12 Multiple Linear Regression

CHAPTER OUTLINE

LEARNING OBJECTIVES

After careful study of this chapter, you should be able to do the following:

1. Use multiple regression techniques to build empirical models to engineering and scientific data
2. Understand how the method of least squares extends to fitting multiple regression models

3. **Assess regression model adequacy**
4. **Test hypotheses and construct confidence intervals on the regression coefficients**
5. **Use the regression model to estimate the mean response and to make predictions and to construct confidence intervals and prediction intervals**
6. **Build regression models with polynomial terms**
7. **Use indicator variables to model categorical regressors**
8. **Use stepwise regression and other model building techniques to select the appropriate set of variables for a regression model**

CD MATERIAL

9. **Understand how ridge regression provides an effective way to estimate model parameters where there is multicollinearity.**
10. **Understand the basic concepts of fitting a nonlinear regression model.**

Answers for many odd numbered exercises are at the end of the book. Answers to exercises whose numbers are surrounded by a box can be accessed in the e-Text by clicking on the box. Complete worked solutions to certain exercises are also available in the e-Text. These are indicated in the Answers to Selected Exercises section by a box around the exercise number. Exercises are also available for some of the text sections that appear on CD only. These exercises may be found within the e-Text immediately following the section they accompany.

12-1 MULTIPLE LINEAR REGRESSION MODEL

12-1.1 Introduction

Many applications of regression analysis involve situations in which there are more than one regressor variable. A regression model that contains more than one regressor variable is called a **multiple regression model.**

As an example, suppose that the effective life of a cutting tool depends on the cutting speed and the tool angle. A multiple regression model that might describe this relationship is

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon \qquad (12\text{-}1)$$

where Y represents the tool life, x_1 represents the cutting speed, x_2 represents the tool angle, and ϵ is a random error term. This is a **multiple linear regression model** with two regressors. The term *linear* is used because Equation 12-1 is a linear function of the unknown parameters β_0, β_1, and β_2.

The regression model in Equation 12-1 describes a plane in the three-dimensional space of Y, x_1, and x_2. Figure 12-1(a) shows this plane for the regression model

$$E(Y) = 50 + 10x_1 + 7x_2$$

where we have assumed that the expected value of the error term is zero; that is $E(\epsilon) = 0$. The parameter β_0 is the **intercept** of the plane. We sometimes call β_1 and β_2 **partial regression coefficients,** because β_1 measures the expected change in Y per unit change in x_1 when x_2 is held constant, and β_2 measures the expected change in Y per unit change in x_2 when x_1 is held constant. Figure 12-1(b) shows a **contour plot** of the regression model—that is, lines of constant $E(Y)$ as a function of x_1 and x_2. Notice that the contour lines in this plot are straight lines.

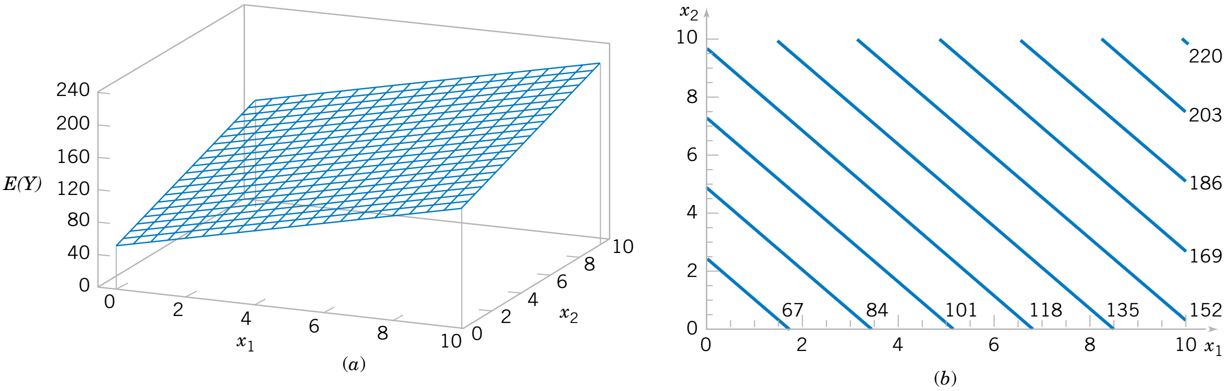

Figure 12-1 (a) The regression plane for the model $E(Y) = 50 + 10x_1 + 7x_2$. (b) The contour plot.

In general, the **dependent variable** or **response** Y may be related to k **independent** or **regressor variables.** The model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon \tag{12-2}$$

is called a multiple linear regression model with k regressor variables. The parameters β_j, $j = 0, 1, \ldots, k$, are called the regression coefficients. This model describes a hyperplane in the k-dimensional space of the regressor variables $\{x_j\}$. The parameter β_j represents the expected change in response Y per unit change in x_j when all the remaining regressors x_i $(i \neq j)$ are held constant.

Multiple linear regression models are often used as approximating functions. That is, the true functional relationship between Y and $x_1, x_2, \ldots, x_k$ is unknown, but over certain ranges of the independent variables the linear regression model is an adequate approximation.

Models that are more complex in structure than Equation 12-2 may often still be analyzed by multiple linear regression techniques. For example, consider the cubic polynomial model in one regressor variable.

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon \tag{12-3}$$

If we let $x_1 = x$, $x_2 = x^2$, $x_3 = x^3$, Equation 12-3 can be written as

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon \tag{12-4}$$

which is a multiple linear regression model with three regressor variables.

Models that include **interaction** effects may also be analyzed by multiple linear regression methods. An interaction between two variables can be represented by a cross-product term in the model, such as

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon \tag{12-5}$$

If we let $x_3 = x_1 x_2$ and $\beta_3 = \beta_{12}$, Equation 12-5 can be written as

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

which is a linear regression model.

Figure 12-2(a) and (b) shows the three-dimensional plot of the regression model

$$Y = 50 + 10x_1 + 7x_2 + 5x_1 x_2$$

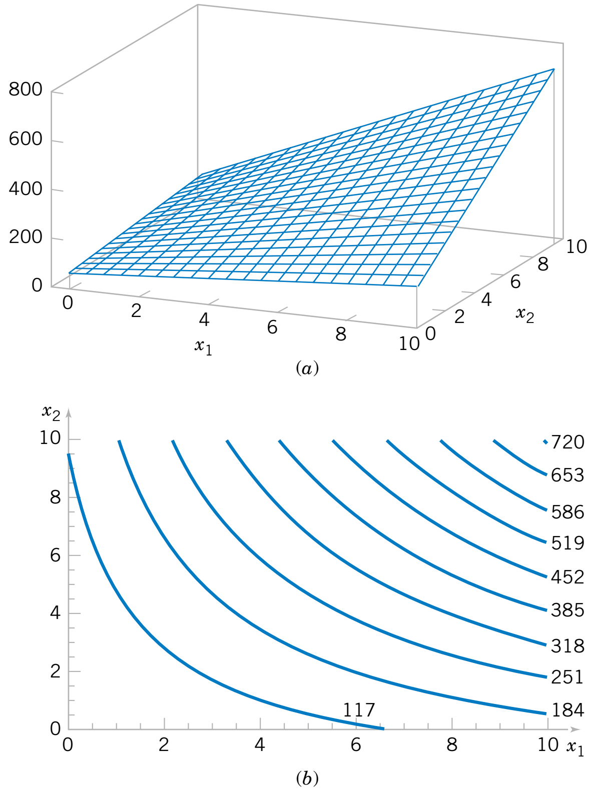

Figure 12-2 (a) Three-dimensional plot of the regression model $E(Y) = 50 + 10x_1 + 7x_2 + 5x_1x_2$. (b) The contour plot.

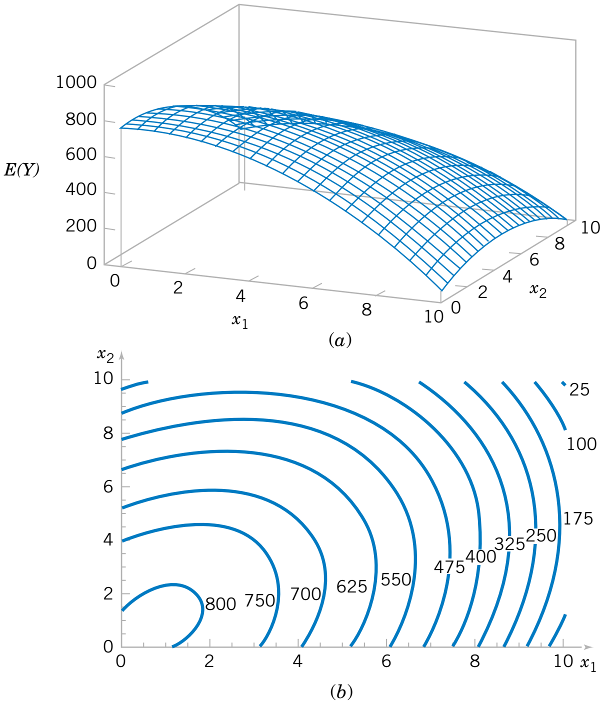

Figure 12-3 (a) Three-dimensional plot of the regression model $E(Y) = 800 + 10x_1 + 7x_2 - 8.5x_1^2 - 5x_2^2 + 4x_1x_2$. (b) The contour plot.

and the corresponding two-dimensional contour plot. Notice that, although this model is a linear regression model, the shape of the surface that is generated by the model is not linear. In general, **any regression model that is linear in parameters** (the β's) **is a linear regression model, regardless of the shape of the surface that it generates.**

Figure 12-2 provides a nice graphical interpretation of an interaction. Generally, interaction implies that the effect produced by changing one variable (x_1, say) depends on the level of the other variable (x_2). For example, Fig. 12-2 shows that changing x_1 from 2 to 8 produces a much smaller change in $E(Y)$ when $x_2 = 2$ than when $x_2 = 10$. Interaction effects occur frequently in the study and analysis of real-world systems, and regression methods are one of the techniques that we can use to describe them.

As a final example, consider the second-order model with interaction

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \epsilon \tag{12-6}$$

If we let $x_3 = x_1^2$, $x_4 = x_2^2$, $x_5 = x_1x_2$, $\beta_3 = \beta_{11}$, $\beta_4 = \beta_{22}$, and $\beta_5 = \beta_{12}$, Equation 12-6 can be written as a multiple linear regression model as follows:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \epsilon$$

Figure 12-3(a) and (b) show the three-dimensional plot and the corresponding contour plot for

$$E(Y) = 800 + 10x_1 + 7x_2 - 8.5x_1^2 - 5x_2^2 + 4x_1x_2$$

These plots indicate that the expected change in Y when x_1 is changed by one unit (say) is a function of *both* x_1 and x_2. The quadratic and interaction terms in this model produce a mound-shaped function. Depending on the values of the regression coefficients, the second-order model with interaction is capable of assuming a wide variety of shapes; thus, it is a very flexible regression model.

12-1.2 Least Squares Estimation of the Parameters

The **method of least squares** may be used to estimate the regression coefficients in the multiple regression model, Equation 12-2. Suppose that $n > k$ observations are available, and let x_{ij} denote the ith observation or level of variable x_j. The observations are

$$(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i), \qquad i = 1, 2, \ldots, n \quad \text{and} \quad n > k$$

It is customary to present the data for multiple regression in a table such as Table 12-1.

Each observation $(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i)$, satisfies the model in Equation 12-2, or

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i \\ &= \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \epsilon_i \qquad i = 1, 2, \ldots, n \end{aligned} \tag{12-7}$$

The least squares function is

$$L = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2 \tag{12-8}$$

We want to minimize L with respect to $\beta_0, \beta_1, \ldots, \beta_k$. The **least squares estimates** of $\beta_0, \beta_1, \ldots, \beta_k$ must satisfy

$$\left.\frac{\partial L}{\partial \beta_0}\right|_{\hat\beta_0, \hat\beta_1, \ldots, \hat\beta_k} = -2 \sum_{i=1}^{n} \left(y_i - \hat\beta_0 - \sum_{j=1}^{k} \hat\beta_j x_{ij} \right) = 0 \tag{12-9a}$$

and

$$\left.\frac{\partial L}{\partial \beta_j}\right|_{\hat\beta_0, \hat\beta_1, \ldots, \hat\beta_k} = -2 \sum_{i=1}^{n} \left(y_i - \hat\beta_0 - \sum_{j=1}^{k} \hat\beta_j x_{ij} \right) x_{ij} = 0 \quad j = 1, 2, \ldots, k \tag{12-9b}$$

Simplifying Equation 12-9, we obtain the **least squares normal Equations**

$$\begin{array}{lllll} n\hat\beta_0 + \hat\beta_1 \sum_{i=1}^{n} x_{i1} & + \hat\beta_2 \sum_{i=1}^{n} x_{i2} & + \cdots + \hat\beta_k \sum_{i=1}^{n} x_{ik} & = \sum_{i=1}^{n} y_i \\ \hat\beta_0 \sum_{i=1}^{n} x_{i1} + \hat\beta_1 \sum_{i=1}^{n} x_{i1}^2 & + \hat\beta_2 \sum_{i=1}^{n} x_{i1} x_{i2} & + \cdots + \hat\beta_k \sum_{i=1}^{n} x_{i1} x_{ik} & = \sum_{i=1}^{n} x_{i1} y_i \\ \vdots \qquad\quad \vdots & \quad \vdots & \quad\vdots \qquad \vdots & \\ \hat\beta_0 \sum_{i=1}^{n} x_{ik} + \hat\beta_1 \sum_{i=1}^{n} x_{ik} x_{i1} & + \hat\beta_2 \sum_{i=1}^{n} x_{ik} x_{i2} & + \cdots + \hat\beta_k \sum_{i=1}^{n} x_{ik}^2 & = \sum_{i=1}^{n} x_{ik} y_i \end{array} \tag{12-10}$$

Note that there are $p = k + 1$ normal Equations, one for each of the unknown regression coefficients. The solution to the normal Equations will be the **least squares estimators** of the

Table 12-1 Data for Multiple Linear Regression

y	x_1	x_2	$\ldots$	x_k
y_1	x_{11}	x_{12}	$\ldots$	x_{1k}
y_2	x_{21}	x_{22}	$\ldots$	x_{2k}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
y_n	x_{n1}	x_{n2}	$\ldots$	x_{nk}

regression coefficients, $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$. The normal Equations can be solved by any method appropriate for solving a system of linear Equations.

EXAMPLE 12-1

In Chapter 1, we used data on pull strength of a wire bond in a semiconductor manufacturing process, wire length, and die height to illustrate building an empirical model. We will use the same data, repeated for convenience in Table 12-2, and show the details of estimating the model parameters. A three-dimensional scatter plot of the data is presented in Fig. 1-13. Figure 12-4 shows a matrix of two-dimensional scatter plots of the data. These displays can be helpful in visualizing the relationships among variables in a multivariable data set.

Specifically, we will fit the multiple linear regression model

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

where Y = pull strength, x_1 = wire length, and x_2 = die height. From the data in Table 12-2 we calculate

$$n = 25, \quad \sum_{i=1}^{25} y_i = 725.82$$

$$\sum_{i=1}^{25} x_{i1} = 206, \quad \sum_{i=1}^{25} x_{i2} = 8{,}294$$

$$\sum_{i=1}^{25} x_{i1}^2 = 2{,}396, \quad \sum_{i=1}^{25} x_{i2}^2 = 3{,}531{,}848$$

$$\sum_{i=1}^{25} x_{i1}x_{i2} = 77{,}177, \quad \sum_{i=1}^{25} x_{i1}y_i = 8{,}008.37, \quad \sum_{i=1}^{25} x_{i2}y_i = 274{,}811.31$$

Table 12-2 Wire Bond Data for Example 11-1

Observation Number	Pull Strength y	Wire Length x_1	Die Height x_2	Observation Number	Pull Strength y	Wire Length x_1	Die Height x_2
1	9.95	2	50	14	11.66	2	360
2	24.45	8	110	15	21.65	4	205
3	31.75	11	120	16	17.89	4	400
4	35.00	10	550	17	69.00	20	600
5	25.02	8	295	18	10.30	1	585
6	16.86	4	200	19	34.93	10	540
7	14.38	2	375	20	46.59	15	250
8	9.60	2	52	21	44.88	15	290
9	24.35	9	100	22	54.12	16	510
10	27.50	8	300	23	56.63	17	590
11	17.08	4	412	24	22.13	6	100
12	37.00	11	400	25	21.15	5	400
13	41.95	12	500				

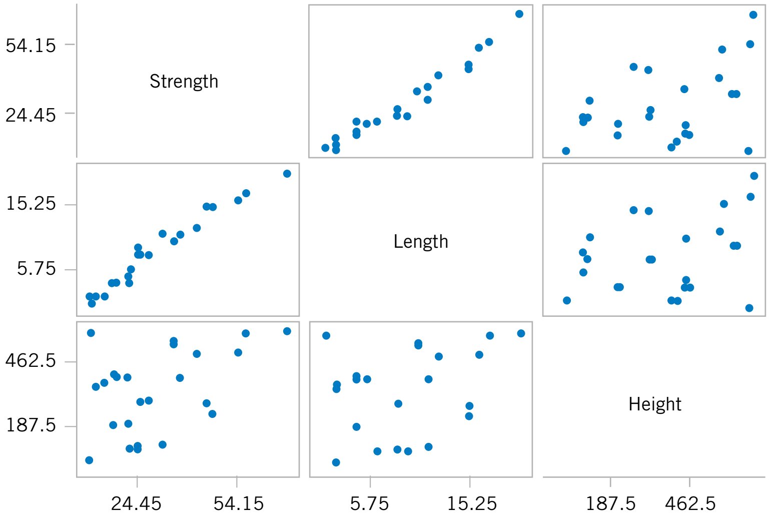

Figure 12-4 Matrix of scatter plots (from Minitab) for the wire bond pull strength data in Table 12-2.

For the model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$, the normal Equations 12-10 are

$$\begin{aligned}
n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} \quad &+ \hat{\beta}_2 \sum_{i=1}^{n} x_{i2} \quad = \sum_{i=1}^{n} y_i \\
\hat{\beta}_0 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}^2 \quad &+ \hat{\beta}_2 \sum_{i=1}^{n} x_{i1}x_{i2} = \sum_{i=1}^{n} x_{i1}y_i \\
\hat{\beta}_0 \sum_{i=1}^{n} x_{i2} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}x_{i2} &+ \hat{\beta}_2 \sum_{i=1}^{n} x_{i2}^2 \quad = \sum_{i=1}^{n} x_{i2}y_i
\end{aligned}$$

Inserting the computed summations into the normal equations, we obtain

$$\begin{aligned}
25\hat{\beta}_0 + \quad 206\hat{\beta}_1 + \quad 8294\hat{\beta}_2 &= 725.82 \\
206\hat{\beta}_0 + \quad 2396\hat{\beta}_1 + \quad 77{,}177\hat{\beta}_2 &= 8{,}008.37 \\
8294\hat{\beta}_0 + 77{,}177\hat{\beta}_1 + 3{,}531{,}848\hat{\beta}_2 &= 274{,}811.31
\end{aligned}$$

The solution to this set of equations is

$$\hat{\beta}_0 = 2.26379, \quad \hat{\beta}_1 = 2.74427, \quad \hat{\beta}_2 = 0.01253$$

Therefore, the fitted regression equation is

$$\hat{y} = 2.26379 + 2.74427x_1 + 0.01253x_2$$

This equation can be used to predict pull strength for pairs of values of the regressor variables wire length (x_1) and die height (x_2). This is essentially the same regression model given in Equation 1-7, Section 1-3. Figure 1-14 shows a three-dimentional plot of the plane of predicted values $\hat{y}$ generated from this equation.

12-1.3 Matrix Approach to Multiple Linear Regression

In fitting a multiple regression model, it is much more convenient to express the mathematical operations using **matrix notation.** Suppose that there are k regressor variables and n observations, $(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i)$, $i = 1, 2, \ldots, n$ and that the model relating the regressors to the response is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i \qquad i = 1, 2, \ldots, n$$

This model is a system of n equations that can be expressed in matrix notation as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{12-11}$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \quad \text{and} \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

In general, $\mathbf{y}$ is an $(n \times 1)$ vector of the observations, $\mathbf{X}$ is an $(n \times p)$ matrix of the levels of the independent variables, $\boldsymbol{\beta}$ is a $(p \times 1)$ vector of the regression coefficients, and $\boldsymbol{\epsilon}$ is a $(n \times 1)$ vector of random errors.

We wish to find the vector of least squares estimators, $\hat{\boldsymbol{\beta}}$, that minimizes

$$L = \sum_{i=1}^{n} \epsilon_i^2 = \boldsymbol{\epsilon}'\boldsymbol{\epsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

The least squares estimator $\hat{\boldsymbol{\beta}}$ is the solution for $\boldsymbol{\beta}$ in the equations

$$\frac{\partial L}{\partial \boldsymbol{\beta}} = \mathbf{0}$$

We will not give the details of taking the derivatives above; however, the resulting equations that must be solved are

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y} \tag{12-12}$$

Equations 12-12 are the least squares normal equations in matrix form. They are identical to the scalar form of the normal equations given earlier in Equations 12-10. To solve the normal equations, multiply both sides of Equations 12-12 by the inverse of $\mathbf{X}'\mathbf{X}$. Therefore, the least squares estimate of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{y} \tag{12-13}$$

Note that there are $p = k + 1$ normal equations in $p = k + 1$ unknowns (the values of $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$). Furthermore, the matrix $\mathbf{X}'\mathbf{X}$ is always nonsingular, as was assumed above, so the methods described in textbooks on determinants and matrices for inverting these matrices can be used to find $(\mathbf{X}'\mathbf{X})^{-1}$. In practice, multiple regression calculations are almost always performed using a computer.

It is easy to see that the matrix form of the normal equations is identical to the scalar form. Writing out Equation 12-12 in detail, we obtain

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i2} & \cdots & \sum_{i=1}^{n} x_{ik} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}^2 & \sum_{i=1}^{n} x_{i1}x_{i2} & \cdots & \sum_{i=1}^{n} x_{i1}x_{ik} \\ \vdots & \vdots & \vdots & & \vdots \\ \sum_{i=1}^{n} x_{ik} & \sum_{i=1}^{n} x_{ik}x_{i1} & \sum_{i=1}^{n} x_{ik}x_{i2} & \cdots & \sum_{i=1}^{n} x_{ik}^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{i1}y_i \\ \vdots \\ \sum_{i=1}^{n} x_{ik}y_i \end{bmatrix}$$

If the indicated matrix multiplication is performed, the scalar form of the normal equations (that is, Equation 12-10) will result. In this form it is easy to see that $\mathbf{X}'\mathbf{X}$ is a $(p \times p)$ symmetric matrix and $\mathbf{X}'\mathbf{y}$ is a $(p \times 1)$ column vector. Note the special structure of the $\mathbf{X}'\mathbf{X}$ matrix. The diagonal elements of $\mathbf{X}'\mathbf{X}$ are the sums of squares of the elements in the columns of $\mathbf{X}$, and the off-diagonal elements are the sums of cross-products of the elements in the columns of $\mathbf{X}$. Furthermore, note that the elements of $\mathbf{X}'\mathbf{y}$ are the sums of cross-products of the columns of $\mathbf{X}$ and the observations $\{y_i\}$.

