara SVONE	91.0	330	47.7

- Using these data, develop an estimated regression equation relating the top speed with the boat's beam and horsepower rating.
 - The Sv fara SV609 has a beam of 85 inches and an engine with a 330 horsepower rating. Use the estimated regression equation developed in part (a) to predict the top speed for the Sv fara SV609.
10. Major League Baseball (MLB) consists of teams that play in the American League and the National League. MLB collects a wide variety of team and player statistics. Some of the statistics often used to evaluate pitching performance are as follows:

ERA: The average number of earned runs given up by the pitcher per nine innings. An earned run is any run that the opponent scores off a particular pitcher except for runs scored as a result of errors.

SO/IP: The average number of strikeouts per inning pitched.

HR/IP: The average number of home runs per inning pitched.

R/IP: The number of runs given up per inning pitched.

The following data show values for these statistics for a random sample of 20 pitchers from the American League for the 2011 season (MLB website, March 1, 2012).

WEB file
MLBPitching

Player	Team	W	L	ERA	SO/IP	HR/IP	R/IP
Verlander, J	DET	24	5	2.40	1.00	.10	.29
Beckett, J	BOS	13	7	2.89	.91	.11	.34
Wilson, C	TEX	16	7	2.94	.92	.07	.40
Sabathia, C	NYN	19	8	3.00	.97	.07	.37
Haren, D	LAA	16	10	3.17	.81	.08	.38
McCarthy, B	OAK	9	9	3.32	.72	.06	.43
Santana, E	LAA	11	12	3.38	.78	.11	.42

(continued)

Player	Team	W	L	ERA	SO/IP	HR/IP	R/IP
Lester, J	BOS	15	9	3.47	.95	.10	.40
Hernandez, F	SEA	14	14	3.47	.95	.08	.42
Buehrle, M	CWS	13	9	3.59	.53	.10	.45
Pineda, M	SEA	9	10	3.74	1.01	.11	.44
Colon, B	NYY	8	10	4.00	.82	.13	.52
Tomlin, J	CLE	12	7	4.25	.54	.15	.48
Pavano, C	MIN	9	13	4.30	.46	.10	.55
Danks, J	CWS	8	12	4.33	.79	.11	.52
Guthrie, J	BAL	9	17	4.33	.63	.13	.54
Lewis, C	TEX	14	10	4.40	.84	.17	.51
Scherzer, M	DET	15	9	4.43	.89	.15	.52
Davis, W	TB	11	10	4.45	.57	.13	.52
Porcello, R	DET	14	9	4.75	.57	.10	.57

- Develop an estimated regression equation that can be used to predict the average number of runs given up per inning given the average number of strikeouts per inning pitched.
- Develop an estimated regression equation that can be used to predict the average number of runs given up per inning given the average number of home runs per inning pitched.
- Develop an estimated regression equation that can be used to predict the average number of runs given up per inning given the average number of strikeouts per inning pitched and the average number of home runs per inning pitched.
- A. J. Burnett, a pitcher for the New York Yankees, had an average number of strikeouts per inning pitched of .91 and an average number of home runs per inning of .16. Use the estimated regression equation developed in part (c) to predict the average number of runs given up per inning for A. J. Burnett. (*Note:* The actual value for R/IP was .6.)
- Suppose a suggestion was made to also use the earned run average as another independent variable in part (c). What do you think of this suggestion?

15.3

Multiple Coefficient of Determination

In simple linear regression we showed that the total sum of squares can be partitioned into two components: the sum of squares due to regression and the sum of squares due to error. The same procedure applies to the sum of squares in multiple regression.

RELATIONSHIP AMONG SST, SSR, AND SSE

$$SST = SSR + SSE \quad (15.7)$$

where

$$SST = \text{total sum of squares} = \sum (y_i - \bar{y})^2$$

$$SSR = \text{sum of squares due to regression} = \sum (\hat{y}_i - \bar{y})^2$$

$$SSE = \text{sum of squares due to error} = \sum (y_i - \hat{y}_i)^2$$

Because of the computational difficulty in computing the three sums of squares, we rely on computer packages to determine those values. The analysis of variance part of the Minitab output in Figure 15.4 shows the three values for the Butler Trucking problem with two independent variables: $SST = 23.900$, $SSR = 21.601$, and $SSE = 2.299$. With only one independent variable (number of miles traveled), the Minitab output in Figure 15.3 shows that $SST = 23.900$, $SSR = 15.871$, and $SSE = 8.029$. The value of SST is the same in both cases because it does not depend on \hat{y} , but SSR increases and SSE decreases when a second independent variable (number of deliveries) is added. The implication is that the estimated multiple regression equation provides a better fit for the observed data.

In Chapter 14, we used the coefficient of determination, $r^2 = SSR/SST$, to measure the goodness of fit for the estimated regression equation. The same concept applies to multiple regression. The term **multiple coefficient of determination** indicates that we are measuring the goodness of fit for the estimated multiple regression equation. The multiple coefficient of determination, denoted R^2 , is computed as follows.

MULTIPLE COEFFICIENT OF DETERMINATION

$$R^2 = \frac{SSR}{SST} \quad (15.8)$$

The multiple coefficient of determination can be interpreted as the proportion of the variability in the dependent variable that can be explained by the estimated multiple regression equation. Hence, when multiplied by 100, it can be interpreted as the percentage of the variability in y that can be explained by the estimated regression equation.

In the two-independent-variable Butler Trucking example, with $SSR = 21.601$ and $SST = 23.900$, we have

$$R^2 = \frac{21.601}{23.900} = .904$$

Therefore, 90.4% of the variability in travel time y is explained by the estimated multiple regression equation with miles traveled and number of deliveries as the independent variables. In Figure 15.4, we see that the multiple coefficient of determination (expressed as a percentage) is also provided by the Minitab output; it is denoted by $R\text{-Sq} = 90.4\%$.

Figure 15.3 shows that the $R\text{-Sq}$ value for the estimated regression equation with only one independent variable, number of miles traveled (x_1), is 66.4%. Thus, the percentage of the variability in travel times that is explained by the estimated regression equation increases from 66.4% to 90.4% when number of deliveries is added as a second independent variable. In general, R^2 always increases as independent variables are added to the model.

Many analysts prefer adjusting R^2 for the number of independent variables to avoid overestimating the impact of adding an independent variable on the amount of variability explained by the estimated regression equation. With n denoting the number of observations and p denoting the number of independent variables, the **adjusted multiple coefficient of determination** is computed as follows.

Adding independent variables causes the prediction errors to become smaller, thus reducing the sum of squares due to error, SSE. Because $SSR = SST - SSE$, when SSE becomes smaller, SSR becomes larger, causing $R^2 = SSR/SST$ to increase.

If a variable is added to the model, R^2 becomes larger even if the variable added is not statistically significant. The adjusted multiple coefficient of determination compensates for the number of independent variables in the model.

ADJUSTED MULTIPLE COEFFICIENT OF DETERMINATION

$$R_a^2 = 1 - (1 - R^2) \frac{n - 1}{n - p - 1} \quad (15.9)$$

For the Butler Trucking example with $n = 10$ and $p = 2$, we have

$$R_a^2 = 1 - (1 - .904) \frac{10 - 1}{10 - 2 - 1} = .88$$

Thus, after adjusting for the two independent variables, we have an adjusted multiple coefficient of determination of .88. This value (expressed as a percentage) is provided by the Minitab output in Figure 15.4 as R-Sq(adj) = 87.6%; the value we calculated differs because we used a rounded value of R^2 in the calculation.

NOTES AND COMMENTS

If the value of R^2 is small and the model contains a large number of independent variables, the adjusted coefficient of determination can take a negative

value; in such cases, Minitab sets the adjusted coefficient of determination to zero.

Exercises

Methods

11. In exercise 1, the following estimated regression equation based on 10 observations was presented.

$$\hat{y} = 29.1270 + .5906x_1 + .4980x_2$$

The values of SST and SSR are 6724.125 and 6216.375, respectively.

- Find SSE.
 - Compute R^2 .
 - Compute R_a^2 .
 - Comment on the goodness of fit.
12. In exercise 2, 10 observations were provided for a dependent variable y and two independent variables x_1 and x_2 ; for these data SST = 15,182.9, and SSR = 14,052.2.
- Compute R^2 .
 - Compute R_a^2 .
 - Does the estimated regression equation explain a large amount of the variability in the data? Explain.
13. In exercise 3, the following estimated regression equation based on 30 observations was presented.

$$\hat{y} = 17.6 + 3.8x_1 - 2.3x_2 + 7.6x_3 + 2.7x_4$$

The values of SST and SSR are 1805 and 1760, respectively.

SELF test

- Compute R^2 .
- Compute R_a^2 .
- Comment on the goodness of fit.

Applications

14. In exercise 4, the following estimated regression equation relating sales to inventory investment and advertising expenditures was given.

$$\hat{y} = 25 + 10x_1 + 8x_2$$

The data used to develop the model came from a survey of 10 stores; for those data, $SST = 16,000$ and $SSR = 12,000$.

- For the estimated regression equation given, compute R^2 .
- Compute R_a^2 .
- Does the model appear to explain a large amount of variability in the data? Explain.

SELF test

15. In exercise 5, the owner of Showtime Movie Theaters, Inc., used multiple regression analysis to predict gross revenue (y) as a function of television advertising (x_1) and newspaper advertising (x_2). The estimated regression equation was

$$\hat{y} = 83.2 + 2.29x_1 + 1.30x_2$$

The computer solution provided $SST = 25.5$ and $SSR = 23.435$.

- Compute and interpret R^2 and R_a^2 .
- When television advertising was the only independent variable, $R^2 = .653$ and $R_a^2 = .595$. Do you prefer the multiple regression results? Explain.

WEB file

Showtime

WEB file

NFLPassing

16. In exercise 6, data were given on the average number of passing yards per attempt (Yds/Att), the number of interceptions thrown per attempt (Int/Att), and the percentage of games won (Win%) for a random sample of 16 National Football League (NFL) teams for the 2011 season (NFL website, February 12, 2012).
- Did the estimated regression equation that uses only the average number of passing yards per attempt as the independent variable to predict the percentage of games won provide a good fit?
 - Discuss the benefit of using both the average number of passing yards per attempt and the number of interceptions thrown per attempt to predict the percentage of games won.

WEB file

Boats

17. In exercise 9, an estimated regression equation was developed relating the top speed for a boat to the boat's beam and horsepower rating.
- Compute and interpret R^2 and R_a^2 .
 - Does the estimated regression equation provide a good fit to the data? Explain.

WEB file

MLBPitching

18. Refer to exercise 10, where Major League Baseball (MLB) pitching statistics were reported for a random sample of 20 pitchers from the American League for the 2011 season (MLB website, March 1, 2012).
- In part (c) of exercise 10, an estimated regression equation was developed relating the average number of runs given up per inning pitched given the average number of strikeouts per inning pitched and the average number of home runs per inning pitched. What are the values of R^2 and R_a^2 ?
 - Does the estimated regression equation provide a good fit to the data? Explain.
 - Suppose the earned run average (ERA) is used as the dependent variable in part (c) instead of the average number of runs given up per inning pitched. Does the estimated regression equation that uses the ERA provide a good fit to the data? Explain.

15.4

Model Assumptions

In Section 15.1 we introduced the following multiple regression model.

MULTIPLE REGRESSION MODEL

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_px_p + \epsilon \quad (15.10)$$

The assumptions about the error term ϵ in the multiple regression model parallel those for the simple linear regression model.

ASSUMPTIONS ABOUT THE ERROR TERM ϵ IN THE MULTIPLE REGRESSION

MODEL $y = \beta_0 + \beta_1x_1 + \cdots + \beta_px_p + \epsilon$

1. The error term ϵ is a random variable with mean or expected value of zero; that is, $E(\epsilon) = 0$.

Implication: For given values of x_1, x_2, \dots, x_p , the expected, or average, value of y is given by

$$E(y) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_px_p \quad (15.11)$$

Equation (15.11) is the multiple regression equation we introduced in Section 15.1. In this equation, $E(y)$ represents the average of all possible values of y that might occur for the given values of x_1, x_2, \dots, x_p .

2. The variance of ϵ is denoted by σ^2 and is the same for all values of the independent variables x_1, x_2, \dots, x_p .

Implication: The variance of y about the regression line equals σ^2 and is the same for all values of x_1, x_2, \dots, x_p .

3. The values of ϵ are independent.

Implication: The value of ϵ for a particular set of values for the independent variables is not related to the value of ϵ for any other set of values.

4. The error term ϵ is a normally distributed random variable reflecting the deviation between the y value and the expected value of y given by $\beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_px_p$.

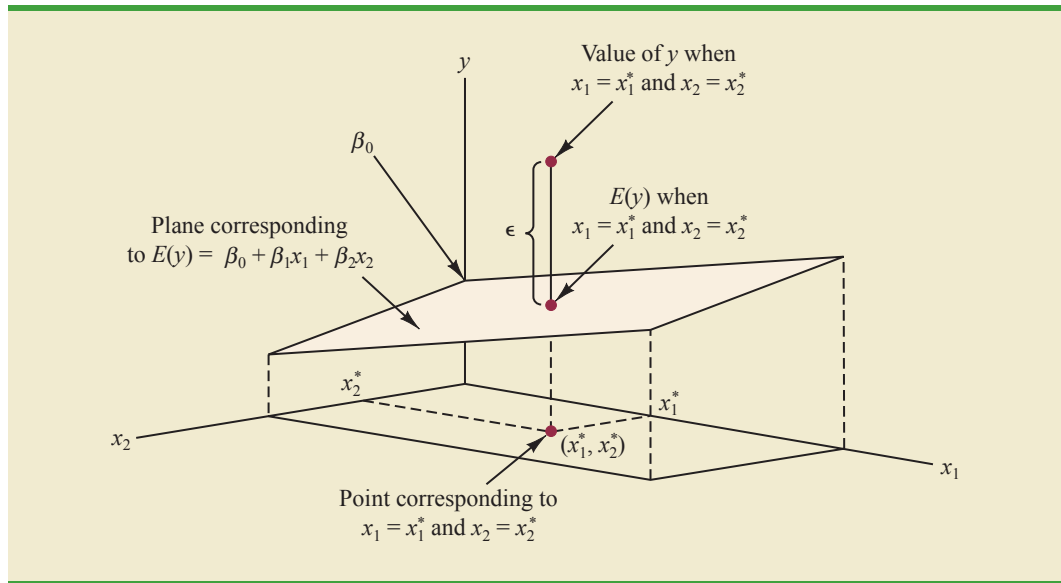
Implication: Because $\beta_0, \beta_1, \dots, \beta_p$ are constants for the given values of x_1, x_2, \dots, x_p , the dependent variable y is also a normally distributed random variable.

