

## CHAPTER 10

# Fitting Regression Models

---

### CHAPTER OUTLINE

- |                                                                                  |                                                                      |
|----------------------------------------------------------------------------------|----------------------------------------------------------------------|
| 10.1 INTRODUCTION                                                                | 10.6 PREDICTION OF NEW RESPONSE OBSERVATIONS                         |
| 10.2 LINEAR REGRESSION MODELS                                                    | 10.7 REGRESSION MODEL DIAGNOSTICS                                    |
| 10.3 ESTIMATION OF THE PARAMETERS<br>IN LINEAR REGRESSION MODELS                 | 10.7.1 Scaled Residuals and PRESS                                    |
| 10.4 HYPOTHESIS TESTING IN MULTIPLE<br>REGRESSION                                | 10.7.2 Influence Diagnostics                                         |
| 10.4.1 Test for Significance of Regression                                       | 10.8 TESTING FOR LACK OF FIT                                         |
| 10.4.2 Tests on Individual Regression Coefficients<br>and Groups of Coefficients | SUPPLEMENTAL MATERIAL FOR CHAPTER 10                                 |
| 10.5 CONFIDENCE INTERVALS IN MULTIPLE<br>REGRESSION                              | S10.1 The Covariance Matrix of the Regression Coefficients           |
| 10.5.1 Confidence Intervals on the Individual<br>Regression Coefficients         | S10.2 Regression Models and Designed Experiments                     |
| 10.5.2 Confidence Interval on the Mean Response                                  | S10.3 Adjusted $R^2$                                                 |
|                                                                                  | S10.4 Stepwise and Other Variable Selection Methods in<br>Regression |
|                                                                                  | S10.5 The Variance of the Predicted Response                         |
|                                                                                  | S10.6 The Variance of Prediction Error                               |
|                                                                                  | S10.7 Leverage in a Regression Model                                 |

---

The supplemental material is on the textbook website [www.wiley.com/college/montgomery](http://www.wiley.com/college/montgomery).

---

### 10.1 Introduction

In many problems two or more variables are related, and it is of interest to model and explore this relationship. For example, in a chemical process the yield of product is related to the operating temperature. The chemical engineer may want to build a model relating yield to temperature and then use the model for prediction, process optimization, or process control.

In general, suppose that there is a single **dependent variable** or **response**  $y$  that depends on  $k$  **independent** or **regressor variables**, for example,  $x_1, x_2, \dots, x_k$ . The relationship between these variables is characterized by a mathematical model called a **regression model**. The regression model is fit to a set of sample data. In some instances, the experimenter knows the exact form of the true functional relationship between  $y$  and  $x_1, x_2, \dots, x_k$ , say  $y = \phi(x_1, x_2, \dots, x_k)$ . However, in most cases, the true functional relationship is unknown, and the experimenter chooses an appropriate function to approximate  $\phi$ . Low-order polynomial models are widely used as approximating functions.

There is a strong interplay between design of experiments and regression analysis. Throughout this book we have emphasized the importance of expressing the results of an experiment quantitatively, in terms of an **empirical model**, to facilitate understanding, interpretation, and implementation. Regression models are the basis for this. On numerous occasions we have shown the regression model that represented the results of an experiment. In this chapter, we present some aspects of fitting these models. More complete presentations of regression are available in Montgomery, Peck, and Vining (2006) and Myers (1990).

Regression methods are frequently used to analyze data from **unplanned experiments**, such as might arise from observation of uncontrolled phenomena or historical records. Regression methods are also very useful in designed experiments where something has “gone wrong.” We will illustrate some of these situations in this chapter.