The fitted regression model is

$$\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j x_{ij} \qquad i = 1, 2, \ldots, n \tag{12-14}$$

In matrix notation, the fitted model is

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

The difference between the observation y_i and the fitted value $\hat{y}_i$ is a **residual,** say, $e_i = y_i - \hat{y}_i$. The $(n \times 1)$ vector of residuals is denoted by

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} \tag{12-15}$$

EXAMPLE 12-2

In Example 12-1, we illustrated fitting the multiple regression model

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

where y is the observed pull strength for a wire bond, x_1 is the wire length, and x_2 is the die height. The 25 observations are in Table 12-2. We will now use the matrix approach to fit the regression model above to these data. The **X** matrix and **y** vector for this model are

$$\mathbf{X} = \begin{bmatrix} 1 & 2 & 50 \\ 1 & 8 & 110 \\ 1 & 11 & 120 \\ 1 & 10 & 550 \\ 1 & 8 & 295 \\ 1 & 4 & 200 \\ 1 & 2 & 375 \\ 1 & 2 & 52 \\ 1 & 9 & 100 \\ 1 & 8 & 300 \\ 1 & 4 & 412 \\ 1 & 11 & 400 \\ 1 & 12 & 500 \\ 1 & 2 & 360 \\ 1 & 4 & 205 \\ 1 & 4 & 400 \\ 1 & 20 & 600 \\ 1 & 1 & 585 \\ 1 & 10 & 540 \\ 1 & 15 & 250 \\ 1 & 15 & 290 \\ 1 & 16 & 510 \\ 1 & 17 & 590 \\ 1 & 6 & 100 \\ 1 & 5 & 400 \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} 9.95 \\ 24.45 \\ 31.75 \\ 35.00 \\ 25.02 \\ 16.86 \\ 14.38 \\ 9.60 \\ 24.35 \\ 27.50 \\ 17.08 \\ 37.00 \\ 41.95 \\ 11.66 \\ 21.65 \\ 17.89 \\ 69.00 \\ 10.30 \\ 34.93 \\ 46.59 \\ 44.88 \\ 54.12 \\ 56.63 \\ 22.13 \\ 21.15 \end{bmatrix}$$

The **X′X** matrix is

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 2 & 8 & \cdots & 5 \\ 50 & 110 & \cdots & 400 \end{bmatrix} \begin{bmatrix} 1 & 2 & 50 \\ 1 & 8 & 110 \\ \vdots & \vdots & \vdots \\ 1 & 5 & 400 \end{bmatrix} = \begin{bmatrix} 25 & 206 & 8{,}294 \\ 206 & 2{,}396 & 77{,}177 \\ 8{,}294 & 77{,}177 & 3{,}531{,}848 \end{bmatrix}$$

and the **X′y** vector is

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 2 & 8 & \cdots & 5 \\ 50 & 110 & \cdots & 400 \end{bmatrix} \begin{bmatrix} 9.95 \\ 24.45 \\ \vdots \\ 21.15 \end{bmatrix} = \begin{bmatrix} 725.82 \\ 8{,}008.37 \\ 274{,}811.31 \end{bmatrix}$$

The least squares estimates are found from Equation 12-13 as

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

or

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} 25 & 206 & 8{,}294 \\ 206 & 2{,}396 & 77{,}177 \\ 8{,}294 & 77{,}177 & 3{,}531{,}848 \end{bmatrix}^{-1} \begin{bmatrix} 725.82 \\ 8{,}008.37 \\ 274{,}811.31 \end{bmatrix}$$

$$= \begin{bmatrix} 0.214653 & -0.007491 & -0.000340 \\ -0.007491 & 0.001671 & -0.000019 \\ -0.000340 & -0.000019 & +0.0000015 \end{bmatrix} \begin{bmatrix} 725.82 \\ 8{,}008.47 \\ 274{,}811.31 \end{bmatrix} = \begin{bmatrix} 2.26379143 \\ 2.74426964 \\ 0.01252781 \end{bmatrix}$$

Therefore, the fitted regression model with the regression coefficients rounded to five decimal places is

$$\hat{y} = 2.26379 + 2.74427x_1 + 0.01253x_2$$

This is identical to the results obtained in Example 12-1.

This regression model can be used to predict values of pull strength for various values of wire length (x_1) and die height (x_2). We can also obtain the **fitted values** $\hat{y}_i$ by substituting each observation (x_{i1}, x_{i2}), $i = 1, 2, \ldots, n$, into the equation. For example, the first observation has $x_{11} = 2$ and $x_{12} = 50$, and the fitted value is

$$\begin{aligned} \hat{y}_1 &= 2.26379 + 2.74427x_{11} + 0.01253x_{12} \\ &= 2.26379 + 2.74427(2) + 0.01253(50) \\ &= 8.38 \end{aligned}$$

The corresponding observed value is $y_1 = 9.95$. The *residual* corresponding to the first observation is

$$\begin{aligned} e_1 &= y_1 - \hat{y}_1 \\ &= 9.95 - 8.38 \\ &= 1.57 \end{aligned}$$

Table 12-3 displays all 25 fitted values $\hat{y}_i$ and the corresponding residuals. The fitted values and residuals are calculated to the same accuracy as the original data.

Table 12-3 Observations, Fitted Values, and Residuals for Example 12-2

Observation Number	y_i	$\hat{y}_i$	$e_i = y_i - \hat{y}_i$	Observation Number	y_i	$\hat{y}_i$	$e_i = y_i - \hat{y}_i$
1	9.95	8.38	1.57	14	11.66	12.26	−0.60
2	24.45	25.60	−1.15	15	21.65	15.81	5.84
3	31.75	33.95	−2.20	16	17.89	18.25	−0.36
4	35.00	36.60	−1.60	17	69.00	64.67	4.33
5	25.02	27.91	−2.89	18	10.30	12.34	−2.04
6	16.86	15.75	1.11	19	34.93	36.47	−1.54
7	14.38	12.45	1.93	20	46.59	46.56	−0.03
8	9.60	8.40	1.20	21	44.88	47.06	−2.18
9	24.35	28.21	−3.86	22	54.12	52.56	1.56
10	27.50	27.98	−0.48	23	56.63	56.31	0.32
11	17.08	18.40	−1.32	24	22.13	19.98	2.15
12	37.00	37.46	−0.46	25	21.15	21.00	0.15
13	41.95	41.46	0.49				

Computers are almost always used in fitting multiple regression models. Table 12-4 presents some annotated output from Minitab for the least squares regression model for wire bond pull strength data. The upper part of the table contains the numerical estimates of the regression coefficients. The computer also calculates several other quantities that reflect important information about the regression model. In subsequent sections, we will define and explain the quantities in this output.

Estimating σ^2

Just as in simple linear regression, it is important to estimate σ^2, the variance of the error term ϵ, in a multiple regression model. Recall that in simple linear regression the estimate of σ^2 was obtained by dividing the sum of the squared residuals by $n - 2$. Now there are two parameters in the simple linear regression model, so in multiple linear regression with p parameters a logical estimator for σ^2 is

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n - p} = \frac{SS_E}{n - p} \tag{12-16}$$

This is an **unbiased estimator** of σ^2. Just as in simple linear regression, the estimate of σ^2 is usually obtained from the **analysis of variance** for the regression model. The numerator of Equation 12-16 is called the **error** or **residual sum of squares**, and the denominator $n - p$ is called the **error** or **residual degrees of freedom.** Table 12-4 shows that the estimate of σ^2 for the wire bond pull strength regression model is $\hat{\sigma}^2 = 115.2/22 = 5.2364$. The Minitab output rounds the estimate to $\hat{\sigma}^2 = 5.2$.

12-1.4 Properties of the Least Squares Estimators

The statistical properties of the least squares estimators $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ may be easily found, under certain assumptions on the error terms $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$, in the regression model. Paralleling the assumptions made in Chapter 11, we assume that the errors ϵ_i are statistically independent with mean zero and variance σ^2. Under these assumptions, the least squares estimators $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ are **unbiased estimators** of the regression coefficients $\beta_0, \beta_1, \ldots, \beta_k$. This property may be shown as follows:

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}) &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] \\ &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon})] \\ &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}] \\ &= \boldsymbol{\beta} \end{aligned}$$

since $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} = \mathbf{I}$, the identity matrix. Thus, $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$.

The variances of the $\hat{\boldsymbol{\beta}}$'s are expressed in terms of the elements of the inverse of the $\mathbf{X}'\mathbf{X}$ matrix. The inverse of $\mathbf{X}'\mathbf{X}$ times the constant σ^2 represents the **covariance matrix** of the regression coefficients $\hat{\boldsymbol{\beta}}$. The diagonal elements of $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ are the variances of $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$, and the off-diagonal elements of this matrix are the covariances. For example, if we have $k = 2$ regressors, such as in the pull-strength problem,

$$\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} C_{00} & C_{01} & C_{02} \\ C_{10} & C_{11} & C_{12} \\ C_{20} & C_{21} & C_{22} \end{bmatrix}$$

Table 12-4 Minitab Multiple Regression Output for the Wire Bond Pull Strength Data

```
Regression Analysis: Strength versus Length, Height
 The regression equation is
 Strength = 2.26 + 2.74 Length + 0.0125 Height

 Predictor              Coef      SE Coef       T        P     VIF
 Constant     β̂0→ 2.264          1.060      2.14    0.044
 Length       β̂1→ 2.74427        0.09352   29.34    0.000     1.2
 Height       β̂2→ 0.012528       0.002798   4.48    0.000     1.2

 S = 2.288              R-Sq = 98.1%            R-Sq (adj) = 97.9%
 PRESS = 156.163        R-Sq (pred) = 97.44%

 Analysis of Variance

 Source           DF       SS        MS            F        P
 Regression        2   5990.8    2995.4       572.17    0.000
 Residual Error   22    115.2       5.2←σ̂²
 Total            24   6105.9

 Source     DF    Seq SS
 Length      1    5885.9
 Height      1     104.9

 Predicted Values for New Observations

 New Obs        Fit     SE Fit        95.0% CI             95.0% PI
    1        27.663      0.482    (26.663, 28.663)     (22.814, 32.512)

 Values of Predictors for New Observations

 New Obs     Length     Height
    1          8.00        275
```

which is symmetric ($C_{10} = C_{01}$, $C_{20} = C_{02}$, and $C_{21} = C_{12}$) because $(\mathbf{X}'\mathbf{X})^{-1}$ is symmetric, and we have

$$V(\hat\beta_j) = \sigma^2 C_{jj}, \qquad j = 0, 1, 2$$
$$\text{cov}(\hat\beta_i, \hat\beta_j) = \sigma^2 C_{ij}, \qquad i \neq j$$

In general, the covariance matrix of $\hat{\boldsymbol\beta}$ is a $(p \times p)$ symmetric matrix whose jjth element is the variance of $\hat\beta_j$ and whose i, jth element is the covariance between $\hat\beta_i$ and $\hat\beta_j$, that is,

$$\text{cov}(\hat{\boldsymbol\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2\mathbf{C}$$

The estimates of the variances of these regression coefficients are obtained by replacing σ^2 with an estimate. When σ^2 is replaced by it's estimate $\hat\sigma^2$, the square root of the estimated variance of the jth regression coefficient is called the **estimated standard error** of $\hat\beta_j$ or $se(\hat\beta_j) = \sqrt{\hat\sigma^2 C_{jj}}$. These standard errors are a useful measure of the **precision of estimation** for the regression coefficients; small standard errors imply good precision.

Multiple regression computer programs usually display these standard errors. For example, the Minitab output in Table 12-4 reports $se(\hat\beta_0) = 1.060$, $se(\hat\beta_1) = 0.09352$, and

$se(\hat{\beta}_1) = 0.002798$. The slope estimate is about twice the magnitude of its standard error, and $\hat{\beta}_1$ and $\hat{\beta}_2$ are considerably larger than $se(\hat{\beta}_1)$ and $se(\hat{\beta}_2)$. This implies reasonable precision of estimation, although the parameters β_1 and β_2 are much more precisely estimated than the intercept (this is not unusual in multiple regression).

EXERCISES FOR SECTION 12-1

12-1. A study was performed to investigate the shear strength of soil (y) as it related to depth in feet (x_1) and moisture content (x_2). Ten observations were collected, and the following summary quantities obtained: $n = 10$, $\sum x_{i1} = 223$, $\sum x_{i2} = 553$, $\sum y_i = 1{,}916$, $\sum x_{i1}^2 = 5{,}200.9$, $\sum x_{i2}^2 = 31{,}729$, $\sum x_{i1}x_{i2} = 12{,}352$, $\sum x_{i1}y_i = 43{,}550.8$, $\sum x_{i2}y_i = 104{,}736.8$, and $\sum y_i^2 = 371{,}595.6$.

(a) Set up the least squares normal equations for the model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$.

(b) Estimate the parameters in the model in part (a).

(c) What is the predicted strength when $x_1 = 18$ feet and $x_2 = 43\%$?

12-2. A regression model is to be developed for predicting the ability of soil to absorb chemical contaminants. Ten observations have been taken on a soil absorption index (y) and two regressors: x_1 = amount of extractable iron ore and x_2 = amount of bauxite. We wish to fit the model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$. Some necessary quantities are:

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 1.17991 & -7.30982\text{ E-3} & 7.3006\text{ E-4} \\ -7.30982\text{ E-3} & 7.9799\text{ E-5} & -1.23713\text{ E-4} \\ 7.3006\text{ E-4} & 1.23713\text{ E-4} & 4.6576\text{ E-4} \end{bmatrix}, \quad \mathbf{X}'\mathbf{y} = \begin{bmatrix} 220 \\ 36{,}768 \\ 9{,}965 \end{bmatrix}$$

(a) Estimate the regression coefficients in the model specified above.

(b) What is the predicted value of the absorption index y when $x_1 = 200$ and $x_2 = 50$?

12-3. A chemical engineer is investigating how the amount of conversion of a product from a raw material (y) depends on reaction temperature (x_1) and the reaction time (x_2). He has developed the following regression models:

1. $\hat{y} = 100 + 2x_1 + 4x_2$
2. $\hat{y} = 95 + 1.5x_1 + 3x_2 + 2x_1x_2$

Both models have been built over the range $0.5 \le x_2 \le 10$.

(a) What is the predicted value of conversion when $x_2 = 2$? Repeat this calculation for $x_2 = 8$. Draw a graph of the predicted values for both conversion models. Comment on the effect of the interaction term in model 2.

(b) Find the expected change in the mean conversion for a unit change in temperature x_1 for model 1 when $x_2 = 5$. Does this quantity depend on the specific value of reaction time selected? Why?

(c) Find the expected change in the mean conversion for a unit change in temperature x_1 for model 2 when $x_2 = 5$. Repeat this calculation for $x_2 = 2$ and $x_2 = 8$. Does the result depend on the value selected for x_2? Why?

12-4. The data in Table 12-5 are the 1976 team performance statistics for the teams in the National Football League (*Source: The Sporting News*).

(a) Fit a multiple regression model relating the number of games won to the teams' passing yardage (x_2), the percent rushing plays (x_7), and the opponents' yards rushing (x_8).

(b) Estimate σ^2.

(c) What are the standard errors of the regression coefficients?

(d) Use the model to predict the number of games won when $x_2 = 2000$ yards, $x_7 = 60\%$, and $x_8 = 1800$.

12-5. Table 12-6 presents gasoline mileage performance for 25 automobiles (*Source: Motor Trend,* 1975).

(a) Fit a multiple regression model relating gasoline mileage to engine displacement (x_1) and number of carburetor barrels (x_6).

(b) Estimate σ^2.

(c) Use the model developed in part (a) to predict mileage performance for a car with displacement $x_1 = 300$ and $x_6 = 2$.

12-6. The electric power consumed each month by a chemical plant is thought to be related to the average ambient temperature (x_1), the number of days in the month (x_2), the average product purity (x_3), and the tons of product produced (x_4). The past year's historical data are available and are presented in the following table:

y	x_1	x_2	x_3	x_4
240	25	24	91	100
236	31	21	90	95
270	45	24	88	110
274	60	25	87	88
301	65	25	91	94
316	72	26	94	99
300	80	25	87	97
296	84	25	86	96
267	75	24	88	110
276	60	25	91	105
288	50	25	90	100
261	38	23	89	98

(a) Fit a multiple linear regression model to these data.

(b) Estimate σ^2.

Table 12-5 National Football League 1976 Team Performance

Team	y	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
Washington	10	2113	1985	38.9	64.7	+4	868	59.7	2205	1917
Minnesota	11	2003	2855	38.8	61.3	+3	615	55.0	2096	1575
New England	11	2957	1737	40.1	60.0	+14	914	65.6	1847	2175
Oakland	13	2285	2905	41.6	45.3	−4	957	61.4	1903	2476
Pittsburgh	10	2971	1666	39.2	53.8	+15	836	66.1	1457	1866
Baltimore	11	2309	2927	39.7	74.1	+8	786	61.0	1848	2339
Los Angeles	10	2528	2341	38.1	65.4	+12	754	66.1	1564	2092
Dallas	11	2147	2737	37.0	78.3	−1	797	58.9	2476	2254
Atlanta	4	1689	1414	42.1	47.6	−3	714	57.0	2577	2001
Buffalo	2	2566	1838	42.3	54.2	−1	797	58.9	2476	2254
Chicago	7	2363	1480	37.3	48.0	+19	984	68.5	1984	2217
Cincinnati	10	2109	2191	39.5	51.9	+6	819	59.2	1901	1686
Cleveland	9	2295	2229	37.4	53.6	−5	1037	58.8	1761	2032
Denver	9	1932	2204	35.1	71.4	+3	986	58.6	1709	2025
Detroit	6	2213	2140	38.8	58.3	+6	819	59.2	1901	1686
Green Bay	5	1722	1730	36.6	52.6	−19	791	54.4	2288	1835
Houston	5	1498	2072	35.3	59.3	−5	776	49.6	2072	1914
Kansas City	5	1873	2929	41.1	55.3	+10	789	54.3	2861	2496
Miami	6	2118	2268	38.6	69.6	+6	582	58.7	2411	2670
New Orleans	4	1775	1983	39.3	78.3	+7	901	51.7	2289	2202
New York Giants	3	1904	1792	39.7	38.1	−9	734	61.9	2203	1988
New York Jets	3	1929	1606	39.7	68.8	−21	627	52.7	2592	2324
Philadelphia	4	2080	1492	35.5	68.8	−8	722	57.8	2053	2550
St. Louis	10	2301	2835	35.3	74.1	+2	683	59.7	1979	2110
San Diego	6	2040	2416	38.7	50.0	0	576	54.9	2048	2628
San Francisco	8	2447	1638	39.9	57.1	−8	848	65.3	1786	1776
Seattle	2	1416	2649	37.4	56.3	−22	684	43.8	2876	2524
Tampa Bay	0	1503	1503	39.3	47.0	−9	875	53.5	2560	2241

y: Games won (per 14 game season)
x_1: Rushing yards (season)
x_2: Passing yards (season)
x_3: Punting yards (yds/punt)
x_4: Field goal percentage (Field goals made/Field goals attempted—season)
x_5: Turnover differential (turnovers acquired—turnovers lost)
x_6: Penalty yards (season)
x_7: Percent rushing (rushing plays/total plays)

Table 12-6 Gasoline Mileage Performance for 25 Automobiles

Automobile	y	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
Apollo	18.90	350	165	260	8.0:1	2.56:1	4	3	200.3	69.9	3910	A
Nova	20.00	250	105	185	8.25:1	2.73:1	1	3	196.7	72.2	3510	A
Monarch	18.25	351	143	255	8.0:1	3.00:1	2	3	199.9	74.0	3890	A
Duster	20.07	225	95	170	8.4:1	2.76:1	1	3	194.1	71.8	3365	M
Jenson Conv.	11.2	440	215	330	8.2:1	2.88:1	4	3	184.5	69	4215	A
Skyhawk	22.12	231	110	175	8.0:1	2.56:1	2	3	179.3	65.4	3020	A
Scirocco	34.70	89.7	70	81	8.2:1	3.90:1	2	4	155.7	64	1905	M
Corolla SR-5	30.40	96.9	75	83	9.0:1	4.30:1	2	5	165.2	65	2320	M
Camaro	16.50	350	155	250	8.5:1	3.08:1	4	3	195.4	74.4	3885	A
Datsun B210	36.50	85.3	80	83	8.5:1	3.89:1	2	4	160.6	62.2	2009	M
Capri II	21.50	171	109	146	8.2:1	3.22:1	2	4	170.4	66.9	2655	M
Pacer	19.70	258	110	195	8.0:1	3.08:1	1	3	171.5	77	3375	A
Granada	17.80	302	129	220	8.0:1	3.0:1	2	3	199.9	74	3890	A
Eldorado	14.39	500	190	360	8.5:1	2.73:1	4	3	224.1	79.8	5290	A
Imperial	14.89	440	215	330	8.2:1	2.71:1	4	3	231.0	79.7	5185	A
Nova LN	17.80	350	155	250	8.5:1	3.08:1	4	3	196.7	72.2	3910	A
Starfire	23.54	231	110	175	8.0:1	2.56:1	2	3	179.3	65.4	3050	A
Cordoba	21.47	360	180	290	8.4:1	2.45:1	2	3	214.2	76.3	4250	A
Trans Am	16.59	400	185	NA	7.6:1	3.08:1	4	3	196	73	3850	A
Corolla E-5	31.90	96.9	75	83	9.0:1	4.30:1	2	5	165.2	61.8	2275	M
Mark IV	13.27	460	223	366	8.0:1	3.00:1	4	3	228	79.8	5430	A
Celica GT	23.90	133.6	96	120	8.4:1	3.91:1	2	5	171.5	63.4	2535	M
Charger SE	19.73	318	140	255	8.5:1	2.71:1	2	3	215.3	76.3	4370	A
Cougar	13.90	351	148	243	8.0:1	3.25:1	2	3	215.5	78.5	4540	A
Corvette	16.50	350	165	255	8.5:1	2.73:1	4	3	185.2	69	3660	A

y: Miles/gallon
x_1: Displacement (cubic inches)
x_2: Horsepower (foot-pounds)
x_3: Torque (foot-pounds)
x_4: Compression ratio
x_5: Rear axle ratio
x_6: Carburetor (barrels)
x_7: No. of transmission speeds
x_8: Overall length (inches)
x_9: Width (inches)
x_{10}: Weight (pounds)
x_{11}: Type of transmission (A—automatic, M—manual)

(c) Compute the standard errors of the regression coefficients.

(d) Predict power consumption for a month in which $x_1 = 75°\text{F}$, $x_2 = 24$ days, $x_3 = 90\%$, and $x_4 = 98$ tons.

12-7. A study was performed on wear of a bearing y and its relationship to x_1 = oil viscosity and x_2 = load. The following data were obtained.

y	x_1	x_2
293	1.6	851
230	15.5	816
172	22.0	1058
91	43.0	1201
113	33.0	1357
125	40.0	1115

(a) Fit a multiple linear regression model to these data.

(b) Estimate σ^2 and the standard errors of the regression coefficients.

(c) Use the model to predict wear when $x_1 = 25$ and $x_2 = 1000$.

(d) Fit a multiple linear regression model with an interaction term to these data.

(e) Estimate σ^2 and $se(\hat{\beta}_j)$ for this new model. How did these quantities change. Does this tell you anything about the value of adding the interaction term to the model?

(f) Use the model in (d) to predict when $x_1 = 25$ and $x_2 = 1000$. Compare this prediction with the predicted value from part (b) above.

12-8. The pull strength of a wire bond is an important characteristic. The following table gives information on pull strength (y), die height (x_1), post height (x_2), loop height (x_3), wire length (x_4), bond width on the die (x_5), and bond width on the post (x_6).

y	x_1	x_2	x_3	x_4	x_5	x_6
8.0	5.2	19.6	29.6	94.9	2.1	2.3
8.3	5.2	19.8	32.4	89.7	2.1	1.8
8.5	5.8	19.6	31.0	96.2	2.0	2.0
8.8	6.4	19.4	32.4	95.6	2.2	2.1
9.0	5.8	18.6	28.6	86.5	2.0	1.8
9.3	5.2	18.8	30.6	84.5	2.1	2.1
9.3	5.6	20.4	32.4	88.8	2.2	1.9
9.5	6.0	19.0	32.6	85.7	2.1	1.9
9.8	5.2	20.8	32.2	93.6	2.3	2.1
10.0	5.8	19.9	31.8	86.0	2.1	1.8
10.3	6.4	18.0	32.6	87.1	2.0	1.6
10.5	6.0	20.6	33.4	93.1	2.1	2.1
10.8	6.2	20.2	31.8	83.4	2.2	2.1
11.0	6.2	20.2	32.4	94.5	2.1	1.9
11.3	6.2	19.2	31.4	83.4	1.9	1.8
11.5	5.6	17.0	33.2	85.2	2.1	2.1
11.8	6.0	19.8	35.4	84.1	2.0	1.8
12.3	5.8	18.8	34.0	86.9	2.1	1.8
12.5	5.6	18.6	34.2	83.0	1.9	2.0

(a) Fit a multiple linear regression model using x_2, x_3, x_4, and x_5 as the regressors.

(b) Estimate σ^2.

(c) Find the $se(\hat{\beta}_j)$. How precisely are the regression coefficients estimated, in your opinion?

(d) Use the model from part (a) to predict pull strength when $x_2 = 20$, $x_3 = 30$, $x_4 = 90$, and $x_5 = 2.0$.