To obtain more insight about the form of the relationship given by equation (15.11), consider the following two-independent-variable multiple regression equation.

$$E(y) = \beta_0 + \beta_1x_1 + \beta_2x_2$$

The graph of this equation is a plane in three-dimensional space. Figure 15.5 provides an example of such a graph. Note that the value of ϵ shown is the difference between the actual y value and the expected value of y , $E(y)$, when $x_1 = x_1^*$ and $x_2 = x_2^*$.

FIGURE 15.5 GRAPH OF THE REGRESSION EQUATION FOR MULTIPLE REGRESSION ANALYSIS WITH TWO INDEPENDENT VARIABLES



In regression analysis, the term *response variable* is often used in place of the term *dependent variable*. Furthermore, since the multiple regression equation generates a plane or surface, its graph is called a *response surface*.

15.5 Testing for Significance

In this section we show how to conduct significance tests for a multiple regression relationship. The significance tests we used in simple linear regression were a *t* test and an *F* test. In simple linear regression, both tests provide the same conclusion; that is, if the null hypothesis is rejected, we conclude that $\beta_1 \neq 0$. In multiple regression, the *t* test and the *F* test have different purposes.

1. The *F* test is used to determine whether a significant relationship exists between the dependent variable and the set of all the independent variables; we will refer to the *F* test as the test for *overall significance*.
2. If the *F* test shows an overall significance, the *t* test is used to determine whether each of the individual independent variables is significant. A separate *t* test is conducted for each of the independent variables in the model; we refer to each of these *t* tests as a test for *individual significance*.

In the material that follows, we will explain the *F* test and the *t* test and apply each to the Butler Trucking Company example.

F Test

The multiple regression model as defined in Section 15.4 is

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_px_p + \epsilon$$

The hypotheses for the *F* test involve the parameters of the multiple regression model.

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

$$H_a: \text{One or more of the parameters is not equal to zero}$$

If H_0 is rejected, the test gives us sufficient statistical evidence to conclude that one or more of the parameters is not equal to zero and that the overall relationship between y and the set of independent variables x_1, x_2, \dots, x_p is significant. However, if H_0 cannot be rejected, we do not have sufficient evidence to conclude that a significant relationship is present.

Before describing the steps of the F test, we need to review the concept of *mean square*. A mean square is a sum of squares divided by its corresponding degrees of freedom. In the multiple regression case, the total sum of squares has $n - 1$ degrees of freedom, the sum of squares due to regression (SSR) has p degrees of freedom, and the sum of squares due to error has $n - p - 1$ degrees of freedom. Hence, the mean square due to regression (MSR) is SSR/p and the mean square due to error (MSE) is $SSE/(n - p - 1)$.

$$\text{MSR} = \frac{\text{SSR}}{p} \quad (15.12)$$

and

$$\text{MSE} = \frac{\text{SSE}}{n - p - 1} \quad (15.13)$$

As discussed in Chapter 14, MSE provides an unbiased estimate of σ^2 , the variance of the error term ϵ . If $H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0$ is true, MSR also provides an unbiased estimate of σ^2 , and the value of MSR/MSE should be close to 1. However, if H_0 is false, MSR overestimates σ^2 and the value of MSR/MSE becomes larger. To determine how large the value of MSR/MSE must be to reject H_0 , we make use of the fact that if H_0 is true and the assumptions about the multiple regression model are valid, the sampling distribution of MSR/MSE is an F distribution with p degrees of freedom in the numerator and $n - p - 1$ in the denominator. A summary of the F test for significance in multiple regression follows.

F TEST FOR OVERALL SIGNIFICANCE

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

H_a : One or more of the parameters is not equal to zero

TEST STATISTIC

$$F = \frac{\text{MSR}}{\text{MSE}} \quad (15.14)$$

REJECTION RULE

p -value approach: Reject H_0 if $p\text{-value} \leq \alpha$

Critical value approach: Reject H_0 if $F \geq F_\alpha$

where F_α is based on an F distribution with p degrees of freedom in the numerator and $n - p - 1$ degrees of freedom in the denominator.

Let us apply the F test to the Butler Trucking Company multiple regression problem. With two independent variables, the hypotheses are written as follows.

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

H_a : β_1 and/or β_2 is not equal to zero

FIGURE 15.6 MINITAB OUTPUT FOR BUTLER TRUCKING WITH TWO INDEPENDENT VARIABLES, MILES TRAVELED (x_1) AND NUMBER OF DELIVERIES (x_2)

```

The regression equation is
Time = - 0.869 + 0.0611 Miles + 0.923 Deliveries

Predictor      Coef      SE Coef      T      p
Constant      -0.8687   0.9515     -0.91  0.392
Miles          0.061135 0.009888    6.18  0.000
Deliveries     0.9234    0.2211    4.18  0.004

S = 0.573142   R-Sq = 90.4%   R-Sq(adj) = 87.6%

Analysis of Variance

SOURCE          DF      SS      MS      F      p
Regression       2    21.601   10.800   32.88  0.000
Residual Error   7     2.299    0.328
Total            9    23.900

```

Figure 15.6 is the Minitab output for the multiple regression model with miles traveled (x_1) and number of deliveries (x_2) as the two independent variables. In the analysis of variance part of the output, we see that $MSR = 10.8$ and $MSE = .328$. Using equation (15.14), we obtain the test statistic.

$$F = \frac{10.8}{.328} = 32.9$$

Note that the F value on the Minitab output is $F = 32.88$; the value we calculated differs because we used rounded values for MSR and MSE in the calculation. Using $\alpha = .01$, the p -value = 0.000 in the last column of the analysis of variance table (Figure 15.6) indicates that we can reject $H_0: \beta_1 = \beta_2 = 0$ because the p -value is less than $\alpha = .01$. Alternatively, Table 4 of Appendix B shows that with two degrees of freedom in the numerator and seven degrees of freedom in the denominator, $F_{.01} = 9.55$. With $32.9 > 9.55$, we reject $H_0: \beta_1 = \beta_2 = 0$ and conclude that a significant relationship is present between travel time y and the two independent variables, miles traveled and number of deliveries.

As noted previously, the mean square error provides an unbiased estimate of σ^2 , the variance of the error term ϵ . Referring to Figure 15.6, we see that the estimate of σ^2 is $MSE = .328$. The square root of MSE is the estimate of the standard deviation of the error term. As defined in Section 14.5, this standard deviation is called the standard error of the estimate and is denoted s . Hence, we have $s = \sqrt{MSE} = \sqrt{.328} = .573$. Note that the value of the standard error of the estimate appears in the Minitab output in Figure 15.6.

Table 15.3 is the general analysis of variance (ANOVA) table that provides the F test results for a multiple regression model. The value of the F test statistic appears in the last column and can be compared to F_α with p degrees of freedom in the numerator and $n - p - 1$ degrees of freedom in the denominator to make the hypothesis test conclusion. By reviewing the Minitab output for Butler Trucking Company in Figure 15.6, we see that Minitab's analysis of variance table contains this information. Moreover, Minitab also provides the p -value corresponding to the F test statistic.

TABLE 15.3 ANOVA TABLE FOR A MULTIPLE REGRESSION MODEL WITH p INDEPENDENT VARIABLES

Source	Sum of Squares	Degrees of Freedom	Mean Square	F
Regression	SSR	p	$MSR = \frac{SSR}{p}$	$F = \frac{MSR}{MSE}$
Error	SSE	$n - p - 1$	$MSE = \frac{SSE}{n - p - 1}$	
Total	SST	$n - 1$		

t Test

If the F test shows that the multiple regression relationship is significant, a t test can be conducted to determine the significance of each of the individual parameters. The t test for individual significance follows.

t TEST FOR INDIVIDUAL SIGNIFICANCE

For any parameter β_i

$$H_0: \beta_i = 0$$

$$H_a: \beta_i \neq 0$$

TEST STATISTIC

$$t = \frac{b_i}{s_{b_i}} \quad (15.15)$$

REJECTION RULE

p -value approach: Reject H_0 if $p\text{-value} \leq \alpha$

Critical value approach: Reject H_0 if $t \leq -t_{\alpha/2}$ or if $t \geq t_{\alpha/2}$

where $t_{\alpha/2}$ is based on a t distribution with $n - p - 1$ degrees of freedom.

In the test statistic, s_{b_i} is the estimate of the standard deviation of b_i . The value of s_{b_i} will be provided by the computer software package.

Let us conduct the t test for the Butler Trucking regression problem. Refer to the section of Figure 15.6 that shows the Minitab output for the t -ratio calculations. Values of b_1 , b_2 , s_{b_1} , and s_{b_2} are as follows.

$$b_1 = .061135 \quad s_{b_1} = .009888$$

$$b_2 = .9234 \quad s_{b_2} = .2211$$

Using equation (15.15), we obtain the test statistic for the hypotheses involving parameters β_1 and β_2 .

$$t = .061135/.009888 = 6.18$$

$$t = .9234/.2211 = 4.18$$

Note that both of these t -ratio values and the corresponding p -values are provided by the Minitab output in Figure 15.6. Using $\alpha = .01$, the p -values of .000 and .004 on the Minitab output indicate that we can reject $H_0: \beta_1 = 0$ and $H_0: \beta_2 = 0$. Hence, both parameters are statistically significant. Alternatively, Table 2 of Appendix B shows that with $n - p - 1 = 10 - 2 - 1 = 7$ degrees of freedom, $t_{.005} = 3.499$. With $6.18 > 3.499$, we reject $H_0: \beta_1 = 0$. Similarly, with $4.18 > 3.499$, we reject $H_0: \beta_2 = 0$.

Multicollinearity

We use the term *independent variable* in regression analysis to refer to any variable being used to predict or explain the value of the dependent variable. The term does not mean, however, that the independent variables themselves are independent in any statistical sense. On the contrary, most independent variables in a multiple regression problem are correlated to some degree with one another. For example, in the Butler Trucking example involving the two independent variables x_1 (miles traveled) and x_2 (number of deliveries), we could treat the miles traveled as the dependent variable and the number of deliveries as the independent variable to determine whether those two variables are themselves related. We could then compute the sample correlation coefficient r_{x_1, x_2} to determine the extent to which the variables are related. Doing so yields $r_{x_1, x_2} = .16$. Thus, we find some degree of linear association between the two independent variables. In multiple regression analysis, **multicollinearity** refers to the correlation among the independent variables.

To provide a better perspective of the potential problems of multicollinearity, let us consider a modification of the Butler Trucking example. Instead of x_2 being the number of deliveries, let x_2 denote the number of gallons of gasoline consumed. Clearly, x_1 (the miles traveled) and x_2 are related; that is, we know that the number of gallons of gasoline used depends on the number of miles traveled. Hence, we would conclude logically that x_1 and x_2 are highly correlated independent variables.

Assume that we obtain the equation $\hat{y} = b_0 + b_1x_1 + b_2x_2$ and find that the F test shows the relationship to be significant. Then suppose we conduct a t test on β_1 to determine whether $\beta_1 \neq 0$, and we cannot reject $H_0: \beta_1 = 0$. Does this result mean that travel time is not related to miles traveled? Not necessarily. What it probably means is that with x_2 already in the model, x_1 does not make a significant contribution to determining the value of y . This interpretation makes sense in our example; if we know the amount of gasoline consumed, we do not gain much additional information useful in predicting y by knowing the miles traveled. Similarly, a t test might lead us to conclude $\beta_2 = 0$ on the grounds that, with x_1 in the model, knowledge of the amount of gasoline consumed does not add much.

To summarize, in t tests for the significance of individual parameters, the difficulty caused by multicollinearity is that it is possible to conclude that none of the individual parameters are significantly different from zero when an F test on the overall multiple regression equation indicates a significant relationship. This problem is avoided when there is little correlation among the independent variables.

Statisticians have developed several tests for determining whether multicollinearity is high enough to cause problems. According to the rule of thumb test, multicollinearity is a potential problem if the absolute value of the sample correlation coefficient exceeds .7 for any two of the independent variables. The other types of tests are more advanced and beyond the scope of this text.

If possible, every attempt should be made to avoid including independent variables that are highly correlated. In practice, however, strict adherence to this policy is rarely possible. When decision makers have reason to believe substantial multicollinearity is present, they must realize that separating the effects of the individual independent variables on the dependent variable is difficult.

A sample correlation coefficient greater than +.7 or less than -.7 for two independent variables is a rule of thumb warning of potential problems with multicollinearity.

When the independent variables are highly correlated, it is not possible to determine the separate effect of any particular independent variable on the dependent variable.

NOTES AND COMMENTS

Ordinarily, multicollinearity does not affect the way in which we perform our regression analysis or interpret the output from a study. However, when multicollinearity is severe—that is, when two or more of the independent variables are highly correlated with one another—we can have difficulty interpreting the results of t tests on the individual parameters. In addition to the type of problem illustrated in this section, severe cases of multicollinearity have been shown to result in least squares estimates that have the wrong sign. That is,

in simulated studies where researchers created the underlying regression model and then applied the least squares technique to develop estimates of β_0 , β_1 , β_2 , and so on, it has been shown that under conditions of high multicollinearity the least squares estimates can have a sign opposite that of the parameter being estimated. For example, β_2 might actually be $+10$ and b_2 , its estimate, might turn out to be -2 . Thus, little faith can be placed in the individual coefficients if multicollinearity is present to a high degree.

Exercises

Methods

SELF test

19. In exercise 1, the following estimated regression equation based on 10 observations was presented.

$$\hat{y} = 29.1270 + .5906x_1 + .4980x_2$$

Here $SST = 6724.125$, $SSR = 6216.375$, $s_{b_1} = .0813$, and $s_{b_2} = .0567$.

- Compute MSR and MSE.
 - Compute F and perform the appropriate F test. Use $\alpha = .05$.
 - Perform a t test for the significance of β_1 . Use $\alpha = .05$.
 - Perform a t test for the significance of β_2 . Use $\alpha = .05$.
20. Refer to the data presented in exercise 2. The estimated regression equation for these data is

$$\hat{y} = -18.37 + 2.01x_1 + 4.74x_2$$

Here $SST = 15,182.9$, $SSR = 14,052.2$, $s_{b_1} = .2471$, and $s_{b_2} = .9484$.