## 10.2 Linear Regression Models

We will focus on fitting linear regression models. To illustrate, suppose that we wish to develop an empirical model relating the viscosity of a polymer to the temperature and the catalyst feed rate. A model that might describe this relationship is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon \quad (10.1)$$

where  $y$  represents the viscosity,  $x_1$  represents the temperature, and  $x_2$  represents the catalyst feed rate. This is a **multiple linear regression model** with two independent variables. We often call the independent variables **predictor variables** or **regressors**. The term **linear** is used because Equation 10.1 is a linear function of the unknown parameters  $\beta_0$ ,  $\beta_1$ , and  $\beta_2$ . The model describes a plane in the two-dimensional  $x_1, x_2$  space. The parameter  $\beta_0$  defines the intercept of the plane. We sometimes call  $\beta_1$  and  $\beta_2$  *partial regression coefficients* because  $\beta_1$  measures the expected change in  $y$  per unit change in  $x_1$  when  $x_2$  is held constant and  $\beta_2$  measures the expected change in  $y$  per unit change in  $x_2$  when  $x_1$  is held constant.

In general, the response variable  $y$  may be related to  $k$  regressor variables. The model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon \quad (10.2)$$

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

Models that are more complex in appearance than Equation 10.2 may often still be analyzed by multiple linear regression techniques. For example, consider adding an interaction term to the first-order model in two variables, say

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon \quad (10.3)$$

If we let  $x_3 = x_1 x_2$  and  $\beta_3 = \beta_{12}$ , then Equation 10.3 can be written as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon \quad (10.4)$$

which is a standard multiple linear regression model with three regressors. Recall that we presented empirical models like Equations 10.2 and 10.4 in several examples in Chapters 6, 7, and 8 to quantitatively express the results of a two-level factorial design. As another example, consider the second-order **response surface model** in two variables:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \epsilon \quad (10.5)$$

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  $\beta_5 = \beta_{12}$ , then this becomes

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_4 + \beta_5x_5 + \epsilon \quad (10.6)$$

which is a linear regression model. We have also seen this model in examples earlier in the text. In general, any regression model that is linear in the parameters (the  $\beta$ 's) is a linear regression model, regardless of the shape of the response surface that it generates.

In this chapter we will summarize methods for estimating the parameters in multiple linear regression models. This is often called **model fitting**. We have used some of these results in previous chapters, but here we give the developments. We will also discuss methods for testing hypotheses and constructing confidence intervals for these models as well as for checking the adequacy of the model fit. Our focus is primarily on those aspects of regression analysis useful in designed experiments. For more complete presentations of regression, refer to Montgomery, Peck, and Vining (2006) and Myers (1990).

## 10.3 Estimation of the Parameters in Linear Regression Models

The method of least squares is typically used to estimate the regression coefficients in a multiple linear regression model. Suppose that  $n > k$  observations on the response variable are available, say  $y_1, y_2, \dots, y_n$ . Along with each observed response  $y_i$ , we will have an observation on each regressor variable and let  $x_{ij}$  denote the  $i$ th observation or level of variable  $x_j$ . The data will appear as in Table 10.1. We assume that the error term  $\epsilon$  in the model has  $E(\epsilon) = 0$  and  $V(\epsilon) = \sigma^2$  and that the  $\{\epsilon_i\}$  are uncorrelated random variables.

We may write the model equation (Equation 10.2) in terms of the observations in Table 10.1 as

$$\begin{aligned} y_i &= \beta_0 + \beta_1x_{i1} + \beta_2x_{i2} + \cdots + \beta_kx_{ik} + \epsilon_i \\ &= \beta_0 + \sum_{j=1}^k \beta_jx_{ij} + \epsilon_i \quad i = 1, 2, \dots, n \end{aligned} \quad (10.7)$$

The method of least squares chooses the  $\beta$ 's in Equation 10.7 so that the sum of the squares of the errors,  $\epsilon_i$ , is minimized. The least squares function is

$$L = \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n \left( y_i - \beta_0 - \sum_{j=1}^k \beta_jx_{ij} \right)^2 \quad (10.8)$$

The function  $L$  is to be minimized with respect to  $\beta_0, \beta_1, \dots, \beta_k$ . The least squares estimators, say  $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ , must satisfy

$$\left. \frac{\partial L}{\partial \beta_0} \right|_{\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k} = -2 \sum_{i=1}^n \left( y_i - \hat{\beta}_0 - \sum_{j=1}^k \hat{\beta}_jx_{ij} \right) = 0 \quad (10.9a)$$