12-9. An engineer at a semiconductor company wants to model the relationship between the device HFE (y) and three parameters: Emitter-RS (x_1), Base-RS (x_2), and Emitter-to-Base RS (x_3). The data are shown in the following table.

x_1 Emitter-RS	x_2 Base-RS	x_3 E-B-RS	y HFE-1M-5V
14.620	226.00	7.000	128.40
15.630	220.00	3.375	52.62
14.620	217.40	6.375	113.90
15.000	220.00	6.000	98.01
14.500	226.50	7.625	139.90
15.250	224.10	6.000	102.60
16.120	220.50	3.375	48.14
15.130	223.50	6.125	109.60
15.500	217.60	5.000	82.68
15.130	228.50	6.625	112.60
15.500	230.20	5.750	97.52
16.120	226.50	3.750	59.06
15.130	226.60	6.125	111.80
15.630	225.60	5.375	89.09
15.380	229.70	5.875	101.00
14.380	234.00	8.875	171.90
15.500	230.00	4.000	66.80
14.250	224.30	8.000	157.10
14.500	240.50	10.870	208.40
14.620	223.70	7.375	133.40

(a) Fit a multiple linear regression model to the data.

(b) Estimate σ^2.

(c) Find the standard errors $se(\hat{\beta}_j)$.

(d) Predict HFE when $x_1 = 14.5$, $x_2 = 220$, and $x_3 = 5.0$.

12-10. Heat treating is often used to carburize metal parts, such as gears. The thickness of the carburized layer is considered a crucial feature of the gear and contributes to the overall reliability of the part. Because of the critical nature of this feature, two different lab tests are performed on each furnace load. One test is run on a sample pin that accompanies each load. The other test is a destructive test, where an actual part is cross-sectioned. This test involves running a carbon analysis on the surface of both the gear pitch (top of the gear tooth) and the gear root (between the gear teeth). Table 12-7 shows the results of the pitch carbon analysis test for 32 parts.

The regressors are furnace temperature (TEMP), carbon concentration and duration of the carburizing cycle

Table 12-7

TEMP	SOAKTIME	SOAKPCT	DIFFTIME	DIFFPCT	PITCH
1650	0.58	1.10	0.25	0.90	0.013
1650	0.66	1.10	0.33	0.90	0.016
1650	0.66	1.10	0.33	0.90	0.015
1650	0.66	1.10	0.33	0.95	0.016
1600	0.66	1.15	0.33	1.00	0.015
1600	0.66	1.15	0.33	1.00	0.016
1650	1.00	1.10	0.50	0.80	0.014
1650	1.17	1.10	0.58	0.80	0.021
1650	1.17	1.10	0.58	0.80	0.018
1650	1.17	1.10	0.58	0.80	0.019
1650	1.17	1.10	0.58	0.90	0.021
1650	1.17	1.10	0.58	0.90	0.019
1650	1.17	1.15	0.58	0.90	0.021
1650	1.20	1.15	1.10	0.80	0.025
1650	2.00	1.15	1.00	0.80	0.025
1650	2.00	1.10	1.10	0.80	0.026
1650	2.20	1.10	1.10	0.80	0.024
1650	2.20	1.10	1.10	0.80	0.025
1650	2.20	1.15	1.10	0.80	0.024
1650	2.20	1.10	1.10	0.90	0.025
1650	2.20	1.10	1.10	0.90	0.027
1650	2.20	1.10	1.50	0.90	0.026
1650	3.00	1.15	1.50	0.80	0.029
1650	3.00	1.10	1.50	0.70	0.030
1650	3.00	1.10	1.50	0.75	0.028
1650	3.00	1.15	1.66	0.85	0.032
1650	3.33	1.10	1.50	0.80	0.033
1700	4.00	1.10	1.50	0.70	0.039
1650	4.00	1.10	1.50	0.70	0.040
1650	4.00	1.15	1.50	0.85	0.035
1700	12.50	1.00	1.50	0.70	0.056
1700	18.50	1.00	1.50	0.70	0.068

(SOAKPCT, SOAKTIME), and carbon concentration and duration of the diffuse cycle (DIFFPCT, DIFFTIME).

(a) Fit a linear regression model relating the results of the pitch carbon analysis test (PITCH) to the five regressor variables.

(b) Estimate σ^2.

(c) Find the standard errors $se(\hat{\beta}_j)$.

(d) Use the model in part (a) to predict PITCH when TEMP = 1650, SOAKTIME = 1.00, SOAKPCT = 1.10, DIFFTIME = 1.00, and DIFFPCT = 0.80.

12-11. Statistics for 21 National Hockey League teams were obtained from the *Hockey Encyclopedia* and are shown in Table 12-8.

The variables and definitions are as follows:

Wins Number of games won in a season.

Pts Points awarded in a season. Two points for winning a game, one point for losing in overtime, zero points for losing in regular time.

GF Goals for. Total goals scored during the season.

GA Goals against. Goals scored against the team during the season.

PPG Power play goals. Points scored while on power play.

PPcT Power play percentage. The number of power play goals divided by the number of power play opportunities.

Table 12-8

Team	Wins	Pts	GF	GA	PPG	PPcT	SHG	PPGA	PKPcT	SHGA
Chicago	47	104	338	268	86	27.2	4	71	76.6	6
Minnesota	40	96	321	290	91	26.4	17	67	80.7	20
Toronto	28	68	23	330	79	22.3	13	83	75	9
St. Louis	25	65	285	316	67	21.2	9	63	81.3	12
Detroit	21	57	263	344	37	19.3	7	80	72.6	9
Edmonton	47	106	424	315	86	29.3	22	89	77.5	6
Calgary	32	78	321	317	90	27	7	59	77.1	6
Vancouver	30	75	303	309	65	23.8	5	56	80.8	13
Winnipeg	33	74	311	333	78	23.6	10	67	72.8	7
Los Angeles	27	66	308	365	81	23.8	10	94	68.2	14
Philadelphia	49	106	326	240	60	21.6	15	61	82	7
NY Islanders	42	96	302	226	69	25.8	10	55	83.4	3
Washington	39	94	306	283	75	20.9	3	53	81.6	11
NY Rangers	35	80	306	287	71	22.4	12	75	76	8
New Jersey	17	48	230	338	66	21.9	74	78	73.5	10
Pittsburgh	18	45	257	394	81	22.6	3	110	72.2	15
Boston	50	110	327	228	67	22.2	8	53	80.7	6
Montreal	42	98	350	286	64	22.2	8	68	73.8	8
Buffalo	38	89	318	285	67	21.5	12	48	82.5	9
Quebec	34	80	343	336	61	20.7	6	92	73.6	6
Hartford	19	45	261	403	51	19.3	6	70	76.1	9

SHG Short-handed goals scored during the season.

PPGA Power play goals against.

PKPcT Penalty killing percentage. Measures a team's ability to prevent goals while its opponent is on a power play. Opponent power play goals divided by opponent's opportunities.

SHGA Short-handed goals against. Fit a multiple linear regression model relating wins to the other variables. Estimate σ^2 and find the standard errors of the regression coefficients.

12-12. Consider the linear regression model

$$Y_i = \beta_0' + \beta_1(x_{i1} - \bar{x}_1) + \beta_2(x_{i2} - \bar{x}_2) + \epsilon_i$$

where $\bar{x}_1 = \sum x_{i1}/n$ and $\bar{x}_2 = \sum x_{i2}/n$.

(a) Write out the least squares normal equations for this model.

(b) Verify that the least squares estimate of the intercept in this model is $\hat{\beta}_0' = \sum y_i/n = \bar{y}$.

(c) Suppose that we use $y_i - \bar{y}$ as the response variable in the model above. What effect will this have on the least squares estimate of the intercept?

12-2 HYPOTHESIS TESTS IN MULTIPLE LINEAR REGRESSION

In multiple linear regression problems, certain tests of hypotheses about the model parameters are useful in measuring model adequacy. In this section, we describe several important hypothesis-testing procedures. As in the simple linear regression case, hypothesis testing requires that the error terms ϵ_i in the regression model are normally and independently distributed with mean zero and variance σ^2.

12-2.1 Test for Significance of Regression

The test for significance of regression is a test to determine whether a linear relationship exists between the response variable y and a subset of the regressor variables $x_1, x_2, \ldots, x_k$. The appropriate hypotheses are

$$H_0\colon \beta_1 = \beta_2 = \cdots = \beta_k = 0$$
$$H_1\colon \beta_j \neq 0 \quad \text{for at least one } j \tag{12-17}$$

Rejection of $H_0\colon \beta_1 = \beta_2 = \cdots = \beta_k = 0$ implies that at least one of the regressor variables $x_1, x_2, \ldots, x_k$ contributes significantly to the model.

The test for significance of regression is a generalization of the procedure used in simple linear regression. The total sum of squares SS_T is partitioned into a sum of squares due to regression and a sum of squares due to error, say,

$$SS_T = SS_R + SS_E$$

Now if $H_0\colon \beta_1 = \beta_2 = \cdots = \beta_k = 0$ is true, SS_R/σ^2 is a chi-square random variable with k degrees of freedom. Note that the number of degrees of freedom for this chi-square random variable is equal to the number of regressor variables in the model. We can also show the SS_E/σ^2 is a chi-square random variable with $n - p$ degrees of freedom, and that SS_E and SS_R are independent. The test statistic for $H_0\colon \beta_1 = \beta_2 = \cdots = \beta_k = 0$ is

$$F_0 = \frac{SS_R/k}{SS_E/(n-p)} = \frac{MS_R}{MS_E} \tag{12-18}$$

We should reject H_0 if the computed value of the test statistic in Equation 12-18, f_0, is greater than $f_{\alpha,k,n-p}$. The procedure is usually summarized in an analysis of variance table such as Table 12-9.

We can find a computing formula for SS_E as follows:

$$SS_E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 = \mathbf{e}'\mathbf{e}$$

Substituting $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ into the above, we obtain

$$SS_E = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} \tag{12-19}$$

Table 12-9 Analysis of Variance for Testing Significance of Regression in Multiple Regression

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Regression	SS_R	k	MS_R	MS_R/MS_E
Error or residual	SS_E	$n - p$	MS_E	
Total	SS_T	$n - 1$		

A computational formula for SS_R may be found easily. Now since $SS_T = \sum_{i=1}^n y_i^2 - (\sum_{i=1}^n y_i)^2/n = \mathbf{y}'\mathbf{y} - (\sum_{i=1}^n y_i)^2/n$, we may rewrite Equation 12-19 as

$$SS_E = \mathbf{y}'\mathbf{y} - \frac{\left(\sum_{i=1}^n y_i\right)^2}{n} - \left[\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \frac{\left(\sum_{i=1}^n y_i\right)^2}{n}\right]$$

or

$$SS_E = SS_T - SS_R$$

Therefore, the regression sum of squares is

$$SS_R = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \frac{\left(\sum_{i=1}^n y_i\right)^2}{n} \tag{12-20}$$

EXAMPLE 12-3

We will test for significance of regression (with $\alpha = 0.05$) using the wire bond pull strength data from Example 12-1. The total sum of squares is

$$\begin{aligned} SS_T &= \mathbf{y}'\mathbf{y} - \frac{\left(\sum_{i=1}^n y_i\right)^2}{n} \\ &= 27{,}177.9510 - \frac{(725.82)^2}{25} = 6105.9447 \end{aligned}$$

The regression sum of squares is computed from Equation 12-20 as follows:

$$\begin{aligned} SS_R &= \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \frac{\left(\sum_{i=1}^n y_i\right)^2}{n} \\ &= 27{,}062.7775 - \frac{(725.82)^2}{25} = 5990.7712 \end{aligned}$$

and by subtraction

$$\begin{aligned} SS_E &= SS_T - SS_R \\ &= \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} = 115.1735 \end{aligned}$$

The analysis of variance is shown in Table 12-10. To test H_0: $\beta_1 = \beta_2 = 0$, we calculate the statistic

$$f_0 = \frac{MS_R}{MS_E} = \frac{2995.3856}{5.2352} = 572.17$$

Since $f_0 > f_{0.05,2,22} = 3.44$ (or since the P-value is considerably smaller than $\alpha = 0.05$), we reject the null hypothesis and conclude that pull strength is linearly related to either wire length or die height, or both. However, we note that this does not necessarily imply that the

Table 12-10 Test for Significance of Regression for Example 12-3

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	f_0	P-value
Regression	5990.7712	2	2995.3856	572.17	1.08E-19
Error or residual	115.1735	22	5.2352		
Total	6105.9447	24			

relationship found is an appropriate model for predicting pull strength as a function of wire length and die height. Further tests of model adequacy are required before we can be comfortable using this model in practice.

Most multiple regression computer programs provide the test for significance of regression in their output display. The middle portion of Table 12-4 is the Minitab output for this example. Compare Tables 12-4 and 12-10 and note their equivalence apart from rounding. The P-value is rounded to zero in the computer output.

R^2 and Adjusted R^2

We may also use the **coefficient of multiple determination** R^2 as a global statistic to assess the fit of the model. Computationally,

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T} \tag{12-21}$$

For the wire bond pull strength data, we find that $R^2 = SS_R/SS_T = 5990.7712/6105.9447 = 0.9811$. Thus the model accounts for about 98% of the variability in the pull strength response (refer to the Minitab output in Table 12-4). The R^2 statistic is somewhat problematic as a measure of the quality of the fit for a multiple regression model because it always increases when a variable is added to a model.

To illustrate, consider the model fit to wire bond pull strength data in Example 11-8. This was a simple linear regression model with x_1 = wire length as the regressor. The value of R^2 for this model is $R^2 = 0.9640$. Therefore, adding x_y = die height to the model increases R^2 by $0.9811 - 0.9640 = 0.0171$, a very small amount. Since R^2 always increases when a regressor is added, it can be difficult to judge whether the increase is telling us anything useful about the new regressor. It is particularly hard to interpret a small increase, such as observed in the pull strength data.

Many regression users prefer to use an **adjusted** R^2 statistic:

$$R^2_{\text{adj}} = 1 - \frac{SS_E/(n-p)}{SS_T/(n-1)} \tag{12-22}$$

Because $SS_E/(n-p)$ is the error or residual mean square and $SS_T/(n-1)$ is a constant, R^2_{adj} will only increase when a variable is added to the model if the new variable reduces the error mean square. Note that for the multiple regression model for the pull strength data $R^2_{\text{adj}} = 0.979$ (see the Minitab output in Table 12-4), whereas in Example 11-8 the adjusted R^2 for the one-variable model is $R^2_{\text{adj}} = 0.962$. Therefore, we would conclude that adding x_2 = die height to the model does result in a meaningful reduction in unexplained variability in the response.

The **adjusted** R^2 statistic essentially penalizes the analyst for adding terms to the model. It is an easy way to guard against **overfitting,** that is, including regressors that are not really useful. Consequently, it is very useful in comparing and evaluating competing regression models. We will use R^2_{adj} for this when we discuss **variable selection** in regression in Section 12-6.3.

12-2.2 Tests on Individual Regression Coefficients and Subsets of Coefficients

We are frequently interested in testing hypotheses on the individual regression coefficients. Such tests would be useful in determining the potential value of each of the regressor variables in the regression model. For example, the model might be more effective with the inclusion of additional variables or perhaps with the deletion of one or more of the regressors presently in the model.

Adding a variable to a regression model always causes the sum of squares for regression to increase and the error sum of squares to decrease (this is why R^2 always increases when a variable is added). We must decide whether the increase in the regression sum of squares is large enough to justify using the additional variable in the model. Furthermore, adding an unimportant variable to the model can actually increase the error mean square, indicating that adding such a variable has actually made the model a poorer fit to the data (this is why R^2_{adj} is a better measure of global model fit then the ordinary R^2).

The hypotheses for testing the significance of any individual regression coefficient, say β_j, are

$$H_0\colon \beta_j = 0 \qquad H_1\colon \beta_j \neq 0 \tag{12-23}$$

If $H_0\colon \beta_j = 0$ is not rejected, this indicates that the regressor x_j can be deleted from the model. The test statistic for this hypothesis is

$$T_0 = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \tag{12-24}$$

where C_{jj} is the diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$ corresponding to $\hat{\beta}_j$. Notice that the denominator of Equation 12-24 is the standard error of the regression coefficient $\hat{\beta}_j$. The null hypothesis H_0: $\beta_j = 0$ is rejected if $|t_0| > t_{\alpha/2,n-p}$. This is called a **partial** or **marginal test** because the regression coefficient $\hat{\beta}_j$ depends on all the other regressor variables $x_i (i \neq j)$ that are in the model. More will be said about this in the following example.

EXAMPLE 12-4 Consider the wire bond pull strength data, and suppose that we want to test the hypothesis that the regression coefficient for x_2 (die height) is zero. The hypotheses are

$$H_0\colon \beta_2 = 0$$
$$H_1\colon \beta_2 \neq 0$$

The main diagonal element of the $(\mathbf{X}'\mathbf{X})^{-1}$ matrix corresponding to $\hat{\beta}_2$ is $C_{22} = 0.0000015$, so the t-statistic in Equation 12-24 is

$$t_0 = \frac{\hat{\beta}_2}{\sqrt{\hat{\sigma}^2 C_{22}}} = \frac{0.01253}{\sqrt{(5.2352)(0.0000015)}} = 4.4767$$

Note that we have used the estimate of σ^2 reported to four decimal places in Table 12-10. Since $t_{0.025,22} = 2.074$, we reject H_0: $\beta_2 = 0$ and conclude that the variable x_2 (die height) contributes significantly to the model. We could also have used a P-value to draw conclusions. The P-value for $t_0 = 4.4767$ is $P = 0.0002$, so with $\alpha = 0.05$ we would reject the null hypothesis. Note that this test measures the marginal or partial contribution of x_2 given that x_1 is in the model. That is, the t-test measures the contribution of adding the variable $x_2 =$ die height to a model that already contains $x_1 =$ wire length. Table 12-4 shows the value of the t-test computed by Minitab. The Minitab t-test statistic is reported to two decimal places. Note that the computer produces a t-test for each regression coefficient in the model. These t-tests indicate that both regressors contribute to the model.

There is another way to test the contribution of an individual regressor variable to the model. This approach determines the increase in the regression sum of squares obtained by adding a variable x_j (say) to the model, given that other variables $x_i (i \neq j)$ are already included in the regression equation.

The procedure used to do this is called the **general regression significance test**, or the **extra sum of squares method.** This procedure can also be used to investigate the contribution of a *subset* of the regressor variables to the model. Consider the regression model with k regressor variables

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{12-25}$$

where $\mathbf{y}$ is $(n \times 1)$, $\mathbf{X}$ is $(n \times p)$, $\boldsymbol{\beta}$ is $(p \times 1)$, $\boldsymbol{\epsilon}$ is $(n \times 1)$, and $p = k + 1$. We would like to determine if the subset of regressor variables $x_1, x_2, \ldots, x_r$ $(r < k)$ as a whole contributes significantly to the regression model. Let the vector of regression coefficients be partitioned as follows:

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} \tag{12-26}$$

where $\boldsymbol{\beta}_1$ is $(r \times 1)$ and $\boldsymbol{\beta}_2$ is $[(p - r) \times 1]$. We wish to test the hypotheses

$$\begin{aligned} H_0&: \boldsymbol{\beta}_1 = \mathbf{0} \\ H_1&: \boldsymbol{\beta}_1 \neq \mathbf{0} \end{aligned} \tag{12-27}$$

where $\mathbf{0}$ denotes a vector of zeroes. The model may be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon} \tag{12-28}$$

where $\mathbf{X}_1$ represents the columns of $\mathbf{X}$ associated with $\boldsymbol{\beta}_1$ and $\mathbf{X}_2$ represents the columns of $\mathbf{X}$ associated with $\boldsymbol{\beta}_2$.

For the **full model** (including both $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$), we know that $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{y}$. In addition, the regression sum of squares for all variables including the intercept is

$$SS_R(\boldsymbol{\beta}) = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} \qquad (p = k + 1 \text{ degrees of freedom})$$

and

$$MS_E = \frac{\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}\mathbf{X}'\mathbf{y}}{n - p}$$

$SS_R(\boldsymbol{\beta})$ is called the regression sum of squares due to $\boldsymbol{\beta}$. To find the contribution of the terms in $\boldsymbol{\beta}_1$ to the regression, fit the model assuming the null hypothesis H_0: $\boldsymbol{\beta}_1 = \mathbf{0}$ to be true. The **reduced model** is found from Equation 12-28 as

$$\mathbf{y} = \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon} \tag{12-29}$$

The least squares estimate of $\boldsymbol{\beta}_2$ is $\hat{\boldsymbol{\beta}}_2 = (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{y}$, and

$$SS_R(\boldsymbol{\beta}_2) = \hat{\boldsymbol{\beta}}_2'\mathbf{X}_2'\mathbf{y} \qquad (p - r \text{ degrees of freedom}) \tag{12-30}$$

The regression sum of squares due to $\boldsymbol{\beta}_1$ given that $\boldsymbol{\beta}_2$ is already in the model is

$$SS_R(\boldsymbol{\beta}_1|\boldsymbol{\beta}_2) = SS_R(\boldsymbol{\beta}) - SS_R(\boldsymbol{\beta}_2) \tag{12-31}$$

This sum of squares has r degrees of freedom. It is sometimes called the extra sum of squares due to $\boldsymbol{\beta}_1$. Note that $SS_R(\boldsymbol{\beta}_1|\boldsymbol{\beta}_2)$ is the increase in the regression sum of squares due to including the variables $x_1, x_2, \ldots, x_r$ in the model. Now $SS_R(\boldsymbol{\beta}_1|\boldsymbol{\beta}_2)$ is independent of MS_E, and the null hypothesis $\boldsymbol{\beta}_1 = \mathbf{0}$ may be tested by the statistic

$$F_0 = \frac{SS_R(\boldsymbol{\beta}_1|\boldsymbol{\beta}_2)/r}{MS_E} \tag{12-32}$$

If the computed value of the test statistic $f_0 > f_{\alpha,r,n-p}$, we reject H_0, concluding that at least one of the parameters in $\boldsymbol{\beta}_1$ is not zero and, consequently, at least one of the variables $x_1, x_2, \ldots, x_r$ in $\mathbf{X}_1$ contributes significantly to the regression model. Some authors call the test in Equation 12-32 a **partial *F*-test.**

The partial F-test is very useful. We can use it to measure the contribution of each individual regressor x_j as if it were the last variable added to the model by computing

$$SS_R(\beta_j|\beta_0, \beta_1, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_k), \qquad j = 1, 2, \ldots, k$$

This is the increase in the regression sum of squares due to adding x_j to a model that already includes $x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k$. The partial F-test is a more general procedure in that we can measure the effect of sets of variables. In Section 12-6.3 we show how the partial F-test plays a major role in *model building*—that is, in searching for the best set of regressor variables to use in the model.

EXAMPLE 12-5 Consider the wire bond pull strength data in Example 12-1. We will investigate the contribution of the variable x_2 (die height) to the model using the partial F-test approach. That is, we wish to test

$$H_0: \beta_2 = 0$$
$$H_1: \beta_2 \neq 0$$

To test this hypothesis, we need the extra sum of squares due to β_2, or

$$\begin{aligned} SS_R(\beta_2 | \beta_1,\beta_0) &= SS_R(\beta_1,\beta_2,\beta_0) - SS_R(\beta_1,\beta_0) \\ &= SS_R(\beta_1,\beta_2 | \beta_0) - SS_R(\beta_1 | \beta_0) \end{aligned}$$

In Example 12-3 we have calculated

$$SS_R(\beta_1,\beta_2 | \beta_0) = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \frac{\left(\sum_{i=1}^{n} y_i\right)}{n} = 5990.7712 \quad \text{(two degrees of freedom)}$$

and from Example 11-8, where we fit the model $Y = \beta_0 + \beta_1 x_1 + \epsilon$, we can calculate

$$\begin{aligned} SS_R(\beta_1 | \beta_0) &= \hat{\beta}_1 S_{xy} = (2.9027)(2027.7132) \\ &= 5885.8521 \quad \text{(one degree of freedom)} \end{aligned}$$

Therefore,

$$\begin{aligned} SS_R(\beta_2 | \beta_1,\beta_0) &= 5990.7712 - 5885.8521 \\ &= 104.9191 \quad \text{(one degree of freedom)} \end{aligned}$$

This is the increase in the regression sum of squares due to adding x_2 to a model already containing x_1. To test H_0: $\beta_2 = 0$, calculate the test statistic

$$f_0 = \frac{SS_R(\beta_2 | \beta_1,\beta_0)/1}{MS_E} = \frac{104.9191/1}{5.2352} = 20.04$$

Note that the MS_E from the full model, using both x_1 and x_2, is used in the denominator of the test statistic. Since $f_{0.05,1,22} = 4.30$, we reject H_0: $\beta_2 = 0$ and conclude that the regressor die height (x_2) contributes significantly to the model.

Table 12-4 shows the Minitab regression output for the wire bond pull strength data. Just below the analysis of variance summary in this table the quantity labeled "Seq SS" shows the sum of squares obtained by fitting x_1 alone (5885.9) and the sum of squares obtained by fitting x_2 after x_1. Notationally, these are referred to above as $SS_R(\beta_1 | \beta_0)$ and $SS_R(\beta_2 | \beta_1,\beta_0)$.

