CHAPTER 15

Multiple Regression

CONTENTS

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 an estimate of the probability that a customer belongs to a particular customer group. The independent variables $x_1, x_2, x_3, \ldots, x_p$ 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 predicting 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.

dunnhumby uses logistic regression to predict customer shopping behavior. © Ariel Skelley/Blend Images/Jupiter Images.

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.

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

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 estimates 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 $x_1, x_2, \ldots, x_p$ 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

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon \tag{15.1}$$

In the multiple regression model, $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ 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 $x_1, x_2, \ldots, x_p$ (the $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$ 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 $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$. The equation that describes how the mean value of y is related to $x_1, x_2, \ldots, x_p$ is called the **multiple regression equation**.

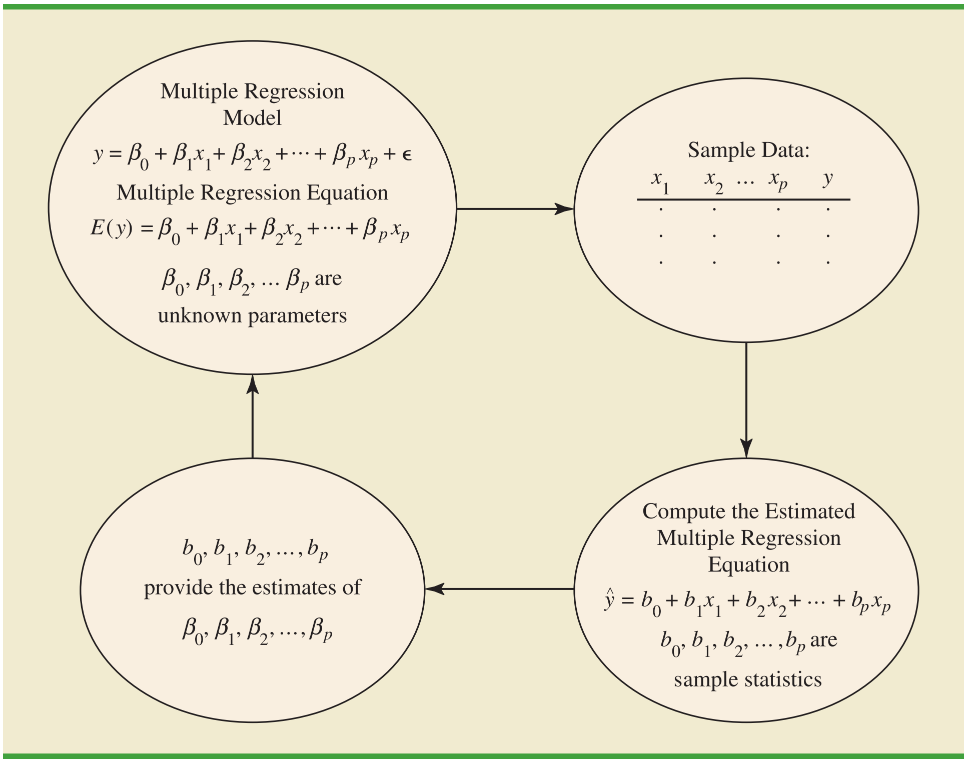
MULTIPLE REGRESSION EQUATION

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \tag{15.2}$$

Estimated Multiple Regression Equation

If the values of $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ were known, equation (15.2) could be used to compute the mean value of y at given values of $x_1, x_2, \ldots, x_p$. 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 $b_0, b_1, b_2, \ldots, b_p$ that are used as the point

FIGURE 15.1 THE ESTIMATION PROCESS FOR MULTIPLE REGRESSION


Multiple Regression Model
$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$
Multiple Regression Equation
$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$
$\beta_0, \beta_1, \beta_2, \ldots \beta_p$ are unknown parameters
Sample Data:
x_1 x_2 … x_p y
Compute the Estimated Multiple Regression Equation
$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$
$b_0, b_1, b_2, \ldots, b_p$ are sample statistics
$b_0, b_1, b_2, \ldots, b_p$
provide the estimates of
$\beta_0, \beta_1, \beta_2, \ldots, \beta_p$

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, $b_1, b_2, \ldots, b_p$ denoting the sample statistics used to estimate the parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$.

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

ESTIMATED MULTIPLE REGRESSION EQUATION

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p \qquad \textbf{(15.3)}$$

where

$$b_0, b_1, b_2, \ldots, b_p \text{ are the estimates of } \beta_0, \beta_1, \beta_2, \ldots, \beta_p$$
$$\hat{y} = \text{estimated value of the dependent variable}$$

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

$$\min \Sigma(y_i - \hat{y}_i)^2 \qquad \textbf{(15.4)}$$

where

y_i = observed value of the dependent variable for the ith observation
$\hat{y}_i$ = estimated value of the dependent variable for the ith observation

The estimated 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 $b_0, b_1, b_2, \ldots, b_p$ that make the sum of squared residuals [the deviations between the observed values of the dependent variable (y_i) and the estimated 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 $b_0, b_1, b_2, \ldots, b_p$ 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 estimate the total daily travel time for their drivers.

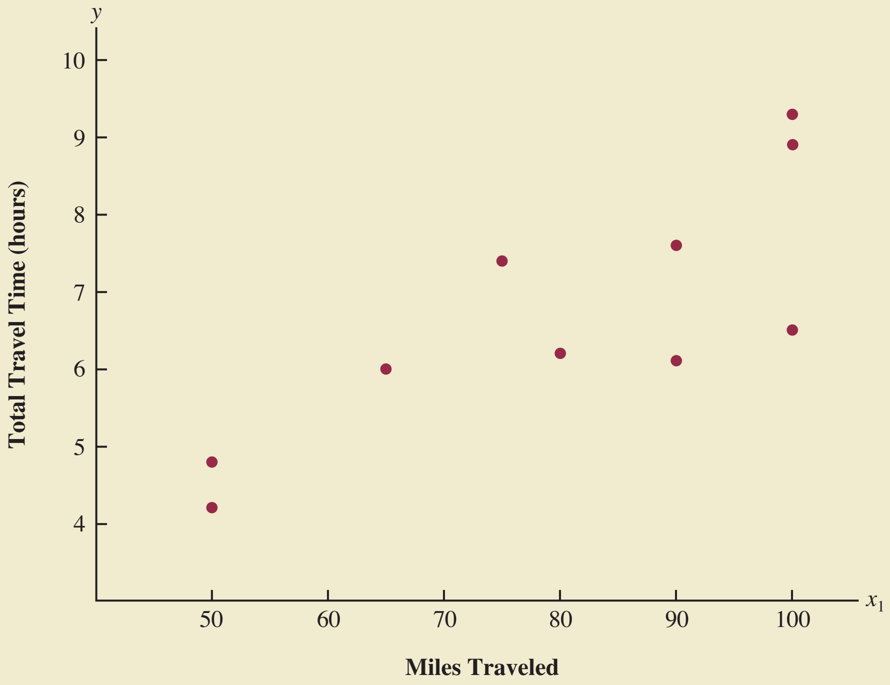
Initially the managers believed that the total daily travel time would be closely related to the number of miles traveled in making the daily deliveries. A simple random sample of 10 driving assignments provided the data shown in Table 15.1 and the scatter diagram shown in Figure 15.2. After reviewing 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

WEB file
Butler

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.

$$\hat{y} = b_0 + b_1x_1 \tag{15.5}$$

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-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

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

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

WEB file

Butler

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

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

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

   a. Interpret b_1 and b_2 in this estimated regression equation.
   b. Estimate y when $x_1 = 180$ and $x_2 = 310$.

SELF test

2. Consider the following data for a dependent variable y and two independent variables, x_1 and x_2.

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

(continued)

x_1	x_2	y
36	12	117
59	13	142
76	16	211

a. Develop an estimated regression equation relating y to x_1. Estimate y if $x_1 = 45$.
b. Develop an estimated regression equation relating y to x_2. Estimate y if $x_2 = 15$.
c. Develop an estimated regression equation relating y to x_1 and x_2. Estimate y if $x_1 = 45$ and $x_2 = 15$.

3. In a regression analysis involving 30 observations, the following estimated regression equation was obtained.

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

a. Interpret b_1, b_2, b_3, and b_4 in this estimated regression equation.
b. Estimate 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

$$x_1 = \text{inventory investment (\$1000s)}$$
$$x_2 = \text{advertising expenditures (\$1000s)}$$
$$y = \text{sales (\$1000s)}$$

a. Estimate sales resulting from a $15,000 investment in inventory and an advertising budget of $10,000.
b. Interpret b_1 and b_2 in this estimated regression equation.

SELF test

5. The owner of Showtime Movie Theaters, Inc., would like to estimate weekly gross revenue as a function of advertising expenditures. Historical data for a sample of eight weeks follow.

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

a. Develop an estimated regression equation with the amount of television advertising as the independent variable.
b. Develop an estimated regression equation with both television advertising and newspaper advertising as the independent variables.
c. 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.

d. What is the estimate of the weekly gross revenue for a week when $3500 is spent on television advertising and $1800 is spent on newspaper advertising?

6. In baseball, a team's success is often thought to be a function of the team's hitting and pitching performance. One measure of hitting performance is the number of home runs the team hits, and one measure of pitching performance is the earned run average for the team's pitching staff. It is generally believed that teams that hit more home runs and have a lower earned run average will win a higher percentage of the games played. The following data show the proportion of games won, the number of team home runs (HR), and the earned run average (ERA) for the 16 teams in the National League for the 2003 Major League Baseball season (USA Today website, January 7, 2004).

WEB file

MLB

	Proportion				Proportion		
Team	**Won**	**HR**	**ERA**	**Team**	**Won**	**HR**	**ERA**
Arizona	.519	152	3.857	Milwaukee	.420	196	5.058
Atlanta	.623	235	4.106	Montreal	.512	144	4.027
Chicago	.543	172	3.842	New York	.410	124	4.517
Cincinnati	.426	182	5.127	Philadelphia	.531	166	4.072
Colorado	.457	198	5.269	Pittsburgh	.463	163	4.664
Florida	.562	157	4.059	San Diego	.395	128	4.904
Houston	.537	191	3.880	San Francisco	.621	180	3.734
Los Angeles	.525	124	3.162	St. Louis	.525	196	4.642

a. Determine the estimated regression equation that could be used to predict the proportion of games won given the number of team home runs.

b. Determine the estimated regression equation that could be used to predict the proportion of games won given the earned run average for the team's pitching staff.

c. Determine the estimated regression equation that could be used to predict the proportion of games won given the number of team home runs and the earned run average for the team's pitching staff.

d. For the 2003 season San Diego won only 39.5% of the games they played, the lowest in the National League. To improve next year's record, the team tried to acquire new players who would increase the number of team home runs to 180 and decrease the earned run average for the team's pitching staff to 4.0. Use the estimated regression equation developed in part (c) to estimate the percentage of games San Diego will win if they have 180 team home runs and have an earned run average of 4.0.

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).

WEB file

Laptop

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. Would you expect more reliable and better performing cars to cost more? *Consumer Reports* provided reliability ratings, overall road-test scores, and prices for affordable family sedans, midpriced family sedans, and large sedans (*Consumer Reports*, February 2008). A portion of the data follows. Reliability was rated on a 5-point scale from poor (1) to excellent (5). The road-test score was rated on a 100-point scale, with higher values indicating better performance. The complete data set is contained in the file named Sedans.