- Test for a significant relationship among x_1 , x_2 , and y . Use $\alpha = .05$.
 - Is β_1 significant? Use $\alpha = .05$.
 - Is β_2 significant? Use $\alpha = .05$.
21. The following estimated regression equation was developed for a model involving two independent variables.

$$\hat{y} = 40.7 + 8.63x_1 + 2.71x_2$$

After x_2 was dropped from the model, the least squares method was used to obtain an estimated regression equation involving only x_1 as an independent variable.

$$\hat{y} = 42.0 + 9.01x_1$$

- Give an interpretation of the coefficient of x_1 in both models.
- Could multicollinearity explain why the coefficient of x_1 differs in the two models? If so, how?

Applications

22. In exercise 4, the following estimated regression equation relating sales to inventory investment and advertising expenditures was given.

$$\hat{y} = 25 + 10x_1 + 8x_2$$

The data used to develop the model came from a survey of 10 stores; for these data $SST = 16,000$ and $SSR = 12,000$.

- Compute SSE, MSE, and MSR.
- Use an F test and a .05 level of significance to determine whether there is a relationship among the variables.

SELF test

- Refer to exercise 5.
 - Use $\alpha = .01$ to test the hypotheses

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

$$H_a: \beta_1 \text{ and/or } \beta_2 \text{ is not equal to zero}$$

for the model $y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \epsilon$, where

x_1 = television advertising (\$1000s)

x_2 = newspaper advertising (\$1000s)

- Use $\alpha = .05$ to test the significance of β_1 . Should x_1 be dropped from the model?
 - Use $\alpha = .05$ to test the significance of β_2 . Should x_2 be dropped from the model?
- The Wall Street Journal* conducted a study of basketball spending at top colleges. A portion of the data showing the revenue (\$ millions), percentage of wins, and the coach's salary (\$ millions) for 39 of the country's top basketball programs follows (*The Wall Street Journal*, March 11–12, 2006).

WEB file

Basketball

School	Revenue	%Wins	Salary
Alabama	6.5	61	1.00
Arizona	16.6	63	.70
Arkansas	11.1	72	.80
Boston College	3.4	80	.53
.	.	.	.
.	.	.	.
Washington	5.0	83	.89
West Virginia	4.9	67	.70
Wichita State	3.1	75	.41
Wisconsin	12.0	66	.70

- Develop the estimated regression equation that can be used to predict the coach's salary given the revenue generated by the program and the percentage of wins.
 - Use the F test to determine the overall significance of the relationship. What is your conclusion at the .05 level of significance?
 - Use the t test to determine the significance of each independent variable. What is your conclusion at the .05 level of significance?
- The *Condé Nast Traveler* Gold List for 2012 provided ratings for the top 20 small cruise ships (*Condé Nast Traveler* website, March 1, 2012). The data shown below are the scores each ship received based upon the results from *Condé Nast Traveler's* annual Readers' Choice Survey. Each score represents the percentage of respondents who rated a ship as excellent or very good on several criteria, including Itineraries/Schedule, Shore Excursions, and Food/Dining. An overall score was also reported and used to rank the ships. The highest ranked ship, the *Seabourn Odyssey*, has an overall score of 94.4, the highest component of which is 97.8 for Food/Dining.



Ship	Overall	Itineraries/ Schedule	Shore Excursions	Food/ Dining
<i>Seabourn Odyssey</i>	94.4	94.6	90.9	97.8
<i>Seabourn Pride</i>	93.0	96.7	84.2	96.7
<i>National Geographic Endeavor</i>	92.9	100.0	100.0	88.5
<i>Seabourn Sojourn</i>	91.3	88.6	94.8	97.1
<i>Paul Gauguin</i>	90.5	95.1	87.9	91.2
<i>Seabourn Legend</i>	90.3	92.5	82.1	98.8
<i>Seabourn Spirit</i>	90.2	96.0	86.3	92.0
<i>Silver Explorer</i>	89.9	92.6	92.6	88.9
<i>Silver Spirit</i>	89.4	94.7	85.9	90.8
<i>Seven Seas Navigator</i>	89.2	90.6	83.3	90.5
<i>Silver Whisperer</i>	89.2	90.9	82.0	88.6
<i>National Geographic Explorer</i>	89.1	93.1	93.1	89.7
<i>Silver Cloud</i>	88.7	92.6	78.3	91.3
<i>Celebrity Xpedition</i>	87.2	93.1	91.7	73.6
<i>Silver Shadow</i>	87.2	91.0	75.0	89.7
<i>Silver Wind</i>	86.6	94.4	78.1	91.6
<i>SeaDream II</i>	86.2	95.5	77.4	90.9
<i>Wind Star</i>	86.1	94.9	76.5	91.5
<i>Wind Surf</i>	86.1	92.1	72.3	89.3
<i>Wind Spirit</i>	85.2	93.5	77.4	91.9

- Determine the estimated regression equation that can be used to predict the overall score given the scores for Itineraries/Schedule, Shore Excursions, and Food/Dining.
 - Use the F test to determine the overall significance of the relationship. What is your conclusion at the .05 level of significance?
 - Use the t test to determine the significance of each independent variable. What is your conclusion at the .05 level of significance?
 - Remove any independent variable that is not significant from the estimated regression equation. What is your recommended estimated regression equation?
26. In exercise 10, data showing the values of several pitching statistics for a random sample of 20 pitchers from the American League of Major League Baseball were provided. In part (c) of this exercise an estimated regression equation was developed to predict the average number of runs given up per inning pitched (R/IP) given the average number of strikeouts per inning pitched (SO/IP) and the average number of home runs per inning pitched (HR/IP).
- Use the F test to determine the overall significance of the relationship. What is your conclusion at the .05 level of significance?
 - Use the t test to determine the significance of each independent variable. What is your conclusion at the .05 level of significance?



15.6

Using the Estimated Regression Equation for Estimation and Prediction

The procedures for estimating the mean value of y and predicting an individual value of y in multiple regression are similar to those in regression analysis involving one independent variable. First, recall that in Chapter 14 we showed that the point estimate of the expected value of y for a given value of x was the same as the point estimate of an individual value of y . In both cases, we used $\hat{y} = b_0 + b_1x$ as the point estimate.

In multiple regression we use the same procedure. That is, we substitute the given values of x_1, x_2, \dots, x_p into the estimated regression equation and use the corresponding value of \hat{y} as the point estimate. Suppose that for the Butler Trucking example we want to use the

TABLE 15.4 THE 95% CONFIDENCE AND PREDICTION INTERVALS FOR BUTLER TRUCKING

Value of x_1	Value of x_2	Confidence Interval		Prediction Interval	
		Lower Limit	Upper Limit	Lower Limit	Upper Limit
50	2	3.146	4.924	2.414	5.656
50	3	4.127	5.789	3.368	6.548
50	4	4.815	6.948	4.157	7.607
100	2	6.258	7.926	5.500	8.683
100	3	7.385	8.645	6.520	9.510
100	4	8.135	9.742	7.362	10.515

estimated regression equation involving x_1 (miles traveled) and x_2 (number of deliveries) to develop two interval estimates:

1. A *confidence interval* of the mean travel time for all trucks that travel 100 miles and make two deliveries
2. A *prediction interval* of the travel time for *one specific* truck that travels 100 miles and makes two deliveries

Using the estimated regression equation $\hat{y} = -.869 + .0611x_1 + .923x_2$ with $x_1 = 100$ and $x_2 = 2$, we obtain the following value of \hat{y} .

$$\hat{y} = -.869 + .0611(100) + .923(2) = 7.09$$

Hence, the point estimate of travel time in both cases is approximately seven hours.

To develop interval estimates for the mean value of y and for an individual value of y , we use a procedure similar to that for regression analysis involving one independent variable. The formulas required are beyond the scope of the text, but computer packages for multiple regression analysis will often provide confidence intervals once the values of x_1, x_2, \dots, x_p are specified by the user. In Table 15.4 we show the 95% confidence and prediction intervals for the Butler Trucking example for selected values of x_1 and x_2 ; these values were obtained using Minitab. Note that the interval estimate for an individual value of y is wider than the interval estimate for the expected value of y . This difference simply reflects the fact that for given values of x_1 and x_2 we can estimate the mean travel time for all trucks with more precision than we can predict the travel time for one specific truck.

Exercises

Methods

27. In exercise 1, the following estimated regression equation based on 10 observations was presented.

$$\hat{y} = 29.1270 + .5906x_1 + .4980x_2$$

- a. Develop a point estimate of the mean value of y when $x_1 = 180$ and $x_2 = 310$.
 - b. Develop a point estimate for an individual value of y when $x_1 = 180$ and $x_2 = 310$.
28. Refer to the data in exercise 2. The estimated regression equation for those data is

$$\hat{y} = -18.4 + 2.01x_1 + 4.74x_2$$

- a. Develop a 95% confidence interval for the mean value of y when $x_1 = 45$ and $x_2 = 15$.
- b. Develop a 95% prediction interval for y when $x_1 = 45$ and $x_2 = 15$.

SELF test

Applications

SELF test

29. In exercise 5, the owner of Showtime Movie Theaters, Inc., used multiple regression analysis to predict gross revenue (y) as a function of television advertising (x_1) and newspaper advertising (x_2). The estimated regression equation was

$$\hat{y} = 83.2 + 2.29x_1 + 1.30x_2$$

- What is the gross revenue expected for a week when \$3500 is spent on television advertising ($x_1 = 3.5$) and \$1800 is spent on newspaper advertising ($x_2 = 1.8$)?
 - Provide a 95% confidence interval for the mean revenue of all weeks with the expenditures listed in part (a).
 - Provide a 95% prediction interval for next week's revenue, assuming that the advertising expenditures will be allocated as in part (a).
30. In exercise 9 an estimated regression equation was developed relating the top speed for a boat to the boat's beam and horsepower rating.
- Develop a 95% confidence interval for the mean top speed of a boat with a beam of 85 inches and an engine with a 330 horsepower rating.
 - The Sv fara SV609 has a beam of 85 inches and an engine with a 330 horsepower rating. Develop a 95% prediction interval for the mean top speed for the Sv fara SV609.
31. The American Association of Individual Investors (AAII) On-Line Discount Broker Survey polls members on their experiences with electronic trades handled by discount brokers. As part of the survey, members were asked to rate their satisfaction with the trade price and the speed of execution, as well as provide an overall satisfaction rating. Possible responses (scores) were no opinion (0), unsatisfied (1), somewhat satisfied (2), satisfied (3), and very satisfied (4). For each broker, summary scores were computed by computing a weighted average of the scores provided by each respondent. A portion of the survey results follows (AAII website, February 7, 2012).

WEB file

Boats

WEB file

Broker

Brokerage	Trade Price	Speed of Execution	Satisfaction Electronic Trades
Scottrade, Inc.	3.4	3.4	3.5
Charles Schwab	3.2	3.3	3.4
Fidelity Brokerage Services	3.1	3.4	3.9
TD Ameritrade	2.9	3.6	3.7
E*Trade Financial	2.9	3.2	2.9
(Not listed)	2.5	3.2	2.7
Vanguard Brokerage Services	2.6	3.8	2.8
USAA Brokerage Services	2.4	3.8	3.6
Thinkorswim	2.6	2.6	2.6
Wells Fargo Investments	2.3	2.7	2.3
Interactive Brokers	3.7	4.0	4.0
Zecco.com	2.5	2.5	2.5
Firstrade Securities	3.0	3.0	4.0
Banc of America Investment Services	4.0	1.0	2.0

- Develop an estimated regression equation using trade price and speed of execution to predict overall satisfaction with the broker.
- Finger Lakes Investments has developed a new electronic trading system and would like to predict overall customer satisfaction assuming they can provide satisfactory levels of service levels (3) for both trade price and speed of execution. Use the estimated regression equation developed in part (a) to predict overall satisfaction level for Finger Lakes Investments if they can achieve these performance levels.

- c. Develop a 95% confidence interval estimate of the overall satisfaction of electronic trades for all brokers that provide satisfactory levels of service for both trade price and speed of execution.
- d. Develop a 95% prediction interval of overall satisfaction for Finger Lakes Investments assuming they achieve service levels of 3 for both trade price and speed of execution.

15.7

Categorical Independent Variables

The independent variables may be categorical or quantitative.

Thus far, the examples we have considered involved quantitative independent variables such as student population, distance traveled, and number of deliveries. In many situations, however, we must work with **categorical independent variables** such as gender (male, female), method of payment (cash, credit card, check), and so on. The purpose of this section is to show how categorical variables are handled in regression analysis. To illustrate the use and interpretation of a categorical independent variable, we will consider a problem facing the managers of Johnson Filtration, Inc.

An Example: Johnson Filtration, Inc.

Johnson Filtration, Inc., provides maintenance service for water-filtration systems throughout southern Florida. Customers contact Johnson with requests for maintenance service on their water-filtration systems. To estimate the service time and the service cost, Johnson's managers want to predict the repair time necessary for each maintenance request. Hence, repair time in hours is the dependent variable. Repair time is believed to be related to two factors, the number of months since the last maintenance service and the type of repair problem (mechanical or electrical). Data for a sample of 10 service calls are reported in Table 15.5.

Let y denote the repair time in hours and x_1 denote the number of months since the last maintenance service. The regression model that uses only x_1 to predict y is

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

Using Minitab to develop the estimated regression equation, we obtained the output shown in Figure 15.7. The estimated regression equation is

$$\hat{y} = 2.15 + .304x_1 \quad (15.16)$$

At the .05 level of significance, the p -value of .016 for the t (or F) test indicates that the number of months since the last service is significantly related to repair time. $R\text{-sq} = 53.4\%$ indicates that x_1 alone explains 53.4% of the variability in repair time.