Since the partial F-test in the above example involves a single variable, it is equivalent to the t-test. To see this, recall from Example 12-5 that the t-test on H_0: $\beta_2 = 0$ resulted in the test statistic $t_0 = 4.4767$. Furthermore, the square of a t-random variable with ν degrees of freedom is an F-random variable with one and ν degrees of freedom, and we note that $t_0^2 = (4.4767)^2 = 20.04 = f_0$.

12-2.3 More About the Extra Sum of Squares Method (CD Only)

EXERCISES FOR SECTION 12-2

12-13. Consider the regression model fit to the soil shear strength data in Exercise 12-1.

(a) Test for significance of regression using $\alpha = 0.05$. What is the *P*-value for this test?

(b) Construct the *t*-test on each regression coefficient. What are your conclusions, using $\alpha = 0.05$?

12-14. Consider the absorption index data in Exercise 12-2. The total sum of squares for y is $SS_T = 742.00$.

(a) Test for significance of regression using $\alpha = 0.01$. What is the *P*-value for this test?

(b) Test the hypothesis H_0: $\beta_1 = 0$ versus H_1: $\beta_1 \neq 0$ using $\alpha = 0.01$. What is the *P*-value for this test?

What conclusion can you draw about the usefulness of x_1 as a regressor in this model?

12-15. Consider the NFL data in Exercise 12-4.

(a) Test for significance of regression using $\alpha = 0.05$. What is the *P*-value for this test?

(b) Conduct the *t*-test for each regression coefficient β_2, β_7, and β_8. Using $\alpha = 0.05$, what conclusions can you draw about the variables in this model?

12-16. Reconsider the NFL data in Exercise 12-4.

(a) Find the amount by which the regressor x_8 (opponents' yards rushing) increases the regression sum of squares.

(b) Use the results from part (a) above and Exercise 12-14 to conduct an *F*-test for H_0: $\beta_8 = 0$ versus H_1: $\beta_8 \neq 0$ using $\alpha = 0.05$. What is the *P*-value for this test? What conclusions can you draw?

12-17. Consider the gasoline mileage data in Exercise 12-5.

(a) Test for significance of regression using $\alpha = 0.05$. What conclusions can you draw?

(b) Find the *t*-test statistic for both regressors. Using $\alpha = 0.05$, what conclusions can you draw? Do both regressors contribute to the model?

12-18. A regression model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$ has been fit to a sample of $n = 25$ observations. The calculated *t*-ratios $\hat{\beta}_j / se(\hat{\beta}_j), j = 1, 2, 3$ are as follows: for β_1, $t_0 = 4.82$, for β_2, $t_0 = 8.21$ and for β_3, $t_0 = 0.98$.

(a) Find *P*-values for each of the *t*-statistics.

(b) Using $\alpha = 0.05$, what conclusions can you draw about the regressor x_3? Does it seem likely that this regressor contributes significantly to the model?

12-19. Consider the electric power consumption data in Exercise 12-6.

(a) Test for significance of regression using $\alpha = 0.05$. What is the *P*-value for this test?

(b) Use the *t*-test to assess the contribution of each regressor to the model. Using $\alpha = 0.05$, what conclusions can you draw?

12-20. Consider the bearing wear data in Exercise 12-7 with no interaction.

(a) Test for significance of regression using $\alpha = 0.05$. What is the *P*-value for this test? What are your conclusions?

(b) Compute the *t*-statistics for each regression coefficient. Using $\alpha = 0.05$, what conclusions can you draw?

(c) Use the extra sum of squares method to investigate the usefulness of adding x_2 = load to a model that already contains x_1 = oil viscosity. Use $\alpha = 0.05$.

12-21. Reconsider the bearing wear data from Exercises 12-7 and 12-20.

(a) Refit the model with an interaction term. Test for significance of regression using $\alpha = 0.05$.

(b) Use the extra sum of squares method to determine whether the interaction term contributes significantly to the model. Use $\alpha = 0.05$.

(c) Estimate σ^2 for the interaction model. Compare this to the estimate of σ^2 from the model in Exercise 12-20.

12-22. Consider the wire bond pull strength data in Exercise 12-8.

(a) Test for significance of regression using $\alpha = 0.05$. Find the *P*-value for this test. What conclusions can you draw?

(b) Calculate the *t*-test statistic for each regression coefficient. Using $\alpha = 0.05$, what conclusions can you draw? Do all variables contribute to the model?

12-23. Reconsider the semiconductor data in Exercise 12-9.

(a) Test for significance of regression using $\alpha = 0.05$. What conclusions can you draw?

(b) Calcuate the *t*-test statistic for each regression coefficient. Using $\alpha = 0.05$, what conclusions can you draw?

12-24. Exercise 12-10 presents data on heat treating gears.

(a) Test the regression model for significance of regression. Using $\alpha = 0.05$, find the *P*-value for the test and draw conclusions.

(b) Evaluate the contribution of each regressor to the model using the *t*-test with $\alpha = 0.05$.

(c) Fit a new model to the response PITCH using new regressors x_1 = SOAKTIME × SOAKPCT and x_2 = DIFFTIME × DIFFPCT.

(d) Test the model in part (c) for significance of regression using $\alpha = 0.05$. Also calculate the *t*-test for each regressor and draw conclusions.

(e) Estimate σ^2 for the model from part (c) and compare this to the estimate of σ^2 for the model in part (a). Which estimate is smaller? Does this offer any insight regarding which model might be preferable?

12-25. Data on National Hockey League team performance was presented in Exercise 12-11.

(a) Test the model from this exercise for significance of regression using $\alpha = 0.05$. What conclusions can you draw?

(b) Use the t-test to evaluate the contribution of each regressor to the model. Does it seem that all regressors are necessary? Use $\alpha = 0.05$.

(c) Fit a regression model relating the number of games won to the number of points scored and the number of power play goals. Does this seem to be a logical choice of regressors, considering your answer to part (b)? Test this new model for significance of regression and evaluate the contribution of each regressor to the model using the t-test. Use $\alpha = 0.05$.

12-3 CONFIDENCE INTERVALS IN MULTIPLE LINEAR REGRESSION

12-3.1 Confidence Intervals on Individual Regression Coefficients

In multiple regression models, it is often useful to construct confidence interval estimates for the regression coefficients $\{\beta_j\}$. The development of a procedure for obtaining these confidence intervals requires that the errors $\{\epsilon_i\}$ are normally and independently distributed with mean zero and variance σ^2. This is the same assumption required in hypothesis testing. Therefore, the observations $\{Y_i\}$ are normally and independently distributed with mean $\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}$ and variance σ^2. Since the least squares estimator $\hat{\boldsymbol{\beta}}$ is a linear combination of the observations, it follows that $\hat{\boldsymbol{\beta}}$ is normally distributed with mean vector $\boldsymbol{\beta}$ and covariance matrix $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. Then each of the statistics

$$T = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} \qquad j = 0, 1, \ldots, k \tag{12-33}$$

has a t distribution with $n - p$ degrees of freedom, where C_{jj} is the jjth element of the $(\mathbf{X}'\mathbf{X})^{-1}$ matrix, and $\hat{\sigma}^2$ is the estimate of the error variance, obtained from Equation 12-16. This leads to the following $100(1 - \alpha)\%$ confidence interval for the regression coefficient $\beta_j, j = 0, 1, \ldots, k$.

Definition

A $100(1 - \alpha)$ % **confidence interval on the regression coefficient** β_j, $j = 0, 1, \ldots, k$ in the multiple linear regression model is given by

$$\hat{\beta}_j - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 C_{jj}} \le \beta_j \le \hat{\beta}_j + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 C_{jj}} \tag{12-34}$$

Because $\sqrt{\hat{\sigma}^2 C_{jj}}$ is the standard error of the regression coefficient $\hat{\beta}_j$, we would also write the CI formula as $\hat{\beta}_j - t_{\alpha/2,n-p}\, se(\hat{\beta}_j) \le \beta_j \le \hat{\beta}_j + t_{\alpha/2,n-p}\, se(\hat{\beta}_j)$.

EXAMPLE 12-6

We will construct a 95% confidence interval on the parameter β_1 in the wire bond pull strength problem. The point estimate of β_1 is $\hat{\beta}_1 = 2.74427$ and the diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$ corresponding to β_1 is $C_{11} = 0.001671$. The estimate of σ^2 is $\hat{\sigma}^2 = 5.2352$, and $t_{0.025,22} = 2.074$. Therefore, the 95% CI on β_1 is computed from Equation 12-34 as

$$2.74427 - (2.074)\sqrt{(5.2352)(.001671)} \le \beta_1 \le 2.74427 + (2.074)\sqrt{(5.2352)(.001671)}$$

which reduces to

$$2.55029 \le \beta_1 \le 2.93825$$

12-3.2 Confidence Interval on the Mean Response

We may also obtain a confidence interval on the mean response at a particular point, say, $x_{01}, x_{02}, \ldots, x_{0k}$. To estimate the mean response at this point, define the vector

$$\mathbf{x}_0 = \begin{bmatrix} 1 \\ x_{01} \\ x_{02} \\ \vdots \\ x_{0k} \end{bmatrix}$$

The mean response at this point is $E(Y|\mathbf{x}_0) = \mu_{Y|\mathbf{x}_0} = \mathbf{x}_0'\boldsymbol{\beta}$, which is estimated by

$$\hat{\mu}_{Y|\mathbf{x}_0} = \mathbf{x}_0'\hat{\boldsymbol{\beta}} \tag{12-35}$$

This estimator is unbiased, since $E(\mathbf{x}_0'\hat{\boldsymbol{\beta}}) = \mathbf{x}_0'\boldsymbol{\beta} = E(Y|\mathbf{x}_0) = \mu_{Y|\mathbf{x}_0}$ and the variance of $\hat{\mu}_{Y|\mathbf{x}_0}$ is

$$V(\hat{\mu}_{Y|\mathbf{x}_0}) = \sigma^2\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0 \tag{12-36}$$

A $100(1 - \alpha)$ % CI on $\mu_{Y|\mathbf{x}_0}$ can be constructed from the statistic

$$\frac{\hat{\mu}_{Y|\mathbf{x}_0} - \mu_{Y|\mathbf{x}_0}}{\sqrt{\hat{\sigma}^2\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}} \tag{12-37}$$

Definition

For the multiple linear regression model, a $100(1 - \alpha)$% **confidence interval on the mean response** at the point $x_{01}, x_{02}, \ldots, x_{0k}$ is

$$\hat{\mu}_{Y|\mathbf{x}_0} - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0} \le \mu_{Y|\mathbf{x}_0} \le \hat{\mu}_{Y|\mathbf{x}_0} + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0} \tag{12-38}$$

Equation 12-38 is a CI about the regression plane (or hyperplane). It is the multiple regression generalization of Equation 11-31.

EXAMPLE 12-7

The engineer in Example 12-1 would like to construct a 95% CI on the mean pull strength for a wire bond with wire length $x_1 = 8$ and die height $x_2 = 275$. Therefore,

$$\mathbf{x}_0 = \begin{bmatrix} 1 \\ 8 \\ 275 \end{bmatrix}$$

The estimated mean response at this point is found from Equation 12-35 as

$$\hat{\mu}_{Y|\mathbf{x}_0} = \mathbf{x}_0'\hat{\boldsymbol{\beta}} = [1 \quad 8 \quad 275] \begin{bmatrix} 2.26379 \\ 2.74427 \\ 0.01253 \end{bmatrix} = 27.66$$

The variance of $\hat{\mu}_{Y|\mathbf{x}_0}$ is estimated by

$$\hat{\sigma}^2 \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0 = 5.2352\,[1\ 8\ 275] \times \begin{bmatrix} .214653 & -.007491 & -.000340 \\ -.007491 & .001671 & -.000019 \\ -.000340 & -.000019 & .0000015 \end{bmatrix} \begin{bmatrix} 1 \\ 8 \\ 275 \end{bmatrix}$$

$$= 5.2352\,(0.04444) = 0.23266$$

Therefore, a 95% CI on the mean pull strength at this point is found from Equation 12-38 as

$$27.66 - 2.074\sqrt{0.23266} \le \mu_{Y|\mathbf{x}_0} \le 27.66 + 2.074\sqrt{0.23266}$$

which reduces to

$$26.66 \le \mu_{Y|\mathbf{x}_0} \le 28.66$$

Some computer software packages will provide estimates of the mean for a point of interest $\mathbf{x}_0$ and the associated CI. Table 12-4 shows the Minitab output for Example 12-7. Both the estimate of the mean and the 95% CI are provided.

12-4 PREDICTION OF NEW OBSERVATIONS

A regression model can be used to predict new or **future observations** on the response variable Y corresponding to particular values of the independent variables, say, $x_{01}, x_{02}, \ldots, x_{0k}$. If $\mathbf{x}_0' = [1, x_{01}, x_{02}, \ldots, x_{0k}]$, a point estimate of the future observation Y_0 at the point $x_{01}, x_{02}, \ldots, x_{0k}$ is

$$\hat{y}_0 = \mathbf{x}_0'\hat{\boldsymbol{\beta}} \qquad (12\text{-}39)$$

A $100(1 - \alpha)\%$ **prediction interval for this future observation** is

$$\hat{y}_0 - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2(1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)}$$
$$\le Y_0 \le \hat{y}_0 + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2(1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)} \qquad (12\text{-}40)$$

This prediction interval is a generalization of the prediction interval given in Equation 11-33 for a future observation in simple linear regression. If you compare the prediction interval Equation 12-40 with the expression for the confidence interval on the mean, Equation 12-38, you will observe that the prediction interval is always wider than the confidence interval. The confidence interval expresses the error in estimating the mean of a distribution, while the prediction interval expresses the error in predicting a future observation from the distribution at

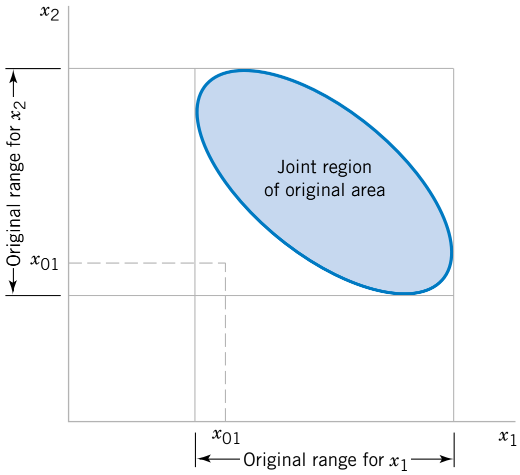

Figure 12-5 An example of extrapolation in multiple regression.

the point $\mathbf{x}_0$. This must include the error in estimating the mean at that point, as well as the inherent variability in the random variable Y at the same value $\mathbf{x} = \mathbf{x}_0$.

In predicting new observations and in estimating the mean response at a given point $x_{01}, x_{02}, \ldots, x_{0k}$, we must be careful about **extrapolating** beyond the region containing the original observations. It is very possible that a model that fits well in the region of the original data will no longer fit well outside of that region. In multiple regression it is often easy to inadvertently extrapolate, since the levels of the variables $(x_{i1}, x_{i2}, \ldots, x_{ik})$, $i = 1, 2, \ldots, n$, jointly define the region containing the data. As an example, consider Fig. 12-5, which illustrates the region containing the observations for a two-variable regression model. Note that the point (x_{01}, x_{02}) lies within the ranges of both regressor variables x_1 and x_2, but it is outside the region that is actually spanned by the original observations. Thus, either predicting the value of a new observation or estimating the mean response at this point is an extrapolation of the original regression model.

EXAMPLE 12-8

Suppose that the engineer in Example 12-1 wishes to construct a 95% prediction interval on the wire bond pull strength when the wire length is $x_1 = 8$ and the die height is $x_2 = 275$. Note that $\mathbf{x}_0' = [1 \quad 8 \quad 275]$, and the point estimate of the pull strength is $\hat{y}_0 = \mathbf{x}_0'\hat{\beta} = 27.66$. Also, in Example 12-7 we calculated $\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0 = 0.04444$. Therefore, from Equation 12-40 we have

$$27.66 - 2.074\sqrt{5.2352(1 + 0.04444)} \leq Y_0 \leq 27.66 + 2.074\sqrt{5.2352(1 + 0.04444)}$$

and the 95% prediction interval is

$$22.81 \leq Y_0 \leq 32.51$$

Notice that the prediction interval is wider than the confidence interval on the mean response at the same point, calculated in Example 12-7. The Minitab output in Table 12-4 also displays this prediction interval.

EXERCISES FOR SECTIONS 12-3 AND 12-4

12-26. Consider the soil absorption data in Exercise 12-2.

(a) Find a 95% confidence interval on the regression coefficient β_1.

(b) Find a 95% confidence interval on mean soil absorption index when $x_1 = 200$ and $x_2 = 50$.

(c) Find a 95% prediction interval on the soil absorption index when $x_1 = 200$ and $x_2 = 50$.

12-27. Consider the NFL data in Exercise 12-4.

(a) Find a 95% confidence interval on β_8.

(b) What is the estimated standard error of $\hat{\mu}_{Y|\mathbf{x}_0}$ when $x_2 =$ 2000 yards, $x_7 = 60\%$, and $x_8 = 1800$ yards?

(c) Find a 95% confidence interval on the mean number of games won when $x_2 = 2000$, $x_7 = 60$, and $x_8 = 1800$.

12-28. Consider the gasoline mileage data in Exercise 12-5.

(a) Find 99% confidence intervals on β_1 and β_6.

(b) Find a 99% confidence interval on the mean of Y when $x_1 = 300$ and $x_6 = 4$.

(c) Fit a new regression model to these data using x_1, x_2, x_6, and x_{10} as the regressors. Find 99% confidence intervals on the regression coefficients in this new model.

(d) Compare the lengths of the confidence intervals on β_1 and β_6 from part (c) with those found in part (a). Which intervals are longer? Does this offer any insight about adding the variables x_2 and x_{10} to the model?

12-29. Consider the electric power consumption data in Exercise 12-6.

(a) Find 95% confidence intervals on β_1, β_2, β_3, and β_4.

(b) Find a 95% confidence interval on the mean of Y when $x_1 = 75$, $x_2 = 24$, $x_3 = 90$, and $x_4 = 98$.

(c) Find a 95% prediction interval on the power consumption when $x_1 = 75$, $x_2 = 24$, $x_3 = 90$, and $x_4 = 98$.

12-30. Consider the bearing wear data in Exercise 12-7.

(a) Find 99% confidence intervals on β_1 and β_2.

(b) Recompute the confidence intervals in part (a) after the interaction term x_1x_2 is added to the model. Compare the lengths of these confidence intervals with those computed in part (a). Do the lengths of these intervals provide any information about the contribution of the interaction term in the model?

12-31. Consider the wire bond pull strength data in Exercise 12-8.

(a) Find 95% confidence interval on the regression coefficients.

(b) Find a 95% confidence interval on mean pull strength when $x_2 = 20$, $x_3 = 30$, $x_4 = 90$ and $x_5 = 2.0$.

(c) Find a 95% prediction interval on pull strength when $x_2 = 20$, $x_3 = 30$, $x_4 = 90$, and $x_5 = 2.0$.

12-32. Consider the semiconductor data in Exercise 12-9.

(a) Find 99% confidence intervals on the regression coefficients.

(b) Find a 99% prediction interval on HFE when $x_1 = 14.5$, $x_2 = 220$, and $x_3 = 5.0$.

(c) Find a 99% confidence interval on mean HFE when $x_1 = 14.5$, $x_2 = 220$, and $x_3 = 5.0$.

12-33. Consider the heat treating data from Exercise 12-10.

(a) Find 95% confidence intervals on the regression coefficients.

(b) Find a 95% confidence interval on mean PITCH when TEMP = 1650, SOAKTIME = 1.00, SOAKPCT = 1.10, DIFFTIME = 1.00, and DIFFPCT = 0.80.

12-34. Reconsider the heat treating data in Exercises 12-10 and 12-24, where we fit a model to PITCH using regressors x_1 = SOAKTIME × SOAKPCT and x_2 = DIFFTIME × DIFFPCT.

(a) Using the model with regressors x_1 and x_2, find a 95% confidence interval on mean PITCH when SOAKTIME = 1.00, SOAKPCT = 1.10, DIFFTIME = 1.00, and DIFFPCT = 0.80.

(b) Compare the length of this confidence interval with the length of the confidence interval on mean PITCH at the same point from Exercise 12-33 part (b), where an additive model in SOAKTIME, SOAKPCT, DIFFTIME, and DIFFPCT was used. Which confidence interval is shorter? Does this tell you anything about which model is preferable?

12-35. Consider the NHL data in Exercise 12-11.

(a) Find a 95% confidence interval on the regression coefficient for the variable "Pts."

(b) Fit a simple linear regression model relating the response variable "wins" to the regressor "Pts."

(c) Find a 95% confidence interval on the slope for the simple linear regression model from part (b).

(d) Compare the lengths of the two confidence intervals computed in parts (a) and (c). Which interval is shorter? Does this tell you anything about which model is preferable?

12-5 MODEL ADEQUACY CHECKING

12-5.1 Residual Analysis

The **residuals** from the multiple regression model, defined by $e_i = y_i - \hat{y}_i$, play an important role in judging model adequacy just as they do in simple linear regression. As noted in Section 11-7.1, several residual plots are often useful; these are illustrated in Example 12-9. It is also

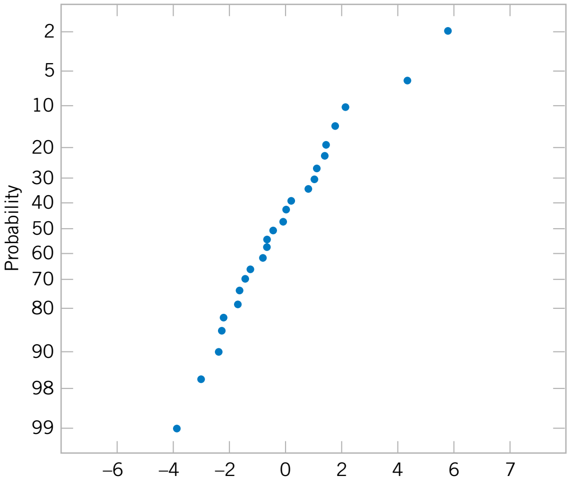

Figure 12-6 Normal probability plot of residuals.

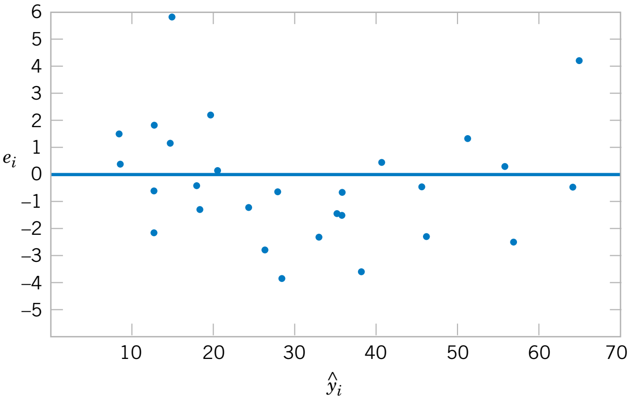

Figure 12-7 Plot of residuals against $\hat{y}$.

helpful to plot the residuals against variables not presently in the model that are possible candidates for inclusion. Patterns in these plots may indicate that the model may be improved by adding the candidate variable.

EXAMPLE 12-9

The residuals for the model from Example 12-1 are shown in Table 12-3. A normal probability plot of these residuals is shown in Fig. 12-6. No severe deviations from normality are obviously apparent, although the two largest residuals ($e_{15} = 5.88$ and $e_{17} = 4.33$) do not fall extremely close to a straight line drawn through the remaining residuals.

The **standardized residuals**

$$d_i = \frac{e_i}{\sqrt{MS_E}} = \frac{e_i}{\sqrt{\hat{\sigma}^2}} \tag{12-41}$$

are often more useful than the ordinary residuals when assessing residual magnitude. The standardized residuals corresponding to e_{15} and e_{17} are $d_{15} = 5.88/\sqrt{5.2352} = 2.57$ and $d_{17} = 4.33/\sqrt{4.2352} = 1.89$, and they do not seem unusually large. Inspection of the data does not reveal any error in collecting observations 15 and 17, nor does it produce any other reason to discard or modify these two points.

