variance (Chapters 14 through 18). As you study these seven chapters, try whenever possible to make the connection back to a general linear model; we'll help you with this connection. For Sections 12.3 through 12.10 of this chapter, we will concentrate on multiple regression, which is a special case of a general linear model.

12.3 Estimating Multiple Regression Coefficients

The multiple regression model relates a response y to a set of quantitative independent variables. For a random sample of n measurements, we can write the ith observation as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \qquad (i = 1, 2, \ldots, n; n > k)$$

where $x_{i1}, x_{i2}, \ldots, x_{ik}$ are the settings of the quantitative independent variables corresponding to the observation y_i.

To find least-squares estimates for $\beta_0, \beta_1, \ldots,$ and β_k in a multiple regression model, we follow the same procedure that we did for a linear regression model in Chapter 11. We obtain a random sample of n observations; we find the least-squares prediction equation

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$$

by choosing $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ to minimize SS(Residual) $= \Sigma_i (y_i - \hat{y}_i)^2$. However, although it was easy to write down the solutions to $\hat{\beta}_0$ and $\hat{\beta}_1$ for the linear regression model,

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

we must find the estimates for $\beta_0, \beta_1, \ldots, \beta_k$ by solving a set of simultaneous equations, called the *normal equations,* shown in Table 12.5.

TABLE 12.5 Normal equations for a multiple regression model

	y_i	$\hat{\beta}_0$	$x_{i1}\hat{\beta}_1$	$\cdots$	$x_{ik}\hat{\beta}_k$
1	$\Sigma y_i =$	$n\hat{\beta}_0$	$+ \Sigma x_{i1}\hat{\beta}_1$	$+ \cdots +$	$\Sigma x_{ik}\hat{\beta}_k$
x_{i1}	$\Sigma x_{i1} y_i =$	$\Sigma x_{i1}\hat{\beta}_0$	$+ \Sigma x_{i1}^2\hat{\beta}_1$	$+ \cdots +$	$\Sigma x_{i1} x_{ik}\hat{\beta}_k$
$\vdots$	$\vdots$				
x_{ik}	$\Sigma x_{ik} y_i =$	$\Sigma x_{ik}\hat{\beta}_0$	$+ \Sigma x_{ik} x_{i1}\hat{\beta}_1$	$+ \cdots +$	$\Sigma x_{ik}^2\hat{\beta}_k$

Note the pattern associated with these equations. By labeling the rows and columns as we have done, we can obtain any term in the normal equations by multiplying the row and column elements and summing. For example, the last term in the second equation is found by multiplying the row element (x_{i1}) by the column element ($x_{ik}\hat{\beta}_k$) and summing; the resulting term is $\Sigma x_{i1} x_{ik} \hat{\beta}_k$. Because all terms in the normal equations can be formed in this way, it is fairly simple to write down the equations to be solved to obtain the least-squares estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$. The solution to these equations is not necessarily trivial; that's why we'll enlist the help of various statistical software packages for their solution.

EXAMPLE 12.5

An experiment was conducted to investigate the weight loss of a compound for different amounts of time the compound was exposed to the air. Additional information was also available on the humidity of the environment during exposure. The complete data are presented in Table 12.6.

TABLE 12.6
, exposure time,
'e humidity data

Weight Loss, y (pounds)	Exposure Time, x_1 (hours)	Relative Humidity, x_2
4.3	4	.20
5.5	5	.20
6.8	6	.20
8.0	7	.20
4.0	4	.30
5.2	5	.30
6.6	6	.30
7.5	7	.30
2.0	4	.40
4.0	5	.40
5.7	6	.40
6.5	7	.40

a. Set up the normal equations for this regression problem if the assumed model is

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

where x_1 is exposure time and x_2 is relative humidity.

b. Use the computer output shown here to determine the least-squares estimates of β_0, β_1, and β_2. Predict weight loss for 6.5 hours of exposure and a relative humidity of .35.

OUTPUT FOR EXAMPLE 12.5

OBS	WT_LOSS	TIME	HUMID
1	4.3	4.0	0.20
2	5.5	5.0	0.20
3	6.8	6.0	0.20
4	8.0	7.0	0.20
5	4.0	4.0	0.30
6	5.2	5.0	0.30
7	6.6	6.0	0.30
8	7.5	7.0	0.30
9	2.0	4.0	0.40
10	4.0	5.0	0.40
11	5.7	6.0	0.40
12	6.5	7.0	0.40
13	.	6.5	0.35