TABLE 15.5 DATA FOR THE JOHNSON FILTRATION EXAMPLE

Service Call	Months Since Last Service	Type of Repair	Repair Time in Hours
1	2	electrical	2.9
2	6	mechanical	3.0
3	8	electrical	4.8
4	3	mechanical	1.8
5	2	electrical	2.9
6	7	electrical	4.9
7	9	mechanical	4.2
8	8	mechanical	4.8
9	4	electrical	4.4
10	6	electrical	4.5

FIGURE 15.7 MINITAB OUTPUT FOR JOHNSON FILTRATION WITH MONTHS SINCE LAST SERVICE (x_1) AS THE INDEPENDENT VARIABLE

The regression equation is
Time = 2.15 + 0.304 Months

Predictor	Coef	SE Coef	T	p
Constant	2.1473	0.6050	3.55	0.008
Months	0.3041	0.1004	3.03	0.016

S = 0.781022 R-Sq = 53.4% R-Sq(adj) = 47.6%

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	1	5.5960	5.5960	9.17	0.016
Residual Error	8	4.8800	0.6100		
Total	9	10.4760			

In the Minitab output the variable names Months and Time were entered as the column headings on the worksheet; thus, $x_1 =$ Months and $y =$ Time.

To incorporate the type of repair into the regression model, we define the following variable.

$$x_2 = \begin{cases} 0 & \text{if the type of repair is mechanical} \\ 1 & \text{if the type of repair is electrical} \end{cases}$$

In regression analysis x_2 is called a **dummy** or **indicator variable**. Using this dummy variable, we can write the multiple regression model as

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

Table 15.6 is the revised data set that includes the values of the dummy variable. Using Minitab and the data in Table 15.6, we can develop estimates of the model parameters. The Minitab output in Figure 15.8 shows that the estimated multiple regression equation is

$$\hat{y} = .93 + .388x_1 + 1.26x_2 \quad (15.17)$$

TABLE 15.6 DATA FOR THE JOHNSON FILTRATION EXAMPLE WITH TYPE OF REPAIR INDICATED BY A DUMMY VARIABLE ($x_2 = 0$ FOR MECHANICAL; $x_2 = 1$ FOR ELECTRICAL)

Customer	Months Since Last Service (x_1)	Type of Repair (x_2)	Repair Time in Hours (y)
1	2	1	2.9
2	6	0	3.0
3	8	1	4.8
4	3	0	1.8
5	2	1	2.9
6	7	1	4.9
7	9	0	4.2
8	8	0	4.8
9	4	1	4.4
10	6	1	4.5

WEB file
Johnson

FIGURE 15.8 MINITAB OUTPUT FOR JOHNSON FILTRATION WITH MONTHS SINCE LAST SERVICE (x_1) AND TYPE OF REPAIR (x_2) AS THE INDEPENDENT VARIABLES

In the Minitab output the variable names Months, Type, and Time were entered as the column headings on the worksheet; thus, x_1 = Months, x_2 = Type, and y = Time.

The regression equation is					
Time = 0.930 + 0.388 Months + 1.26 Type					
Predictor	Coef	SE Coef	T	p	
Constant	0.9305	0.4670	1.99	0.087	
Months	0.38762	0.06257	6.20	0.000	
Type	1.2627	0.3141	4.02	0.005	
S = 0.459048 R-Sq = 85.9% R-Sq (adj) = 81.9%					
Analysis of Variance					
SOURCE	DF	SS	MS	F	p
Regression	2	9.0009	4.5005	21.36	0.001
Residual Error	7	1.4751	0.2107		
Total	9	10.4760			

At the .05 level of significance, the p -value of .001 associated with the F test ($F = 21.36$) indicates that the regression relationship is significant. The t test part of the printout in Figure 15.8 shows that both months since last service (p -value = .000) and type of repair (p -value = .005) are statistically significant. In addition, $R\text{-Sq} = 85.9\%$ and $R\text{-Sq (adj)} = 81.9\%$ indicate that the estimated regression equation does a good job of explaining the variability in repair times. Thus, equation (15.17) should prove helpful in predicting the repair time necessary for the various service calls.

Interpreting the Parameters

The multiple regression equation for the Johnson Filtration example is

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \quad (15.18)$$

To understand how to interpret the parameters β_0 , β_1 , and β_2 when a categorical variable is present, consider the case when $x_2 = 0$ (mechanical repair). Using $E(y \mid \text{mechanical})$ to denote the mean or expected value of repair time *given* a mechanical repair, we have

$$E(y \mid \text{mechanical}) = \beta_0 + \beta_1 x_1 + \beta_2(0) = \beta_0 + \beta_1 x_1 \quad (15.19)$$

Similarly, for an electrical repair ($x_2 = 1$), we have

$$\begin{aligned} E(y \mid \text{electrical}) &= \beta_0 + \beta_1 x_1 + \beta_2(1) = \beta_0 + \beta_1 x_1 + \beta_2 \\ &= (\beta_0 + \beta_2) + \beta_1 x_1 \end{aligned} \quad (15.20)$$

Comparing equations (15.19) and (15.20), we see that the mean repair time is a linear function of x_1 for both mechanical and electrical repairs. The slope of both equations is β_1 , but the y -intercept differs. The y -intercept is β_0 in equation (15.19) for mechanical repairs and $(\beta_0 + \beta_2)$ in equation (15.20) for electrical repairs. The interpretation of β_2 is that it indicates the difference between the mean repair time for an electrical repair and the mean repair time for a mechanical repair.

If β_2 is positive, the mean repair time for an electrical repair will be greater than that for a mechanical repair; if β_2 is negative, the mean repair time for an electrical repair will be less than that for a mechanical repair. Finally, if $\beta_2 = 0$, there is no difference in the mean repair time between electrical and mechanical repairs and the type of repair is not related to the repair time.

Using the estimated multiple regression equation $\hat{y} = .93 + .388x_1 + 1.26x_2$, we see that .93 is the estimate of β_0 and 1.26 is the estimate of β_2 . Thus, when $x_2 = 0$ (mechanical repair)

$$\hat{y} = .93 + .388x_1 \quad (15.21)$$

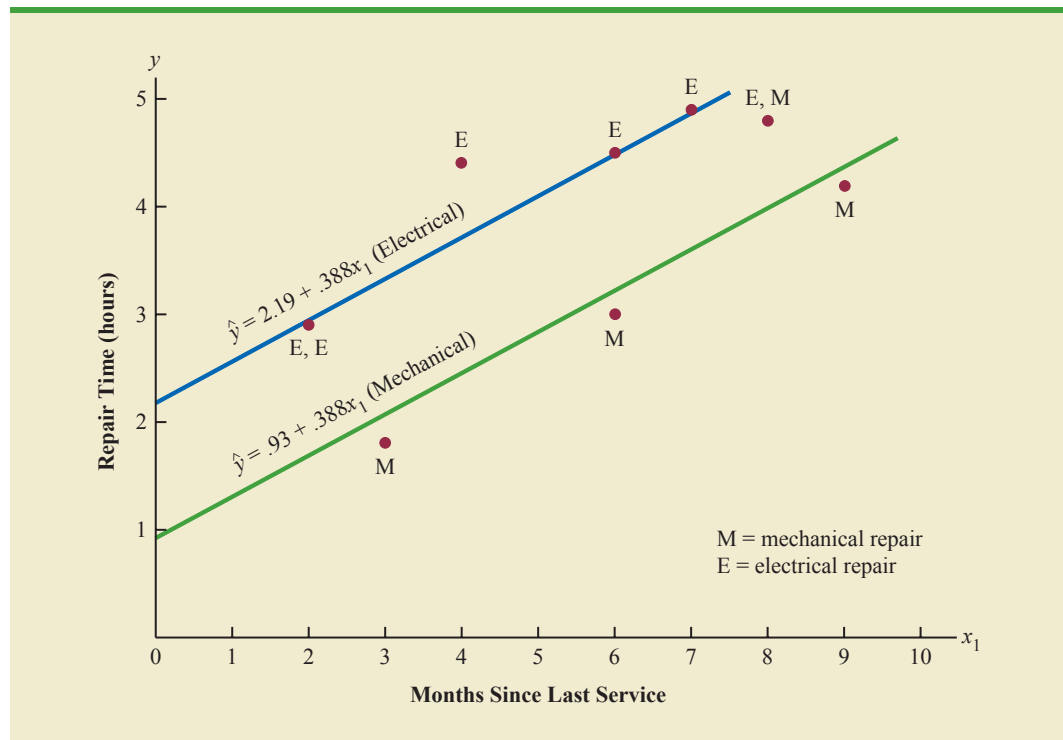
and when $x_2 = 1$ (electrical repair)

$$\begin{aligned} \hat{y} &= .93 + .388x_1 + 1.26(1) \\ &= 2.19 + .388x_1 \end{aligned} \quad (15.22)$$

In effect, the use of a dummy variable for type of repair provides two estimated regression equations that can be used to predict the repair time, one corresponding to mechanical repairs and one corresponding to electrical repairs. In addition, with $b_2 = 1.26$, we learn that, on average, electrical repairs require 1.26 hours longer than mechanical repairs.

Figure 15.9 is the plot of the Johnson data from Table 15.6. Repair time in hours (y) is represented by the vertical axis and months since last service (x_1) is represented by the horizontal axis. A data point for a mechanical repair is indicated by an M and a data point for an electrical repair is indicated by an E. Equations (15.21) and (15.22) are plotted on the graph to show graphically the two equations that can be used to predict the repair time, one corresponding to mechanical repairs and one corresponding to electrical repairs.

FIGURE 15.9 SCATTER DIAGRAM FOR THE JOHNSON FILTRATION REPAIR DATA FROM TABLE 15.6



More Complex Categorical Variables

A categorical variable with k levels must be modeled using $k - 1$ dummy variables. Care must be taken in defining and interpreting the dummy variables.

Because the categorical variable for the Johnson Filtration example had two levels (mechanical and electrical), defining a dummy variable with zero indicating a mechanical repair and one indicating an electrical repair was easy. However, when a categorical variable has more than two levels, care must be taken in both defining and interpreting the dummy variables. As we will show, if a categorical variable has k levels, $k - 1$ dummy variables are required, with each dummy variable being coded as 0 or 1.

For example, suppose a manufacturer of copy machines organized the sales territories for a particular state into three regions: A, B, and C. The managers want to use regression analysis to help predict the number of copiers sold per week. With the number of units sold as the dependent variable, they are considering several independent variables (the number of sales personnel, advertising expenditures, and so on). Suppose the managers believe sales region is also an important factor in predicting the number of copiers sold. Because sales region is a categorical variable with three levels, A, B and C, we will need $3 - 1 = 2$ dummy variables to represent the sales region. Each variable can be coded 0 or 1 as follows.

$$x_1 = \begin{cases} 1 & \text{if sales region B} \\ 0 & \text{otherwise} \end{cases}$$

$$x_2 = \begin{cases} 1 & \text{if sales region C} \\ 0 & \text{otherwise} \end{cases}$$

With this definition, we have the following values of x_1 and x_2 .

Region	x_1	x_2
A	0	0
B	1	0
C	0	1

Observations corresponding to region A would be coded $x_1 = 0, x_2 = 0$; observations corresponding to region B would be coded $x_1 = 1, x_2 = 0$; and observations corresponding to region C would be coded $x_1 = 0, x_2 = 1$.

The regression equation relating the expected value of the number of units sold, $E(y)$, to the dummy variables would be written as

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

To help us interpret the parameters β_0, β_1 , and β_2 , consider the following three variations of the regression equation.

$$E(y \text{ region A}) = \beta_0 + \beta_1(0) + \beta_2(0) = \beta_0$$

$$E(y \text{ region B}) = \beta_0 + \beta_1(1) + \beta_2(0) = \beta_0 + \beta_1$$

$$E(y \text{ region C}) = \beta_0 + \beta_1(0) + \beta_2(1) = \beta_0 + \beta_2$$

Thus, β_0 is the mean or expected value of sales for region A; β_1 is the difference between the mean number of units sold in region B and the mean number of units sold in region A; and β_2 is the difference between the mean number of units sold in region C and the mean number of units sold in region A.

Two dummy variables were required because sales region is a categorical variable with three levels. But the assignment of $x_1 = 0, x_2 = 0$ to indicate region A, $x_1 = 1, x_2 = 0$ to

indicate region B, and $x_1 = 0, x_2 = 1$ to indicate region C was arbitrary. For example, we could have chosen $x_1 = 1, x_2 = 0$ to indicate region A, $x_1 = 0, x_2 = 0$ to indicate region B, and $x_1 = 0, x_2 = 1$ to indicate region C. In that case, β_1 would have been interpreted as the mean difference between regions A and B and β_2 as the mean difference between regions C and B.

The important point to remember is that when a categorical variable has k levels, $k - 1$ dummy variables are required in the multiple regression analysis. Thus, if the sales region example had a fourth region, labeled D, three dummy variables would be necessary. For example, the three dummy variables can be coded as follows.

$$x_1 = \begin{cases} 1 & \text{if sales region B} \\ 0 & \text{otherwise} \end{cases} \quad x_2 = \begin{cases} 1 & \text{if sales region C} \\ 0 & \text{otherwise} \end{cases} \quad x_3 = \begin{cases} 1 & \text{if sales region D} \\ 0 & \text{otherwise} \end{cases}$$

Exercises

Methods

SELF test

32. Consider a regression study involving a dependent variable y , a quantitative independent variable x_1 , and a categorical independent variable with two levels (level 1 and level 2).
 - a. Write a multiple regression equation relating x_1 and the categorical variable to y .
 - b. What is the expected value of y corresponding to level 1 of the categorical variable?
 - c. What is the expected value of y corresponding to level 2 of the categorical variable?
 - d. Interpret the parameters in your regression equation.
33. Consider a regression study involving a dependent variable y , a quantitative independent variable x_1 , and a categorical independent variable with three possible levels (level 1, level 2, and level 3).
 - a. How many dummy variables are required to represent the categorical variable?
 - b. Write a multiple regression equation relating x_1 and the categorical variable to y .
 - c. Interpret the parameters in your regression equation.

Applications

SELF test

34. Management proposed the following regression model to predict sales at a fast-food outlet.

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

where

x_1 = number of competitors within one mile

x_2 = population within one mile (1000s)

$x_3 = \begin{cases} 1 & \text{if drive-up window present} \\ 0 & \text{otherwise} \end{cases}$

y = sales (\$1000s)

The following estimated regression equation was developed after 20 outlets were surveyed.

$$\hat{y} = 10.1 - 4.2x_1 + 6.8x_2 + 15.3x_3$$

- a. What is the expected amount of sales attributable to the drive-up window?
- b. Predict sales for a store with two competitors, a population of 8000 within one mile, and no drive-up window.
- c. Predict sales for a store with one competitor, a population of 3000 within one mile, and a drive-up window.