The residuals are plotted against $\hat{y}$ in Fig. 12-7, and against x_1 and x_2 in Figs. 12-8 and 12-9, respectively.* The two largest residuals, e_{15} and e_{17}, are apparent. Figure 12-8 gives some indication that the model underpredicts the pull strength for assemblies with short wire length ($x_1 \leq 6$) and long wire length ($x_1 \geq 15$) and overpredicts the strength for assemblies with intermediate wire length ($7 \leq x_1 \leq 14$). The same impression is obtained from Fig. 12-7.

*There are other methods, described in Montgomery, Peck, and Vining (2001) and Myers (1990), that plot a modified version of the residual, called a **partial residual,** against each regressor. These partial residual plots are useful in displaying the relationship between the response y and each individual regressor.

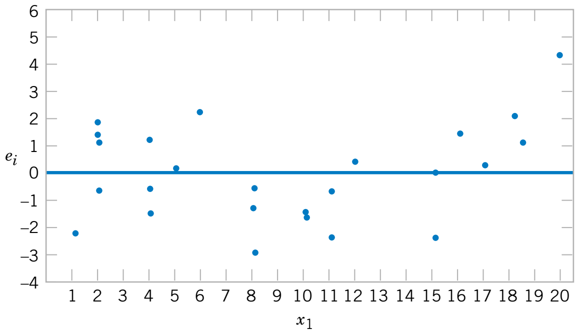

Figure 12-8 Plot of residuals against x_1.

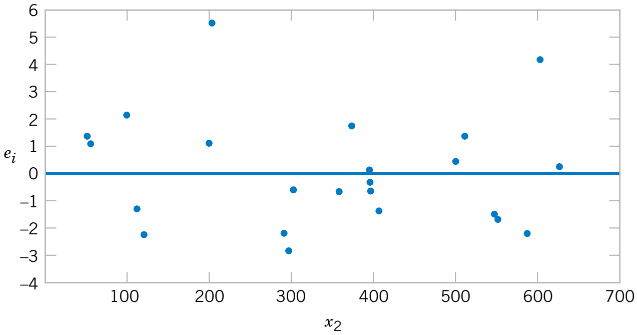

Figure 12-9 Plot of residuals against x_2.

Either the relationship between strength and wire length is not linear (requiring that a term involving x_1^2, say, be added to the model), or other regressor variables not presently in the model affected the response.

In Example 12-9 we used the standardized residuals $d_i = e_i/\sqrt{\hat{\sigma}^2}$ as a measure of residual magnitude. Some analysts prefer to plot standardized residuals instead of ordinary residuals, because the standardized residuals are scaled so that their standard deviation is approximately unity. Consequently, large residuals (that may indicate possible outliers or unusual observations) will be more obvious from inspection of the residual plots.

Many regression computer programs compute other types of scaled residuals. One of the most popular is the **studentized residual**

$$r_i = \frac{e_i}{\sqrt{\hat{\sigma}^2(1 - h_{ii})}} \qquad i = 1, 2, \ldots, n \qquad (12\text{-}42)$$

where h_{ii} is the ith diagonal element of the matrix

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

The **H** matrix is sometimes called the **"hat" matrix,** since

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}$$

Thus **H** transforms the observed values of **y** into a vector of fitted values $\hat{\mathbf{y}}$.

Since each row of the matrix **X** corresponds to a vector, say $\mathbf{x}_i' = [1, x_{i1}, x_{i2}, \ldots, x_{ik}]$, another way to write the diagonal elements of the hat matrix is

$$h_{ii} = \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i \qquad (12\text{-}43)$$

Note that apart from σ^2, h_{ii} is the variance of the fitted value $\hat{y}_i$. The quantities h_{ii} were used in the computation of the confidence interval on the mean response in Section 12-3.2.

Under the usual assumptions that the model errors are independently distributed with mean zero and variance σ^2, we can show that the variance of the ith residual e_i is

$$V(e_i) = \sigma^2(1 - h_{ii}), \qquad i = 1, 2, \ldots, n$$

Furthermore, the h_{ii} elements must fall in the interval $0 < h_{ii} \leq 1$. This implies that the standardized residuals understate the true residual magnitude; thus, the studentized residuals would be a better statistic to examine in evaluating potential **outliers.**

To illustrate, consider the two observations identified in Example 12-9 as having residuals that might be unusually large, observations 15 and 17. The standardized residuals are

$$d_{15} = \frac{e_{15}}{\sqrt{\hat{\sigma}^2}} = \frac{5.88}{\sqrt{5.2352}} = 2.57 \qquad \text{and} \qquad d_{17} = \frac{e_{17}}{\sqrt{MS_E}} = \frac{4.33}{\sqrt{5.2352}} = 1.89$$

Now $h_{15,15} = 0.0737$ and $h_{17,17} = 0.2593$, so the studentized residuals are

$$r_{15} = \frac{e_{15}}{\sqrt{\hat{\sigma}^2(1 - h_{15,15})}} = \frac{5.88}{\sqrt{5.2352(1 - 0.0737)}} = 2.67$$

and

$$r_{17} = \frac{e_{17}}{\sqrt{\hat{\sigma}^2(1 - h_{17,17})}} = \frac{4.33}{\sqrt{5.2352(1 - 0.2593)}} = 2.20$$

Notice that the studentized residuals are larger than the corresponding standardized residuals. However, the studentized residuals are still not so large as to cause us serious concern about possible outliers.

12-5.2 Influential Observations

When using multiple regression, we occasionally find that some subset of the observations is unusually influential. Sometimes these influential observations are relatively far away from the vicinity where the rest of the data were collected. A hypothetical situation for two variables is depicted in Fig. 12-10, where one observation in x-space is remote from the rest of the data. The disposition of points in the x-space is important in determining the properties of the model. For example, point (x_{i1}, x_{i2}) in Fig. 12-10 may be very influential in determining R^2, the estimates of the regression coefficients, and the magnitude of the error mean square.

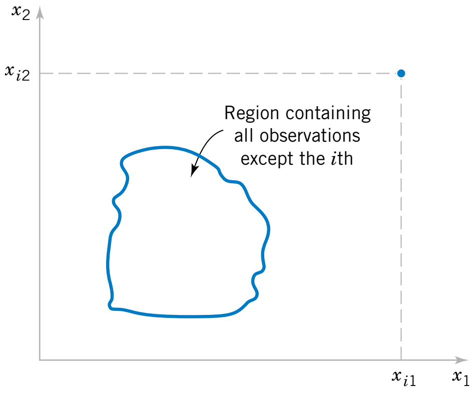

Figure 12-10 A point that is remote in x-space.

We would like to examine the influential points to determine whether they control many model properties. If these influential points are "bad" points, or erroneous in any way, they should be eliminated. On the other hand, there may be nothing wrong with these points, but at least we would like to determine whether or not they produce results consistent with the rest of the data. In any event, even if an influential point is a valid one, if it controls important model properties, we would like to know this, since it could have an impact on the use of the model.

Montgomery, Peck, and Vining (2001) and Myers (1990) describe several methods for detecting influential observations. An excellent diagnostic is the **distance measure** developed by Dennis R. Cook. This is a measure of the squared distance between the usual least squares estimate of $\boldsymbol{\beta}$ based on all n observations and the estimate obtained when the ith point is removed, say, $\hat{\boldsymbol{\beta}}_{(i)}$. The **Cook distance measure** is

$$D_i = \frac{(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})}{p\hat{\sigma}^2} \qquad i = 1, 2, \ldots, n$$

Clearly, if the ith point is influential, its removal will result in $\hat{\boldsymbol{\beta}}_{(i)}$ changing considerably from the value $\hat{\boldsymbol{\beta}}$. Thus, a large value of D_i implies that the ith point is influential. The statistic D_i is actually computed using

$$D_i = \frac{r_i^2}{p}\frac{h_{ii}}{(1 - h_{ii})} \qquad i = 1, 2, \ldots, n \qquad (12\text{-}44)$$

From Equation 12-44 we see that D_i consists of the squared studentized residual, which reflects how well the model fits the ith observation y_i [recall that $r_i = e_i/\sqrt{\hat{\sigma}^2(1 - h_{ii})}$] and a component that measures how far that point is from the rest of the data $[h_{ii}/(1 - h_{ii})$ is a measure of the distance of the ith point from the centroid of the remaining $n - 1$ points]. A value of $D_i > 1$ would indicate that the point is influential. Either component of D_i (or both) may contribute to a large value.

EXAMPLE 12-10 Table 12-11 lists the values of the hat matrix diagonals h_{ii} and Cook's distance measure D_i for the wire bond pull strength data in Example 12-1. To illustrate the calculations, consider the first observation:

$$\begin{aligned} D_1 &= \frac{r_1^2}{p} \cdot \frac{h_{11}}{(1 - h_{11})} \\ &= -\frac{[e_1/\sqrt{MS_E(1 - h_{11})}]^2}{p} \cdot \frac{h_{11}}{(1 - h_{11})} \\ &= \frac{[1.57/\sqrt{5.2352(1 - 0.1573)}]^2}{3} \cdot \frac{0.1573}{(1 - 0.1573)} \\ &= 0.035 \end{aligned}$$

The Cook distance measure D_i does not identify any potentially influential observations in the data, for no value of D_i exceeds unity.

Table 12-11 Influence Diagnostics for the Wire Bond Pull Strength Data 2

Observations		Cook's Distance Measure	Observations		Cook's Distance Measure
i	h_{ii}	D_i	i	h_{ii}	D_i
1	0.1573	0.035	14	0.1129	0.003
2	0.1116	0.012	15	0.0737	0.187
3	0.1419	0.060	16	0.0879	0.001
4	0.1019	0.021	17	0.2593	0.565
5	0.0418	0.024	18	0.2929	0.155
6	0.0749	0.007	19	0.0962	0.018
7	0.1181	0.036	20	0.1473	0.000
8	0.1561	0.020	21	0.1296	0.052
9	0.1280	0.160	22	0.1358	0.028
10	0.0413	0.001	23	0.1824	0.002
11	0.0925	0.013	24	0.1091	0.040
12	0.0526	0.001	25	0.0729	0.000
13	0.0820	0.001			

EXERCISES FOR SECTION 12-5

12-36. Consider the regression model for the NFL data in Exercise 12-4.

(a) What proportion of total variability is explained by this model?

(b) Construct a normal probability plot of the residuals. What conclusion can you draw from this plot?

(c) Plot the residuals versus $\hat{y}$ and versus each regressor, and comment on model adequacy.

(d) Are there any influential points in these data?

12-37. Consider the gasoline mileage data in Exercise 12-5.

(a) What proportion of total variability is explained by this model?

(b) Construct a normal probability plot of the residuals and comment on the normality assumption.

(c) Plot residuals versus $\hat{y}$ and versus each regressor. Discuss these residual plots.

(d) Calculate Cook's distance for the observations in this data set. Are any observations influential?

12-38. Consider the electric power consumption data in Exercise 12-6.

(a) Calculate R^2 for this model. Interpret this quantity.

(b) Plot the residuals versus $\hat{y}$. Interpret this plot.

(c) Construct a normal probability plot of the residuals and comment on the normality assumption.

12-39. Consider the wear data in Exercise 12-7.

(a) Find the value of R^2 when the model uses the regressors x_1 and x_2.

(b) What happens to the value of R^2 when an interaction term x_1x_2 is added to the model? Does this necessarily imply that adding the interaction term is a good idea?

12-40. For the regression model for the wire bond pull strength data in Exercise 12-8.

(a) Plot the residuals versus $\hat{y}$ and versus the regressors used in the model. What information do these plots provide?

(b) Construct a normal probability plot of the residuals. Are there reasons to doubt the normality assumption for this model?

(c) Are there any indications of influential observations in the data?

12-41. Consider the semiconductor HFE data in Exercise 12-9.

(a) Plot the residuals from this model versus $\hat{y}$. Comment on the information in this plot.

(b) What is the value of R^2 for this model?

(c) Refit the model using log HFE as the response variable.

(d) Plot the residuals versus predicted log HFE for the model in part (c). Does this give any information about which model is preferable?

(e) Plot the residuals from the model in part (d) versus the regressor x_3. Comment on this plot.

(f) Refit the model to log HFE using x_1, x_2, and $1/x_3$, as the regressors. Comment on the effect of this change in the model.

12-42. Consider the regression model for the heat treating data in Exercise 12-10.

(a) Calculate the percent of variability explained by this model.

(b) Construct a normal probability plot for the residuals. Comment on the normality assumption.

(c) Plot the residuals versus $\hat{y}$ and interpret the display.

(d) Calculate Cook's distance for each observation and provide an interpretation of this statistic.

12-43. In Exercise 12-24 we fit a model to the response PITCH in the heat treating data of Exercise 12-10 using new regressors x_1 = SOAKTIME × SOAKPCT and x_2 = DIFFTIME × DIFFPCT.

(a) Calculate the R^2 for this model and compare it to the value of R^2 from the original model in Exercise 12-10. Does this provide some information about which model is preferable?

(b) Plot the residuals from this model versus $\hat{y}$ and on a normal probability scale. Comment on model adequacy.

(c) Find the values of Cook's distance measure. Are any observations unusually influential?

12-44. Consider the regression model for the NHL data from Exercise 12-11.

(a) Fit a model using "pts" as the only regressor.

(b) How much variability is explained by this model?

(c) Plot the residuals versus $\hat{y}$ and comment on model adequacy.

(d) Plot the residuals versus "PPG," the points scored while in power play. Does this indicate that the model would be better if this variable were included?

12-45. The diagonal elements of the hat matrix are often used to denote **leverage**—that is, a point that is unusual in its location in the x-space and that may be influential. Generally, the ith point is called a **leverage point** if its hat diagonal h_{ii} exceeds $2p/n$, which is twice the average size of all the hat diagonals. Recall that $p = k + 1$.

(a) Table 12-11 contains the hat diagonal for the wire bond pull strength data used in Example 12-1. Find the average size of these elements.

(b) Based on the criterion above, are there any observations that are leverage points in the data set?

12-6 ASPECTS OF MULTIPLE REGRESSION MODELING

In this section we briefly discuss several other aspects of building multiple regression models. For more extensive presentations of these topics and additional examples refer to Montgomery, Peck, and Vining (2001) and Myers (1990).

12-6.1 Polynomial Regression Models

The linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ is a general model that can be used to fit any relationship that is **linear in the unknown parameters $\boldsymbol{\beta}$.** This includes the important class of **polynomial regression models.** For example, the second-degree polynomial in one variable

$$Y = \beta_0 + \beta_1 x + \beta_{11} x^2 + \epsilon \qquad (12\text{-}45)$$

and the second-degree polynomial in two variables

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \epsilon \qquad (12\text{-}46)$$

are linear regression models.

Polynomial regression models are widely used when the response is curvilinear, because the general principles of multiple regression can be applied. The following example illustrates some of the types of analyses that can be performed.

EXAMPLE 12-11

Sidewall panels for the interior of an airplane are formed in a 1500-ton press. The unit manufacturing cost varies with the production lot size. The data shown below give the average cost per unit (in hundreds of dollars) for this product (y) and the production lot size (x). The scatter diagram, shown in Fig. 12-11, indicates that a second-order polynomial may be appropriate.

y	1.81	1.70	1.65	1.55	1.48	1.40	1.30	1.26	1.24	1.21	1.20	1.18
x	20	25	30	35	40	50	60	65	70	75	80	90

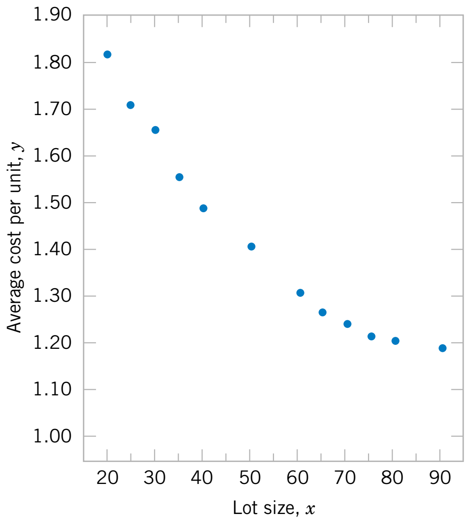

Figure 12-11 Data for Example 12-11.

We will fit the model

$$Y = \beta_0 + \beta_1 x + \beta_{11} x^2 + \epsilon$$

The **y** vector, **X** matrix, and $\boldsymbol{\beta}$ vector are as follows:

$$\mathbf{y} = \begin{bmatrix} 1.81 \\ 1.70 \\ 1.65 \\ 1.55 \\ 1.48 \\ 1.40 \\ 1.30 \\ 1.26 \\ 1.24 \\ 1.21 \\ 1.20 \\ 1.18 \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & 20 & 400 \\ 1 & 25 & 625 \\ 1 & 30 & 900 \\ 1 & 35 & 1225 \\ 1 & 40 & 1600 \\ 1 & 50 & 2500 \\ 1 & 60 & 3600 \\ 1 & 65 & 4225 \\ 1 & 70 & 4900 \\ 1 & 75 & 5625 \\ 1 & 80 & 6400 \\ 1 & 90 & 8100 \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_{11} \end{bmatrix}$$

Solving the normal equations $\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$ gives the fitted model

$$\hat{y} = 2.19826629 - 0.02252236x + 0.00012507x^2$$

The test for significance of regression is shown in Table 12-12. Since $f_0 = 2171.07$ is significant at 1%, we conclude that at least one of the parameters β_1 and β_{11} is not zero. Furthermore, the standard tests for model adequacy do not reveal any unusual behavior, and we would conclude that this is a reasonable model for the sidewall panel cost data.

Table 12-12 Test for Significance of Regression for the Second-Order Model in Example 12-11

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	f_0	P-value
Regression	0.5254	2	0.262700	2171.07	5.18E-15
Error	0.0011	9	0.000121		
Total	0.5265	11			

In fitting polynomials, we generally like to use the **lowest-degree model** consistent with the data. In this example, it would seem logical to investigate the possibility of dropping the quadratic term from the model. That is, we would like to test

$$H_0\text{: } \beta_{11} = 0$$
$$H_1\text{: } \beta_{11} \neq 0$$

The general regression significance test can be used to test this hypothesis. We need to determine the "extra sum of squares" due to β_{11}, or

$$SS_R(\beta_{11}|\beta_1,\beta_0) = SS_R(\beta_1,\beta_{11}|\beta_0) - SS_R(\beta_1|\beta_0)$$

The sum of squares $SS_R(\beta_1,\beta_{11}|\beta_0) = 0.5254$ from Table 12-12. To find $SS_R(\beta_1|\beta_0)$, we fit a simple linear regression model to the original data, yielding

$$\hat{y} = 1.90036320 - 0.00910056x$$

It can be easily verified that the regression sum of squares for this model is

$$SS_R(\beta_1|\beta_0) = 0.4942$$

Therefore, the extra sum of the squares due to β_{11}, given that β_1 and β_0 are in the model, is

$$\begin{aligned} SS_R(\beta_{11}|\beta_1,\beta_0) &= SS_R(\beta_1,\beta_{11}|\beta_0) - SS_R(\beta_1|\beta_0) \\ &= 0.5254 - 0.4942 \\ &= 0.0312 \end{aligned}$$

The analysis of variance, with the test of H_0: $\beta_{11} = 0$ incorporated into the procedure, is displayed in Table 12-13. Note that the quadratic term contributes significantly to the model.

Table 12-13 Analysis of Variance for Example 12-11, Showing the Test for H_0: $\beta_{11} = 0$

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	f_0	P-value
Regression	$SS_R(\beta_1,\beta_{11}|\beta_0) = 0.5254$	2	0.262700	2171.07	5.18E-15
Linear	$SS_R(\beta_1|\beta_0) = 0.4942$	1	0.494200	4084.30	1.17E-15
Quadratic	$SS_R(\beta_{11}|\beta_0,\beta_1) = 0.0312$	1	0.031200	258.18	5.51E-9
Error	0.0011	9	0.00121		
Total	0.5265	11			

12-6.2 Categorical Regressors and Indicator Variables

The regression models presented in previous sections have been based on **quantitative** variables, that is, variables that are measured on a numerical scale. For example, variables such as temperature, pressure, distance, and voltage are quantitative variables. Occasionally, we need to incorporate **categorical,** or **qualitative,** variables in a regression model. For example, suppose that one of the variables in a regression model is the operator who is associated with each observation y_i. Assume that only two operators are involved. We may wish to assign different levels to the two operators to account for the possibility that each operator may have a different effect on the response.

The usual method of accounting for the different levels of a qualitative variable is to use **indicator variables.** For example, to introduce the effect of two different operators into a regression model, we could define an indicator variable as follows:

$$x = \begin{cases} 0 \text{ if the observation is from operator 1} \\ 1 \text{ if the observation is from operator 2} \end{cases}$$

In general, a qualitative variable with r-levels can be modeled by $r - 1$ indicator variables, which are assigned the value of either zero or one. Thus, if there are *three* operators, the different levels will be accounted for by the *two* indicator variables defined as follows:

x_1	x_2	
0	0	if the observation is from operator 1
1	0	if the observation is from operator 2
0	1	if the observation is from operator 3

Indicator variables are also referred to as **dummy** variables. The following example [from Montgomery, Peck, and Vining (2001)] illustrates some of the uses of indicator variables; for other applications, see Montgomery, Peck, and Vining (2001).

EXAMPLE 12-12 A mechanical engineer is investigating the surface finish of metal parts produced on a lathe and its relationship to the speed (in revolutions per minute) of the lathe. The data are shown in Table 12-14. Note that the data have been collected using two different types of cutting tools. Since the type of cutting tool likely affects the surface finish, we will fit the model

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

where Y is the surface finish, x_1 is the lathe speed in revolutions per minute, and x_2 is an indicator variable denoting the type of cutting tool used; that is,

$$x_2 = \begin{cases} 0, \text{ for tool type 302} \\ 1, \text{ for tool type 416} \end{cases}$$

The parameters in this model may be easily interpreted. If $x_2 = 0$, the model becomes

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

which is a straight-line model with slope β_1 and intercept β_0. However, if $x_2 = 1$, the model becomes

$$Y = \beta_0 + \beta_1 x_1 + \beta_2(1) + \epsilon = (\beta_0 + \beta_2) + \beta_1 x_1 + \epsilon$$

Table 12-14 Surface Finish Data for Example 12-13

Observation Number, i	Surface Finish y_i	RPM	Type of Cutting Tool	Observation Number, i	Surface Finish y_i	RPM	Type of Cutting Tool
1	45.44	225	302	11	33.50	224	416
2	42.03	200	302	12	31.23	212	416
3	50.10	250	302	13	37.52	248	416
4	48.75	245	302	14	37.13	260	416
5	47.92	235	302	15	34.70	243	416
6	47.79	237	302	16	33.92	238	416
7	52.26	265	302	17	32.13	224	416
8	50.52	259	302	18	35.47	251	416
9	45.58	221	302	19	33.49	232	416
10	44.78	218	302	20	32.29	216	416

which is a straight-line model with slope β_1 and intercept $\beta_0 + \beta_2$. Thus, the model $Y = \beta_0 + \beta_1 x + \beta_2 x_2 + \epsilon$ implies that surface finish is linearly related to lathe speed and that the slope β_1 does not depend on the type of cutting tool used. However, the type of cutting tool does affect the intercept, and β_2 indicates the change in the intercept associated with a change in tool type from 302 to 416.

The **X** matrix and **y** vector for this problem are as follows:

$$\mathbf{X} = \begin{bmatrix} 1 & 225 & 0 \\ 1 & 200 & 0 \\ 1 & 250 & 0 \\ 1 & 245 & 0 \\ 1 & 235 & 0 \\ 1 & 237 & 0 \\ 1 & 265 & 0 \\ 1 & 259 & 0 \\ 1 & 221 & 0 \\ 1 & 218 & 0 \\ 1 & 224 & 1 \\ 1 & 212 & 1 \\ 1 & 248 & 1 \\ 1 & 260 & 1 \\ 1 & 243 & 1 \\ 1 & 238 & 1 \\ 1 & 224 & 1 \\ 1 & 251 & 1 \\ 1 & 232 & 1 \\ 1 & 216 & 1 \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} 45.44 \\ 42.03 \\ 50.10 \\ 48.75 \\ 47.92 \\ 47.79 \\ 52.26 \\ 50.52 \\ 45.58 \\ 44.78 \\ 33.50 \\ 31.23 \\ 37.52 \\ 37.13 \\ 34.70 \\ 33.92 \\ 32.13 \\ 35.47 \\ 33.49 \\ 32.29 \end{bmatrix}$$

The fitted model is

$$\hat{y} = 14.27620 + 0.14115x_1 - 13.28020x_2$$

Table 12-15 Analysis of Variance of Example 12-12

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	f_0	P-value
Regression	1012.0595	2	506.0297	1103.69	1.02E-18
$SS_R(\beta_1 \| \beta_0)$	130.6091	1	130.6091	284.87	4.70E-12
$SS_R(\beta_2 \| \beta_1,\beta_0)$	881.4504	1	881.4504	1922.52	6.24E-19
Error	7.7943	17	0.4508		
Total	1019.8538	19			

The analysis of variance for this model is shown in Table 12-15. Note that the hypothesis H_0: $\beta_1 = \beta_2 = 0$ (significance of regression) would be rejected at any reasonable level of significance because the P-value is very small. This table also contains the sums of squares

$$\begin{aligned} SS_R &= SS_R(\beta_1,\beta_2|\beta_0) \\ &= SS_R(\beta_1|\beta_0) + SS_R(\beta_2|\beta_1,\beta_0) \end{aligned}$$

so a test of the hypothesis H_0: $\beta_2 = 0$ can be made. Since this hypothesis is also rejected, we conclude that tool type has an effect on surface finish.

