In Chapter 12 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.

13.1

Μ

# **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.

ULTIPLE REGRESSION MODEL  

$$v = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \epsilon$$
(13.1)

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

In Section 13.4 we will discuss the assumptions for the multiple regression model and  $\epsilon$ . One of the assumptions is that the mean or expected value of  $\epsilon$  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

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n$$
 (13.2)

# **Estimated Multiple Regression Equation**

If the values of  $\beta_0$ ,  $\beta_1$ ,  $\beta_2$ , ...,  $\beta_p$  were known, equation (13.2) could be used to compute the mean value of y at given values of  $x_1, x_2, ..., x_p$ . 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  $b_0, b_1, b_2, ..., b_p$  that are used as the point



FIGURE 13.1 THE ESTIMATION PROCESS FOR MULTIPLE REGRESSION

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



where

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

The estimation process for multiple regression is shown in Figure 13.1.

13.2

# Least Squares Method

In Chapter 12, 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

 $\min \Sigma (y_i - \hat{y}_i)^2$  (13.4)

#### where

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

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

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

As expression (13.4) 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 12 we presented formulas for computing the least squares estimators  $b_0$  and  $b_1$  for the estimated simple linear regression equation  $\hat{y} = b_0 + b_1 x$ . 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 13.1 and the scatter diagram shown in Figure 13.2. After reviewing this scatter diagram, the managers hypothesized that the simple linear regression model  $y = \beta_0 + \beta_1 x_1 + \epsilon$  could be used to describe the relationship between the total travel time (y) and the number of miles traveled (x<sub>1</sub>). To estimate

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

#### TABLE 13.1 PRELIMINARY DATA FOR BUTLER TRUCKING

WEB IILE

Butler

FIGURE 13.2 SCATTER DIAGRAM OF PRELIMINARY DATA FOR BUTLER TRUCKING



the parameters  $\beta_0$  and  $\beta_1$ , the least squares method was used to develop the estimated regression equation.

$$\hat{y} = b_0 + b_1 x_1 \tag{13.5}$$

In Figure 13.3, we show the Minitab computer output from applying simple linear regression to the data in Table 13.1. 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 13.2. The Minitab computer solution with both miles traveled  $(x_1)$  and number of deliveries  $(x_2)$  as independent variables is shown in Figure 13.4. The estimated regression equation is

$$\hat{y} = -.869 + .0611x_1 + .923x_2 \tag{13.6}$$

#### 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.

#### FIGURE 13.3 MINITAB OUTPUT FOR BUTLER TRUCKING WITH ONE INDEPENDENT VARIABLE

The regression equation is Time = 1.27 + 0.0678 Miles						
Predictor C Constant 1.	Coef         SE         Coef           274         1.401           2782         0.01706	T p 0.91 0.390				
S = 1.00179 R	1.83 = 0.01700 1.8q = 66.4%	R-sq(adj) =	62.2%			
Analysis of Variance						
SOURCE	DF SS	MS I	α			
Regression	1 15.871	15.871 15.82	0.004			
Residual Error	8 8.029	1.004				
Total	9 23.900					

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 (13.6).

# **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, represents an estimate of the change in y corresponding to a one-unit change in x, when all other independent variables are held constant. In the Butler Trucking example involving two independent variables,  $b_1 = .0611$ . Thus,

### **TABLE 13.2** DATA FOR BUTLER TRUCKING WITH MILES TRAVELED $(x_1)$ AND NUMBER OF DELIVERIES $(x_2)$ AS THE INDEPENDENT VARIABLES

	Driving Assignment	$x_1 = Miles$ Traveled	x <sub>2</sub> = 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
ər	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

WEB

# FIGURE 13.4 MINITAB OUTPUT FOR BUTLER TRUCKING WITH TWO INDEPENDENT VARIABLES

The regression equation is Time = - 0.869 + 0.0611 Miles + 0.923 Deliveries						
Predictor	Coef	SE Co	bef	Т	р	
Constant -0	.8687	0.9	515 -0.	91 0.3	92	
Miles 0.0	61135	0.0098	888 6.	18 0.0	00	
Deliveries 0	.9234	0.22	211 4.	18 0.0	04	
S = 0.573142 R-sq = 90.4% R-sq(adj) = 87.6%						
Analysis of Variance						
SOURCE	DF	SS	MS	F	р	
Regression	2 2	1.601	10.800	32.88	0.000	
Residual Error	7	2.299	0.328			
Total	9 2	3.900				

.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  $x_2 = 310$ .
- 2. Consider the following data for a dependent variable y and two independent variables,  $x_1$  and  $x_2$ .

WEB	file	
	Exer2	

SELF LC

	У	$x_2$	$x_1$	
	94	12	30	
	108	10	47	
	112	17	25	
	178	16	51	
	94	5	40	
	175	19	51	
	170	7	74	
(continued)				

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.

Copyright 2010 Cengage Learning. All Rights Reserved. May not be copied, scanned, or duplicated, in whole or in part. Due to electronic rights, some third party content may be suppressed from the eBook and/or eChapter(s). Editorial review has deemed that any suppressed content does not materially affect the overall learning experience. Cengage Learning reserves the right to remove additional content at any time if subsequent rights restrictions require it.