35. Refer to the Johnson Filtration problem introduced in this section. Suppose that in addition to information on the number of months since the machine was serviced and whether a mechanical or an electrical repair was necessary, the managers obtained a list showing which repairperson performed the service. The revised data follow.



Repair Time in Hours	Months Since Last Service	Type of Repair	Repairperson
2.9	2	Electrical	Dave Newton
3.0	6	Mechanical	Dave Newton
4.8	8	Electrical	Bob Jones
1.8	3	Mechanical	Dave Newton
2.9	2	Electrical	Dave Newton
4.9	7	Electrical	Bob Jones
4.2	9	Mechanical	Bob Jones
4.8	8	Mechanical	Bob Jones
4.4	4	Electrical	Bob Jones
4.5	6	Electrical	Dave Newton

- a. Ignore for now the months since the last maintenance service (x_1) and the repairperson who performed the service. Develop the estimated simple linear regression equation to predict the repair time (y) given the type of repair (x_2). Recall that $x_2 = 0$ if the type of repair is mechanical and 1 if the type of repair is electrical.
- b. Does the equation that you developed in part (a) provide a good fit for the observed data? Explain.
- c. Ignore for now the months since the last maintenance service and the type of repair associated with the machine. Develop the estimated simple linear regression equation to predict the repair time given the repairperson who performed the service. Let $x_3 = 0$ if Bob Jones performed the service and $x_3 = 1$ if Dave Newton performed the service.
- d. Does the equation that you developed in part (c) provide a good fit for the observed data? Explain.
36. This problem is an extension of the situation described in exercise 35.
- a. Develop the estimated regression equation to predict the repair time given the number of months since the last maintenance service, the type of repair, and the repairperson who performed the service.
- b. At the .05 level of significance, test whether the estimated regression equation developed in part (a) represents a significant relationship between the independent variables and the dependent variable.
- c. Is the addition of the independent variable x_3 , the repairperson who performed the service, statistically significant? Use $\alpha = .05$. What explanation can you give for the results observed?
37. The *Consumer Reports* Restaurant Customer Satisfaction Survey is based upon 148,599 visits to full-service restaurant chains (*Consumer Reports* website, February 11, 2009). Assume the following data are representative of the results reported. The variable type indicates whether the restaurant is an Italian restaurant or a seafood/steakhouse. Price indicates the average amount paid per person for dinner and drinks, minus the tip. Score reflects diners' overall satisfaction, with higher values indicating greater overall satisfaction. A score of 80 can be interpreted as very satisfied.



Restaurant	Type	Price (\$)	Score
Bertucci's	Italian	16	77
Black Angus Steakhouse	Seafood/Steakhouse	24	79
Bonefish Grill	Seafood/Steakhouse	26	85

(continued)

Restaurant	Type	Price (\$)	Score
Bravo! Cucina Italiana	Italian	18	84
Buca di Beppo	Italian	17	81
Bugaboo Creek Steak House	Seafood/Steakhouse	18	77
Carrabba's Italian Grill	Italian	23	86
Charlie Brown's Steakhouse	Seafood/Steakhouse	17	75
Il Fornaio	Italian	28	83
Joe's Crab Shack	Seafood/Steakhouse	15	71
Johnny Carino's Italian	Italian	17	81
Lone Star Steakhouse & Saloon	Seafood/Steakhouse	17	76
LongHorn Steakhouse	Seafood/Steakhouse	19	81
Maggiano's Little Italy	Italian	22	83
McGrath's Fish House	Seafood/Steakhouse	16	81
Olive Garden	Italian	19	81
Outback Steakhouse	Seafood/Steakhouse	20	80
Red Lobster	Seafood/Steakhouse	18	78
Romano's Macaroni Grill	Italian	18	82
The Old Spaghetti Factory	Italian	12	79
Uno Chicago Grill	Italian	16	76

- Develop the estimated regression equation to show how overall customer satisfaction is related to the independent variable average meal price.
 - At the .05 level of significance, test whether the estimated regression equation developed in part (a) indicates a significant relationship between overall customer satisfaction and average meal price.
 - Develop a dummy variable that will account for the type of restaurant (Italian or seafood/steakhouse).
 - Develop the estimated regression equation to show how overall customer satisfaction is related to the average meal price and the type of restaurant.
 - Is type of restaurant a significant factor in overall customer satisfaction?
 - Predict the *Consumer Reports* customer satisfaction score for a seafood/steakhouse that has an average meal price of \$20. How much would the predicted score have changed for an Italian restaurant?
38. A 10-year study conducted by the American Heart Association provided data on how age, blood pressure, and smoking relate to the risk of strokes. Assume that the following data are from a portion of this study. Risk is interpreted as the probability (times 100) that the patient will have a stroke over the next 10-year period. For the smoking variable, define a dummy variable with 1 indicating a smoker and 0 indicating a nonsmoker.

WEB file
Stroke

Risk	Age	Pressure	Smoker
12	57	152	No
24	67	163	No
13	58	155	No
56	86	177	Yes
28	59	196	No
51	76	189	Yes
18	56	155	Yes
31	78	120	No
37	80	135	Yes
15	78	98	No
22	71	152	No
36	70	173	Yes

Risk	Age	Pressure	Smoker
15	67	135	Yes
48	77	209	Yes
15	60	199	No
36	82	119	Yes
8	66	166	No
34	80	125	Yes
3	62	117	No
37	59	207	Yes

- Develop an estimated regression equation that relates risk of a stroke to the person's age, blood pressure, and whether the person is a smoker.
- Is smoking a significant factor in the risk of a stroke? Explain. Use $\alpha = .05$.
- What is the probability of a stroke over the next 10 years for Art Speen, a 68-year-old smoker who has blood pressure of 175? What action might the physician recommend for this patient?

15.8

Residual Analysis

In Chapter 14 we pointed out that standardized residuals are frequently used in residual plots and in the identification of outliers. The general formula for the standardized residual for observation i follows.

STANDARDIZED RESIDUAL FOR OBSERVATION i

$$\frac{y_i - \hat{y}_i}{s_{y_i - \hat{y}_i}} \quad (15.23)$$

where

$$s_{y_i - \hat{y}_i} = \text{the standard deviation of residual } i$$

The general formula for the standard deviation of residual i is defined as follows.

STANDARD DEVIATION OF RESIDUAL i

$$s_{y_i - \hat{y}_i} = s\sqrt{1 - h_i} \quad (15.24)$$

where

s = standard error of the estimate

h_i = leverage of observation i

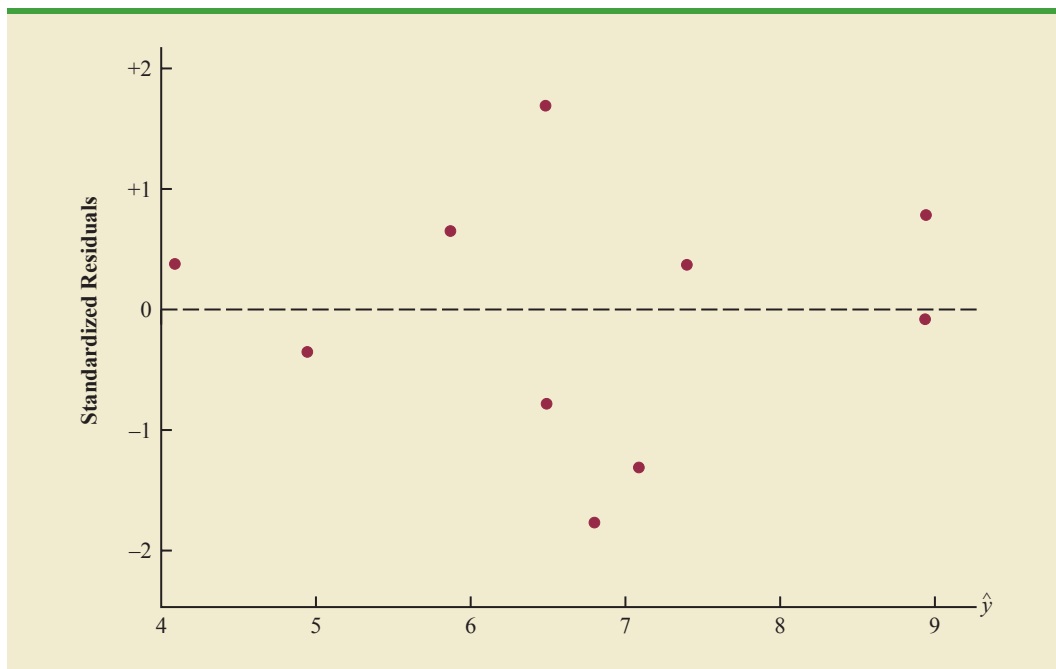
As we stated in Chapter 14, the **leverage** of an observation is determined by how far the values of the independent variables are from their means. The computation of h_i , $s_{y_i - \hat{y}_i}$ and hence the standardized residual for observation i in multiple regression analysis is too complex to be

TABLE 15.7 RESIDUALS AND STANDARDIZED RESIDUALS FOR THE BUTLER TRUCKING REGRESSION ANALYSIS

Miles Traveled (x_1)	Deliveries (x_2)	Travel Time (y)	Predicted Value (\hat{y})	Residual ($y - \hat{y}$)	Standardized Residual
100	4	9.3	8.93846	.361541	.78344
50	3	4.8	4.95830	-.158304	-.34962
100	4	8.9	8.93846	-.038460	-.08334
100	2	6.5	7.09161	-.591609	-1.30929
50	2	4.2	4.03488	.165121	.38167
80	2	6.2	5.86892	.331083	.65431
75	3	7.4	6.48667	.913331	1.68917
65	4	6.0	6.79875	-.798749	-1.77372
90	3	7.6	7.40369	.196311	.36703
90	2	6.1	6.48026	-.380263	-.77639

done by hand. However, the standardized residuals can be easily obtained as part of the output from statistical software packages. Table 15.7 lists the predicted values, the residuals, and the standardized residuals for the Butler Trucking example presented previously in this chapter; we obtained these values by using the Minitab statistical software package. The predicted values in the table are based on the estimated regression equation $\hat{y} = -.869 + .0611x_1 + .923x_2$.

The standardized residuals and the predicted values of y from Table 15.7 are used in Figure 15.10, the standardized residual plot for the Butler Trucking multiple regression example. This standardized residual plot does not indicate any unusual abnormalities. Also, all the standardized residuals are between -2 and $+2$; hence, we have no reason to question the assumption that the error term ϵ is normally distributed. We conclude that the model assumptions are reasonable.

FIGURE 15.10 STANDARDIZED RESIDUAL PLOT FOR BUTLER TRUCKING

A normal probability plot also can be used to determine whether the distribution of ϵ appears to be normal. The procedure and interpretation for a normal probability plot were discussed in Section 14.8. The same procedure is appropriate for multiple regression. Again, we would use a statistical software package to perform the computations and provide the normal probability plot.

Detecting Outliers

An **outlier** is an observation that is unusual in comparison with the other data; in other words, an outlier does not fit the pattern of the other data. In Chapter 14 we showed an example of an outlier and discussed how standardized residuals can be used to detect outliers. Minitab classifies an observation as an outlier if the value of its standardized residual is less than -2 or greater than $+2$. Applying this rule to the standardized residuals for the Butler Trucking example (see Table 15.7), we do not detect any outliers in the data set.

In general, the presence of one or more outliers in a data set tends to increase s , the standard error of the estimate, and hence increase $s_{y-\hat{y}_i}$, the standard deviation of residual i . Because $s_{y-\hat{y}_i}$ appears in the denominator of the formula for the standardized residual (15.23), the size of the standardized residual will decrease as s increases. As a result, even though a residual may be unusually large, the large denominator in expression (15.23) may cause the standardized residual rule to fail to identify the observation as being an outlier. We can circumvent this difficulty by using a form of the standardized residuals called **studentized deleted residuals**.

Studentized Deleted Residuals and Outliers

Suppose the i th observation is deleted from the data set and a new estimated regression equation is developed with the remaining $n - 1$ observations. Let $s_{(i)}$ denote the standard error of the estimate based on the data set with the i th observation deleted. If we compute the standard deviation of residual i using $s_{(i)}$ instead of s , and then compute the standardized residual for observation i using the revised $s_{y_i-\hat{y}_i}$ value, the resulting standardized residual is called a studentized deleted residual. If the i th observation is an outlier, $s_{(i)}$ will be less than s . The absolute value of the i th studentized deleted residual therefore will be larger than the absolute value of the standardized residual. In this sense, studentized deleted residuals may detect outliers that standardized residuals do not detect.

Many statistical software packages provide an option for obtaining studentized deleted residuals. Using Minitab, we obtained the studentized deleted residuals for the Butler Trucking example; the results are reported in Table 15.8. The t distribution can be used to

TABLE 15.8 STUDENTIZED DELETED RESIDUALS FOR BUTLER TRUCKING

Miles Traveled (x_1)	Deliveries (x_2)	Travel Time (y)	Standardized Residual	Studentized Deleted Residual
100	4	9.3	.78344	.75939
50	3	4.8	-.34962	-.32654
100	4	8.9	-.08334	-.07720
100	2	6.5	-1.30929	-1.39494
50	2	4.2	.38167	.35709
80	2	6.2	.65431	.62519
75	3	7.4	1.68917	2.03187
65	4	6.0	-1.77372	-2.21314
90	3	7.6	.36703	.34312
90	2	6.1	-.77639	-.75190

TABLE 15.9 LEVERAGE AND COOK'S DISTANCE MEASURES FOR BUTLER TRUCKING

Miles Traveled (x_1)	Deliveries (x_2)	Travel Time (y)	Leverage (h_i)	Cook's D (D_i)
100	4	9.3	.351704	.110994
50	3	4.8	.375863	.024536
100	4	8.9	.351704	.001256
100	2	6.5	.378451	.347923
50	2	4.2	.430220	.036663
80	2	6.2	.220557	.040381
75	3	7.4	.110009	.117562
65	4	6.0	.382657	.650029
90	3	7.6	.129098	.006656
90	2	6.1	.269737	.074217

determine whether the studentized deleted residuals indicate the presence of outliers. Recall that p denotes the number of independent variables and n denotes the number of observations. Hence, if we delete the i th observation, the number of observations in the reduced data set is $n - 1$; in this case the error sum of squares has $(n - 1) - p - 1$ degrees of freedom. For the Butler Trucking example with $n = 10$ and $p = 2$, the degrees of freedom for the error sum of squares with the i th observation deleted is $9 - 2 - 1 = 6$. At a .05 level of significance, the t distribution (Table 2 of Appendix B) shows that with six degrees of freedom, $t_{.025} = 2.447$. If the value of the i th studentized deleted residual is less than -2.447 or greater than $+2.447$, we can conclude that the i th observation is an outlier. The studentized deleted residuals in Table 15.8 do not exceed those limits; therefore, we conclude that outliers are not present in the data set.