It is also possible to use indicator variables to investigate whether tool type affects both the slope and intercept. Let the model be

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon$$

where x_2 is the indicator variable. Now if tool type 302 is used, $x_2 = 0$, and the model is

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

If tool type 416 is used, $x_2 = 1$, and the model becomes

$$\begin{aligned} Y &= \beta_0 + \beta_1 x_1 + \beta_2 + \beta_3 x_1 + \epsilon \\ &= (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_1 + \epsilon \end{aligned}$$

Note that β_2 is the change in the intercept and that β_3 is the change in slope produced by a change in tool type.

Another method of analyzing these data is to fit separate regression models to the data for each tool type. However, the indicator variable approach has several advantages. First, only one regression model must be fit. Second, by pooling the data on both tool types, more degrees of freedom for error are obtained. Third, tests of both hypotheses on the parameters β_2 and β_3 are just special cases of the extra sum of squares method.

12-6.3 Selection of Variables and Model Building

An important problem in many applications of regression analysis involves selecting the set of regressor variables to be used in the model. Sometimes previous experience or underlying theoretical considerations can help the analyst specify the set or regressor variables to use in a particular situation. Usually, however, the problem consists of selecting an appropriate set of regressors from a set that quite likely includes all the important variables, but we are sure that not all these candidate regressors are necessary to adequately model the response Y.

In such a situation, we are interested in **variable selection;** that is, screening the candidate variables to obtain a regression model that contains the "best" subset of regressor variables. We would like the final model to contain enough regressor variables so that in the intended use of the model (prediction, for example) it will perform satisfactorily. On the other hand, to keep model maintenance costs to a minimum and to make the model easy to use, we would like the model to use as few regressor variables as possible. The compromise between these conflicting objectives is often called finding the "best" regression equation. However, in most problems, no single regression model is "best" in terms of the various evaluation criteria that have been proposed. A great deal of judgment and experience with the system being modeled is usually necessary to select an appropriate set of regressor variables for a regression equation.

No single algorithm will always produce a good solution to the variable selection problem. Most of the currently available procedures are search techniques, and to perform satisfactorily, they require interaction with judgment by the analyst. We now briefly discuss some of the more popular variable selection techniques. We assume that there are K candidate regressors, $x_1, x_2, \ldots, x_K$, and a single response variable y. All models will include an intercept term β_0, so the model with *all* variables included would have $K + 1$ terms. Furthermore, the functional form of each candidate variable (for example, $x_1 = 1/x$, $x_2 = \ln x$, etc.) is correct.

All Possible Regressions

This approach requires that the analyst fit all the regression equations involving one candidate variable, all regression equations involving two candidate variables, and so on. Then these equations are evaluated according to some suitable criteria to select the "best" regression model. If there are K candidate regressors, there are 2^K total equations to be examined. For example, if $K = 4$, there are $2^4 = 16$ possible regression equations; while if $K = 10$, there are $2^{10} = 1024$ possible regression equations. Hence, the number of equations to be examined increases rapidly as the number of candidate variables increases. However, there are some very efficient computing algorithms for all possible regressions available and they are widely implemented in statistical software, so it is a very practical procedure unless the number of candidate regressors is fairly large.

Several criteria may be used for evaluating and comparing the different regression models obtained. A commonly used criterion is based on the value of R^2 or the value of the adjusted R^2, R^2_{adj}. Basically, the analyst continues to increase the number of variables in the model until the increase in R^2 or the adjusted R^2_{adj} is small. Often, we will find that the R^2_{adj} will stabilize and actually begin to decrease as the number of variables in the model increases. Usually, the model that maximizes R^2_{adj} is considered to be a good candidate for the best regression equation. Because we can write $R^2_{\text{adj}} = 1 - \{MS_E/[SS_E/(n - 1)]\}$ and $SS_E/(n - 1)$ is a constant, the model that maximizes the R^2_{adj} value also minimizes the mean square error, so this is a very attractive criterion.

Another criterion used to evaluate regression models is the C_p statistic, which is a measure of the total mean square error for the regression model. We define the total standardized mean square error for the regression model as

$$\begin{aligned}\Gamma_p &= \frac{1}{\sigma^2}\sum_{i=1}^{n} E[\hat{Y}_i - E(Y_i)]^2 \\ &= \frac{1}{\sigma^2}\left\{\sum_{i=1}^{n}[E(Y_i) - E(\hat{Y}_i)]^2 + \sum_{i=1}^{n} V(\hat{Y}_i)\right\} \\ &= \frac{1}{\sigma^2}[(\text{bias})^2 + \text{variance}]\end{aligned}$$

We use the mean square error from the *full* $K + 1$ term model as an estimate of σ^2; that is, $\hat{\sigma}^2 = MS_E(K + 1)$. Then an estimator of Γ_p is [see Montgomery, Peck, and Vining (2001) or Myers (1990) for the details]:

$$C_p = \frac{SS_E(p)}{\hat{\sigma}^2} - n + 2p \qquad (12\text{-}47)$$

If the p-term model has negligible bias, it can be shown that

$$E(C_p \mid \text{zero bias}) = p$$

Therefore, the values of C_p for each regression model under consideration should be evaluated relative to p. The regression equations that have negligible bias will have values of C_p that are close to p, while those with significant bias will have values of C_p that are significantly greater than p. We then choose as the "best" regression equation either a model with *minimum* C_p or a model with a slightly larger C_p, that does not contain as much bias (i.e., $C_p \cong p$).

The PRESS statistic can also be used to evaluate competing regression models. PRESS is an acronym for Prediction Error Sum of Squares, and it is defined as the sum of the squares of the differences between each observation y_i and the corresponding predicted value based on a model fit to the *remaining* $n - 1$ points, say $\hat{y}_{(i)}$. So PRESS provides a measure of how well the model is likely to perform when predicting **new data**, or data that was not used to fit the regression model. The computing formula for PRESS is

$$\text{PRESS} = \sum_{i=1}^{n} (y_i - \hat{y}_{(i)})^2 = \sum_{i=1}^{n} \left(\frac{e_i}{1 - h_{ii}} \right)^2$$

where $e_i = y_i - \hat{y}_i$ is the usual residual. Thus PRESS is easy to calculate from the standard least squares regression results. Models that have small values of PRESS are preferred.

EXAMPLE 12-13 Table 12-16 presents data on taste-testing 38 brands of pinot noir wine (the data were first reported in an article by Kwan, Kowalski, and Skogenboe in an article in the *Journal of Agricultural and Food Chemistry*, Vol. 27, 1979, and it also appears as one of the default data sets in Minitab). The response variable is y = quality, and we wish to find the "best" regression equation that relates quality to the other five parameters.

Figure 12-12 is the matrix of scatter plots for the wine quality data, as constructed by Minitab. We notice that there are some indications of possible linear relationships between quality and the regressors, but there is no obvious visual impression of which regressors would be appropriate. Table 12-16 lists the all possible regressions output from Minitab. In this analysis, we asked Minitab to present the best three equations for each subset size. Note that Minitab reports the values of R^2, R^2_{adj}, C_p, and $S = \sqrt{MS_E}$ for each model. From Table 12-17 we see that the three-variable equation with x_2 = aroma, x_4 = flavor, and x_5 = oakiness produces the minimum C_p equation, whereas the four-variable model, which adds

Table 12-16 Wine Quality Data

	x_1	x_2	x_3	x_4	x_5	y
	Clarity	Aroma	Body	Flavor	Oakiness	Quality
1	1.0	3.3	2.8	3.1	4.1	9.8
2	1.0	4.4	4.9	3.5	3.9	12.6
3	1.0	3.9	5.3	4.8	4.7	11.9
4	1.0	3.9	2.6	3.1	3.6	11.1
5	1.0	5.6	5.1	5.5	5.1	13.3
6	1.0	4.6	4.7	5.0	4.1	12.8
7	1.0	4.8	4.8	4.8	3.3	12.8
8	1.0	5.3	4.5	4.3	5.2	12.0
9	1.0	4.3	4.3	3.9	2.9	13.6
10	1.0	4.3	3.9	4.7	3.9	13.9
11	1.0	5.1	4.3	4.5	3.6	14.4
12	0.5	3.3	5.4	4.3	3.6	12.3
13	0.8	5.9	5.7	7.0	4.1	16.1
14	0.7	7.7	6.6	6.7	3.7	16.1
15	1.0	7.1	4.4	5.8	4.1	15.5
16	0.9	5.5	5.6	5.6	4.4	15.5
17	1.0	6.3	5.4	4.8	4.6	13.8
18	1.0	5.0	5.5	5.5	4.1	13.8
19	1.0	4.6	4.1	4.3	3.1	11.3
20	0.9	3.4	5.0	3.4	3.4	7.9
21	0.9	6.4	5.4	6.6	4.8	15.1
22	1.0	5.5	5.3	5.3	3.8	13.5
23	0.7	4.7	4.1	5.0	3.7	10.8
24	0.7	4.1	4.0	4.1	4.0	9.5
25	1.0	6.0	5.4	5.7	4.7	12.7
26	1.0	4.3	4.6	4.7	4.9	11.6
27	1.0	3.9	4.0	5.1	5.1	11.7
28	1.0	5.1	4.9	5.0	5.1	11.9
29	1.0	3.9	4.4	5.0	4.4	10.8
30	1.0	4.5	3.7	2.9	3.9	8.5
31	1.0	5.2	4.3	5.0	6.0	10.7
32	0.8	4.2	3.8	3.0	4.7	9.1
33	1.0	3.3	3.5	4.3	4.5	12.1
34	1.0	6.8	5.0	6.0	5.2	14.9
35	0.8	5.0	5.7	5.5	4.8	13.5
36	0.8	3.5	4.7	4.2	3.3	12.2
37	0.8	4.3	5.5	3.5	5.8	10.3
38	0.8	5.2	4.8	5.7	3.5	13.2

x_1 = clarity to the previous three regressors, results in maximum R^2_{adj} (or minimum MS_E). The three-variable model is

$$\hat{y} = 6.47 + 0.580x_2 + 1.20x_4 - 0.602x_5$$

and the four-variable model is

$$\hat{y} = 4.99 + 1.79x_1 + 0.530x_2 + 1.26x_4 - 0.659x_5$$

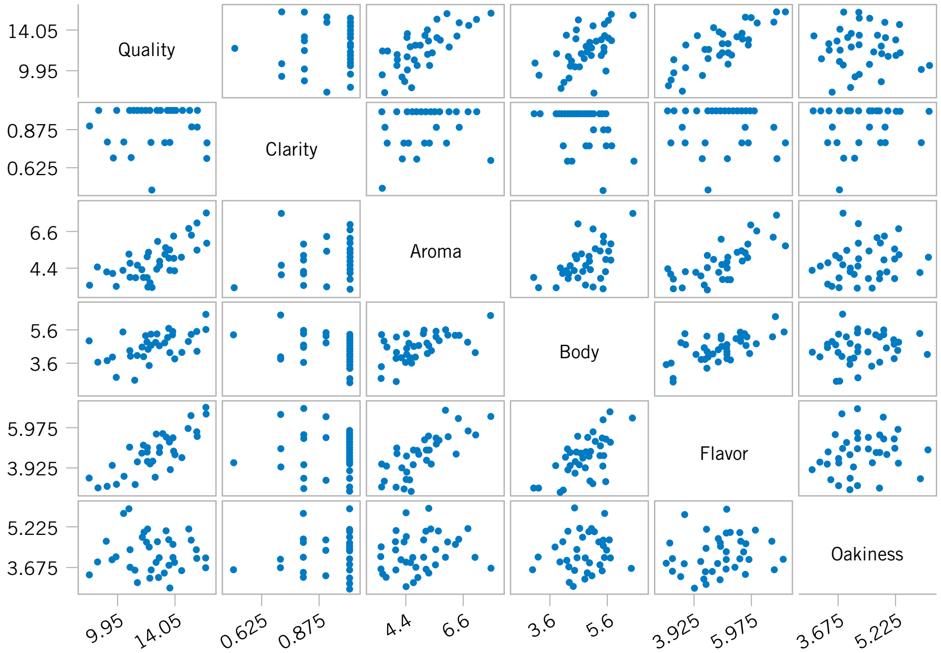

Figure 12-12 A Matrix of Scatter Plots from Minitab for the Wine Quality Data.

Table 12-17 Minitab All Possible Regressions Output for the Wine Quality Data

```
Best Subsets Regression: Quality versus Clarity, Aroma, . . .
Response is Quality

                                              O
                                      C       a
                                      l     F k
                                      a A   l i
                                      r r B a n
                                      i o o v e
                                      t m d o s
Vars  R-Sq  R-Sq (adj)   C–p       S  y a y r s
   1  62.4        61.4   9.0  1.2712        X
   1  50.0        48.6  23.2  1.4658    X
   1  30.1        28.2  46.0  1.7335      X
   2  66.1        64.2   6.8  1.2242        X X
   2  65.9        63.9   7.1  1.2288    X   X
   2  63.3        61.2  10.0  1.2733  X     X
   3  70.4        67.8   3.9  1.1613    X   X X
   3  68.0        65.2   6.6  1.2068  X     X X
   3  66.5        63.5   8.4  1.2357      X X X
   4  71.5        68.0   4.7  1.1568  X X   X X
   4  70.5        66.9   5.8  1.1769    X X X X
   4  69.3        65.6   7.1  1.1996  X   X X X
   5  72.1        67.7   6.0  1.1625  X X X X X
```

These models should now be evaluated further using residuals plots and the other techniques discussed earlier in the chapter, to see if either model is satisfactory with respect to the underlying assumptions and to determine if one of them is preferable. It turns out that the residual plots do not reveal any major problems with either model. The value of PRESS for the three-variable model is 56.0525 and for the four-variable model it is 60.3927. Since PRESS is smaller in the model with three regressors, and since it is the model with the fewest predictors, it would likely be the preferred choice.

Stepwise Regression

Stepwise regression is probably the most widely used variable selection technique. The procedure iteratively constructs a sequence of regression models by adding or removing variables at each step. The criterion for adding or removing a variable at any step is usually expressed in terms of a partial F-test. Let f_{in} be the value of the F-random variable for adding a variable to the model, and let f_{out} be the value of the F-random variable for removing a variable from the model. We must have $f_{\text{in}} \geq f_{\text{out}}$, and usually $f_{\text{in}} = f_{\text{out}}$.

Stepwise regression begins by forming a one-variable model using the regressor variable that has the highest correlation with the response variable Y. This will also be the regressor producing the largest F-statistic. For example, suppose that at this step, x_1 is selected. At the second step, the remaining $K - 1$ candidate variables are examined, and the variable for which the partial F-statistic

$$F_j = \frac{SS_R(\beta_j | \beta_1, \beta_0)}{MS_E(x_j, x_1)} \tag{12-48}$$

is a maximum is added to the equation, provided that $f_j > f_{\text{in}}$. In equation 12-48, $MS_E\ (x_j, x_1)$ denotes the mean square for error for the model containing both x_1 and x_j. Suppose that this procedure indicates that x_2 should be added to the model. Now the stepwise regression algorithm determines whether the variable x_1 added at the first step should be removed. This is done by calculating the F-statistic

$$F_1 = \frac{SS_R(\beta_1 | \beta_2, \beta_0)}{MS_E(x_1, x_2)} \tag{12-49}$$

If the calculated value $f_1 < f_{\text{out}}$, the variable x_1 is removed; otherwise it is retained, and we would attempt to add a regressor to the model containing both x_1 and x_2.

In general, at each step the set of remaining candidate regressors is examined, and the regressor with the largest partial F-statistic is entered, provided that the observed value of f exceeds f_{in}. Then the partial F-statistic for each regressor in the model is calculated, and the regressor with the smallest observed value of F is deleted if the observed $f < f_{\text{out}}$. The procedure continues until no other regressors can be added to or removed from the model.

Stepwise regression is almost always performed using a computer program. The analyst exercises control over the procedure by the choice of f_{in} and f_{out}. Some stepwise regression computer programs require that numerical values be specified for f_{in} and f_{out}. Since the number of degrees of freedom on MS_E depends on the number of variables in the model, which changes from step to step, a fixed value of f_{in} and f_{out} causes the type I and type II error rates to vary. Some computer programs allow the analyst to specify the type I error levels for f_{in} and f_{out}. However, the "advertised" significance level is not the true level, because the variable selected is the one that maximizes (or minimizes) the partial F-statistic at that stage. Sometimes it is useful to experiment with different values of f_{in} and f_{out} (or different advertised

Table 12-18 Minitab Stepwise Regression Output for the Wine Quality Data

```
Stepwise Regression: Quality versus Clarity, Aroma, . . .
Alpha-to-Enter: 0.15 Alpha-to-Remove: 0.15

Response is Quality on 5 predictors, with N = 38

    Step              1          2          3
Constant          4.941      6.912      6.467

Flavor             1.57       1.64       1.20
T-Value            7.73       8.25       4.36
P-Value           0.000      0.000      0.000

Oakiness                     −0.54      −0.60
T-Value                      −1.95      −2.28
P-Value                      0.059      0.029

Aroma                                    0.58
T-Value                                  2.21
P-Value                                 0.034

S                  1.27       1.22       1.16
R-Sq              62.42      66.11      70.38
R-Sq(adj)         61.37      64.17      67.76
C–p                 9.0        6.8        3.9
```

type I error rates) in several different runs to see if this substantially affects the choice of the final model.

EXAMPLE 12-14 Table 12-18 gives the Minitab stepwise regression output for the wine quality data. Minitab uses fixed values of α for entering and removing variables. The default level is $\alpha = 0.15$ for both decisions. The output in Table 12-18 uses the default value. Notice that the variables were entered in the order Flavor (step 1), Oakiness (step 2), and Aroma (step 3) and that no variables were removed. No other variable could be entered, so the algorithm terminated. This is the three-variable model found by all possible regressions that results in a minimum value of C_p.

Forward Selection

The **forward selection** procedure is a variation of stepwise regression and is based on the principle that regressors should be added to the model one at a time until there are no remaining candidate regressors that produce a significant increase in the regression sum of squares. That is, variables are added one at a time as long as their partial F-value exceeds f_{in}. Forward selection is a simplification of stepwise regression that omits the partial F-test for deleting variables from the model that have been added at previous steps. This is a potential weakness of forward selection; that is, the procedure does not explore the effect that adding a regressor at the current step has on regressor variables added at earlier steps. Notice that if we were to apply forward selection to the wine quality data, we would obtain exactly the same results as we did with stepwise regression in Example 12-14, since stepwise regression terminated without deleting a variable.

Table 12-19 Minitab Backward Elimination Output for the Wine Quality Data

```
Stepwise Regression: Quality versus Clarity, Aroma, . . .
Backward elimination. Alpha-to-Remove: 0.1

Response is Quality on 5 predictors, with N = 38

    Step            1          2          3
Constant        3.997      4.986      6.467

Clarity           2.3        1.8
T-Value          1.35       1.12
P-Value         0.187      0.269

Aroma            0.48       0.53       0.58
T-Value          1.77       2.00       2.21
P-Value         0.086      0.054      0.034

Body             0.27
T-Value          0.82
P-Value         0.418

Flavor           1.17       1.26       1.20
T-Value          3.84       4.52       4.36
P-Value         0.001      0.000      0.000

Oakiness        −0.68      −0.66      −0.60
T-Value         −2.52      −2.46      −2.28
P-Value         0.017      0.019      0.029

S                1.16       1.16       1.16
R-Sq            72.06      71.47      70.38
R-Sq(adj)       67.69      68.01      67.76
C–p               6.0        4.7        3.9
```

Backward Elimination

The **backward elimination** algorithm begins with all K candidate regressors in the model. Then the regressor with the smallest partial F-statistic is deleted if this F-statistic is insignificant, that is, if $f < f_{\text{out}}$. Next, the model with $K - 1$ regressors is fit, and the next regressor for potential elimination is found. The algorithm terminates when no further regressor can be deleted.

Table 12-19 shows the Minitab output for backward elimination applied to the wine quality data. The α value for removing a variable is $\alpha = 0.10$. Notice that this procedure removes Body at step 1 and then Clarity at step 2, terminating with the three-variable model found previously.

Some Comments on Final Model Selection

We have illustrated several different approaches to the selection of variables in multiple linear regression. The final model obtained from any model-building procedure should be subjected to the usual adequacy checks, such as residual analysis, lack-of-fit testing, and examination of the effects of influential points. The analyst may also consider augmenting the original set of candidate variables with cross-products, polynomial terms, or other transformations of the original variables that might improve the model. A major criticism of variable selection methods such as stepwise regression is that the analyst may conclude there is one “best” regression equation. Generally, this is not the case, because several equally good regression models can often be used. One way to avoid this problem is to use several different model-building techniques and see if different models result. For example, we have found the same model for the wine quality data using stepwise regression, forward selection, and backward elimination. The same model was also one of the two best found from all possible regressions. The results from variable selection methods frequently do not agree, so this is a good indication that the three-variable model is the best regression equation.

If the number of candidate regressors is not too large, the all-possible regressions method is recommended. We usually recommend using the minimum MS_E and C_p evaluation criteria in conjunction with this procedure. The all-possible regressions approach can find the "best" regression equation with respect to these criteria, while stepwise-type methods offer no such assurance. Furthermore, the all-possible regressions procedure is not distorted by dependencies among the regressors, as stepwise-type methods are.

12-6.4 Multicollinearity

In multiple regression problems, we expect to find dependencies between the response variable Y and the regressors x_j. In most regression problems, however, we find that there are also dependencies among the regressor variables x_j. In situations where these dependencies are strong, we say that **multicollinearity** exists. Multicollinearity can have serious effects on the estimates of the regression coefficients and on the general applicability of the estimated model.

The effects of multicollinearity may be easily demonstrated. The diagonal elements of the matrix $\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}$ can be written as

$$C_{jj} = \frac{1}{(1 - R_j^2)} \qquad j = 1, 2, \ldots, k$$

where R_j^2 is the coefficient of multiple determination resulting from regressing x_j on the other $k - 1$ regressor variables. Clearly, the stronger the linear dependency of x_j on the remaining regressor variables, and hence the stronger the multicollinearity, the larger the value of R_j^2 will be. Recall that $V(\hat{\beta}_j) = \sigma^2 C_{jj}$. Therefore, we say that the variance of $\hat{\beta}_j$ is "inflated" by the quantity $(1 - R_j^2)^{-1}$. Consequently, we define the **variance inflation factor** for β_j as

$$VIF(\beta_j) = \frac{1}{(1 - R_j^2)} \qquad j = 1, 2, \ldots, k \qquad \text{(12-50)}$$

These factors are an important measure of the extent to which multicollinearity is present.

Although the estimates of the regression coefficients are very imprecise when multicollinearity is present, the fitted model equation may still be useful. For example, suppose we wish to predict new observations on the response. If these predictions are interpolations in the original region of the x-space where the multicollinearity is in effect, satisfactory predictions will often be obtained, because while individual β_j may be poorly estimated, the function $\sum_{j=1}^{k} \beta_j x_{ij}$ may be estimated quite well. On the other hand, if the prediction of new observations requires extrapolation beyond the original region of the x-space where the data were collected, generally we would expect to obtain poor results. Extrapolation usually requires good estimates of the individual model parameters.

Multicollinearity arises for several reasons. It will occur when the analyst collects data such that a linear constraint holds approximately among the columns of the **X** matrix. For example, if four regressor variables are the components of a mixture, such a constraint will always exist because the sum of the components is always constant. Usually, these constraints do not hold exactly, and the analyst might not know that they exist.

The presence of multicollinearity can be detected in several ways. Two of the more easily understood of these will be discussed briefly.