WEB file

Sedans

Make and Model	Road-Test Score	Reliability	Price ($)
Nissan Altima 2.5 S	85	4	22705
Honda Accord LX-P	79	4	22795
Kia Optima EX (4-cyl.)	78	4	22795
Toyota Camry LE	77	4	21080
Hyundai Sonata SE	76	3	22995
.	.	.	.
.	.	.	.
.	.	.	.
Chrysler 300 Touring	60	2	30255
Dodge Charger SXT	58	4	28860

a. Develop an estimated regression equation that can be used to predict the price of the car given the reliability rating. Test for significance using $\alpha = .05$.

b. Consider the addition of the independent variable overall road-test score. Develop the estimated regression equation that can be used to predict the price of the car given the road-test score and the reliability rating.

c. Estimate the price for a car with a road-test score of 80 and a reliability rating of 4.

9. Waterskiing and wakeboarding are two popular water-sports. Finding a model that best suits your intended needs, whether it is waterskiing, wakeboading, 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	330	45.3
Correct Craft Air Nautique 210	91	330	47.3
Correct Craft Air Nautique SV-211	93	375	46.9
Correct Craft Ski Nautique 206 Limited	91	330	46.7
Gekko GTR 22	96	375	50.1
Gekko GTS 20	83	375	52.2
Malibu Response LXi	93.5	340	47.2
Malibu Sunsettter LXi	98	400	46
Malibu Sunsetter 21 XTi	98	340	44

Make and Model	Beam	HP	TopSpeed
Malibu Sunscape 21 LSV	98	400	47.5
Malibu Wakesetter 21 XTi	98	340	44.9
Malibu Wakesetter VLX	98	400	47.3
Malibu vRide	93.5	340	44.5
Malibu Ride XTi	93.5	320	44.5
Mastercraft ProStar 209	96	350	42.5
Mastercraft X-1	90	310	45.8
Mastercraft X-2	94	310	42.8
Mastercraft X-9	96	350	43.2
MB Sports 190 Plus	92	330	45.3
Svfara SVONE	91	330	47.7

a. Using these data, develop an estimated regression equation relating the top speed with the boat's beam and horsepower rating.

b. The Svfara 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 estimate the top speed for the Svfara SV609.

10. The National Basketball Association (NBA) records a variety of statistics for each team. Four of these statistics are the proportion of games won (PCT), the proportion of field goals made by the team (FG%), the proportion of three-point shots made by the team's opponent (Opp 3 Pt%), and the number of turnovers committed by the team's opponent (Opp TO). The following data show the values of these statistics for the 29 teams in the NBA for a portion of the 2004 season (NBA website, January 3, 2004).

WEB file

NBA

			Opp 3	Opp				Opp 3	Opp
Team	PCT	FG%	Pt%	TO	Team	PCT	FG%	Pt%	TO
Atlanta	0.265	0.435	0.346	13.206	Minnesota	0.677	0.473	0.348	13.839
Boston	0.471	0.449	0.369	16.176	New Jersey	0.563	0.435	0.338	17.063
Chicago	0.313	0.417	0.372	15.031	New Orleans	0.636	0.421	0.330	16.909
Cleveland	0.303	0.438	0.345	12.515	New York	0.412	0.442	0.330	13.588
Dallas	0.581	0.439	0.332	15.000	Orlando	0.242	0.417	0.360	14.242
Denver	0.606	0.431	0.366	17.818	Philadelphia	0.438	0.428	0.364	16.938
Detroit	0.606	0.423	0.262	15.788	Phoenix	0.364	0.438	0.326	16.515
Golden State	0.452	0.445	0.384	14.290	Portland	0.484	0.447	0.367	12.548
Houston	0.548	0.426	0.324	13.161	Sacramento	0.724	0.466	0.327	15.207
Indiana	0.706	0.428	0.317	15.647	San Antonio	0.688	0.429	0.293	15.344
L.A. Clippers	0.464	0.424	0.326	14.357	Seattle	0.533	0.436	0.350	16.767
L.A. Lakers	0.724	0.465	0.323	16.000	Toronto	0.516	0.424	0.314	14.129
Memphis	0.485	0.432	0.358	17.848	Utah	0.531	0.456	0.368	15.469
Miami	0.424	0.410	0.369	14.970	Washington	0.300	0.411	0.341	16.133
Milwaukee	0.500	0.438	0.349	14.750					

a. Determine the estimated regression equation that can be used to predict the proportion of games won given the proportion of field goals made by the team.

b. Provide an interpretation for the slope of the estimated regression equation developed in part (a).

c. Determine the estimated regression equation that can be used to predict the proportion of games won given the proportion of field goals made by the team, the proportion of three-point shots made by the team's opponent, and the number of turnovers committed by the team's opponent.

d. Discuss the practical implications of the estimated regression equation developed in part (c).

e. Estimate the proportion of games won for a team with the following values for the three independent variables: FG% = .45, Opp 3 Pt% = .34, and Opp TO = 17.

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

$$\text{SST} = \text{SSR} + \text{SSE} \tag{15.7}$$

where

$$\begin{aligned}\text{SST} &= \text{total sum of squares} = \Sigma(y_i - \bar{y})^2\\ \text{SSR} &= \text{sum of squares due to regression} = \Sigma(\hat{y}_i - \bar{y})^2\\ \text{SSE} &= \text{sum of squares due to error} = \Sigma(y_i - \hat{y}_i)^2\end{aligned}$$

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, $r^2 = \text{SSR/SST}$, 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

$$R^2 = \frac{\text{SSR}}{\text{SST}} \tag{15.8}$$

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-sq = 90.4%.

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.*

Figure 15.3 shows that the R-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.

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

$$R_{\mathrm{a}}^2 = 1 - (1 - R^2)\frac{n - 1}{n - p - 1} \tag{15.9}$$

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

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

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.

a. Find SSE.
b. Compute R^2.
c. Compute R_{a}^2.
d. Comment on the goodness of fit.

SELF test

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.

a. Compute R^2.
b. Compute R_{a}^2.
c. 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.

a. Compute R^2.
b. Compute R_a^2.
c. 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.

a. For the estimated regression equation given, compute R^2.
b. Compute R_a^2.
c. 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

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

WEB file Showtime

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

a. Compute and interpret R^2 and R_a^2.
b. 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 MLB

16. In exercise 6, data were given on the proportion of games won, the number of team home runs, and the earned run average for the team's pitching staff for the 16 teams in the National League for the 2003 Major League Baseball season (USA Today website, January 7, 2004).

a. Did the estimated regression equation that uses only the number of home runs as the independent variable to predict the proportion of games won provide a good fit? Explain.
b. Discuss the benefits of using both the number of home runs and the earned run average to predict the proportion 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.

a. Compute and interpret and R^2 and R_a^2.
b. Does the estimated regression equation provide a good fit to the data? Explain.

WEB file NBA

18. Refer to exercise 10, where data were reported on a variety of statistics for the 29 teams in the National Basketball Association for a portion of the 2004 season (NBA website, January 3, 2004).

a. In part (c) of exercise 10, an estimated regression equation was developed relating the proportion of games won given the percentage of field goals made by the team, the proportion of three-point shots made by the team's opponent, and the number of turnovers committed by the team's opponent. What are the values of R^2 and R_a^2?
b. Does the estimated regression equation 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

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon \qquad \textbf{(15.10)}$$

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 $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$

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 $x_1, x_2, \ldots, x_p$, the expected, or average, value of y is given by

   $$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \qquad \textbf{(15.11)}$$

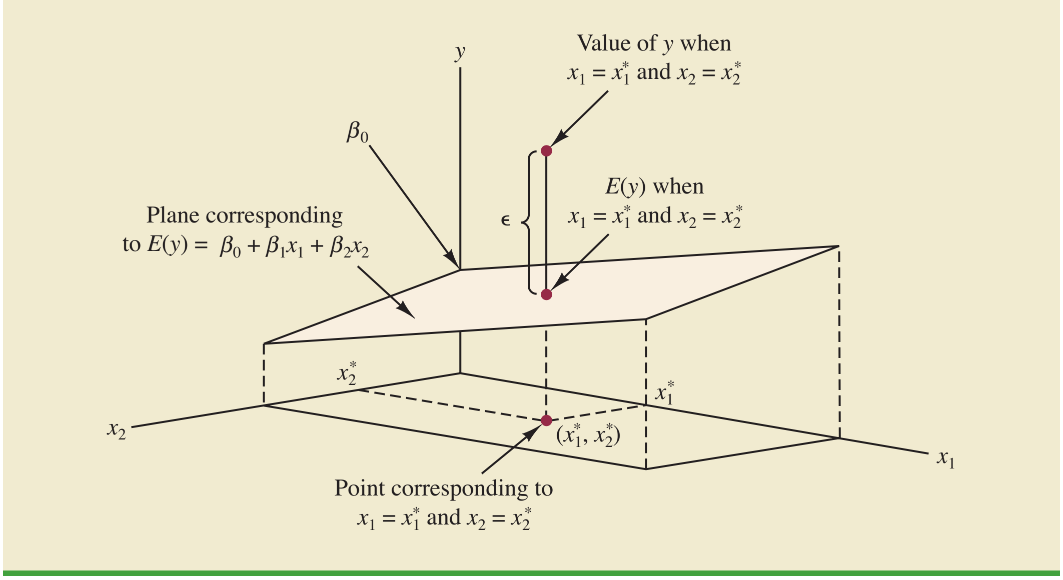
   Equation (15.11) is the multiple regression equation we introduced in Section 15.1. In this equation, $E(y)$ represents the average of all possible values of y that might occur for the given values of $x_1, x_2, \ldots, x_p$.
2. The variance of ϵ is denoted by σ^2 and is the same for all values of the independent variables $x_1, x_2, \ldots, x_p$.
   Implication: The variance of y about the regression line equals σ^2 and is the same for all values of $x_1, x_2, \ldots, x_p$.
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 $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$.
   Implication: Because $\beta_0, \beta_1, \ldots, \beta_p$ are constants for the given values of $x_1, x_2, \ldots, x_p$, 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.

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

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


y
Value of y when
$x_1 = x_1^*$ and $x_2 = x_2^*$
β_0
$E(y)$ when
$x_1 = x_1^*$ and $x_2 = x_2^*$
Plane corresponding
to $E(y) = \beta_0 + \beta_1x_1 + \beta_2x_2$
ϵ
x_2^*
x_1^*
x_2
(x_1^*, x_2^*)
x_1
Point corresponding to
$x_1 = x_1^*$ and $x_2 = x_2^*$

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 $x_1, x_2, \ldots, x_p$ 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/$(n - p - 1)$.

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

and

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

As discussed in Chapter 14, MSE provides an unbiased estimate of σ^2, the variance of the error term ϵ. If H_0: $\beta_1 = \beta_2 = \cdots = \beta_p = 0$ 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

$$H_0\colon \beta_1 = \beta_2 = \cdots = \beta_p = 0$$
$$H_a\colon \text{One or more of the parameters is not equal to zero}$$

TEST STATISTIC

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

REJECTION RULE

p-value approach:	Reject H_0 if p-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.

$$H_0\colon \beta_1 = \beta_2 = 0$$
$$H_a\colon \beta_1 \text{ and/or } \beta_2 \text{ is not equal to zero}$$

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 $s = \sqrt{\text{MSE}} = \sqrt{.328} = .573$. 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	$\text{MSR} = \dfrac{\text{SSR}}{p}$	$F = \dfrac{\text{MSR}}{\text{MSE}}$
Error	SSE	$n - p - 1$	$\text{MSE} = \dfrac{\text{SSE}}{n - p - 1}$	
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

$$H_0\colon \beta_i = 0$$
$$H_a\colon \beta_i \neq 0$$

TEST STATISTIC

$$t = \frac{b_i}{s_{b_i}} \tag{15.15}$$

REJECTION RULE

p-value approach:	Reject H_0 if p-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 $r_{x_1x_2}$ to determine the extent to which the variables are related. Doing so yields $r_{x_1x_2} = .16$. 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.

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.

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.

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.

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.