Dependent Variable: WT_LOSS   WEIGHT LOSS

P638

```
Analysis of Variance

                         Sum of         Mean
Source          DF      Squares       Square      F Value       Prob>F

Model            2     31.12417     15.56208      104.133       0.0001
Error            9      1.34500      0.14944
C Total         11     32.46917

     Root MSE       0.38658      R-square      0.9586
     Dep Mean       5.50833      Adj R-sq      0.9494
     C.V.           7.01810

Parameter Estimates

                    Parameter       Standard     T for H0:
Variable    DF      Estimate           Error     Parameter=0     Prob > |T|

INTERCEP     1      0.666667      0.69423219           0.960         0.3620
TIME         1      1.316667      0.09981464          13.191         0.0001
HUMID        1     -8.000000      1.36676829          -5.853         0.0002

OBS   WT_LOSS      PRED       RESID     L95MEAN    U95MEAN

  1     4.3      4.33333   -0.03333    3.80985    4.85682
  2     5.5      5.65000   -0.15000    5.23519    6.06481
  3     6.8      6.96667   -0.16667    6.55185    7.38148
  4     8.0      8.28333   -0.28333    7.75985    8.80682
  5     4.0      3.53333    0.46667    3.11091    3.95576
  6     5.2      4.85000    0.35000    4.57346    5.12654
  7     6.6      6.16667    0.43333    5.89012    6.44321
  8     7.5      7.48333    0.01667    7.06091    7.90576
  9     2.0      2.73333   -0.73333    2.20985    3.25682
 10     4.0      4.05000   -0.05000    3.63519    4.46481
 11     5.7      5.36667    0.33333    4.95185    5.78148
 12     6.5      6.68333   -0.18333    6.15985    7.20682
 13      .       6.42500     .         6.05269    6.79731

Sum of Residuals                         0
Sum of Squared Residuals            1.3450
Predicted Resid SS (Press)          2.6123
```

Solution

a. The three normal equations for this model are shown in Table 12.7.

12.7
tions
12.5

	y_i		$\hat{\beta}_0$		$x_{i1}\hat{\beta}_1$		$x_{i2}\hat{\beta}_2$
1	$\sum y_i$	$=$	$n\hat{\beta}_0$	$+$	$\sum x_{i1}\hat{\beta}_1$	$+$	$\sum x_{i2}\hat{\beta}_2$
x_{i1}	$\sum x_{i1}y_i$	$=$	$\sum x_{i1}\hat{\beta}_0$	$+$	$\sum x_{i1}^2\hat{\beta}_1$	$+$	$\sum x_{i1}x_{i2}\hat{\beta}_2$
x_{i2}	$\sum x_{i2}y_i$	$=$	$\sum x_{i2}\hat{\beta}_0$	$+$	$\sum x_{i2}x_{i1}\hat{\beta}_1$	$+$	$\sum x_{i2}^2\hat{\beta}_2$

For these data, we have

$$\sum y_i = 66.10 \qquad \sum x_{i1} = 66 \qquad \sum x_{i2} = 3.60$$

$$\sum x_{i1}y_i = 383.3 \qquad \sum x_{i2}y_i = 19.19 \qquad \sum x_{i1}x_{i2} = 19.8$$

$$\sum x_{i1}^2 = 378 \qquad \sum x_{i2}^2 = 1.16$$

P 639

Substituting these values into the normal equation yields the result shown here:

$$66.1 = 12\hat{\beta}_0 + 66\hat{\beta}_1 + 3.6\hat{\beta}_2$$
$$383.3 = 66\hat{\beta}_0 + 378\hat{\beta}_1 + 19.8\hat{\beta}_2$$
$$19.19 = 3.6\hat{\beta}_0 + 19.8\hat{\beta}_1 + 1.16\hat{\beta}_2$$

b. The normal equations of part (a) could be solved to determine $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$. The solution would agree with that shown here in the output. The least-squares prediction equation is

$$\hat{y} = 0.667 + 1.317x_1 - 8.000x_2$$

where x_1 is exposure time and x_2 is relative humidity. Substituting $x_1 = 6.5$ and $x_2 = .35$, we have

$$\hat{y} = 0.667 + 1.317(6.5) - 8.000(.35) = 6.428$$

This value agrees with the predicted value shown as observation 13 in the output, except for rounding errors.

There are many software programs that provide the calculations to obtain least-squares estimates for parameters in the general linear model (and hence for multiple regression). The output of such programs typically has a list of variable names, together with the estimated partial slopes, labeled COEFFICIENTS (or ESTIMATES or PARAMETERS). The intercept term $\hat{\beta}_0$ is usually called INTERCEPT (or CONSTANT); sometimes it is shown along with the slopes but with no variable name.

EXAMPLE 12.6

A kinesiologist is investigating measures of the physical fitness of persons entering 10-kilometer races. A major component of overall fitness is cardiorespiratory capacity as measured by maximal oxygen uptake. Direct measurement of maximal oxygen is expensive, and thus is difficult to apply to large groups of individuals in a timely fashion. The researcher wanted to determine if a prediction of maximal oxygen uptake can be obtained from a prediction equation using easily measured explanatory variables from the runners. In a preliminary study, the kinesiologist randomly selects 50 males and obtains the following data for the variables:

y = maximal oxygen uptake (in liters per minute)

x_1 = weight (in kilograms)

x_2 = age (in years)

x_3 = time necessary to walk 1 mile (in minutes)

x_4 = heart rate at end of the walk (in beats per minute)

The data shown in Table 12.8 were simulated from a model that is consistent with information given in the article "Validation of the Rockport Fitness Walking Test in College Males and Females," *Research Quarterly for Exercise and Sport* (1994) 152–158.

TABL

Fitness walking t