1. The **variance inflation factors,** defined in equation 12-50, are very useful measures of multicollinearity. The larger the variance inflation factor, the more severe the multicollinearity. Some authors have suggested that if any variance inflation factor exceeds 10, multicollinearity is a problem. Other authors consider this value too liberal and suggest that the variance inflation factors should not exceed 4 or 5. Minitab will calculate the variance inflation factors. Table 12-4 presents the Minitab multiple regression output for the wire bond pull strength data. Since both VIF_1 and VIF_2 are small, there is no problem with multicollinearity.
2. If the F-test for significance of regression is significant, but tests on the individual regression coefficients are not significant, multicollinearity may be present.

Several remedial measures have been proposed for solving the problem of multicollinearity. Augmenting the data with new observations specifically designed to break up the approximate linear dependencies that currently exist is often suggested. However, this is sometimes impossible because of economic reasons or because of the physical constraints that relate the x_j. Another possibility is to delete certain variables from the model, but this approach has the disadvantage of discarding the information contained in the deleted variables.

Since multicollinearity primarily affects the stability of the regression coefficients, it would seem that estimating these parameters by some method that is less sensitive to multicollinearity than ordinary least squares would be helpful. Several methods have been suggested. One alternative to ordinary least squares, **ridge regression,** can be useful in combating multicollinearity. For more details on ridge regression, see Section 12-6.5 on the CD material or the more extensive presentations in Montgomery, Peck, and Vining (2001) and Myers (1990).

12-6.5 Ridge Regression (CD Only)

12-6.6 Nonlinear Regression (CD Only)

EXERCISES FOR SECTION 12-6

12-46. An article entitled "A Method for Improving the Accuracy of Polynomial Regression Analysis" in the *Journal of Quality Technology* (1971, pp. 149–155) reported the following data on y = ultimate shear strength of a rubber compound (psi) and x = cure temperature (°F).

y	770	800	840	810
x	280	284	292	295
y	735	640	590	560
x	298	305	308	315

(a) Fit a second-order polynomial to these data.
(b) Test for significance of regression using $\alpha = 0.05$.
(c) Test the hypothesis that $\beta_{11} = 0$ using $\alpha = 0.05$.
(d) Compute the residuals from part (a) and use them to evaluate model adequacy.

12-47. Consider the following data, which result from an experiment to determine the effect of x = test time in hours at a particular temperature on y = change in oil viscosity:

y	−1.42	−1.39	−1.55	−1.89	−2.43
x	.25	.50	.75	1.00	1.25
y	−3.15	−4.05	−5.15	−6.43	−7.89
x	1.50	1.75	2.00	2.25	2.50

(a) Fit a second-order polynomial to the data.
(b) Test for significance of regression using $\alpha = 0.05$.
(c) Test the hypothesis that $\beta_{11} = 0$ using $\alpha = 0.05$.

(d) Compute the residuals from part (a) and use them to evaluate model adequacy.

12-48. When fitting polynomial regression models, we often subtract $\bar{x}$ from each x value to produce a "centered" regressor $x' = x - \bar{x}$. This reduces the effects of dependencies among the model terms and often leads to more accurate estimates of the regression coefficients. Using the data from Exercise 12-46, fit the model $Y = \beta_0^* + \beta_1^* x' + \beta_{11}^*(x')^2 + \epsilon$. Use the results to estimate the coefficients in the uncentered model $Y = \beta_0 + \beta_1 x + \beta_{11} x^2 + \epsilon$.

12-49. Suppose that we use a standardized variable $x' = (x - \bar{x})/s_x$, where s_x is the standard deviation of x, in constructing a polynomial regression model. Using the data in Exercise 12-46 and the standardized variable approach, fit the model $Y = \beta_0^* + \beta_1^* x' + \beta_{11}^*(x')^2 + \epsilon$.

(a) What value of y do you predict when $x = 285$°F?

(b) Estimate the regression coefficients in the unstandardized model $Y = \beta_0 + \beta_1 x + \beta_{11} x^2 + \epsilon$.

(c) What can you say about the relationship between SS_E and R^2 for the standardized and unstandardized models?

(d) Suppose that $y' = (y - \bar{y})/s_y$ is used in the model along with x'. Fit the model and comment on the relationship between SS_E and R^2 in the standardized model and the unstandardized model.

12-50. The following data shown were collected during an experiment to determine the change in thrust efficiency (y, in percent) as the divergence angle of a rocket nozzle (x) changes:

y	24.60	24.71	23.90	39.50	39.60	57.12
x	4.0	4.0	4.0	5.0	5.0	6.0
y	67.11	67.24	67.15	77.87	80.11	84.67
x	6.5	6.5	6.75	7.0	7.1	7.3

(a) Fit a second-order model to the data.

(b) Test for significance of regression and lack of fit using $\alpha = 0.05$.

(c) Test the hypothesis that $\beta_{11} = 0$, using $\alpha = 0.05$.

(d) Plot the residuals and comment on model adequacy.

(e) Fit a cubic model, and test for the significance of the cubic term using $\alpha = 0.05$.

12-51. An article in the *Journal of Pharmaceuticals Sciences* (Vol. 80, 1991, pp. 971–977) presents data on the observed mole fraction solubility of a solute at a constant temperature and the dispersion, dipolar, and hydrogen bonding Hansen partial solubility parameters. The data are as shown in the following table, where y is the negative logarithm of the mole fraction solubility, x_1 is the dispersion partial solubility, x_2 is the dipolar partial solubility, and x_3 is the hydrogen bonding partial solubility.

(a) Fit the model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{33} x_3^2 + \epsilon$.

(b) Test for significance of regression using $\alpha = 0.05$.

(c) Plot the residuals and comment on model adequacy.

Observation Number	y	x_1	x_2	x_3
1	0.22200	7.3	0.0	0.0
2	0.39500	8.7	0.0	0.3
3	0.42200	8.8	0.7	1.0
4	0.43700	8.1	4.0	0.2
5	0.42800	9.0	0.5	1.0
6	0.46700	8.7	1.5	2.8
7	0.44400	9.3	2.1	1.0
8	0.37800	7.6	5.1	3.4
9	0.49400	10.0	0.0	0.3
10	0.45600	8.4	3.7	4.1
11	0.45200	9.3	3.6	2.0
12	0.11200	7.7	2.8	7.1
13	0.43200	9.8	4.2	2.0
14	0.10100	7.3	2.5	6.8
15	0.23200	8.5	2.0	6.6
16	0.30600	9.5	2.5	5.0
17	0.09230	7.4	2.8	7.8
18	0.11600	7.8	2.8	7.7
19	0.07640	7.7	3.0	8.0
20	0.43900	10.3	1.7	4.2
21	0.09440	7.8	3.3	8.5
22	0.11700	7.1	3.9	6.6
23	0.07260	7.7	4.3	9.5
24	0.04120	7.4	6.0	10.9
25	0.25100	7.3	2.0	5.2
26	0.00002	7.6	7.8	20.7

(d) Use the extra sum of squares method to test the contribution of the second-order terms using $\alpha = 0.05$.

12-52. Consider the gasoline mileage data in Exercise 12-5.

(a) Discuss how you would model the information about the type of transmission in the car.

(b) Fit a regression model to the gasoline mileage using engine displacement, horsepower, and the type of transmission in the car as the regressors.

(c) Is there evidence that the type of transmission affects gasoline mileage performance?

12-53. Consider the tool life data in Example 12-12. Test the hypothesis that two different regression models (with different slopes and intercepts) are required to adequately model the data. Use indicator variables in answering this question.

12-54. Use the National Football League Team Performance data in Exercise 12-4 to build regression models using the following techniques:

(a) All possible regressions. Find the equations that minimize MS_E and that minimize C_p.

(b) Stepwise regression.

(c) Forward selection.
(d) Backward elimination.
(e) Comment on the various models obtained. Which model seems "best," and why?

12-55. Use the gasoline mileage data in Exercise 12-5 to build regression models using the following techniques:
(a) All possible regressions. Find the minimum C_p and minimum MS_E equations.
(b) Stepwise regression.
(c) Forward selection.
(d) Backward elimination.
(e) Comment on the various models obtained.

12-56. Consider the electric power data in Exercise 12-6. Build regression models for the data using the following techniques:
(a) All possible regressions.
(b) Stepwise regression.
(c) Forward selection.
(d) Backward elimination.
(e) Comment on the models obtained. Which model would you prefer?

12-57. Consider the wire bond pull strength data in Exercise 12-8. Build regression models for the data using the following methods:
(a) Stepwise regression.
(b) Forward selection.
(c) Backward elimination.
(d) Comment on the models obtained. Which model would you prefer?

12-58. Consider the NHL data in Exercise 12-11. Build regression models for these data using the following methods:
(a) Stepwise regression.
(b) Forward selection.
(c) Backward elimination.
(d) Which model would you prefer?

12-59. Consider the data in Exercise 12-51. Use all the terms in the full quadratic model as the candidate regressors.
(a) Use forward selection to identify a model.
(b) Use backward elimination to identify a model.
(c) Compare the two models obtained in parts (a) and (b). Which model would you prefer and why?

12-60. Find the minimum C_p equation and the equation that maximizes the adjusted R^2 statistic for the wire bond pull strength data in Exercise 12-8. Does the same equation satisfy both criteria?

12-61. For the NHL data in Exercise 12-11.
(a) Find the equation that minimizes C_p.
(b) Find the equation that minimizes MS_E.
(c) Find the equation that maximizes the adjusted R^2. Is this the same equation you found in part (b)?

12-62. We have used a sample of 30 observations to fit a regression model. The full model has nine regressors, the variance estimate is $\hat{\sigma}^2 = MS_E = 100$, and $R^2 = 0.92$.
(a) Calculate the F-statistic for testing significance of regression. Using $\alpha = 0.05$, what would you conclude?
(b) Suppose that we fit another model using only four of the original regressors and that the error sum of squares for this new model is 2200. Find the estimate of σ^2 for this new reduced model. Would you conclude that the reduced model is superior to the old one? Why?
(c) Find the value of C_p for the reduced model in part (b). Would you conclude that the reduced model is better than the old model?

12-63. A sample of 25 observations is used to fit a regression model in seven variables. The estimate of σ^2 for this full model is $MS_E = 10$.
(a) A forward selection algorithm has put three of the original seven regressors in the model. The error sum of squares for the three-variable model is $SS_E = 300$. Based on C_p, would you conclude that the three-variable model has any remaining bias?
(b) After looking at the forward selection model in part (a), suppose you could add one more regressor to the model. This regressor will reduce the error sum of squares to 275. Will the addition of this variable improve the model? Why?

Supplemental Exercises

12-64. The data shown in the table on page 464 represent the thrust of a jet-turbine engine (y) and six candidate regressors: x_1 = primary speed of rotation, x_2 = secondary speed of rotation, x_3 = fuel flow rate, x_4 = pressure, x_5 = exhaust temperature, and x_6 = ambient temperature at time of test.
(a) Fit a multiple linear regression model using x_3 = fuel flow rate, x_4 = pressure, and x_5 = exhaust temperature as the regressors.
(b) Test for significance of regression using $\alpha = 0.01$. Find the P-value for this test. What are your conclusions?
(c) Find the t-test statistic for each regressor. Using $\alpha = 0.01$, explain carefully the conclusion you can draw from these statistics.
(d) Find R^2 and the adjusted statistic for this model.
(e) Construct a normal probability plot of the residuals and interpret this graph.
(f) Plot the residuals versus $\hat{y}$. Are there any indications of inequality of variance or nonlinearity?
(g) Plot the residuals versus x_3. Is there any indication of nonlinearity?
(h) Predict the thrust for an engine for which $x_3 = 1670$, $x_4 = 170$, and $x_5 = 1589$.

12-65. Consider the engine thrust data in Exercise 12-64. Refit the model using $y^* = \ln y$ as the response variable and $x_3^* = \ln x_3$ as the regressor (along with x_4 and x_5).
(a) Test for significance of regression using $\alpha = 0.01$. Find the P-value for this test and state your conclusions.

Observation Number	x_1	x_2	x_3	x_4	x_5	y
1	3	3	3	3	0	0.787
2	8	30	8	8	0	0.293
3	3	6	6	6	0	1.710
4	4	4	4	12	0	0.203
5	8	7	6	5	0	0.806
6	10	20	5	5	0	4.713
7	8	6	3	3	25	0.607
8	6	24	4	4	25	9.107
9	4	10	12	4	25	9.210
10	16	12	8	4	25	1.365
11	3	10	8	8	25	4.554
12	8	3	3	3	25	0.293
13	3	6	3	3	50	2.252
14	3	8	8	3	50	9.167
15	4	8	4	8	50	0.694
16	5	2	2	2	50	0.379
17	2	2	2	3	50	0.485
18	10	15	3	3	50	3.345
19	15	6	2	3	50	0.208
20	15	6	2	3	75	0.201
21	10	4	3	3	75	0.329
22	3	8	2	2	75	4.966
23	6	6	6	4	75	1.362
24	2	3	8	6	75	1.515
25	3	3	8	8	75	0.751

(b) Use the t-statistic to test H_0: $\beta_j = 0$ versus H_1: $\beta_j \neq 0$ for each variable in the model. If $\alpha = 0.01$, what conclusions can you draw?

(c) Plot the residuals versus $\hat{y}^*$ and versus x_3^*. Comment on these plots. How do they compare with their counterparts obtained in Exercise 12-64 parts (f) and (g)?

12-66. The transient points of an electronic inverter are influenced by many factors. Table 12-20 gives data on the transient point (y, in volts) of PMOS-NMOS inverters and five candidate regressors:, x_1 = width of the NMOS device, x_2 = length of the NMOS device, x_3 = width of the PMOS device, x_4 = length of the PMOS device, and x_5 = temperature (°C).

(a) Fit the multiple linear regression model to these data. Test for significance of regression using $\alpha = 0.01$. Find the P-value for this test and use it to draw your conclusions.

(b) Test the contribution of each variable to the model using the t-test with $\alpha = 0.05$. What are your conclusions?

(c) Delete x_5 from the model. Test the new model for significance of regression. Also test the relative contribution of each regressor to the new model with the t-test. Using $\alpha = 0.05$, what are your conclusions?

(d) Notice that the MS_E for the model in part (c) is smaller than the MS_E for the full model in part (a). Explain why this has occurred.

(e) Calculate the studentized residuals. Do any of these seem unusually large?

(f) Suppose that you learn that the second observation was incorrectly recorded. Delete this observation and refit the model using x_1, x_2, x_3, and x_4 as the regressors. Notice that the R^2 for this model is considerably higher than the R^2 for either of the models fitted previously. Explain why the R^2 for this model has increased.

(g) Test the model from part (f) for significance of regression using $\alpha = 0.05$. Also investigate the contribution of each regressor to the model using the t-test with $\alpha = 0.05$. What conclusions can you draw?

(h) Plot the residuals from the model in part (f) versus $\hat{y}$ and versus each of the regressors x_1, x_2, x_3, and x_4. Comment on the plots.

12-67. Consider the inverter data in Exercise 12-66. Delete observation 2 from the original data. Define new variables as follows: $y^* = \ln y$, $x_1^* = 1/\sqrt{x_1}$, $x_2^* = \sqrt{x_2}$, $x_3^* = 1/\sqrt{x_3}$, and $x_4^* = \sqrt{x_4}$.

(a) Fit a regression model using these transformed regressors (do not use x_5).

(b) Test the model for significance of regression using $\alpha = 0.05$. Use the t-test to investigate the contribution of each variable to the model ($\alpha = 0.05$). What are your conclusions?

(c) Plot the residuals versus $\hat{y}^*$ and versus each of the transformed regressors. Comment on the plots.

12-68. Following are data on y = green liquor (g/l) and x = paper machine speed (feet per minute) from a Kraft paper machine. (The data were read from a graph in an article in the *Tappi Journal,* March 1986.)

y	16.0	15.8	15.6	15.5	14.8
x	1700	1720	1730	1740	1750
y	14.0	13.5	13.0	12.0	11.0
x	1760	1770	1780	1790	1795

(a) Fit the model $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$ using least squares.

(b) Test for significance of regression using $\alpha = 0.05$. What are your conclusions?

(c) Test the contribution of the quadratic term to the model, over the contribution of the linear term, using an F-statistic. If $\alpha = 0.05$, what conclusion can you draw?

(d) Plot the residuals from the model in part (a) versus $\hat{y}$. Does the plot reveal any inadequacies?

(e) Construct a normal probability plot of the residuals. Comment on the normality assumption.

Table 12-20

Observation Number	y	x_1	x_2	x_3	x_4	x_5	x_6
1	4540	2140	20640	30250	205	1732	99
2	4315	2016	20280	30010	195	1697	100
3	4095	1905	19860	29780	184	1662	97
4	3650	1675	18980	29330	164	1598	97
5	3200	1474	18100	28960	144	1541	97
6	4833	2239	20740	30083	216	1709	87
7	4617	2120	20305	29831	206	1669	87
8	4340	1990	19961	29604	196	1640	87
9	3820	1702	18916	29088	171	1572	85
10	3368	1487	18012	28675	149	1522	85
11	4445	2107	20520	30120	195	1740	101
12	4188	1973	20130	29920	190	1711	100
13	3981	1864	19780	29720	180	1682	100
14	3622	1674	19020	29370	161	1630	100
15	3125	1440	18030	28940	139	1572	101
16	4560	2165	20680	30160	208	1704	98
17	4340	2048	20340	29960	199	1679	96
18	4115	1916	19860	29710	187	1642	94
19	3630	1658	18950	29250	164	1576	94
20	3210	1489	18700	28890	145	1528	94
21	4330	2062	20500	30190	193	1748	101
22	4119	1929	20050	29960	183	1713	100
23	3891	1815	19680	29770	173	1684	100
24	3467	1595	18890	29360	153	1624	99
25	3045	1400	17870	28960	134	1569	100
26	4411	2047	20540	30160	193	1746	99
27	4203	1935	20160	29940	184	1714	99
28	3968	1807	19750	29760	173	1679	99
29	3531	1591	18890	29350	153	1621	99
30	3074	1388	17870	28910	133	1561	99
31	4350	2071	20460	30180	198	1729	102
32	4128	1944	20010	29940	186	1692	101
33	3940	1831	19640	29750	178	1667	101
34	3480	1612	18710	29360	156	1609	101
35	3064	1410	17780	28900	136	1552	101
36	4402	2066	20520	30170	197	1758	100
37	4180	1954	20150	29950	188	1729	99
38	3973	1835	19750	29740	178	1690	99
39	3530	1616	18850	29320	156	1616	99
40	3080	1407	17910	28910	137	1569	100

12-69. Consider the jet engine thrust data in Exercise 12-64.

(a) Use all possible regressions to select the best regression equation, where the model with the minimum value of MS_E is to be selected as "best."

(b) Repeat part (a) using the C_P criterion to identify the best equation.

(c) Use stepwise regression to select a subset regression model.

(d) Compare the models obtained in parts (a), (b), and (c) above.

12-70. Consider the electronic inverter data in Exercise 12-66 and 12-67. Define the response and regressors variables as in Exercise 12-67, and delete the second observation in the sample.

(a) Use all possible regressions to find the equation that minimizes C_p.

(b) Use all possible regressions to find the equation that minimizes MS_E.

(c) Use stepwise regression to select a subset regression model.

(d) Compare the models you have obtained.

12-71. Consider the three-variable regression model for the jet engine thrust data in Exercise 12-65. Calculate the variance inflation factors for this model. Would you conclude that multicollinearity is a problem in this model?

12-72. A multiple regression model was used to relate y = viscosity of a chemical product to x_1 = temperature and x_2 = reaction time. The data set consisted of $n = 15$ observations.

(a) The estimated regression coefficients were $\hat{\beta}_0 = 300.00$, $\hat{\beta}_1 = 0.85$, and $\hat{\beta}_2 = 10.40$. Calculate an estimate of mean viscosity when $x_1 = 100$°F and $x_2 = 2$ hours.

(b) The sums of squares were $SS_T = 1230.50$ and $SS_E = 120.30$. Test for significance of regression using $\alpha = 0.05$. What conclusion can you draw?

(c) What proportion of total variability in viscosity is accounted for by the variables in this model?

(d) Suppose that another regressor, x_3 = stirring rate, is added to the model. The new value of the error sum of squares is $SS_E = 117.20$. Has adding the new variable resulted in a smaller value of MS_E? Discuss the significance of this result.

(e) Calculate an F-statistic to assess the contribution of x_3 to the model. Using $\alpha = 0.05$, what conclusions do you reach?

12-73. An article in the *Journal of the American Ceramics Society* (Vol. 75, 1992, pp. 112–116) describes a process for immobilizing chemical or nuclear wastes in soil by dissolving the contaminated soil into a glass block. The authors mix CaO and Na_2O with soil and model viscosity and electrical conductivity. The electrical conductivity model involves six regressors, and the sample consists of $n = 14$ observations.

(a) For the six-regressor model, suppose that $SS_T = 0.50$ and $R^2 = 0.94$. Find SS_E and SS_R, and use this information to test for significance of regression with $\alpha = 0.05$. What are your conclusions?

(b) Suppose that one of the original regressors is deleted from the model, resulting in $R^2 = 0.92$. What can you conclude about the contribution of the variable that was removed? Answer this question by calculating an F-statistic.

(c) Does deletion of the regressor variable in part (b) result in a smaller value of MS_E for the five-variable model, in comparison to the original six-variable model? Comment on the significance of your answer.

MIND-EXPANDING EXERCISES

12-74. Consider a multiple regression model with k regressors. Show that the test statistic for significance of regression can be written as

$$F_0 = \frac{R^2/k}{(1 - R^2)/(n - k - 1)}$$

Suppose that $n = 20$, $k = 4$, and $R^2 = 0.90$. If $\alpha = 0.05$, what conclusion would you draw about the relationship between y and the four regressors?

12-75. A regression model is used to relate a response y to $k = 4$ regressors with $n = 20$. What is the smallest value of R^2 that will result in a significant regression if $\alpha = 0.05$? Use the results of the previous exercise. Are you surprised by how small the value of R^2 is?

12-76. Show that we can express the residuals from a multiple regression model as $\boldsymbol{e} = (\mathbf{I} - \mathbf{H})\mathbf{y}$, where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

12-77. Show that the variance of the ith residual e_i in a multiple regression model is $\sigma(1 - h_{ii})$ and that the covariance between e_i and e_j is $-\sigma^2 h_{ij}$, where the h's are the elements of $\mathbf{H} = \mathbf{X}(\mathbf{X}\ \mathbf{X})^{-1}\mathbf{X}'$.

12-78. Consider the multiple linear regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. If $\hat{\boldsymbol{\beta}}$ denotes the least squares estimator of $\boldsymbol{\beta}$, show that $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \mathbf{R}\boldsymbol{\epsilon}$, where $\mathbf{R} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

12-79. **Constrained Least Squares.** Suppose we wish to find the least squares estimator of $\boldsymbol{\beta}$ in the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ subject to a set of equality constraints, say, $\mathbf{T}\boldsymbol{\beta} = \mathbf{c}$.

(a) Show that the estimator is

$$\hat{\boldsymbol{\beta}}_c = \hat{\boldsymbol{\beta}} + (\mathbf{X}'\mathbf{X})^{-1} \times \mathbf{T}'[\mathbf{T}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{T}']^{-1}(\mathbf{c} - \mathbf{T}\hat{\boldsymbol{\beta}})$$

where $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

(b) Discuss situations where this model might be appropriate.

12-80. **Piecewise Linear Regression (I).** Suppose that y is piecewise linearly related to x. That is, different linear relationships are appropriate over the intervals $-\infty < x \le x^*$ and $x^* < x < \infty$. Show how indicator variables can be used to fit such a piecewise linear regression model, assuming that the point x^* is known.

12-81. **Piecewise Linear Regression (II).** Consider the piecewise linear regression model described in Exercise 12-79. Suppose that at the point x^* a discontinuity occurs in the regression function. Show how indicator variables can be used to incorporate the discontinuity into the model.

12-82. **Piecewise Linear Regression (III).** Consider the piecewise linear regression model described in Exercise 12-79. Suppose that the point x^* is not known with certainty and must be estimated. Suggest an approach that could be used to fit the piecewise linear regression model.

IMPORTANT TERMS AND CONCEPTS

In the E-book, click on any term or concept below to go to that subject.