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, b_2 might actually be $+10$ and β_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 $s_{b_2} = .0567$.

a. Compute MSR and MSE.
b. Compute *F* and perform the appropriate *F* test. Use $\alpha = .05$.
c. Perform a *t* test for the significance of β_1. Use $\alpha = .05$.
d. 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 $s_{b_2} = .9484$.

a. Test for a significant relationship among x_1, x_2, and y. Use $\alpha = .05$.
b. Is β_1 significant? Use $\alpha = .05$.
c. 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.

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

a. Give an interpretation of the coefficient of x_1 in both models.
b. 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.

a. Compute SSE, MSE, and MSR.

b. Use an F test and a .05 level of significance to determine whether there is a relationship among the variables.

SELF test

23. Refer to exercise 5.

a. 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 $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$, where

$$x_1 = \text{television advertising (\$1000s)}$$
$$x_2 = \text{newspaper advertising (\$1000s)}$$

b. Use $\alpha = .05$ to test the significance of β_1. Should x_1 be dropped from the model?

c. Use $\alpha = .05$ to test the significance of β_2. Should x_2 be dropped from the model?

24. *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	0.70
Arkansas	11.1	72	0.80
Boston College	3.4	80	0.53
⋮	⋮	⋮	⋮
Washington	5.0	83	0.89
West Virginia	4.9	67	0.70
Wichita State	3.1	75	0.41
Wisconsin	12.0	66	0.70

a. 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.

b. Use the F test to determine the overall significance of the relationship. What is your conclusion at the .05 level of significance?

c. Use the t test to determine the significance of each independent variable. What is your conclusion at the .05 level of significance?

25. *Barron's* conducts an annual review of online brokers, including both brokers who can be accessed via a Web browser, as well as direct-access brokers who connect customers directly with the broker's network server. Each broker's offerings and performance are evaluated in six areas, using a point value of 0–5 in each category. The results are weighted to obtain an overall score, and a final star rating, ranging from zero to five stars, is assigned to each broker. Trade

execution, ease of use, and range of offerings are three of the areas evaluated. A point value of 5 in the trade execution area means the order entry and execution process flowed easily from one step to the next. A value of 5 in the ease of use area means that the site was easy to use and can be tailored to show what the user wants to see. A value of 5 in the range offerings area means that all the investment transactions can be executed online. The following data show the point values for trade execution, ease of use, range of offerings, and the star rating for a sample of 10 of the online brokers that *Barron's* evaluated (*Barron's,* March 10, 2003).

WEB file

Brokers

Broker	Trade Execution	Use	Range	Rating
Wall St. Access	3.7	4.5	4.8	4.0
E*TRADE (Power)	3.4	3.0	4.2	3.5
E*TRADE (Standard)	2.5	4.0	4.0	3.5
Preferred Trade	4.8	3.7	3.4	3.5
my Track	4.0	3.5	3.2	3.5
TD Waterhouse	3.0	3.0	4.6	3.5
Brown & Co.	2.7	2.5	3.3	3.0
Brokerage America	1.7	3.5	3.1	3.0
Merrill Lynch Direct	2.2	2.7	3.0	2.5
Strong Funds	1.4	3.6	2.5	2.0

a. Determine the estimated regression equation that can be used to predict the star rating given the point values for execution, ease of use, and range of offerings.

b. Use the *F* test to determine the overall significance of the relationship. What is the conclusion at the .05 level of significance?

c. Use the *t* test to determine the significance of each independent variable. What is your conclusion at the .05 level of significance?

d. Remove any independent variable that is not significant from the estimated regression equation. What is your recommended estimated regression equation? Compare the R^2 with the value of R^2 from part (a). Discuss the differences.

WEB file

NBA

26. In exercise 10 an estimated regression equation was developed relating the proportion of games won given the proportion of field goals made by the team, the proportion of three-point shots made by the team's opponent, and the number of turnovers committed by the team's opponent.

a. Use the *F* test to determine the overall significance of the relationship. What is your conclusion at the .05 level of significance?

b. 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 $x_1, x_2, \ldots, x_p$ 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	Value of	Confidence Interval		Prediction Interval	
x_1	x_2	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, $x_2, \ldots, x_p$ 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 $x_2 = 310$.

SELF test

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 $x_2 = 15$.

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

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

a. 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 ($x_2 = 1.8$)?
b. Provide a 95% confidence interval for the mean revenue of all weeks with the expenditures listed in part (a).
c. Provide a 95% prediction interval for next week's revenue, assuming that the advertising expenditures will be allocated as in part (a).

WEB file
Boats

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.
a. 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.
b. The Svfara SV609 has a beam of 85 inches and an engine with a 330 horsepower rating. Develop a 95% confidence interval for the mean top speed for the Svfara SV609.

31. The Buyer's Guide section of the Web site for *Car and Driver* magazine provides reviews and road tests for cars, trucks, SUVs, and vans. The average ratings of overall quality, vehicle styling, braking, handling, fuel economy, interior comfort, acceleration, dependability, fit and finish, transmission, and ride are summarized for each vehicle using a scale ranging from 1 (worst) to 10 (best). A portion of the data for 14 Sports/GT cars is shown here (Car and Driver website, January 7, 2004).

WEB file
SportsCar

Sports/GT	Overall	Handling	Dependability	Fit and Finish
Acura 3.2CL	7.80	7.83	8.17	7.67
Acura RSX	9.02	9.46	9.35	8.97
Audi TT	9.00	9.58	8.74	9.38
BMW 3-Series/M3	8.39	9.52	8.39	8.55
Chevrolet Corvette	8.82	9.64	8.54	7.87
Ford Mustang	8.34	8.85	8.70	7.34
Honda Civic Si	8.92	9.31	9.50	7.93
Infiniti G35	8.70	9.34	8.96	8.07
Mazda RX-8	8.58	9.79	8.96	8.12
Mini Cooper	8.76	10.00	8.69	8.33
Mitsubishi Eclipse	8.17	8.95	8.25	7.36
Nissan 350Z	8.07	9.35	7.56	8.21
Porsche 911	9.55	9.91	8.86	9.55
Toyota Celica	8.77	9.29	9.04	7.97

a. Develop an estimated regression equation using handling, dependability, and fit and finish to predict overall quality.
b. Another Sports/GT car rated by *Car and Driver* is the Honda Accord. The ratings for handling, dependability, and fit and finish for the Honda Accord were 8.28, 9.06, and 8.07, respectively. Estimate the overall rating for this car.
c. Provide a 95% confidence interval for overall quality for all sports and GT cars with the characteristics listed in part (b).

d. Provide a 95% prediction interval for overall quality for the Honda Accord described in part (b).
e. The overall rating reported by *Car and Driver* for the Honda Accord was 8.65. How does this rating compare to the estimates you developed in parts (b) and (d)?

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

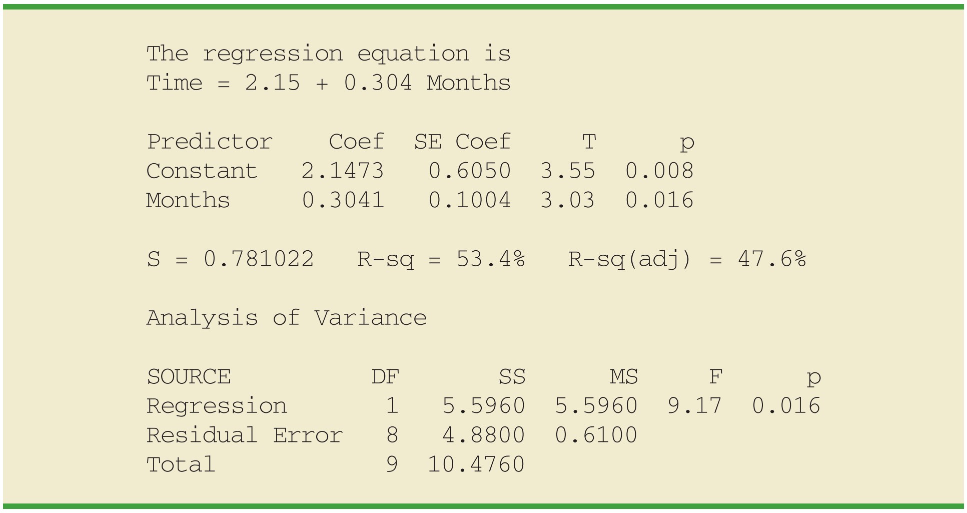
$$\hat{y} = 2.15 + .304x_1 \tag{15.16}$$

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-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

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

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)

WEB file

Johnson

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

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-sq = 85.9% and R-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 estimating the repair time necessary for the various service calls.

Interpreting the Parameters

The multiple regression equation for the Johnson Filtration example is

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \tag{15.18}$$

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

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

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

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

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)

$$\hat{y} = .93 + .388x_1 \tag{15.21}$$

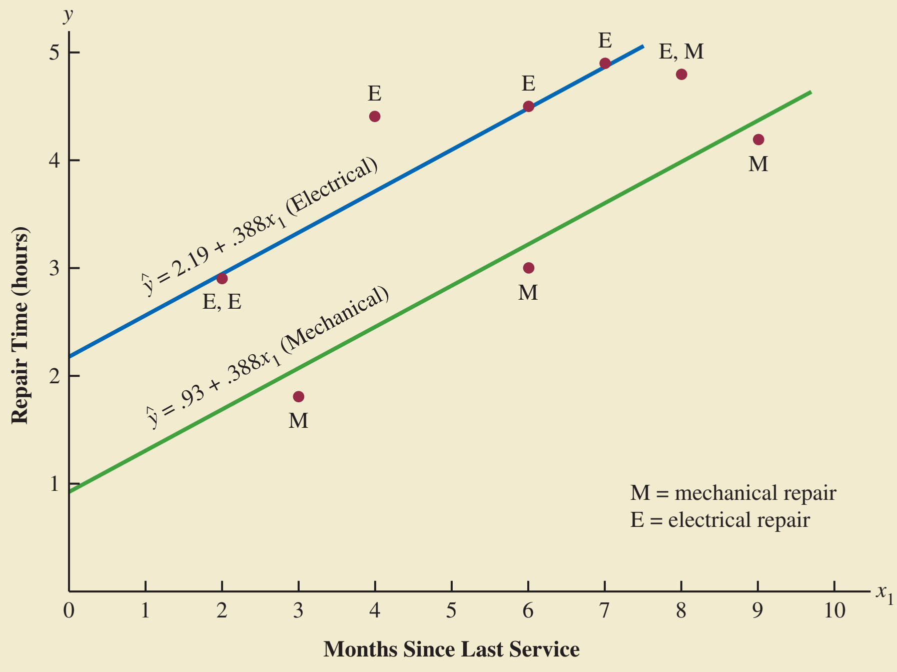
and when $x_2 = 1$ (electrical repair)

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

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.

$$\begin{aligned} E(y \mid \text{region A}) &= \beta_0 + \beta_1(0) + \beta_2(0) = \beta_0 \\ E(y \mid \text{region B}) &= \beta_0 + \beta_1(1) + \beta_2(0) = \beta_0 + \beta_1 \\ E(y \mid \text{region C}) &= \beta_0 + \beta_1(0) + \beta_2(1) = \beta_0 + \beta_2 \end{aligned}$$

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.

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

Exercises

Methods

SELF test

32. Consider a regression study involving a dependent variable y, a categorical independent variable x_1, and a categorical 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

$$\begin{aligned} x_1 &= \text{number of competitors within one mile} \\ x_2 &= \text{population within one mile (1000s)} \\ x_3 &= \begin{cases} 1 \text{ if drive-up window present} \\ 0 \text{ otherwise} \end{cases} \\ y &= \text{sales (\$1000s)} \end{aligned}$$

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.

WEB file

Repair

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.