Influential Observations

In Section 14.9 we discussed how the leverage of an observation can be used to identify observations for which the value of the independent variable may have a strong influence on the regression results. As we indicated in the discussion of standardized residuals, the leverage of an observation, denoted h_i , measures how far the values of the independent variables are from their mean values. The leverage values are easily obtained as part of the output from statistical software packages. Minitab computes the leverage values and uses the rule of thumb $h_i > 3(p + 1)/n$ to identify **influential observations**. For the Butler Trucking example with $p = 2$ independent variables and $n = 10$ observations, the critical value for leverage is $3(2 + 1)/10 = .9$. The leverage values for the Butler Trucking example obtained by using Minitab are reported in Table 15.9. Because h_i does not exceed .9, we do not detect influential observations in the data set.

TABLE 15.10

DATA SET
ILLUSTRATING
POTENTIAL
PROBLEM USING
THE LEVERAGE
CRITERION

x_i	y_i	Leverage h_i
1	18	.204170
1	21	.204170
2	22	.164205
3	21	.138141
4	23	.125977
4	24	.125977
5	26	.127715
15	39	.909644

Using Cook's Distance Measure to Identify Influential Observations

A problem that can arise in using leverage to identify influential observations is that an observation can be identified as having high leverage and not necessarily be influential in terms of the resulting estimated regression equation. For example, Table 15.10 is a data set consisting of eight observations and their corresponding leverage values (obtained by using Minitab). Because the leverage for the eighth observation is $.91 > .75$ (the critical leverage value), this observation is identified as influential. Before reaching any final conclusions, however, let us consider the situation from a different perspective.

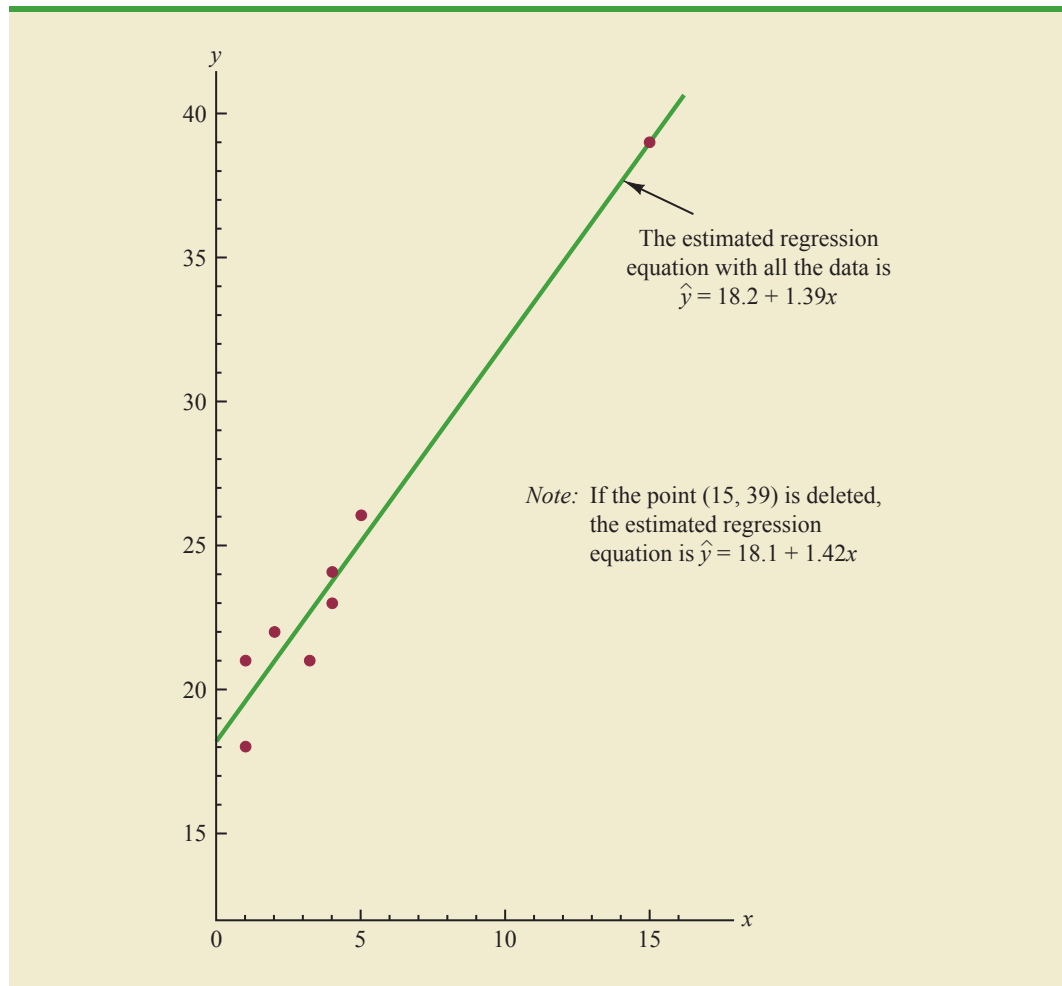
FIGURE 15.11 SCATTER DIAGRAM FOR THE DATA SET IN TABLE 15.10

Figure 15.11 shows the scatter diagram corresponding to the data set in Table 15.10. We used Minitab to develop the following estimated regression equation for these data.

$$\hat{y} = 18.2 + 1.39x$$

The straight line in Figure 15.11 is the graph of this equation. Now, let us delete the observation $x = 15, y = 39$ from the data set and fit a new estimated regression equation to the remaining seven observations; the new estimated regression equation is

$$\hat{y} = 18.1 + 1.42x$$

We note that the y -intercept and slope of the new estimated regression equation are very close to the values obtained using all the data. Although the leverage criterion identified the eighth observation as influential, this observation clearly had little influence on the results obtained. Thus, in some situations using only leverage to identify influential observations can lead to wrong conclusions.

Cook's distance measure uses both the leverage of observation i , h_i , and the residual for observation i , $(y_i - \hat{y}_i)$, to determine whether the observation is influential.

COOK'S DISTANCE MEASURE

$$D_i = \frac{(y_i - \hat{y}_i)^2}{(p + 1)s^2} \left[\frac{h_i}{(1 - h_i)^2} \right] \quad (15.25)$$

where

- D_i = Cook's distance measure for observation i
- $y_i - \hat{y}_i$ = the residual for observation i
- h_i = the leverage for observation i
- p = the number of independent variables
- s = the standard error of the estimate

The value of Cook's distance measure will be large and indicate an influential observation if the residual or the leverage is large. As a rule of thumb, values of $D_i > 1$ indicate that the i th observation is influential and should be studied further. The last column of Table 15.9 provides Cook's distance measure for the Butler Trucking problem as given by Minitab. Observation 8 with $D_i = .650029$ has the most influence. However, applying the rule $D_i > 1$, we should not be concerned about the presence of influential observations in the Butler Trucking data set.

NOTES AND COMMENTS

- The procedures for identifying outliers and influential observations provide warnings about the potential effects some observations may have on the regression results. Each outlier and influential observation warrants careful examination. If data errors are found, the errors can be corrected and the regression analysis repeated. In general, outliers and influential observations should not be removed from the data set unless clear evidence shows that they are not based on elements of the population being studied and should not have been included in the original data set.
- To determine whether the value of Cook's distance measure D_i is large enough to conclude that the i th observation is influential, we can also compare the value of D_i to the 50th percentile of an F distribution (denoted $F_{.50}$) with $p + 1$ numerator degrees of freedom and $n - p - 1$ denominator degrees of freedom. F tables corresponding to a .50 level of significance must be available to carry out the test. The rule of thumb we provided ($D_i > 1$) is based on the fact that the table value is close to one for a wide variety of cases.

Exercises

Methods

39. Data for two variables, x and y , follow.

x_i	1	2	3	4	5
y_i	3	7	5	11	14

- Develop the estimated regression equation for these data.
- Plot the standardized residuals versus \hat{y} . Do there appear to be any outliers in these data? Explain.
- Compute the studentized deleted residuals for these data. At the .05 level of significance, can any of these observations be classified as an outlier? Explain.

SELF test

40. Data for two variables, x and y , follow.

x_i	22	24	26	28	40
y_i	12	21	31	35	70

- Develop the estimated regression equation for these data.
- Compute the studentized deleted residuals for these data. At the .05 level of significance, can any of these observations be classified as an outlier? Explain.
- Compute the leverage values for these data. Do there appear to be any influential observations in these data? Explain.
- Compute Cook's distance measure for these data. Are any observations influential? Explain.

Applications

SELF test

41. Exercise 5 gave the following data on weekly gross revenue, television advertising, and newspaper advertising for Showtime Movie Theaters.

WEB file

Showtime

Weekly Gross Revenue (\$1000s)	Television Advertising (\$1000s)	Newspaper Advertising (\$1000s)
96	5.0	1.5
90	2.0	2.0
95	4.0	1.5
92	2.5	2.5
95	3.0	3.3
94	3.5	2.3
94	2.5	4.2
94	3.0	2.5

- Find an estimated regression equation relating weekly gross revenue to television and newspaper advertising.
 - Plot the standardized residuals against \hat{y} . Does the residual plot support the assumptions about ϵ ? Explain.
 - Check for any outliers in these data. What are your conclusions?
 - Are there any influential observations? Explain.
42. The following data show the curb weight, horsepower, and $\frac{1}{4}$ -mile speed for 16 popular sports and GT cars. Suppose that the price of each sports and GT car is also available. The complete data set is as follows:

WEB file

Auto2

Sports & GT Car	Price (\$1000s)	Curb Weight (lb.)	Horsepower	Speed at $\frac{1}{4}$ Mile (mph)
Acura Integra Type R	25.035	2577	195	90.7
Acura NSX-T	93.758	3066	290	108.0
BMW Z3 2.8	40.900	2844	189	93.2
Chevrolet Camaro Z28	24.865	3439	305	103.2
Chevrolet Corvette Convertible	50.144	3246	345	102.1
Dodge Viper RT/10	69.742	3319	450	116.2
Ford Mustang GT	23.200	3227	225	91.7
Honda Prelude Type SH	26.382	3042	195	89.7
Mercedes-Benz CLK320	44.988	3240	215	93.0
Mercedes-Benz SLK230	42.762	3025	185	92.3
Mitsubishi 3000GT VR-4	47.518	3737	320	99.0

(continued)

Sports & GT Car	Price (\$1000s)	Curb Weight (lb.)	Horsepower	Speed at ¼ Mile (mph)
Nissan 240SX SE	25.066	2862	155	84.6
Pontiac Firebird Trans Am	27.770	3455	305	103.2
Porsche Boxster	45.560	2822	201	93.2
Toyota Supra Turbo	40.989	3505	320	105.0
Volvo C70	41.120	3285	236	97.0

- Find the estimated regression equation that uses price and horsepower to predict ¼-mile speed.
 - Plot the standardized residuals against \hat{y} . Does the residual plot support the assumption about ϵ ? Explain.
 - Check for any outliers. What are your conclusions?
 - Are there any influential observations? Explain.
43. The Ladies Professional Golfers Association (LPGA) maintains statistics on performance and earnings for members of the LPGA Tour. Year-end performance statistics for the 30 players who had the highest total earnings in LPGA Tour events for 2005 appear in the file named LPGA (LPGA website, 2006). Earnings (\$1000s) is the total earnings in thousands of dollars; Scoring Avg. is the average score for all events; Greens in Reg. is the percentage of time a player is able to hit the green in regulation; and Putting Avg. is the average number of putts taken on greens hit in regulation. A green is considered hit in regulation if any part of the ball is touching the putting surface and the difference between the value of par for the hole and the number of strokes taken to hit the green is at least 2.
- Develop an estimated regression equation that can be used to predict the average score for all events given the percentage of time a player is able to hit the green in regulation and the average number of putts taken on greens hit in regulation.
 - Plot the standardized residuals against \hat{y} . Does the residual plot support the assumption about ϵ ? Explain.
 - Check for any outliers. What are your conclusions?
 - Are there any influential observations? Explain.



15.9 Logistic Regression

In many regression applications the dependent variable may only assume two discrete values. For instance, a bank might like to develop an estimated regression equation for predicting whether a person will be approved for a credit card. The dependent variable can be coded as $y = 1$ if the bank approves the request for a credit card and $y = 0$ if the bank rejects the request for a credit card. Using logistic regression we can estimate the probability that the bank will approve the request for a credit card given a particular set of values for the chosen independent variables.

Let us consider an application of logistic regression involving a direct mail promotion being used by Simmons Stores. Simmons owns and operates a national chain of women's apparel stores. Five thousand copies of an expensive four-color sales catalog have been printed, and each catalog includes a coupon that provides a \$50 discount on purchases of \$200 or more. The catalogs are expensive and Simmons would like to send them to only those customers who have the highest probability of using the coupon.

Management thinks that annual spending at Simmons Stores and whether a customer has a Simmons credit card are two variables that might be helpful in predicting whether a customer who receives the catalog will use the coupon. Simmons conducted a pilot

study using a random sample of 50 Simmons credit card customers and 50 other customers who do not have a Simmons credit card. Simmons sent the catalog to each of the 100 customers selected. At the end of a test period, Simmons noted whether the customer used the coupon. The sample data for the first 10 catalog recipients are shown in Table 15.11. The amount each customer spent last year at Simmons is shown in thousands of dollars and the credit card information has been coded as 1 if the customer has a Simmons credit card and 0 if not. In the Coupon column, a 1 is recorded if the sampled customer used the coupon and 0 if not.

We might think of building a multiple regression model using the data in Table 15.11 to help Simmons estimate whether a catalog recipient will use the coupon. We would use Annual Spending (\$1000) and Simmons Card as independent variables and Coupon as the dependent variable. Because the dependent variable may only assume the values of 0 or 1, however, the ordinary multiple regression model is not applicable. This example shows the type of situation for which logistic regression was developed. Let us see how logistic regression can be used to help Simmons estimate which type of customer is most likely to take advantage of their promotion.