All possible regressions
Analysis of variance test in multiple regression
Categorical variables as regressors
Confidence interval on the mean response
Extra sum of squares method
Inference (test and intervals) on individual model parameters
Influential observations
Model parameters and their interpretation in multiple regression
Outliers
Polynomial terms in a regression model
Prediction interval on a future observation
Residual analysis and model adequacy checking
Significance of regression
Stepwise regression and related methods

CD MATERIAL

Ridge regression
Nonlinear regression models

12-2.3 More about the Extra Sum of Squares Method (CD Only)

The extra sum of squares method for evaluating the contribution of one or more terms to a model is a very useful technique. Basically, one considers how much the **regression** or **model** sum of squares increases upon adding terms to a basic model. The expanded model is called the **full** model, and the basic model is called the **reduced** model. Although the development in the text is quite general, Example 12-5 illustrates the simplest case, one in which there is only one additional parameter in the full model. In this case, the partial F-test based on the extra sum of squares is equivalent to a t-test. When there is more than one additional parameter in the full model, the partial F-test is not equivalent to a t-test.

The extra sum of squares method is often used sequentially when fitting a polynomial model, such as

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

Here $SS_R(\beta_1 \mid \beta_0)$ would measure the contribution of the linear term over and above a model containing only a mean (β_0), $SS_R(\beta_2 \mid \beta_0, \beta_1)$ would measure the contribution of the quadratic terms over and above the linear, and $SS_R(\beta_3 \mid \beta_0, \beta_1, \beta_2)$ would measure the contribution of the cubic terms over and above the linear and the quadratic. This can be very useful in selecting the **order** of the polynomial to fit. Notice from Table 12-4 that Minitab automatically produces this sequential computation. Also, note that in a sequential partition of the model or regression sum of squares,

$$SS_R(\beta_1, \beta_2, \beta_3, \mid \beta_0) = SS_R(\beta_1 \mid \beta_0) + SS_R(\beta_2 \mid \beta_0, \beta_1) + SS_R(\beta_3 \mid \beta_0, \beta_1, \beta_2).$$

However, if we consider each variable as if it were the last to be added,

$$SS_R(\beta_1, \beta_2, \beta_3 \mid \beta_0) \neq SS_R(\beta_1 \mid \beta_0, \beta_2, \beta_3) + SS_R(\beta_2 \mid \beta_0, \beta_1, \beta_3)\, SS_R(\beta_3 \mid \beta_0, \beta_1, \beta_2)$$

As another illustration of the extra sum of squares method, consider the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \epsilon$$

Suppose that we are uncertain about the contribution of the second-order terms. We could evaluate this with a partial F-test by fitting the reduced model

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

and computing

$$SS_R(\beta_{12}, \beta_{11}, \beta_{22} \mid \beta_0, \beta_1, \beta_2) = SS_R(\beta_1, \beta_2, \beta_{12}, \beta_{11}, \beta_{22} \mid \beta_0) - SS_R(\beta_1, \beta_2 \mid \beta_0)$$

Finally, note that we have expressed the extra sum of squares as the difference in the regression sum of squares between the full model and the reduced model:

$$SS_R(\text{Extra}) = SS_R(\text{Full Model}) - SS_R(\text{Reduced Model})$$

Some authors write SS_R(Extra) as the difference between error or residual sums of squares for the two model.

Now

$$SS_T = SS_R(\text{Full Model}) + SS_E(\text{Full Model})$$

and

$$SS_T = SS_R(\text{Reduced Model}) + SS_E(\text{Reduced Model})$$

so

$$\begin{aligned} SS_R(\text{Extra}) &= SS_R(\text{Full Model}) - SS_R(\text{Reduced Model}) \\ &= SS_T - SS_E(\text{Full Model}) - [SS_T - SS_E(\text{Reduced Model})] \\ &= SS_E(\text{Reduced Model}) - SS_E(\text{Full Model}) \end{aligned}$$

Therefore, this is an equivalent way to write the extra sum of squares.

12-6.5 Ridge Regression (CD Only)

Since multicollinearity primarily affects the stability of the regression coefficients, it would seem that estimating these parameters by some method that is less sensitive to multicollinearity than ordinary least squares would be helpful. Several methods have been suggested. One alternative to ordinary least squares, **ridge regression,** can be useful in combating multicollinearity. In ridge regression, the parameter estimates are obtained from

$$\boldsymbol{\beta}^*(\theta) = (\mathbf{X}'\mathbf{X} + \theta\mathbf{I})^{-1}\mathbf{X}'\mathbf{y} \qquad \text{(S12-1)}$$

where $\theta \geq 0$ is a constant. Generally, values of θ in the interval $0 \leq \theta \leq 1$ are appropriate. The ridge estimator $\boldsymbol{\beta}^*(\theta)$ is not an unbiased estimator of $\boldsymbol{\beta}$, as is the ordinary least squares estimator $\hat{\boldsymbol{\beta}}$. Thus, ridge regression seeks to find a set of regression coefficients that are more "stable," in the sense that they have a small mean square error. Since multicollinearity usually results in ordinary least squares estimators that may have extremely large variances, ridge regression is suitable for situations where the multicollinearity problem exists.

To obtain the ridge regression estimator from Equation S12-1, we must specify a value for the constant θ. Generally, there is an "optimum" θ for any problem, but the simplest approach is to solve Equation S12-1 for several values of θ in the interval $0 \leq \theta \leq 1$. Then a plot of the value of $\boldsymbol{\beta}^*(\theta)$ is constructed. This display is called a **ridge trace.** The approximate value for θ is chosen subjectively by inspection of the ridge trace. Typically, its value is chosen to obtain stable parameter estimates. Generally, the variance of $\boldsymbol{\beta}^*(\theta)$ is a decreasing function of θ, while the squared bias $[\boldsymbol{\beta} - E(\boldsymbol{\beta}^*(\theta))]^2$ is an increasing function of θ. Choosing the value of θ involves trading off these two properties of $\boldsymbol{\beta}^*(\theta)$.

Extensive practical discussions of the use of ridge regression are in Montgomery, Peck, and Vining (2001) and Myers (1990). In addition, several other biased estimation techniques have been proposed for dealing with multicollinearity. Many regression computer packages incorporate ridge regression capability. SAS PROC REG will fit ridge regression models.

To illustrate ridge regression, consider the data in Table S12-1, which shows the heat generated in calories per gram for a particular type of cement as a function of the quantities of four additives (w_1, w_2, w_3, and w_4). We wish to fit a multiple linear regression model to these data. This is some very "classical" regression data, first analyzed by Anders Hald. Refer to Montgomery, Peck, and Vining (2001) for sources and more details.

Table S12-1 Cement Data

Observation Number	y	w_1	w_2	w_3	w_4
1	28.25	10	31	5	45
2	24.80	12	35	5	52
3	11.86	5	15	3	24
4	36.60	17	42	9	65
5	15.80	8	6	5	19
6	16.23	6	17	3	25
7	29.50	12	36	6	55
8	28.75	10	34	5	50
9	43.20	18	40	10	70
10	38.47	23	50	10	80
11	10.14	16	37	5	61
12	38.92	20	40	11	70
13	36.70	15	45	8	68
14	15.31	7	22	2	30
15	8.40	9	12	3	24

The data will be coded by defining a new set of scaled regressor variables as

$$x_{ij} = \frac{w_{ij} - \overline{w}_j}{\sqrt{S_{jj}}} \qquad i = 1, 2, \ldots, 15 \qquad j = 1, 2, 3, 4$$

where $S_{jj} = \sum_{i=1}^{n}(w_{ij} - \overline{w}_j)^2$ is the corrected sum of squares of the levels of w_j. The coded data are shown in Table S12-2. This transformation makes the column of ones in **X** orthogonal to the

Table S12-2 Coded Cement Data

Observation Number	y	x_1	x_2	x_3	x_4
1	28.25	−.12515	.00405	−.09206	−.05538
2	24.80	−.02635	.08495	−.09206	.03692
3	11.86	−.37217	−.31957	−.27614	−.03692
4	36.60	.22066	.22653	.27617	−.33226
5	15.80	−.22396	−.50161	−.09206	−.39819
6	16.23	−.32276	−.27912	−.27617	−.31907
7	29.50	−.02635	.10518	.00000	.07647
8	28.75	−.12515	.06472	−.09206	.01055
9	43.20	.27007	.18608	.36823	.27425
10	38.47	.51709	.38834	.36823	.40609
11	10.14	.17126	.12500	−.09206	.15558
12	38.92	.36887	.18608	.46029	.27425
13	36.70	.12186	.28721	.18411	.24788
14	15.31	−.27336	−.17799	−.36823	−.25315
15	8.40	−.17456	−.38025	−.27617	−.33226

other columns, so the intercept in this model will always be estimated by $\bar{y}$. The (4 × 4) $\mathbf{X}'\mathbf{X}$ matrix for the four coded variables is the correlation matrix

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1.00000 & 0.84894 & 0.91412 & 0.93367 \\ 0.84894 & 1.00000 & 0.76899 & 0.97567 \\ 0.91412 & 0.76899 & 1.00000 & 0.86784 \\ 0.93367 & 0.97567 & 0.86784 & 1.00000 \end{bmatrix}$$

This matrix contains several large correlation coefficients, and this may indicate significant multicollinearity. The inverse of $\mathbf{X}'\mathbf{X}$ is

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 20.769 & 25.813 & -0.608 & -44.042 \\ 25.813 & 74.486 & 12.597 & -107.710 \\ -0.608 & 12.597 & 8.274 & -18.903 \\ -44.042 & -107.710 & -18.903 & 163.620 \end{bmatrix}$$

The variance inflation factors are the main diagonal elements of this matrix. Note that three of the variance inflation factors exceed 10, a good indication that multicollinearity is present.

Equation S12-1 was solved for various values of θ, and the results are summarized in Table S12-3. The ridge trace is shown in Fig. S12-1. The instability of the least squares estimates $\beta_j^*(\theta = 0)$ is evident from inspection of the ridge trace. It is often difficult to choose a value of θ from the ridge trace that simultaneously stabilizes the estimates of all regression coefficients. Because the bias in the coefficients increases as θ increases, it is usually best to choose θ as small as possible; yet we want θ to be large enough to provide reasonable stability in the coefficients. In our example, most of the change in the regression coefficients has occurred when $0.05 < \theta < 0.1$. We will choose $\theta = 0.064$, which implies that the regression model is

$$\hat{y} = 25.53 - 18.0566x_1 + 17.2202x_2 + 36.0743x_3 + 4.7242x_4$$

using $\hat{\beta}_0 = \bar{y} = 25.53$. Converting the model to the original variables w_j, we have

$$\hat{y} = 2.9913 - 0.8920w_1 + 0.3483w_2 + 3.3209w_3 - 0.0623w_4$$

This is the ridge regression model for the cement data.

Table S12-3 Ridge Regression Estimates for the Cement Data

θ	$\beta_1^*(\theta)$	$\beta_2^*(\theta)$	$\beta_3^*(\theta)$	$\beta_4^*(\theta)$
.000	−28.3318	65.9996	64.0479	−57.2491
.001	−31.0360	57.0244	61.9645	−44.0901
.002	−32.6441	50.9649	60.3899	−35.3088
.004	−34.1071	43.2358	58.0266	−24.3241
.008	−34.3195	35.1426	54.7018	−13.3348
.016	−31.9710	27.9534	50.0949	−4.5489
.032	−26.3451	22.0347	43.8309	1.2950
.064	−18.0566	17.2202	36.0743	4.7242
.128	−9.1786	13.4944	27.9363	6.5914
.256	−1.9896	10.9160	20.8028	7.5076
.512	2.4922	9.2014	15.3197	7.7224

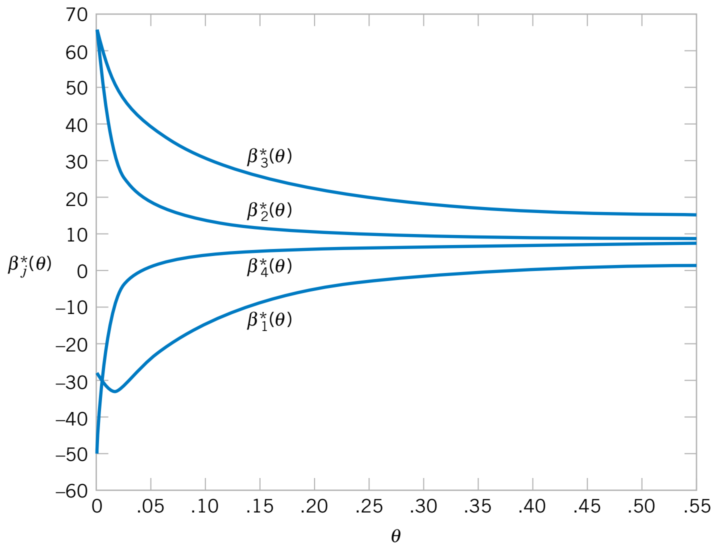

Figure S12-1 Ridge trace for the cement data.

12-6.6 Nonlinear Regression

Linear regression models provide a rich and flexible framework that work extremely well in many problems in engineering and science. However, linear regression models are not appropriate for all situations. There are many problems where the response variable and the predictor variables are related through a known **nonlinear** function. This leads to a **nonlinear regression model.** When the method of least squares is applied to such models, the resulting normal equations are nonlinear and, in general, difficult to solve. The usual approach is to directly minimize the residual sum of squares by an iterative procedure. We now give a very brief introduction to nonlinear regression models.

Linear or Nonlinear Models

We have focused in Chapter 12 on the **linear regression model**

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon \qquad \text{(S12-2)}$$

These models can include not only the first-order relationships, such as Equation S12-2, but polynomial models, and other more complex relationships as well. In fact, we could write the linear regression model as

$$Y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \cdots + \beta_r z_r + \epsilon \qquad \text{(S12-3)}$$

where z_i represents any **function** of the original regressors $x_1, x_2, \ldots, x_k$, including transformations such as $\exp(x_i)$, $\sqrt{x_i}$, and $\sin(x_i)$. These models are called **linear** regression models because they are **linear in the unknown parameters,** the $\beta_j, j = 1, 2, \ldots, k$.

We may write the linear regression model (Equation S12-2) in a general form as

$$\begin{aligned} Y &= \mathbf{x}'\boldsymbol{\beta} + \epsilon \\ &= f(\mathbf{x}, \boldsymbol{\beta}) + \epsilon \end{aligned} \qquad \text{(S12-4)}$$

where $\mathbf{x}' = [1, x_1, x_2, \ldots, x_k]$. Since the expected value of the model errors is zero, the expected value of the response variable is

$$\begin{aligned} E(Y) &= E[f(\mathbf{x}, \boldsymbol{\beta}) + \epsilon] \\ &= f(\mathbf{x}, \boldsymbol{\beta}) \end{aligned}$$

We usually refer to $f(\mathbf{x}, \boldsymbol{\beta})$ as the **expectation function** for the model. Obviously, the expectation function here is just a linear function of the unknown parameters.

Any model that is not linear in the unknown parameters is a **nonlinear regression model.** For example, the model

$$Y = \theta_1 e^{\theta_2 x} + \epsilon \tag{S12-5}$$

is not linear in the unknown parameters θ_1 and θ_2. We will use the symbol θ to represent a parameter in a nonlinear model to emphasize the difference between the linear and the nonlinear case. Nonlinear models often arise in cases where the relationship between the response and the regressors is a differential equation or the solution to a differential equation.

In general, we will write the nonlinear regression model as

$$Y = f(\mathbf{x}, \boldsymbol{\theta}) + \epsilon \tag{S12-6}$$

where $\boldsymbol{\theta}$ is a $p \times 1$ vector of unknown parameters, and ϵ is an uncorrelated random error term with $E(\epsilon) = 0$ and $\text{Var}(\epsilon) = \sigma^2$. We also typically assume that the errors are normally distributed, as in linear regression. Since

$$\begin{aligned} E(Y) &= E[f(\mathbf{x}, \boldsymbol{\theta}) + \epsilon] \\ &= f(\mathbf{x}, \boldsymbol{\theta}) \end{aligned} \tag{S12-7}$$

we call $f(\mathbf{x}, \boldsymbol{\theta})$ the **expectation function** for the nonlinear regression model. This is very similar to the linear regression case, except that now the expectation function is a **nonlinear** function of the parameters.

In a nonlinear regression model, at least one of the derivatives of the expectation function with respect to the parameters depends on at least one of the parameters. In linear regression, these derivatives are **not** functions of the unknown parameters. To illustrate these points, consider a linear regression model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$$

with expectation function $f(\mathbf{x}, \boldsymbol{\beta}) = \beta_0 + \sum_{j=1}^{k} \beta_j x_j$. Now

$$\frac{\partial f(\mathbf{x}, \boldsymbol{\beta})}{\partial \beta_j} = x_j, \qquad j = 0, 1, \ldots, k$$

where $x_0 \equiv 1$. Notice that in the linear case the derivatives are **not** functions of the β's.

Now consider the nonlinear model

$$\begin{aligned} y &= f(x, \boldsymbol{\theta}) + \epsilon \\ &= \theta_1 e^{\theta_2 x} + \epsilon \end{aligned}$$

The derivatives of the expectation function with respect to θ_1 and θ_2 are

$$\frac{\partial f(x, \boldsymbol{\theta})}{\partial \theta_1} = e^{\theta_2 x}$$

and

$$\frac{\partial f(x, \boldsymbol{\theta})}{\partial \theta_2} = \theta_1 x e^{\theta_2 x}$$

Since the derivatives are a function of the unknown parameters θ_1 and θ_2, the model is nonlinear.

Parameter Estimation in a Nonlinear Model

Suppose that we have a sample of n observations on the response and the regressors, say y_i, $x_{i1}, x_{i2}, \ldots, x_{ik}$, for $i = 1, 2, \ldots, n$. We have observed previously that the method of least squares in linear regression involves minimizing the least-squares function

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i - \left(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} \right) \right]^2$$

Because this is a linear regression model, when we differentiate $S(\boldsymbol{\beta})$ with respect to the unknown parameters and equate the derivatives to zero, the resulting normal equations are **linear** equations, and consequently, they are easy to solve.

Now consider the nonlinear regression situation. The model is

$$Y_i = f(\mathbf{x}_i, \boldsymbol{\theta}) + \epsilon_i \qquad i = 1, 2, \ldots, n$$

where now $\mathbf{x}_i' = [1, x_{i1}, x_{i2}, \ldots, x_{ik}]$ for $i = 1, 2, \ldots, n$. The least squares function is

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{n} [y_i - f(\mathbf{x}_i, \boldsymbol{\theta})]^2 \qquad \text{(S12-8)}$$

To find the least squares estimates we must differentiate Equation S12-8 with respect to each element of $\boldsymbol{\theta}$. This will provide a set of p normal equations for the nonlinear regression situation. The normal equations are

$$\sum_{i=1}^{n} [y_i - f(\mathbf{x}_i, \boldsymbol{\theta})] \left[\frac{\partial f(\mathbf{x}_i, \boldsymbol{\theta})}{\partial \theta_j} \right]_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}} = 0 \qquad \text{for } j = 1, 2, \ldots, p \qquad \text{(S12-9)}$$

In a nonlinear regression model the derivatives in the large square brackets will be functions of the unknown parameters. Furthermore, the expectation function is also a nonlinear function, so the normal equations can be very difficult to solve.

To illustrate this point, consider the nonlinear regression model in Equation S12-5:

$$Y = \theta_1 e^{\theta_2 x} + \epsilon$$

The least squares normal equations for this model are

$$\sum_{i=1}^{n} [y_i - \hat{\theta}_1 e^{\hat{\theta}_2 x_i}]^{\hat{\theta}_2 x_i} = 0$$

$$\sum_{i=1}^{n} [y_i - \hat{\theta}_1 e^{\hat{\theta}_2 x_i}] \hat{\theta}_1 x_i e^{\hat{\theta}_2 x_i} = 0$$

After simplification, the normal equations are

$$\sum_{i=1}^{n} y_i e^{\hat{\theta}_2 x_i} - \hat{\theta}_1 \sum_{i=1}^{n} e^{2\hat{\theta}_2 x_i} = 0$$
$$\sum_{i=1}^{n} y_i x_i e^{\hat{\theta}_2 x_i} - \hat{\theta}_1 \sum_{i=1}^{n} x_i e^{2\hat{\theta}_2 x_i} = 0$$

These equations are not linear in $\hat{\theta}_1$ and $\hat{\theta}_2$, and no simple closed-form solution exists. In general, **iterative methods** must be used to find the values of $\hat{\theta}_1$ and $\hat{\theta}_2$. To further complicate the problem, sometimes there are multiple solutions to the normal equations. That is, there are multiple stationary values for the residual sum of squares function $L(\boldsymbol{\theta})$.

A widely used method for nonlinear regression is **linearization** of the nonlinear function followed by the Gauss-Newton iteration method of parameter estimation. Linearization is accomplished by a **Taylor series expansion** of $f(\mathbf{x}_i, \boldsymbol{\theta})$ about the point $\boldsymbol{\theta}_0' = [\theta_{10}, \theta_{20}, \dots, \theta_{p0}]$ with only the linear terms retained. This yields

$$f(\mathbf{x}_i, \boldsymbol{\theta}) = f(\mathbf{x}_i, \boldsymbol{\theta}_0) + \sum_{j=1}^{p} \left[\frac{\partial f(\mathbf{x}_i, \boldsymbol{\theta})}{\partial \theta_j} \right]_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} (\theta_j - \theta_{j0}) \qquad \text{(S12-10)}$$

If we set

$$f_i^0 = f(\mathbf{x}_i, \boldsymbol{\theta}_0)$$
$$\beta_j^0 = \theta_j - \theta_{j0}$$
$$Z_{ij}^0 = \left[\frac{\partial f(\mathbf{x}_i, \boldsymbol{\theta})}{\partial \theta_j} \right]_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$$

we note that the nonlinear regression model can be written as

$$y_i - f_i^0 = \sum_{j=1}^{p} \beta_j^0 Z_{ij}^0 + \epsilon_i, \qquad i = 1, 2, \dots, n \qquad \text{(S12-11)}$$

That is, we now have a linear regression model. We usually call $\boldsymbol{\theta}_0$ the starting values for the parameters.

We may write Equation S12-10 as

$$\mathbf{y}_0 = \mathbf{Z}_0 \boldsymbol{\beta}_0 + \boldsymbol{\epsilon} \qquad \text{(S12-12)}$$

so the estimate of $\boldsymbol{\beta}_0$ is

$$\begin{aligned} \hat{\boldsymbol{\beta}}_0 &= (\mathbf{Z}_0'\mathbf{Z}_0)^{-1}\mathbf{Z}_0'\mathbf{y}_0 \\ &= (\mathbf{Z}_0'\mathbf{Z}_0)^{-1}\mathbf{Z}_0'(\mathbf{y} - \mathbf{f}_0) \end{aligned}$$

Now since $\boldsymbol{\beta}_0 = \boldsymbol{\theta} - \boldsymbol{\theta}_0$, we could define

$$\hat{\boldsymbol{\theta}}_1 = \hat{\boldsymbol{\beta}}_0 + \boldsymbol{\theta}_0$$

as revised estimates of $\boldsymbol{\theta}$. Sometimes $\hat{\boldsymbol{\beta}}_0$ is called the **vector of increments.** We may now replace the revised estimates $\hat{\boldsymbol{\theta}}_1$ in Equation S12-10 (in the same roles played by the initial estimates $\boldsymbol{\theta}_0$) and then produce another set of revised estimates, say $\hat{\boldsymbol{\theta}}_2$ and so forth.

In general, we have at the kth iteration

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{k+1} &= \hat{\boldsymbol{\theta}}_k + \hat{\boldsymbol{\beta}}_k \\ &= \hat{\boldsymbol{\theta}}_k + (\mathbf{Z}_k'\mathbf{Z}_k)^{-1}\mathbf{Z}_k'(\mathbf{y} - \mathbf{f}_k)\end{aligned}$$

where

$$\begin{aligned}\mathbf{Z}_k &= [Z_{ij}^k] \\ \mathbf{f}_k &= [f_1^k, f_2^k, \ldots, f_n^k]' \\ \hat{\boldsymbol{\theta}}_k &= [\theta_{1k}, \theta_{2k}, \ldots, \theta_{pk}]'\end{aligned}$$

This iterative process continues until convergence, that is, until

$$|(\hat{\theta}_{j,k+1} - \hat{\theta}_{jk})/\hat{\theta}_{jk}| < \delta, \quad j = 1, 2, \ldots, p$$

where δ is some small number, say, 1.0×10^{-6}. At each iteration the residual sum of squares $L(\hat{\boldsymbol{\theta}}_k)$ should be evaluated to ensure that a reduction in its value has been obtained. This scheme is available in many software packages. The SAS procedure is PROC NLIN. There are also several widely used variations of this procedure. For details and examples refer to Montgomery, Peck, and Vining (2001) and Myers (1990).