WEB file

RestaurantRatings

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

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

a. Develop the estimated regression equation to show how overall customer satisfaction is related to the independent variable average meal price.

b. 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.

c. Develop a dummy variable that will account for the type of restaurant (Italian or seafood/steakhouse).

d. Develop the estimated regression equation to show how overall customer satisfaction is related to the average meal price and the type of restaurant.

e. Is type of restaurant a significant factor in overall customer satisfaction?

f. Estimate the *Consumer Reports* customer satisfaction score for a seafood/steakhouse that has an average meal price of $20. How much would the estimated 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

(*continued*)

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

a. 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.

b. Is smoking a significant factor in the risk of a stroke? Explain. Use $\alpha = .05$.

c. 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

$$\frac{y_i - \hat{y}_i}{s_{y_i - \hat{y}_i}} \tag{15.23}$$

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

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

where

$$s = \text{standard error of the estimate}$$
$$h_i = \text{leverage of observation } i$$

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

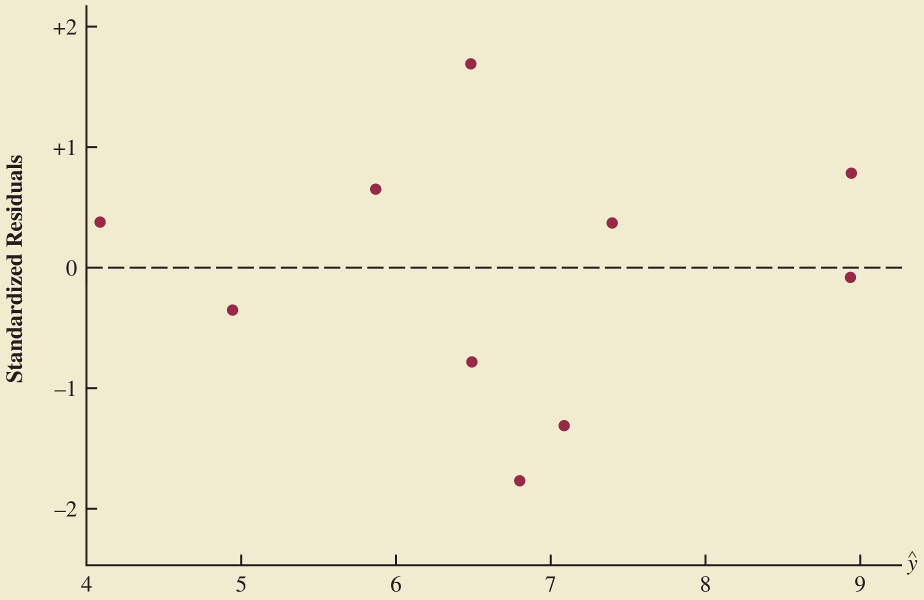
TABLE 15.7 RESIDUALS AND STANDARDIZED RESIDUALS FOR THE BUTLER TRUCKING REGRESSION ANALYSIS

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	0.361541	0.78344
50	3	4.8	4.95830	−0.158304	−0.34962
100	4	8.9	8.93846	−0.038460	−0.08334
100	2	6.5	7.09161	−0.591609	−1.30929
50	2	4.2	4.03488	0.165121	0.38167
80	2	6.2	5.86892	0.331083	0.65431
75	3	7.4	6.48667	0.913331	1.68917
65	4	6.0	6.79875	−0.798749	−1.77372
90	3	7.6	7.40369	0.196311	0.36703
90	2	6.1	6.48026	−0.380263	−0.77639

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 $s_{y_i-\hat{y}_i}$ 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 ith 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 ith 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 ith observation is an outlier, $s_{(i)}$ will be less than s. The absolute value of the ith 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	0.78344	0.75939
50	3	4.8	−0.34962	−0.32654
100	4	8.9	−0.08334	−0.07720
100	2	6.5	−1.30929	−1.39494
50	2	4.2	0.38167	0.35709
80	2	6.2	0.65431	0.62519
75	3	7.4	1.68917	2.03187
65	4	6.0	−1.77372	−2.21314
90	3	7.6	0.36703	0.34312
90	2	6.1	−0.77639	−0.75190

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 ith 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 ith 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 ith studentized deleted residual is less than -2.447 or greater than $+2.447$, we can conclude that the ith 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.

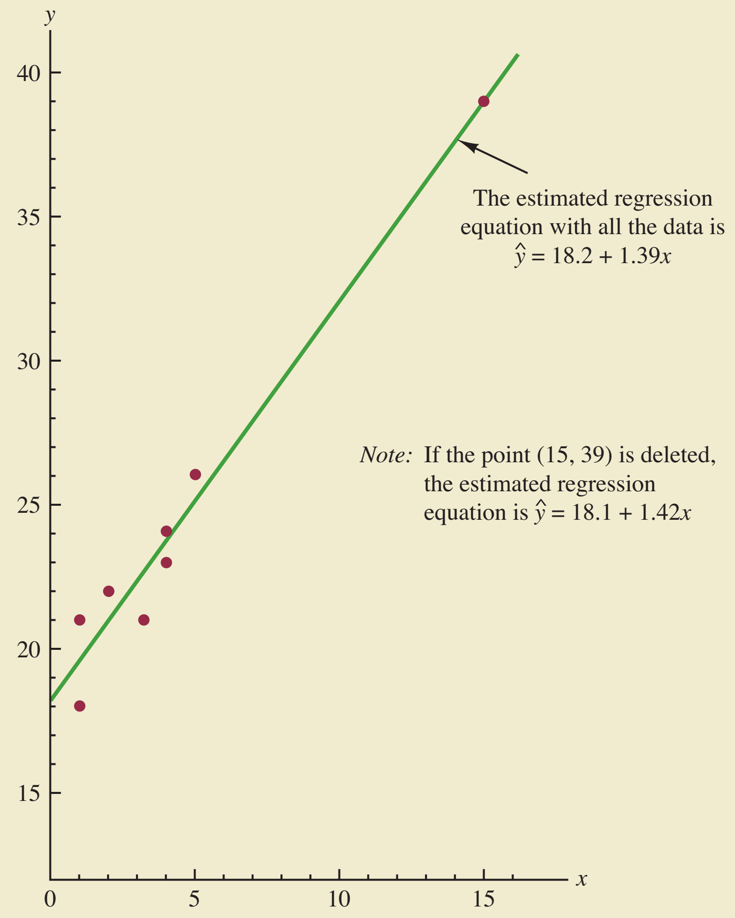
Using Cook's Distance Measure to Identify Influential Observations

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.

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

FIGURE 15.11 SCATTER DIAGRAM FOR THE DATA SET IN TABLE 15.10



The estimated regression equation with all the data is $\hat{y} = 18.2 + 1.39x$

Note: If the point (15, 39) is deleted, the estimated regression equation is $\hat{y} = 18.1 + 1.42x$

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 not significantly different from the values obtained by 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

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

where

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

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 ith 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

1. 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.

2. To determine whether the value of Cook's distance measure D_i is large enough to conclude that the ith 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

SELF test

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

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

a. Develop the estimated regression equation for these data.
b. Plot the standardized residuals versus $\hat{y}$. Do there appear to be any outliers in these data? Explain.
c. 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.

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

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

a. Develop the estimated regression equation for these data.
b. 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.
c. Compute the leverage values for these data. Do there appear to be any influential observations in these data? Explain.
d. 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

a. Find an estimated regression equation relating weekly gross revenue to television and newspaper advertising.
b. Plot the standardized residuals against $\hat{y}$. Does the residual plot support the assumptions about ϵ? Explain.
c. Check for any outliers in these data. What are your conclusions?
d. Are there any influential observations? Explain.

42. The following data show the curb weight, horsepower, and ¼-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 ¼ 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

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

a. Find the estimated regression equation, which uses price and horsepower to predict ¼-mile speed.
b. Plot the standardized residuals against $\hat{y}$. Does the residual plot support the assumption about ϵ? Explain.
c. Check for any outliers. What are your conclusions?
d. Are there any influential observations? Explain.

WEB file
LPGA

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.
a. 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.
b. Plot the standardized residuals against $\hat{y}$. Does the residual plot support the assumption about ϵ? Explain.
c. Check for any outliers. What are your conclusions?
d. 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 predict whether a catalog recipient will use the coupon. We would use Annual Spending 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 predict 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.

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \qquad (15.26)$$

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

LOGISTIC REGRESSION EQUATION

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

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

WEB file

Simmons

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 $x_1, x_2, \ldots, x_p$. 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

$$E(y) = P(y = 1 | x_1, x_2, \ldots, x_p) \tag{15.28}$$

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

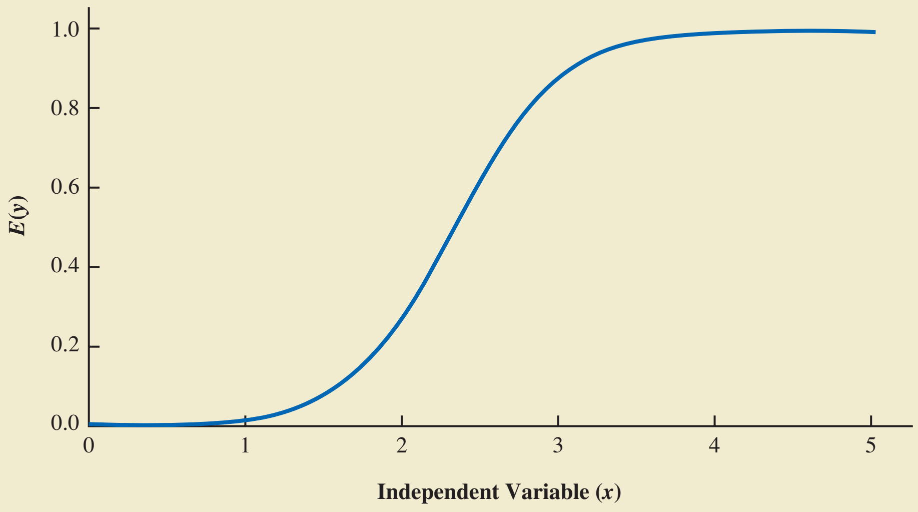
$$E(y) = P(y = 1 | x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} = \frac{e^{-7+3x}}{1 + e^{-7+3x}} \tag{15.29}$$

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 $b_0, b_1, \ldots, b_p$ as estimates of the model parameters $(\beta_0, \beta_1, \ldots, \beta_p)$. 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

$$\hat{y} = \text{estimate of } P(y = 1 \mid x_1, x_2, \ldots, x_p) = \frac{e^{b_0 + b_1x_1 + b_2x_2 + \cdots + b_px_p}}{1 + e^{b_0 + b_1x_1 + b_2x_2 + \cdots + b_px_p}} \qquad \textbf{(15.30)}$$

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.

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

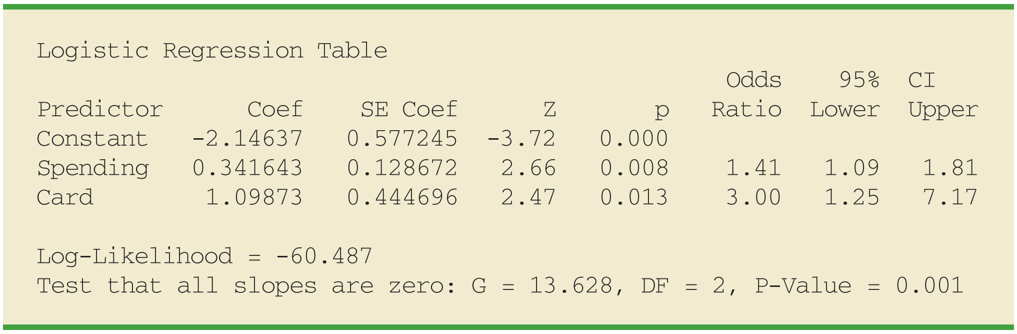
In Appendix 15.3 we show how Minitab is used to generate the output in Figure 15.13.