Logistic Regression Equation

In many ways logistic regression is like ordinary regression. It requires a dependent variable, y , and one or more independent variables. In multiple regression analysis, the mean or expected value of y is referred to as the multiple regression equation.

$$E(y) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_px_p \quad (15.26)$$

In logistic regression, statistical theory as well as practice has shown that the relationship between $E(y)$ and x_1, x_2, \dots, x_p is better described by the following nonlinear equation.

LOGISTIC REGRESSION EQUATION

$$E(y) = \frac{e^{\beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_px_p}}{1 + e^{\beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_px_p}} \quad (15.27)$$

If the two values of the dependent variable y are coded as 0 or 1, the value of $E(y)$ in equation (15.27) provides the *probability* that $y = 1$ given a particular set of values for the

TABLE 15.11 PARTIAL SAMPLE DATA FOR THE SIMMONS STORES EXAMPLE

Customer	Annual Spending (\$1000)	Simmons Card	Coupon
1	2.291	1	0
2	3.215	1	0
3	2.135	1	0
4	3.924	0	0
5	2.528	1	0
6	2.473	0	1
7	2.384	0	0
8	7.076	0	0
9	1.182	1	1
10	3.345	0	0



independent variables x_1, x_2, \dots, x_p . Because of the interpretation of $E(y)$ as a probability, the **logistic regression equation** is often written as follows.

INTERPRETATION OF $E(y)$ AS A PROBABILITY IN LOGISTIC REGRESSION

$$E(y) = P(y = 1 | x_1, x_2, \dots, x_p) \quad (15.28)$$

To provide a better understanding of the characteristics of the logistic regression equation, suppose the model involves only one independent variable x and the values of the model parameters are $\beta_0 = -7$ and $\beta_1 = 3$. The logistic regression equation corresponding to these parameter values is

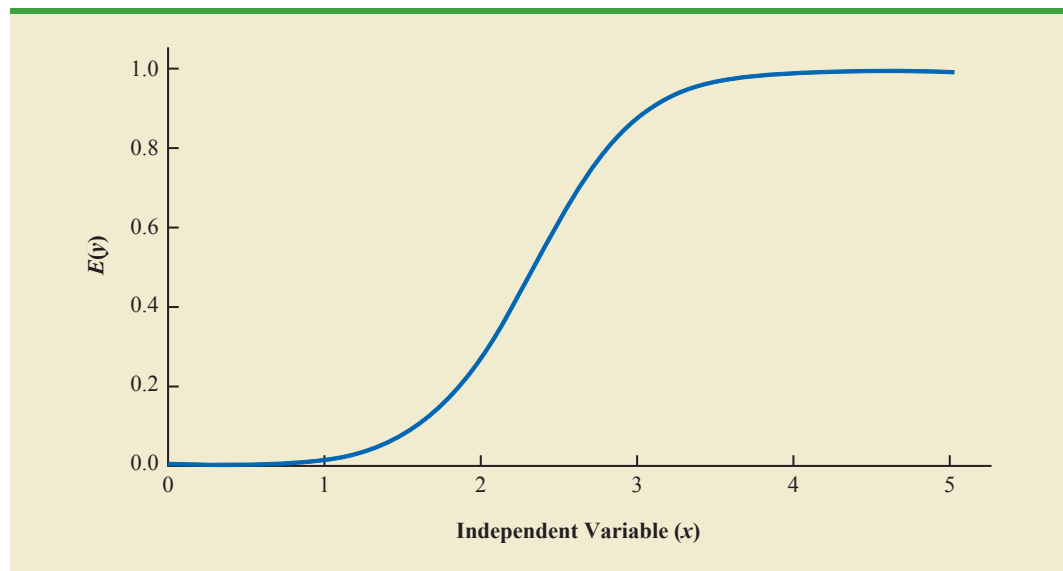
$$E(y) = P(y = 1 | x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} = \frac{e^{-7+3x}}{1 + e^{-7+3x}} \quad (15.29)$$

Figure 15.12 shows a graph of equation (15.29). Note that the graph is S-shaped. The value of $E(y)$ ranges from 0 to 1, with the value of $E(y)$ gradually approaching 1 as the value of x becomes larger and the value of $E(y)$ approaching 0 as the value of x becomes smaller. Note also that the values of $E(y)$, representing probability, increase fairly rapidly as x increases from 2 to 3. The fact that the values of $E(y)$ range from 0 to 1 and that the curve is S-shaped makes equation (15.29) ideally suited to model the probability the dependent variable is equal to 1.

Estimating the Logistic Regression Equation

In simple linear and multiple regression the least squares method is used to compute b_0, b_1, \dots, b_p as estimates of the model parameters $(\beta_0, \beta_1, \dots, \beta_p)$. The nonlinear form of the logistic regression equation makes the method of computing estimates more complex and beyond the scope of this text. We will use computer software to provide the estimates. The **estimated logistic regression equation** is

FIGURE 15.12 LOGISTIC REGRESSION EQUATION FOR $\beta_0 = -7$ AND $\beta_1 = 3$



ESTIMATED LOGISTIC REGRESSION EQUATION

$$\hat{y} = \text{estimate of } P(y = 1 | x_1, x_2, \dots, x_p) = \frac{e^{b_0 + b_1x_1 + b_2x_2 + \dots + b_px_p}}{1 + e^{b_0 + b_1x_1 + b_2x_2 + \dots + b_px_p}} \quad (15.30)$$

Here, \hat{y} provides an estimate of the probability that $y = 1$ given a particular set of values for the independent variables.

Let us now return to the Simmons Stores example. The variables in the study are defined as follows:

$$y = \begin{cases} 0 & \text{if the customer did not use the coupon} \\ 1 & \text{if the customer used the coupon} \end{cases}$$

$$x_1 = \text{annual spending at Simmons Stores (\$1000s)}$$

$$x_2 = \begin{cases} 0 & \text{if the customer does not have a Simmons credit card} \\ 1 & \text{if the customer has a Simmons credit card} \end{cases}$$

Thus, we choose a logistic regression equation with two independent variables.

$$E(y) = \frac{e^{\beta_0 + \beta_1x_1 + \beta_2x_2}}{1 + e^{\beta_0 + \beta_1x_1 + \beta_2x_2}} \quad (15.31)$$

In Appendix 15.3 we show how Minitab is used to generate the output in Figure 15.13.

Using the sample data (see Table 15.11), Minitab's binary logistic regression procedure was used to compute estimates of the model parameters β_0 , β_1 , and β_2 . A portion of the output obtained is shown in Figure 15.13. We see that $b_0 = -2.14637$, $b_1 = .341643$, and $b_2 = 1.09873$. Thus, the estimated logistic regression equation is

$$\hat{y} = \frac{e^{b_0 + b_1x_1 + b_2x_2}}{1 + e^{b_0 + b_1x_1 + b_2x_2}} = \frac{e^{-2.14637 + .341643x_1 + 1.09873x_2}}{1 + e^{-2.14637 + .341643x_1 + 1.09873x_2}} \quad (15.32)$$

We can now use equation (15.32) to estimate the probability of using the coupon for a particular type of customer. For example, to estimate the probability of using the coupon for customers who spend \$2000 annually and do not have a Simmons credit card, we substitute $x_1 = 2$ and $x_2 = 0$ into equation (15.32).

FIGURE 15.13 PARTIAL LOGISTIC REGRESSION OUTPUT FOR THE SIMMONS STORES EXAMPLE

Logistic Regression Table					Odds	95%	CI
Predictor	Coef	SE Coef	Z	p	Ratio	Lower	Upper
Constant	-2.14637	0.577245	-3.72	0.000			
Spending	0.341643	0.128672	2.66	0.008	1.41	1.09	1.81
Card	1.09873	0.444696	2.47	0.013	3.00	1.25	7.17
Log-Likelihood = -60.487							
Test that all slopes are zero: G = 13.628, DF = 2, P-Value = 0.001							

In the Minitab output, x_1 = Spending and x_2 = Card.

$$\hat{y} = \frac{e^{-2.14637+.341643(2)+1.09873(0)}}{1 + e^{-2.14637+.341643(2)+1.09873(0)}} = \frac{e^{-1.4631}}{1 + e^{-1.4631}} = \frac{.2315}{1.2315} = .1880$$

Thus, an estimate of the probability of using the coupon for this particular group of customers is approximately 0.19. Similarly, to estimate the probability of using the coupon for customers who spent \$2000 last year and have a Simmons credit card, we substitute $x_1 = 2$ and $x_2 = 1$ into equation (15.32).

$$\hat{y} = \frac{e^{-2.14637+.341643(2)+1.09873(1)}}{1 + e^{-2.14637+.341643(2)+1.09873(1)}} = \frac{e^{-.3644}}{1 + e^{-.3644}} = \frac{.6946}{1.6946} = .4099$$

Thus, for this group of customers, the probability of using the coupon is approximately 0.41. It appears that the probability of using the coupon is much higher for customers with a Simmons credit card. Before reaching any conclusions, however, we need to assess the statistical significance of our model.

Testing for Significance

Testing for significance in logistic regression is similar to testing for significance in multiple regression. First we conduct a test for overall significance. For the Simmons Stores example, the hypotheses for the test of overall significance follow:

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

$$H_a: \text{One or both of the parameters is not equal to zero}$$

The test for overall significance is based upon the value of a G test statistic. If the null hypothesis is true, the sampling distribution of G follows a chi-square distribution with degrees of freedom equal to the number of independent variables in the model. Although the computation of G is beyond the scope of the book, the value of G and its corresponding p -value are provided as part of Minitab's binary logistic regression output. Referring to the last line in Figure 15.13, we see that the value of G is 13.628, its degrees of freedom are 2, and its p -value is .001. Thus, at any level of significance $\alpha \geq .001$, we would reject the null hypothesis and conclude that the overall model is significant.

If the G test shows an overall significance, a z test can be used to determine whether each of the individual independent variables is making a significant contribution to the overall model. For the independent variables x_i , the hypotheses are

$$H_0: \beta_i = 0$$

$$H_a: \beta_i \neq 0$$

If the null hypothesis is true, the value of the estimated coefficient divided by its standard error follows a standard normal probability distribution. The column labeled Z in the Minitab output contains the values of $z_i = b_i/s_{b_i}$ for each of the estimated coefficients and the column labeled p contains the corresponding p -values. Suppose we use $\alpha = .05$ to test for the significance of the independent variables in the Simmons model. For the independent variable x_1 the z value is 2.66 and the corresponding p -value is .008. Thus, at the .05 level of significance we can reject $H_0: \beta_1 = 0$. In a similar fashion we can also reject $H_0: \beta_2 = 0$ because the p -value corresponding to $z = 2.47$ is .013. Hence, at the .05 level of significance, both independent variables are statistically significant.

Managerial Use

We described how to develop the estimated logistic regression equation and how to test it for significance. Let us now use it to make a decision recommendation concerning the Simmons Stores catalog promotion. For Simmons Stores, we already computed $P(y = 1|x_1 = 2, x_2 = 1) = .4099$ and $P(y = 1|x_1 = 2, x_2 = 0) = .1880$. These probabilities indicate that for customers with annual spending of \$2000 the presence of a Simmons credit card increases the probability of using the coupon. In Table 15.12 we show estimated probabilities for values of annual spending ranging from \$1000 to \$7000 for both customers who have a Simmons credit card and customers who do not have a Simmons credit card. How can Simmons use this information to better target customers for the new promotion? Suppose Simmons wants to send the promotional catalog only to customers who have a 0.40 or higher probability of using the coupon. Using the estimated probabilities in Table 15.12, Simmons promotion strategy would be:

Customers who have a Simmons credit card: Send the catalog to every customer who spent \$2000 or more last year.

Customers who do not have a Simmons credit card: Send the catalog to every customer who spent \$6000 or more last year.

Looking at the estimated probabilities further, we see that the probability of using the coupon for customers who do not have a Simmons credit card but spend \$5000 annually is .3922. Thus, Simmons may want to consider revising this strategy by including those customers who do not have a credit card, as long as they spent \$5000 or more last year.

Interpreting the Logistic Regression Equation

Interpreting a regression equation involves relating the independent variables to the business question that the equation was developed to answer. With logistic regression, it is difficult to interpret the relation between the independent variables and the probability that $y = 1$ directly because the logistic regression equation is nonlinear. However, statisticians have shown that the relationship can be interpreted indirectly using a concept called the odds ratio.

The **odds in favor of an event occurring** is defined as the probability the event will occur divided by the probability the event will not occur. In logistic regression the event of interest is always $y = 1$. Given a particular set of values for the independent variables, the odds in favor of $y = 1$ can be calculated as follows:

$$\text{odds} = \frac{P(y = 1|x_1, x_2, \dots, x_p)}{P(y = 0|x_1, x_2, \dots, x_p)} = \frac{P(y = 1|x_1, x_2, \dots, x_p)}{1 - P(y = 1|x_1, x_2, \dots, x_p)} \quad (15.33)$$

The **odds ratio** measures the impact on the odds of a one-unit increase in only one of the independent variables. The odds ratio is the odds that $y = 1$ given that one of the

TABLE 15.12 ESTIMATED PROBABILITIES FOR SIMMONS STORES

		Annual Spending						
		\$1000	\$2000	\$3000	\$4000	\$5000	\$6000	\$7000
Credit Card	Yes	.3305	.4099	.4943	.5791	.6594	.7315	.7931
	No	.1413	.1880	.2457	.3144	.3922	.4759	.5610

independent variables has been increased by one unit (odds₁) divided by the odds that $y = 1$ given no change in the values for the independent variables (odds₀).