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 $b_0 = -2.14637$, $b_1 = 0.341643$, and $b_2 = 1.09873$. Thus, the estimated logistic regression equation is

$$\hat{y} = \frac{e^{b_0 + b_1x_1 + b_2x_2}}{1 + e^{b_0 + b_1x_1 + b_2x_2}} = \frac{e^{-2.14637 + 0.341643x_1 + 1.09873x_2}}{1 + e^{-2.14637 + 0.341643x_1 + 1.09873x_2}} \qquad \textbf{(15.32)}$$

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.

$$\hat{y} = \frac{e^{-2.14637+0.341643(2)+1.09873(0)}}{1 + e^{-2.14637+0.341643(2)+1.09873(0)}} = \frac{e^{-1.4631}}{1 + e^{-1.4631}} = \frac{.2315}{1.2315} = 0.1880$$

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).

$$\hat{y} = \frac{e^{-2.14637+0.341643(2)+1.09873(1)}}{1 + e^{-2.14637+0.341643(2)+1.09873(1)}} = \frac{e^{-0.3644}}{1 + e^{-0.3644}} = \frac{.6946}{1.6946} = 0.4099$$

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:

$$H_0\colon \beta_1 = \beta_2 = 0$$
$$H_{\mathrm{a}}\colon \text{One or both of the parameters is not equal to zero}$$

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 0.001. 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

$$H_0\colon \beta_i = 0$$
$$H_{\mathrm{a}}\colon \beta_i \neq 0$$

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 0.3922. 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:

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

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	0.3305	0.4099	0.4943	0.5791	0.6594	0.7315	0.7931
	No	0.1413	0.1880	0.2457	0.3144	0.3922	0.4759	0.5610

independent variables has been increased by one unit (odds_1) divided by the odds that $y = 1$ given no change in the values for the independent variables (odds_0).

ODDS RATIO

$$\text{Odds Ratio} = \frac{\text{odds}_1}{\text{odds}_0} \tag{15.34}$$

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 estimates of the 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 $g(x_1, x_2, \ldots, x_p)$ to denote the logit.

LOGIT

$$g(x_1, x_2, \ldots, x_p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \tag{15.35}$$

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

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

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

ESTIMATED LOGIT

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

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

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

For the Simmons Stores example, the estimated logit is

$$\hat{g}(x_1, x_2) = -2.14637 + 0.341643 x_1 + 1.09873 x_2$$

and the estimated regression equation is

$$\hat{y} = \frac{e^{\hat{g}(x_1, x_2)}}{1 + e^{\hat{g}(x_1, x_2)}} = \frac{e^{-2.14637 + 0.341643 x_1 + 1.09873 x_2}}{1 + e^{-2.14637 + 0.341643 x_1 + 1.09873 x_2}}$$

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

WEB file

Simmons

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 compute an estimate of the probability of using the coupon for customers who do not have a Simmons credit card and an estimate of the probability of using the coupon for customers who have a Simmons credit card.
    e. What is the estimate of the odds ratio? What is its interpretation?
45. In Table 15.12 we provided estimates of the probability 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

a. Write the logistic regression equation relating x to y.

b. For the Community Bank data, use Minitab to compute the estimated logistic regression equation.

c. Conduct a test of significance using the G test statistic. Use $\alpha = .05$.

d. Estimate the probability that customers with an average monthly balance of $1000 will sign up for direct payroll deposit.

e. 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?

f. What is the estimate of the 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:

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

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

a. Write the logistic regression equation relating x_1 and x_2 to y.
b. What is the interpretation of $E(y)$ when $x_2 = 0$?
c. Use both independent variables and Minitab to compute the estimated logit.
d. Conduct a test for overall significance using $\alpha = .05$.
e. Use $\alpha = .05$ to determine whether each of the independent variables is significant.
f. Use the estimated logit computed in part (c) to compute an estimate of 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?
g. What is the estimate of the odds ratio for the orientation program? Interpret it.
h. Would you recommend making the orientation program a required activity? Why or why not?

48. *Consumer Reports* conducted a taste test on 19 brands of boxed chocolates. The following data show the price per serving, based on the FDA serving size of 1.4 ounces, and the quality rating for the 19 chocolates tested (*Consumer Reports,* February 2002).

WEB file
Chocolate

Manufacturer	Price	Rating
Bernard Callebaut	3.17	Very Good
Candinas	3.58	Excellent
Fannie May	1.49	Good
Godiva	2.91	Very Good
Hershey's	0.76	Good
L.A. Burdick	3.70	Very Good
La Maison du Chocolate	5.08	Excellent
Leonidas	2.11	Very Good
Lindt	2.20	Good
Martine's	4.76	Excellent
Michael Recchiuti	7.05	Very Good
Neuchatel	3.36	Good
Neuchatel Sugar Free	3.22	Good
Richard Donnelly	6.55	Very Good
Russell Stover	0.70	Good
See's	1.06	Very Good
Teuscher Lake of Zurich	4.66	Very Good
Whitman's	0.70	Fair
Whitman's Sugar Free	1.21	Fair

Suppose that you would like to determine whether products that cost more rate higher in quality. For the purpose of this exercise, use the following binary dependent variable:

$$y = 1 \text{ if the quality rating is very good or excellent and } 0 \text{ if good or fair}$$

a. Write the logistic regression equation relating $x =$ price per serving to y.
b. Use Minitab to compute the estimated logit.
c. Use the estimated logit computed in part (b) to compute an estimate of the probability a chocolate that has a price per serving of \$4.00 will have a quality rating of very good or excellent.
d. What is the estimate of the odds ratio? What is its interpretation?

Summary

In this chapter, we introduced multiple regression analysis as an extension of simple linear regression analysis presented in Chapter 14. Multiple regression analysis enables us to understand how a dependent variable is related to two or more independent variables. The mulitple regression equation $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$ shows that the mean or expected value of the dependent variable y, denoted $E(y)$, is related to the values of the independent variables $x_1, x_2, \ldots, x_p$. Sample data and the least squares method are used to develop the estimated multiple regression equation $\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p$. In effect $b_0, b_1, b_2, \ldots, b_p$ are sample statistics used to estimate the unknown model parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$. Computer printouts were used throughout the chapter to emphasize the fact that statistical software packages are the only realistic means of performing the numerous computations required in multiple regression analysis.

The multiple coefficient of determination was presented as a measure of the goodness of fit of the estimated regression equation. It determines the proportion of the variation of y that can be explained by the estimated regression equation. The adjusted multiple coefficient of determination is a similar measure of goodness of fit that adjusts for the number of independent variables and thus avoids overestimating the impact of adding more independent variables.

An F test and a t test were presented as ways to determine statistically whether the relationship among the variables is significant. The F test is used to determine whether there is a significant overall relationship between the dependent variable and the set of all independent variables. The t test is used to determine whether there is a significant relationship between the dependent variable and an individual independent variable given the other independent variables in the regression model. Correlation among the independent variables, known as multicollinearity, was discussed.

The section on categorical independent variables showed how dummy variables can be used to incorporate categorical data into multiple regression analysis. The section on residual analysis showed how residual analysis can be used to validate the model assumptions, detect outliers, and identify influential observations. Standardized residuals, leverage, studentized deleted residuals, and Cook's distance measure were discussed. The chapter concluded with a section on how logistic regression can be used to model situations in which the dependent variable may only assume two values.

Glossary

Multiple regression analysis Regression analysis involving two or more independent variables.

Multiple regression model The mathematical equation that describes how the dependent variable y is related to the independent variables $x_1, x_2, \ldots, x_p$ and an error term ϵ.

Multiple regression equation The mathematical equation relating the expected value or mean value of the dependent variable to the values of the independent variables; that is, $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$.

Estimated multiple regression equation The estimate of the multiple regression equation based on sample data and the least squares method; it is $\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$.

Least squares method The method used to develop the estimated regression equation. It minimizes the sum of squared residuals (the deviations between the observed values of the dependent variable, y_i, and the estimated values of the dependent variable, $\hat{y}_i$).

Multiple coefficient of determination A measure of the goodness of fit of the estimated multiple regression equation. It can be interpreted as the proportion of the variability in the dependent variable that is explained by the estimated regression equation.

Adjusted multiple coefficient of determination A measure of the goodness of fit of the estimated multiple regression equation that adjusts for the number of independent variables in the model and thus avoids overestimating the impact of adding more independent variables.

Multicollinearity The term used to describe the correlation among the independent variables.

Categorical independent variable An independent variable with categorical data.

Dummy variable A variable used to model the effect of categorical independent variables. A dummy variable may take only the value zero or one.

Leverage A measure of how far the values of the independent variables are from their mean values.

Outlier An observation that does not fit the pattern of the other data.

Studentized deleted residuals Standardized residuals that are based on a revised standard error of the estimate obtained by deleting observation *i* from the data set and then performing the regression analysis and computations.

Influential observation An observation that has a strong influence on the regression results.

Cook's distance measure A measure of the influence of an observation based on both the leverage of observation *i* and the residual for observation *i*.

Logistic regression equation The mathematical equation relating $E(y)$, the probability that $y = 1$, to the values of the independent variables; that is, $E(y) = P(y = 1|x_1, x_2, \ldots, x_p) = \dfrac{e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p}}{1 + e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p}}$.

Estimated logistic regression equation The estimate of the logistic regression equation based on sample data; that is, $\hat{y} =$ estimate of $P(y = 1|x_1, x_2, \ldots, x_p) = \dfrac{e^{b_0+b_1x_1+b_2x_2+\cdots+b_px_p}}{1 + e^{b_0+b_1x_1+b_2x_2+\cdots+b_px_p}}$.

Odds in favor of an event occurring The probability the event will occur divided by the probability the event will not occur.

Odds ratio The odds that $y = 1$ given that one of the independent variables increased by one unit (odds_1) divided by the odds that $y = 1$ given no change in the values for the independent variables (odds_0); that is, Odds ratio $= \text{odds}_1/\text{odds}_0$.

Logit The natural logarithm of the odds in favor of $y = 1$; that is, $g(x_1, x_2, \ldots, x_p) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_px_p$.

Estimated logit An estimate of the logit based on sample data; that is, $\hat{g}(x_1, x_2, \ldots, x_p) = b_0 + b_1x_1 + b_2x_2 + \cdots + b_px_p$.

Key Formulas

Multiple Regression Model

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_px_p + \epsilon \tag{15.1}$$

Multiple Regression Equation

$$E(y) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_px_p \tag{15.2}$$

Estimated Multiple Regression Equation

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \cdots + b_px_p \tag{15.3}$$

Least Squares Criterion

$$\min \Sigma(y_i - \hat{y}_i)^2 \tag{15.4}$$

Relationship Among SST, SSR, and SSE

$$\text{SST} = \text{SSR} + \text{SSE} \tag{15.7}$$

Multiple Coefficient of Determination

$$R^2 = \frac{\text{SSR}}{\text{SST}} \tag{15.8}$$

Adjusted Multiple Coefficient of Determination

$$R_{\text{a}}^2 = 1 - (1 - R^2)\frac{n - 1}{n - p - 1} \tag{15.9}$$

Mean Square Due to Regression

$$\text{MSR} = \frac{\text{SSR}}{p} \tag{15.12}$$

Mean Square Due to Error

$$\text{MSE} = \frac{\text{SSE}}{n - p - 1} \tag{15.13}$$

F Test Statistic

$$F = \frac{\text{MSR}}{\text{MSE}} \tag{15.14}$$

t Test Statistic

$$t = \frac{b_i}{s_{b_i}} \tag{15.15}$$

Standardized Residual for Observation *i*

$$\frac{y_i - \hat{y}_i}{s_{y_i - \hat{y}_i}} \tag{15.23}$$

Standard Deviation of Residual *i*

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

Cook's Distance Measure

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

Logistic Regression Equation

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

Estimated Logistic Regression Equation

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

Odds Ratio

$$\text{Odds ratio} = \frac{\text{odds}_1}{\text{odds}_0} \tag{15.34}$$

Logit

$$g(x_1, x_2, \ldots, x_p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \tag{15.35}$$

Estimated Logit

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

Supplementary Exercises

49. The admissions officer for Clearwater College developed the following estimated regression equation relating the final college GPA to the student's SAT mathematics score and high-school GPA.

$$\hat{y} = -1.41 + .0235x_1 + .00486x_2$$

where

$$\begin{aligned} x_1 &= \text{high-school grade point average} \\ x_2 &= \text{SAT mathematics score} \\ y &= \text{final college grade point average} \end{aligned}$$

a. Interpret the coefficients in this estimated regression equation.
b. Estimate the final college GPA for a student who has a high-school average of 84 and a score of 540 on the SAT mathematics test.

50. The personnel director for Electronics Associates developed the following estimated regression equation relating an employee's score on a job satisfaction test to his or her length of service and wage rate.

$$\hat{y} = 14.4 - 8.69x_1 + 13.5x_2$$

where

$$\begin{aligned} x_1 &= \text{length of service (years)} \\ x_2 &= \text{wage rate (dollars)} \\ y &= \text{job satisfaction test score (higher scores} \\ &\quad\ \text{indicate greater job satisfaction)} \end{aligned}$$

a. Interpret the coefficients in this estimated regression equation.
b. Develop an estimate of the job satisfaction test score for an employee who has four years of service and makes \$6.50 per hour.

51. A partial computer output from a regression analysis follows.

```
The regression equation is
Y = 8.103 + 7.602 X1 + 3.111 X2

Predictor            Coef       SE Coef         T
Constant          _______         2.667     _____
X1                _______         2.105     _____
X2                _______         0.613     _____

S = 3.335     R-sq = 92.3%     R-sq(adj) = _____%

Analysis of Variance

SOURCE              DF        SS         MS       F
Regression       ______      1612     ______   _____
Residual Error      12    ______      ______
Total            ______   ______
```

a. Compute the missing entries in this output.
b. Use the F test and $\alpha = .05$ to see whether a significant relationship is present.
c. Use the t test and $\alpha = .05$ to test H_0: $\beta_1 = 0$ and H_0: $\beta_2 = 0$.
d. Compute R_a^2.

52. Recall that in exercise 49, the admissions officer for Clearwater College developed the following estimated regression equation relating final college GPA to the student's SAT mathematics score and high-school GPA.

$$\hat{y} = -1.41 + .0235x_1 + .00486x_2$$

where

x_1 = high-school grade point average
x_2 = SAT mathematics score
y = final college grade point average

A portion of the Minitab computer output follows.

```
The regression equation is
Y = -1.41 + .0235 X1 + .00486 X2

Predictor            Coef       SE Coef         T
Constant          -1.4053        0.4848     _____
X1               0.023467      0.008666     _____
X2                ______       0.001077     _____

S = 0.1298     R-sq = ______    R-sq(adj) = ______

Analysis of Variance

SOURCE              DF        SS          MS       F
Regression       _____     1.76209     _____   _____
Residual Error   _____     _______     _____
Total                9     1.88000
```

a. Complete the missing entries in this output.
b. Use the F test and a .05 level of significance to see whether a significant relationship is present.
c. Use the t test and $\alpha = .05$ to test H_0: $\beta_1 = 0$ and H_0: $\beta_2 = 0$.
d. Did the estimated regression equation provide a good fit to the data? Explain.

53. Recall that in exercise 50 the personnel director for Electronics Associates developed the following estimated regression equation relating an employee's score on a job satisfaction test to length of service and wage rate.

$$\hat{y} = 14.4 - 8.69x_1 + 13.5x_2$$

where

$$\begin{aligned} x_1 &= \text{length of service (years)} \\ x_2 &= \text{wage rate (dollars)} \\ y &= \text{job satisfaction test score (higher scores} \\ &\quad\ \text{indicate greater job satisfaction)} \end{aligned}$$

A portion of the Minitab computer output follows.

```
The regression equation is
Y = 14.4 - 8.69 X1 + 13.52 X2

Predictor             Coef         SE Coef           T
Constant            14.448           8.191        1.76
X1                  ______           1.555       _____
X2                  13.517           2.085       _____

S = 3.773       R-sq = ______%  R-sq(adj) = ______%

Analysis of Variance

SOURCE            DF          SS        MS         F
Regression         2      ______     _____     _____
Residual Error _____       71.17     _____
Total              7       720.0
```

a. Complete the missing entries in this output.
b. Compute F and test using $\alpha = .05$ to see whether a significant relationship is present.
c. Did the estimated regression equation provide a good fit to the data? Explain.
d. Use the t test and $\alpha = .05$ to test H_0: $\beta_1 = 0$ and H_0: $\beta_2 = 0$.

54. The Tire Rack, America's leading online distributor of tires and wheels, conducts extensive testing to provide customers with products that are right for their vehicle, driving style, and driving conditions. In addition, the Tire Rack maintains an independent consumer survey to help drivers help each other by sharing their long-term tire experiences. The following data show survey ratings (1 to 10 scale with 10 the highest rating) for 18 maximum performance summer tires (Tire Rack website, February 3, 2009). The variable Steering rates the tire's steering responsiveness, Tread Wear rates quickness of wear based on the driver's expectations, and Buy Again rates the driver's overall tire satisfaction and desire to purchase the same tire again.

TireRack

Tire	Steering	Tread Wear	Buy Again
Goodyear Assurance TripleTred	8.9	8.5	8.1
Michelin HydroEdge	8.9	9.0	8.3
Michelin Harmony	8.3	8.8	8.2
Dunlop SP 60	8.2	8.5	7.9
Goodyear Assurance ComforTred	7.9	7.7	7.1
Yokohama Y372	8.4	8.2	8.9
Yokohama Aegis LS4	7.9	7.0	7.1
Kumho Power Star 758	7.9	7.9	8.3
Goodyear Assurance	7.6	5.8	4.5
Hankook H406	7.8	6.8	6.2
Michelin Energy LX4	7.4	5.7	4.8
Michelin MX4	7.0	6.5	5.3
Michelin Symmetry	6.9	5.7	4.2
Kumho 722	7.2	6.6	5.0
Dunlop SP 40 A/S	6.2	4.2	3.4
Bridgestone Insignia SE200	5.7	5.5	3.6
Goodyear Integrity	5.7	5.4	2.9
Dunlop SP20 FE	5.7	5.0	3.3

a. Develop an estimated regression equation that can be used to predict the Buy Again rating given based on the Steering rating. At the .05 level of significance, test for a significant relationship.

b. Did the estimated regression equation developed in part (a) provide a good fit to the data? Explain.

c. Develop an estimated regression equation that can be used to predict the Buy Again rating given the Steering rating and the Tread Wear rating.

d. Is the addition of the Tread Wear independent variable significant? Use $\alpha = .05$.

55. *Consumer Reports* provided extensive testing and ratings for 24 treadmills. An overall score, based primarily on ease of use, ergonomics, exercise range, and quality, was developed for each treadmill tested. In general, a higher overall score indicates better performance. The following data show the price, the quality rating, and overall score for the 24 treadmills (*Consumer Reports*, February 2006).

Treadmills

Brand & Model	Price	Quality	Score
Landice L7	2900	Excellent	86
NordicTrack S3000	3500	Very good	85
SportsArt 3110	2900	Excellent	82
Precor	3500	Excellent	81
True Z4 HRC	2300	Excellent	81
Vision Fitness T9500	2000	Excellent	81
Precor M 9.31	3000	Excellent	79
Vision Fitness T9200	1300	Very good	78
Star Trac TR901	3200	Very good	72
Trimline T350HR	1600	Very good	72
Schwinn 820p	1300	Very good	69
Bowflex 7-Series	1500	Excellent	83
NordicTrack S1900	2600	Very good	83
Horizon Fitness PST8	1600	Very good	82
Horizon Fitness 5.2T	1800	Very good	80
Evo by Smooth Fitness FX30	1700	Very good	75
ProForm 1000S	1600	Very good	75
Horizon Fitness CST4.5	1000	Very good	74

(*continued*)

Brand & Model	Price	Quality	Score
Keys Fitness 320t	1200	Very good	73
Smooth Fitness 7.1HR Pro	1600	Very good	73
NordicTrack C2300	1000	Good	70
Spirit Inspire	1400	Very good	70
ProForm 750	1000	Good	67
Image 19.0 R	600	Good	66

a. Use these data to develop an estimated regression equation that could be used to estimate the overall score given the price.

b. Use $\alpha = .05$ to test for overall significance.

c. To incorporate the effect of quality, a categorical variable with three levels, we used two dummy variables: Quality-E and Quality-VG. Each variable was coded 0 or 1 as follows.

$$\text{Quality-E} = \begin{cases} 1 \text{ if quality rating is excellent} \\ 0 \text{ otherwise} \end{cases}$$

$$\text{Quality-VG} = \begin{cases} 1 \text{ if quality rating is very good} \\ 0 \text{ otherwise} \end{cases}$$

Develop an estimated regression equation that could be used to estimate the overall score given the price and the quality rating.

d. For the estimated regression equation developed in part (c), test for overall significance using $\alpha = .10$.

e. For the estimated regression equation developed in part (c), use the *t* test to determine the significance of each independent variable. Use $\alpha = .10$.

f. Develop a standardized residual plot. Does the pattern of the residual plot appear to be reasonable?

g. Do the data contain any outliers or influential observations?

h. Estimate the overall score for a treadmill with a price of $2000 and a good quality rating. How much would the estimate change if the quality rating were very good? Explain.

56. A portion of a data set containing information for 45 mutual funds that are part of the *Morningstar Funds 500* for 2008 follows. The complete data set is available in the file named MutualFunds. The data set includes the following five variables:

Type: The type of fund, labeled DE (Domestic Equity), IE (International Equity), and FI (Fixed Income).

Net Asset Value ($): The closing price per share on December 31, 2007.

5-Year Average Return (%): The average annual return for the fund over the past 5 years.

Expense Ratio (%): The percentage of assets deducted each fiscal year for fund expenses.

Morningstar Rank: The risk adjusted star rating for each fund; Morningstar ranks go from a low of 1-Star to a high of 5-Stars.

WEB file

MutualFunds

		Net	5-Year		
		Asset	Average	Expense	
	Fund	Value	Return	Ratio	Morningstar
Fund Name	Type	($)	(%)	(%)	Rank
Amer Cent Inc & Growth Inv	DE	28.88	12.39	0.67	2-Star
American Century Intl. Disc	IE	14.37	30.53	1.41	3-Star
American Century Tax-Free Bond	FI	10.73	3.34	0.49	4-Star

Fund Name	Fund Type	Net Asset Value ($)	5-Year Average Return (%)	Expense Ratio (%)	Morningstar Rank
American Century Ultra	DE	24.94	10.88	0.99	3-Star
Ariel	DE	46.39	11.32	1.03	2-Star
Artisan Intl Val	IE	25.52	24.95	1.23	3-Star
Artisan Small Cap	DE	16.92	15.67	1.18	3-Star
Baron Asset	DE	50.67	16.77	1.31	5-Star
Brandywine	DE	36.58	18.14	1.08	4-Star
⋮	⋮	⋮	⋮	⋮	⋮

a. Develop an estimated regression equation that can be used to predict the 5-year average return given fund type. At the .05 level of significance, test for a significant relationship.

b. Did the estimated regression equation developed in part (a) provide a good fit to the data? Explain.

c. Develop the estimated regression equation that can be used to predict the 5-year average return given the type of fund, the net asset value, and the expense ratio. At the .05 level of significance, test for a significant relationship. Do you think any variables should be deleted from the estimated regression equation? Explain.

d. Morningstar Rank is a categorical variable. Because the data set contains only funds with four ranks (2-Star through 5-Star), use the following dummy variables: 3StarRank = 1 for a 3-Star fund, 0 otherwise; 4StarRank = 1 for a 4-Star fund, 0 otherwise; and 5StarRank = 1 for a 5-Star fund, 0 otherwise. Develop an estimated regression equation that can be used to predict the 5-year average return given the type of fund, the expense ratio, and the Morningstar Rank. Using $\alpha = .05$, remove any independent variables that are not significant.

e. Use the estimated regression equation developed in part (d) to estimate the 5-year average return for a domestic equity fund with an expense ratio of 1.05% and a 3-Star Morningstar Rank.