ODDS RATIO

$$\text{Odds Ratio} = \frac{\text{odds}_1}{\text{odds}_0} \quad (15.34)$$

For example, suppose we want to compare the odds of using the coupon for customers who spend \$2000 annually and have a Simmons credit card ($x_1 = 2$ and $x_2 = 1$) to the odds of using the coupon for customers who spend \$2000 annually and do not have a Simmons credit card ($x_1 = 2$ and $x_2 = 0$). We are interested in interpreting the effect of a one-unit increase in the independent variable x_2 . In this case

$$\text{odds}_1 = \frac{P(y = 1|x_1 = 2, x_2 = 1)}{1 - P(y = 1|x_1 = 2, x_2 = 1)}$$

and

$$\text{odds}_0 = \frac{P(y = 1|x_1 = 2, x_2 = 0)}{1 - P(y = 1|x_1 = 2, x_2 = 0)}$$

Previously we showed that an estimate of the probability that $y = 1$ given $x_1 = 2$ and $x_2 = 1$ is .4099, and an estimate of the probability that $y = 1$ given $x_1 = 2$ and $x_2 = 0$ is .1880. Thus,

$$\text{estimate of odds}_1 = \frac{.4099}{1 - .4099} = .6946$$

and

$$\text{estimate of odds}_0 = \frac{.1880}{1 - .1880} = .2315$$

The estimated odds ratio is

$$\text{Estimated odds ratio} = \frac{.6946}{.2315} = 3.00$$

Thus, we can conclude that the estimated odds in favor of using the coupon for customers who spent \$2000 last year and have a Simmons credit card are 3 times greater than the estimated odds in favor of using the coupon for customers who spent \$2000 last year and do not have a Simmons credit card.

The odds ratio for each independent variable is computed while holding all the other independent variables constant. But it does not matter what constant values are used for the other independent variables. For instance, if we computed the odds ratio for the Simmons credit card variable (x_2) using \$3000, instead of \$2000, as the value for the annual spending variable (x_1), we would still obtain the same value for the estimated odds ratio (3.00). Thus, we can conclude that the estimated odds of using the coupon for customers who have a Simmons credit card are 3 times greater than the estimated odds of using the coupon for customers who do not have a Simmons credit card.

The odds ratio is standard output for logistic regression software packages. Refer to the Minitab output in Figure 15.13. The column with the heading Odds Ratio contains the

estimated odds ratios for each of the independent variables. The estimated odds ratio for x_1 is 1.41 and the estimated odds ratio for x_2 is 3.00. We already showed how to interpret the estimated odds ratio for the binary independent variable x_2 . Let us now consider the interpretation of the estimated odds ratio for the continuous independent variable x_1 .

The value of 1.41 in the Odds Ratio column of the Minitab output tells us that the estimated odds in favor of using the coupon for customers who spent \$3000 last year is 1.41 times greater than the estimated odds in favor of using the coupon for customers who spent \$2000 last year. Moreover, this interpretation is true for any one-unit change in x_1 . For instance, the estimated odds in favor of using the coupon for someone who spent \$5000 last year is 1.41 times greater than the odds in favor of using the coupon for a customer who spent \$4000 last year. But suppose we are interested in the change in the odds for an increase of more than one unit for an independent variable. Note that x_1 can range from 1 to 7. The odds ratio given by the Minitab output does not answer this question. To answer this question we must explore the relationship between the odds ratio and the regression coefficients.

A unique relationship exists between the odds ratio for a variable and its corresponding regression coefficient. For each independent variable in a logistic regression equation it can be shown that

$$\text{Odds ratio} = e^{\beta_i}$$

To illustrate this relationship, consider the independent variable x_1 in the Simmons example. The estimated odds ratio for x_1 is

$$\text{Estimated odds ratio} = e^{b_1} = e^{.341643} = 1.41$$

Similarly, the estimated odds ratio for x_2 is

$$\text{Estimated odds ratio} = e^{b_2} = e^{1.09873} = 3.00$$

This relationship between the odds ratio and the coefficients of the independent variables makes it easy to compute estimated odds ratios once we develop estimates of the model parameters. Moreover, it also provides us with the ability to investigate changes in the odds ratio of more than or less than one unit for a continuous independent variable.

The odds ratio for an independent variable represents the change in the odds for a one-unit change in the independent variable holding all the other independent variables constant. Suppose that we want to consider the effect of a change of more than one unit, say c units. For instance, suppose in the Simmons example that we want to compare the odds of using the coupon for customers who spend \$5000 annually ($x_1 = 5$) to the odds of using the coupon for customers who spend \$2000 annually ($x_1 = 2$). In this case $c = 5 - 2 = 3$ and the corresponding estimated odds ratio is

$$e^{cb_1} = e^{3(.341643)} = e^{1.0249} = 2.79$$

This result indicates that the estimated odds of using the coupon for customers who spend \$5000 annually is 2.79 times greater than the estimated odds of using the coupon for customers who spend \$2000 annually. In other words, the estimated odds ratio for an increase of \$3000 in annual spending is 2.79.

In general, the odds ratio enables us to compare the odds for two different events. If the value of the odds ratio is 1, the odds for both events are the same. Thus, if the independent variable we are considering (such as Simmons credit card status) has a positive impact on the probability of the event occurring, the corresponding odds ratio will be greater than 1. Most logistic regression software packages provide a confidence interval for the odds ratio. The Minitab output in Figure 15.13 provides a 95% confidence interval for each of the odds

ratios. For example, the point estimate of the odds ratio for x_1 is 1.41 and the 95% confidence interval is 1.09 to 1.81. Because the confidence interval does not contain the value of 1, we can conclude that x_1 has a significant effect on the estimated odds ratio. Similarly, the 95% confidence interval for the odds ratio for x_2 is 1.25 to 7.17. Because this interval does not contain the value of 1, we can also conclude that x_2 has a significant effect on the odds ratio.

Logit Transformation

An interesting relationship can be observed between the odds in favor of $y = 1$ and the exponent for e in the logistic regression equation. It can be shown that

$$\ln(\text{odds}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

This equation shows that the natural logarithm of the odds in favor of $y = 1$ is a linear function of the independent variables. This linear function is called the **logit**. We will use the notation $g(x_1, x_2, \dots, x_p)$ to denote the logit.

LOGIT

$$g(x_1, x_2, \dots, x_p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \quad (15.35)$$

Substituting $g(x_1, x_2, \dots, x_p)$ for $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$ in equation (15.27), we can write the logistic regression equation as

$$E(y) = \frac{e^{g(x_1, x_2, \dots, x_p)}}{1 + e^{g(x_1, x_2, \dots, x_p)}} \quad (15.36)$$

Once we estimate the parameters in the logistic regression equation, we can compute an estimate of the logit. Using $\hat{g}(x_1, x_2, \dots, x_p)$ to denote the **estimated logit**, we obtain

ESTIMATED LOGIT

$$\hat{g}(x_1, x_2, \dots, x_p) = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p \quad (15.37)$$

Thus, in terms of the estimated logit, the estimated regression equation is

$$\hat{y} = \frac{e^{b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p}}{1 + e^{b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p}} = \frac{e^{\hat{g}(x_1, x_2, \dots, x_p)}}{1 + e^{\hat{g}(x_1, x_2, \dots, x_p)}} \quad (15.38)$$

For the Simmons Stores example, the estimated logit is

$$\hat{g}(x_1, x_2) = -2.14637 + .341643x_1 + 1.09873x_2$$

and the estimated regression equation is

$$\hat{y} = \frac{e^{\hat{g}(x_1, x_2)}}{1 + e^{\hat{g}(x_1, x_2)}} = \frac{e^{-2.14637 + .341643x_1 + 1.09873x_2}}{1 + e^{-2.14637 + .341643x_1 + 1.09873x_2}}$$

Thus, because of the unique relationship between the estimated logit and the estimated logistic regression equation, we can compute the estimated probabilities for Simmons Stores by dividing $e^{\hat{g}(x_1, x_2)}$ by $1 + e^{\hat{g}(x_1, x_2)}$.

NOTES AND COMMENTS

1. Because of the unique relationship between the estimated coefficients in the model and the corresponding odds ratios, the overall test for significance based upon the G statistic is also a test of overall significance for the odds ratios. In addition, the z test for the individual significance of a model parameter also provides a statistical test of significance for the corresponding odds ratio.
2. In simple and multiple regression, the coefficient of determination is used to measure the goodness of fit. In logistic regression, no single measure provides a similar interpretation. A discussion of goodness of fit is beyond the scope of our introductory treatment of logistic regression.

Exercises

Applications



44. Refer to the Simmons Stores example introduced in this section. The dependent variable is coded as $y = 1$ if the customer used the coupon and 0 if not. Suppose that the only information available to help predict whether the customer will use the coupon is the customer's credit card status, coded as $x = 1$ if the customer has a Simmons credit card and $x = 0$ if not.
 - a. Write the logistic regression equation relating x to y .
 - b. What is the interpretation of $E(y)$ when $x = 0$?
 - c. For the Simmons data in Table 15.11, use Minitab to compute the estimated logit.
 - d. Use the estimated logit computed in part (c) to estimate the probability of using the coupon for customers who do not have a Simmons credit card and to estimate the probability of using the coupon for customers who have a Simmons credit card.
 - e. What is the estimated odds ratio? What is its interpretation?
45. In Table 15.12 we provided estimates of the probability of using the coupon in the Simmons Stores catalog promotion. A different value is obtained for each combination of values for the independent variables.
 - a. Compute the odds in favor of using the coupon for a customer with annual spending of \$4000 who does not have a Simmons credit card ($x_1 = 4, x_2 = 0$).
 - b. Use the information in Table 15.12 and part (a) to compute the odds ratio for the Simmons credit card variable $x_2 = 0$, holding annual spending constant at $x_1 = 4$.
 - c. In the text, the odds ratio for the credit card variable was computed using the information in the \$2000 column of Table 15.12. Did you get the same value for the odds ratio in part (b)?
46. Community Bank would like to increase the number of customers who use payroll direct deposit. Management is considering a new sales campaign that will require each branch manager to call each customer who does not currently use payroll direct deposit. As an incentive to sign up for payroll direct deposit, each customer contacted will be offered free checking for two years. Because of the time and cost associated with the new campaign, management would like to focus their efforts on customers who have the highest probability of signing up for payroll direct deposit. Management believes that the average monthly balance in a customer's checking account may be a useful predictor of whether the customer will sign up for direct payroll deposit. To investigate the relationship between these two variables, Community Bank tried the new campaign using a sample of 50 checking account customers who do not currently use payroll direct deposit. The sample data show the average monthly checking account balance (in hundreds of dollars) and whether the customer contacted signed up for payroll direct deposit (coded 1 if the customer signed up for payroll direct deposit and 0 if not). The data are contained in the data set named Bank; a portion of the data follows.

WEB file
Bank

Customer	x = Monthly Balance	y = Direct Deposit
1	1.22	0
2	1.56	0
3	2.10	0
4	2.25	0
5	2.89	0
6	3.55	0
7	3.56	0
8	3.65	1
⋮	⋮	⋮
48	18.45	1
49	24.98	0
50	26.05	1

- Write the logistic regression equation relating x to y .
 - For the Community Bank data, use Minitab to compute the estimated logistic regression equation.
 - Conduct a test of significance using the G test statistic. Use $\alpha = .05$.
 - Estimate the probability that customers with an average monthly balance of \$1000 will sign up for direct payroll deposit.
 - Suppose Community Bank only wants to contact customers who have a .50 or higher probability of signing up for direct payroll deposit. What is the average monthly balance required to achieve this level of probability?
 - What is the estimated odds ratio? What is its interpretation?
47. Over the past few years the percentage of students who leave Lakeland College at the end of the first year has increased. Last year Lakeland started a voluntary one-week orientation program to help first-year students adjust to campus life. If Lakeland is able to show that the orientation program has a positive effect on retention, they will consider making the program a requirement for all first-year students. Lakeland's administration also suspects that students with lower GPAs have a higher probability of leaving Lakeland at the end of the first year. In order to investigate the relation of these variables to retention, Lakeland selected a random sample of 100 students from last year's entering class. The data are contained in the data set named Lakeland; a portion of the data follows.

WEB file
Lakeland

Student	GPA	Program	Return
1	3.78	1	1
2	2.38	0	1
3	1.30	0	0
4	2.19	1	0
5	3.22	1	1
6	2.68	1	1
⋮	⋮	⋮	⋮
98	2.57	1	1
99	1.70	1	1
100	3.85	1	1

The dependent variable was coded as $y = 1$ if the student returned to Lakeland for the sophomore year and $y = 0$ if not. The two independent variables are:

$$x_1 = \text{GPA at the end of the first semester}$$

$$x_2 = \begin{cases} 0 & \text{if the student did not attend the orientation program} \\ 1 & \text{if the student attended the orientation program} \end{cases}$$

- Write the logistic regression equation relating x_1 and x_2 to y .
- What is the interpretation of $E(y)$ when $x_2 = 0$?
- Use both independent variables and Minitab to compute the estimated logit.
- Conduct a test for overall significance using $\alpha = .05$.
- Use $\alpha = .05$ to determine whether each of the independent variables is significant.
- Use the estimated logit computed in part (c) to estimate the probability that students with a 2.5 grade point average who did not attend the orientation program will return to Lakeland for their sophomore year. What is the estimated probability for students with a 2.5 grade point average who attended the orientation program?
- What is the estimated odds ratio for the orientation program? Interpret it.
- Would you recommend making the orientation program a required activity? Why or why not?



48. The Tire Rack maintains an independent consumer survey to help drivers help each other by sharing their long-term tire experiences. The data contained in the file named TireRatings show survey results for 68 all-season tires (Tire Rack website, March 21, 2012). Performance traits are rated using the following 10-point scale.

Superior		Excellent		Good		Fair		Unacceptable	
10	9	8	7	6	5	4	3	2	1

The values for the variable labeled Wet are the average of the ratings for each tire's wet traction performance and the values for the variable labeled Noise are the average of the ratings for the noise level generated by each tire. Respondents were also asked whether they would buy the tire again using the following 10-point scale:

Definitely		Probably		Possibly		Probably Not		Definitely Not	
10	9	8	7	6	5	4	3	2	1

The values for the variable labeled Buy Again are the average of the buy-again responses. For the purposes of this exercise, we created the following binary dependent variable:

$$\text{Purchase} = \begin{cases} 1 & \text{if the value of the Buy-Again variable is 7 or greater} \\ 0 & \text{if the value of the Buy-Again variable is less than 7} \end{cases}$$

Thus, if Purchase = 1, the respondent would probably or definitely buy the tire again.

- Write the logistic regression equation relating $x_1 =$ Wet performance rating and $x_2 =$ Noise performance rating to $y =$ Purchase.
- Use Minitab to compute the estimated logit.
- Use the estimated logit to compute an estimate of the probability that a customer will probably or definitely purchase a particular tire again with a Wet performance rating of 8 and a Noise performance rating of 8.
- Suppose that the Wet and Noise performance ratings were 7. How does that affect the probability that a customer will probably or definitely purchase a particular tire again with these performance ratings?
- If you were the CEO of a tire company, what do the results for parts (c) and (d) tell you?