57. The U.S. Department of Energy's Fuel Economy Guide provides fuel efficiency data for cars and trucks (U.S. Department of Energy website, February 22, 2008). A portion of the data for 311 compact, midsize, and large cars follows. The column labeled Class identifies the size of the car; Compact, Midsize, or Large. The column labeled Displacement shows the engine's displacement in liters. The column labeled Fuel Type shows whether the car uses premium (P) or regular (R) fuel, and the column labeled Hwy MPG shows the fuel efficiency rating for highway driving in terms of miles per gallon. The complete data set is contained in the file named FuelData.

WEB file

FuelData

Car	Class	Displacement	Fuel Type	Hwy MPG
1	Compact	3.1	P	25
2	Compact	3.1	P	25
3	Compact	3	P	25
⋮	⋮	⋮	⋮	⋮
161	Midsize	2.4	R	30
162	Midsize	2	P	29
⋮	⋮	⋮	⋮	⋮
310	Large	3	R	25
311	Large	3	R	25

a. Develop an estimated regression equation that can be used to predict the fuel efficiency for highway driving given the engine's displacement. Test for significance using $\alpha = .05$.

b. Consider the addition of the dummy variables ClassMidsize and ClassLarge. The value of ClassMidsize is 1 if the car is a midsize car and 0 otherwise; the value of ClassLarge is 1 if the car is a large car and 0 otherwise. Thus, for a compact car, the value of ClassMidsize and the value of ClassLarge is 0. Develop the estimated regression equation that can be used to predict the fuel efficiency for highway driving given the engine's displacement and the dummy variables ClassMidsize and ClassLarge.

c. Use $\alpha = .05$ to determine whether the dummy variables added in part (b) are significant.

d. Consider the addition of the dummy variable FuelPremium, where the value of FuelPremium is 1 if the car uses premium fuel and 0 if the car uses regular fuel. Develop the estimated regression equation that can be used to predict the fuel efficiency for highway driving given the engine's displacement, the dummy variables ClassMidsize and ClassLarge, and the dummy variable FuelPremium.

e. For the estimated regression equation developed in part (d), test for overall significance and individual significance using $\alpha = .05$.

Case Problem 1 Consumer Research, Inc.

Consumer Research, Inc., is an independent agency that conducts research on consumer attitudes and behaviors for a variety of firms. In one study, a client asked for an investigation of consumer characteristics that can be used to predict the amount charged by credit card users. Data were collected on annual income, household size, and annual credit card charges for a sample of 50 consumers. The following data are contained in the file named Consumer.

WEB file

Consumer

Income ($1000s)	Household Size	Amount Charged ($)	Income ($1000s)	Household Size	Amount Charged ($)
54	3	4016	54	6	5573
30	2	3159	30	1	2583
32	4	5100	48	2	3866
50	5	4742	34	5	3586
31	2	1864	67	4	5037
55	2	4070	50	2	3605
37	1	2731	67	5	5345
40	2	3348	55	6	5370
66	4	4764	52	2	3890
51	3	4110	62	3	4705
25	3	4208	64	2	4157
48	4	4219	22	3	3579
27	1	2477	29	4	3890
33	2	2514	39	2	2972
65	3	4214	35	1	3121
63	4	4965	39	4	4183
42	6	4412	54	3	3730
21	2	2448	23	6	4127
44	1	2995	27	2	2921
37	5	4171	26	7	4603
62	6	5678	61	2	4273
21	3	3623	30	2	3067
55	7	5301	22	4	3074
42	2	3020	46	5	4820
41	7	4828	66	4	5149

Managerial Report

1. Use methods of descriptive statistics to summarize the data. Comment on the findings.
2. Develop estimated regression equations, first using annual income as the independent variable and then using household size as the independent variable. Which variable is the better predictor of annual credit card charges? Discuss your findings.
3. Develop an estimated regression equation with annual income and household size as the independent variables. Discuss your findings.
4. What is the predicted annual credit card charge for a three-person household with an annual income of $40,000?
5. Discuss the need for other independent variables that could be added to the model. What additional variables might be helpful?

Case Problem 2 Alumni Giving

Alumni donations are an important source of revenue for colleges and universities. If administrators could determine the factors that could lead to increases in the percentage of alumni who make a donation, they might be able to implement policies that could lead to increased revenues. Research shows that students who are more satisfied with their contact with teachers are more likely to graduate. As a result, one might suspect that smaller class sizes and lower student-faculty ratios might lead to a higher percentage of satisfied graduates, which in turn might lead to increases in the percentage of alumni who make a donation. Table 15.13 shows data for 48 national universities (*America's Best Colleges,* Year 2000 ed.). The column labeled Graduation Rate is the percentage of students who initially enrolled at the university and graduated. The column labeled % of Classes Under 20 shows the percentage of classes offered with fewer than 20 students. The column labeled Student-Faculty Ratio is the number of students enrolled divided by the total number of faculty. Finally, the column labeled Alumni Giving Rate is the percentage of alumni who made a donation to the university.

Managerial Report

1. Use methods of descriptive statistics to summarize the data.
2. Develop an estimated regression equation that can be used to predict the alumni giving rate given the number of students who graduate. Discuss your findings.
3. Develop an estimated regression equation that could be used to predict the alumni giving rate using the data provided.
4. What conclusions and recommendations can you derive from your analysis?

Case Problem 3 PGA Tour Statistics

The Professional Golfers Association (PGA) maintains data on performance and earnings for members of the PGA Tour. The top 125 players based on total earnings in PGA Tour events are exempt for the following season. Making the top 125 money list is important because a player who is "exempt" has qualified to be a full-time member of the PGA tour for the following season.

Scoring average is generally considered the most important statistic in terms of success on the PGA Tour. To investigate the relationship between variables such as driving distance, driving accuracy, greens in regulation, sand saves, and average putts per round have on average score, year-end performance data for the 125 players who had the highest total

TABLE 15.13 DATA FOR 48 NATIONAL UNIVERSITIES

WEB file
Alumni

	State	Graduation Rate	% of Classes Under 20	Student-Faculty Ratio	Alumni Giving Rate
Boston College	MA	85	39	13	25
Brandeis University	MA	79	68	8	33
Brown University	RI	93	60	8	40
California Institute of Technology	CA	85	65	3	46
Carnegie Mellon University	PA	75	67	10	28
Case Western Reserve Univ.	OH	72	52	8	31
College of William and Mary	VA	89	45	12	27
Columbia University	NY	90	69	7	31
Cornell University	NY	91	72	13	35
Dartmouth College	NH	94	61	10	53
Duke University	NC	92	68	8	45
Emory University	GA	84	65	7	37
Georgetown University	DC	91	54	10	29
Harvard University	MA	97	73	8	46
Johns Hopkins University	MD	89	64	9	27
Lehigh University	PA	81	55	11	40
Massachusetts Inst. of Technology	MA	92	65	6	44
New York University	NY	72	63	13	13
Northwestern University	IL	90	66	8	30
Pennsylvania State Univ.	PA	80	32	19	21
Princeton University	NJ	95	68	5	67
Rice University	TX	92	62	8	40
Stanford University	CA	92	69	7	34
Tufts University	MA	87	67	9	29
Tulane University	LA	72	56	12	17
U. of California–Berkeley	CA	83	58	17	18
U. of California–Davis	CA	74	32	19	7
U. of California–Irvine	CA	74	42	20	9
U. of California–Los Angeles	CA	78	41	18	13
U. of California–San Diego	CA	80	48	19	8
U. of California–Santa Barbara	CA	70	45	20	12
U. of Chicago	IL	84	65	4	36
U. of Florida	FL	67	31	23	19
U. of Illinois–Urbana Champaign	IL	77	29	15	23
U. of Michigan–Ann Arbor	MI	83	51	15	13
U. of North Carolina–Chapel Hill	NC	82	40	16	26
U. of Notre Dame	IN	94	53	13	49
U. of Pennsylvania	PA	90	65	7	41
U. of Rochester	NY	76	63	10	23
U. of Southern California	CA	70	53	13	22
U. of Texas–Austin	TX	66	39	21	13
U. of Virginia	VA	92	44	13	28
U. of Washington	WA	70	37	12	12
U. of Wisconsin–Madison	WI	73	37	13	13
Vanderbilt University	TN	82	68	9	31
Wake Forest University	NC	82	59	11	38
Washington University–St. Louis	MO	86	73	7	33
Yale University	CT	94	77	7	50

WEB file PGATour

earnings in PGA Tour events for 2008 are contained in the file named PGATour (PGA Tour website, 2009). Each row of the data set corresponds to a PGA Tour player, and the data have been sorted based upon total earnings. Descriptions for the variables in the data set follow.

Money: Total earnings in PGA Tour events.

Scoring Average: The average number of strokes per completed round.

DrDist (Driving Distance): DrDist is the average number of yards per measured drive. On the PGA Tour driving distance is measured on two holes per round. Care is taken to select two holes which face in opposite directions to counteract the effect of wind. Drives are measured to the point at which they come to rest regardless of whether they are in the fairway or not.

DrAccu (Driving Accuracy): The percentage of time a tee shot comes to rest in the fairway (regardless of club). Driving accuracy is measured on every hole, excluding par 3s.

GIR (Greens in Regulation): The percentage of time a player was able to hit the green in regulation. A green is considered hit in regulation if any portion of the ball is touching the putting surface after the GIR stroke has been taken. The GIR stroke is determined by subtracting 2 from par (1st stroke on a par 3, 2nd on a par 4, 3rd on a par 5). In other words, a green is considered hit in regulation if the player has reached the putting surface in par minus two strokes.

Sand Saves: The percentage of time a player was able to get "up and down" once in a greenside sand bunker (regardless of score). "Up and down" indicates it took the player 2 shots or less to put the ball in the hole from a greenside sand bunker.

PPR (Putts Per Round): The average number of putts per round.

Scrambling: The percentage of time a player missed the green in regulation but still made par or better.

Managerial Report

1. To predict Scoring Average, develop estimated regression equations, first using DrDist as the independent variable and then using DrAccu as the independent variable. Which variable is the better predictor of Scoring Average? Discuss your findings.
2. Develop an estimated regression equation with GIR as the independent variable. Compare your findings with the results obtained using DrDist and DrAccu.
3. Develop an estimated regression equation with GIR and Sand Saves as the independent variables. Discuss your findings.
4. Develop an estimated regression equation with GIR and PPR as the independent variables. Discuss your findings.
5. Develop an estimated regression equation with GIR and Scrambling as the independent variables. Discuss your findings.
6. Compare the results obtained for the estimated regression equations that use GIR and Sand Saves, GIR and PPR, and GIR and Scrambling as the two independent variables. If you had to select one of these two-independent variable estimated regression equations to predict Scoring Average, which estimated regression equation would you use? Explain.
7. Develop the estimated regression equation that uses GIR, Sand Saves, and PPR to predict Scoring Average. Compare the results to an estimated regression equation that uses GIR, PPR, and Scrambling as the independent variables.
8. Develop an estimated regression equation that uses GIR, Sand Saves, PPR, and Scrambling to predict Scoring Average. Discuss your results.

Case Problem 4 Predicting Winning Percentage for the NFL

WEB file NFLStats

The National Football League (NFL) records a variety of performance data for individuals and teams. Some of the year-end performance data for the 2005 season are contained in the file named NFLStats (NFL website). Each row of the data set corresponds to an NFL team, and the teams are ranked by winning percentage. Descriptions for the data follow:

WinPct	Percentage of games won
TakeInt	Takeaway interceptions; the total number of interceptions made by the team's defense
TakeFum	Takeaway fumbles; the total number of fumbles recovered by the team's defense
GiveInt	Giveaway interceptions; the total number of interceptions made by the team's offense
GiveFum	Giveaway fumbles; the total number of fumbles made by the team's offense
DefYds/G	Average number of yards per game given up on defense
RushYds/G	Average number of rushing yards per game
PassYds/G	Average number of passing yards per game
FGPct	Percentage of field goals

Managerial Report

1. Use methods of descriptive statistics to summarize the data. Comment on the findings.
2. Develop an estimated regression equation that can be used to predict WinPct using DefYds/G, RushYds/G, PassYds/G, and FGPct. Discuss your findings.
3. Starting with the estimated regression equation developed in part (2), delete any independent variables that are not significant and develop a new estimated regression equation that can be used to predict WinPct. Use $\alpha = .05$.
4. Some football analysts believe that turnovers are one of the most important factors in determining a team's success. With Takeaways = TakeInt + TakeFum and Giveaways = GiveInt + GiveFum, let NetDiff = Takeaways − Giveaways. Develop an estimated regression equation that can be used to predict WinPct using NetDiff. Compare your results with the estimated regression equation developed in part (3).
5. Develop an estimated regression equation that can be used to predict WinPct using all the data provided.

Appendix 15.1 Multiple Regression with Minitab

WEB file Butler

In Section 15.2 we discussed the computer solution of multiple regression problems by showing Minitab's output for the Butler Trucking Company problem. In this appendix we describe the steps required to generate the Minitab computer solution. First, the data must be entered in a Minitab worksheet. The miles traveled are entered in column C1, the number of deliveries are entered in column C2, and the travel times (hours) are entered in column C3. The variable names Miles, Deliveries, and Time were entered as the column headings on the worksheet. In subsequent steps, we refer to the data by using the variable names Miles, Deliveries, and Time or the column indicators C1, C2, and C3. The following steps describe how to use Minitab to produce the regression results shown in Figure 15.4.

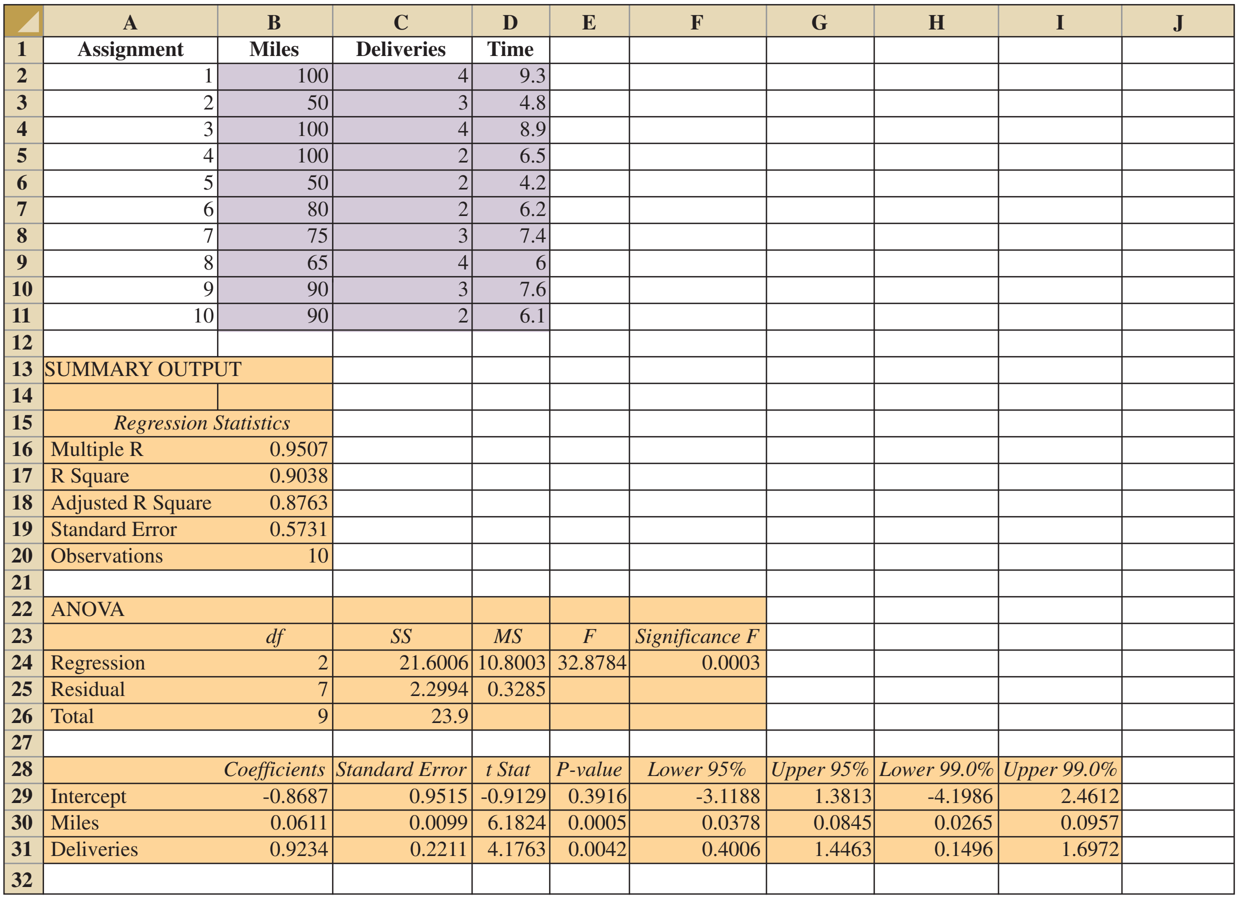
Step 1. Select the **Stat** menu
Step 2. Select the **Regression** menu
Step 3. Choose **Regression**
Step 4. When the **Regression** dialog box appears:
Enter Time in the **Response** box
Enter Miles and Deliveries in the **Predictors** box
Click **OK**

Appendix 15.2 Multiple Regression with Excel

WEB file
Butler

In Section 15.2 we discussed the computer solution of multiple regression problems by showing Minitab's output for the Butler Trucking Company problem. In this appendix we describe how to use Excel's Regression tool to develop the estimated multiple regression equation for the Butler Trucking problem. Refer to Figure 15.14 as we describe the tasks involved. First, the labels Assignment, Miles, Deliveries, and Time are entered into cells A1:D1 of the worksheet, and the sample data into cells B2:D11. The numbers 1–10 in cells A2:A11 identify each observation.

FIGURE 15.14 EXCEL OUTPUT FOR BUTLER TRUCKING WITH TWO INDEPENDENT VARIABLES



	A	B	C	D	E	F	G	H	I	J
1	Assignment	Miles	Deliveries	Time						
2	1	100	4	9.3						
3	2	50	3	4.8						
4	3	100	4	8.9						
5	4	100	2	6.5						
6	5	50	2	4.2						
7	6	80	2	6.2						
8	7	75	3	7.4						
9	8	65	4	6						
10	9	90	3	7.6						
11	10	90	2	6.1						
12										
13	SUMMARY OUTPUT									
14										
15	*Regression Statistics*									
16	Multiple R	0.9507								
17	R Square	0.9038								
18	Adjusted R Square	0.8763								
19	Standard Error	0.5731								
20	Observations	10								
21										
22	ANOVA									
23		*df*	*SS*	*MS*	*F*	*Significance F*				
24	Regression	2	21.6006	10.8003	32.8784	0.0003				
25	Residual	7	2.2994	0.3285						
26	Total	9	23.9							
27										
28		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*	
29	Intercept	-0.8687	0.9515	-0.9129	0.3916	-3.1188	1.3813	-4.1986	2.4612	
30	Miles	0.0611	0.0099	6.1824	0.0005	0.0378	0.0845	0.0265	0.0957	
31	Deliveries	0.9234	0.2211	4.1763	0.0042	0.4006	1.4463	0.1496	1.6972	
32										

The following steps describe how to use the Regression tool for the multiple regression analysis.

Step 1. Click the **Data** tab on the Ribbon
Step 2. In the **Analysis** group, click **Data Analysis**
Step 3. Choose **Regression** from the list of Analysis Tools
Step 4. When the Regression dialog box appears:
Enter D1:D11 in the **Input Y Range** box
Enter B1:C11 in the **Input X Range** box
Select **Labels**
Select **Confidence Level**
Enter 99 in the **Confidence Level** box
Select **Output Range**
Enter A13 in the **Output Range** box (to identify the upper left corner of the section of the worksheet where the output will appear)
Click **OK**

In the Excel output shown in Figure 15.14 the label for the independent variable x_1 is Miles (see cell A30), and the label for the independent variable x_2 is Deliveries (see cell A31). The estimated regression equation is

$$\hat{y} = -.8687 + .0611x_1 + .9234x_2$$

Note that using Excel's Regression tool for multiple regression is almost the same as using it for simple linear regression. The major difference is that in the multiple regression case a larger range of cells is required in order to identify the independent variables.

Appendix 15.3 Logistic Regression with Minitab

WEB file

Simmons

Minitab calls logistic regression with a dependent variable that can only assume the values 0 and 1 Binary Logistic Regression. In this appendix we describe the steps required to use Minitab's Binary Logistic Regression procedure to generate the computer output for the Simmons Stores problem shown in Figure 15.13. First, the data must be entered in a Minitab worksheet. The amounts customers spent last year at Simmons (in thousands of dollars) are entered into column C2, the credit card data (1 if a Simmons card; 0 otherwise) are entered into column C3, and the coupon use data (1 if the customer used the coupon; 0 otherwise) are entered in column C4. The variable names Spending, Card, and Coupon are entered as the column headings on the worksheet. In subsequent steps, we refer to the data by using the variable names Spending, Card, and Coupon or the column indicators C2, C3, and C4. The following steps will generate the logistic regression output.

Step 1. Select the **Stat** menu
Step 2. Select the **Regression** menu
Step 3. Choose **Binary Logistic Regression**
Step 4. When the **Binary Logistic Regression** dialog box appears:
Enter Coupon in the **Response** box
Enter Spending and Card in the **Model** box
Click **OK**

The information in Figure 15.13 will now appear as a portion of the output.

Appendix 15.4 Multiple Regression Analysis Using StatTools

WEB file

Butler

In this appendix we show how StatTools can be used to perform the regression analysis computations for the Butler Trucking problem. Begin by using the Data Set Manager to create a StatTools data set for these data using the procedure described in the appendix in Chapter 1. The following steps describe how StatTools can be used to provide the regression results.

Step 1. Click the **StatTools** tab on the Ribbon
Step 2. In the **Analyses** group, click **Regression and Classification**
Step 3. Choose the **Regression** option
Step 4. When the StatTools—Regression dialog box appears:
Select **Multiple** in the **Regression Type** box
In the **Variables** section:
Click the **Format** button and select **Unstacked**
In the column labeled **I** select **Miles**
In the column labeled **I** select **Deliveries**
In the column labeled **D** select **Time**
Click **OK**

The regression analysis output will appear in a new worksheet.

The StatTools—Regression dialog box contains a number of more advanced options for developing prediction interval estimates and producing residual plots. The StatTools Help facility provides information on using all of these options.