■ **TABLE 10.1**  
Data for Multiple Linear Regression

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

and

$$\left. \frac{\partial L}{\partial \beta_j} \right|_{\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k} = -2 \sum_{i=1}^n \left( y_i - \hat{\beta}_0 - \sum_{j=1}^k \hat{\beta}_j x_{ij} \right) x_{ij} = 0 \quad j = 1, 2, \dots, k \quad (10.9b)$$

Simplifying Equation 10.9, we obtain

$$\begin{aligned} n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^n x_{i1} + \hat{\beta}_2 \sum_{i=1}^n x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^n x_{ik} &= \sum_{i=1}^n y_i \\ \hat{\beta}_0 \sum_{i=1}^n x_{i1} + \hat{\beta}_1 \sum_{i=1}^n x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^n x_{i1}x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^n x_{i1}x_{ik} &= \sum_{i=1}^n x_{i1}y_i \\ \vdots & \\ \hat{\beta}_0 \sum_{i=1}^n x_{ik} + \hat{\beta}_1 \sum_{i=1}^n x_{ik}x_{i1} + \hat{\beta}_2 \sum_{i=1}^n x_{ik}x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^n x_{ik}^2 &= \sum_{i=1}^n x_{ik}y_i \end{aligned} \quad (10.10)$$

These equations are called the **least squares normal equations**. 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 regression coefficients  $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ .

It is simpler to solve the normal equations if they are expressed in matrix notation. We now give a matrix development of the normal equations that parallels the development of Equation 10.10. The model in terms of the observations, Equation 10.7, may be written in matrix notation as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where

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

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 an  $(n \times 1)$  vector of random errors.

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

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

Note that  $L$  may be expressed as

$$\begin{aligned} L &= \mathbf{y}'\mathbf{y} - \boldsymbol{\beta}'\mathbf{X}'\mathbf{y} - \mathbf{y}'\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{y}'\mathbf{y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \end{aligned} \quad (10.11)$$

because  $\boldsymbol{\beta}'\mathbf{X}'\mathbf{y}$  is a  $(1 \times 1)$  matrix, or a scalar, and its transpose  $(\boldsymbol{\beta}'\mathbf{X}'\mathbf{y})' = \mathbf{y}'\mathbf{X}\boldsymbol{\beta}$  is the same scalar. The least squares estimators must satisfy

$$\left. \frac{\partial L}{\partial \boldsymbol{\beta}} \right|_{\hat{\boldsymbol{\beta}}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{0}$$

which simplifies to

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y} \quad (10.12)$$

Equation 10.12 is the matrix form of the least squares normal equations. It is identical to Equation 10.10. To solve the normal equations, multiply both sides of Equation 10.12 by the inverse of  $\mathbf{X}'\mathbf{X}$ . Thus, the least squares estimator of  $\boldsymbol{\beta}$  is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \quad (10.13)$$

It is easy to see that the matrix form of the normal equations is identical to the scalar form. Writing out Equation 10.12 in detail, we obtain

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

If the indicated matrix multiplication is performed, the scalar form of the normal equations (i.e., Equation 10.10) 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

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} \quad (10.14)$$

In scalar notation, the fitted model is

$$\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^k \hat{\beta}_j x_{ij} \quad i = 1, 2, \dots, n$$

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

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} \quad (10.15)$$

**Estimating  $\sigma^2$ .** It is also usually necessary to estimate  $\sigma^2$ . To develop an estimator of this parameter, consider the sum of squares of the residuals, say

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

Substituting  $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ , we have

$$\begin{aligned} SS_E &= (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \mathbf{y}'\mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \mathbf{y}'\mathbf{y} - 2\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \end{aligned}$$

Because  $\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$ , this last equation becomes

$$SS_E = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} \quad (10.16)$$

Equation 10.16 is called the **error** or **residual sum of squares**, and it has  $n - p$  degrees of freedom associated with it. It can be shown that

$$E(SS_E) = \sigma^2(n - p)$$

so an unbiased estimator of  $\sigma^2$  is given by

$$\hat{\sigma}^2 = \frac{SS_E}{n - p} \quad (10.17)$$

**Properties of the Estimators.** The method of least squares produces an unbiased estimator of the parameter  $\beta$  in the linear regression model. This may be easily demonstrated by taking the expected value of  $\hat{\beta}$  as follows:

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

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

The variance property of  $\hat{\beta}$  is expressed in the **covariance matrix**:

$$\text{Cov}(\hat{\beta}) \equiv E[(\hat{\beta} - E(\hat{\beta}))(\hat{\beta} - E(\hat{\beta}))'] \quad (10.18)$$

which is just a symmetric matrix whose  $i$ th main diagonal element is the variance of the individual regression coefficient  $\hat{\beta}_i$  and whose  $(ij)$ th element is the covariance between  $\hat{\beta}_i$  and  $\hat{\beta}_j$ . The covariance matrix of  $\hat{\beta}$  is

$$\text{Cov}(\hat{\beta}) = \sigma^2(X'X)^{-1} \quad (10.19)$$

If  $\sigma^2$  in Equation 10.19 is replaced with the estimate  $\hat{\sigma}^2$  from Equation 10.12, we obtain an estimate of the covariance matrix of  $\hat{\beta}$ . The square roots of the main diagonal elements of this matrix are the **standard errors** of the model parameters.

## EXAMPLE 10.1

Sixteen observations on the viscosity of a polymer ( $y$ ) and two process variables—reaction temperature ( $x_1$ ) and catalyst feed rate ( $x_2$ )—are shown in Table 10.2. We will fit a multiple linear regression model

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

to these data. The  $X$  matrix and  $y$  vector are

$$X = \begin{bmatrix} 1 & 80 & 8 \\ 1 & 93 & 9 \\ 1 & 100 & 10 \\ 1 & 82 & 12 \\ 1 & 90 & 11 \\ 1 & 99 & 8 \\ 1 & 81 & 8 \\ 1 & 96 & 10 \\ 1 & 94 & 12 \\ 1 & 93 & 11 \\ 1 & 97 & 13 \\ 1 & 95 & 11 \\ 1 & 100 & 8 \\ 1 & 85 & 12 \\ 1 & 86 & 9 \\ 1 & 87 & 12 \end{bmatrix} \quad y = \begin{bmatrix} 2256 \\ 2340 \\ 2426 \\ 2293 \\ 2330 \\ 2368 \\ 2250 \\ 2409 \\ 2364 \\ 2379 \\ 2440 \\ 2364 \\ 2404 \\ 2317 \\ 2309 \\ 2328 \end{bmatrix}$$

The  $X'X$  matrix is

$$\begin{aligned} X'X &= \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 80 & 93 & \cdots & 87 \\ 8 & 9 & \cdots & 12 \end{bmatrix} \begin{bmatrix} 1 & 80 & 8 \\ 1 & 93 & 9 \\ \vdots & \vdots & \vdots \\ 1 & 87 & 12 \end{bmatrix} \\ &= \begin{bmatrix} 16 & 1458 & 164 \\ 1458 & 133,560 & 14,946 \\ 164 & 14,946 & 1,726 \end{bmatrix} \end{aligned}$$

and the  $X'y$  vector is

$$X'y = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 80 & 93 & \cdots & 87 \\ 8 & 9 & \cdots & 12 \end{bmatrix} \begin{bmatrix} 2256 \\ 2340 \\ \vdots \\ 2328 \end{bmatrix} = \begin{bmatrix} 37,577 \\ 3,429,550 \\ 385,562 \end{bmatrix}$$

The least squares estimate of  $\beta$  is

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

or

$$\hat{\beta} = \begin{bmatrix} 14.176004 & -0.129746 & -0.223453 \\ -0.129746 & 1.429184 \times 10^{-3} & -4.763947 \times 10^{-5} \\ -0.223453 & -4.763947 \times 10^{-5} & 2.222381 \times 10^{-2} \end{bmatrix} \begin{bmatrix} 37,577 \\ 3,429,550 \\ 385,562 \end{bmatrix} = \begin{bmatrix} 1566.07777 \\ 7.62129 \\ 8.58485 \end{bmatrix}$$

■ **TABLE 10.2**  
Viscosity Data for Example 10.1 (viscosity in centistokes @ 100°C)

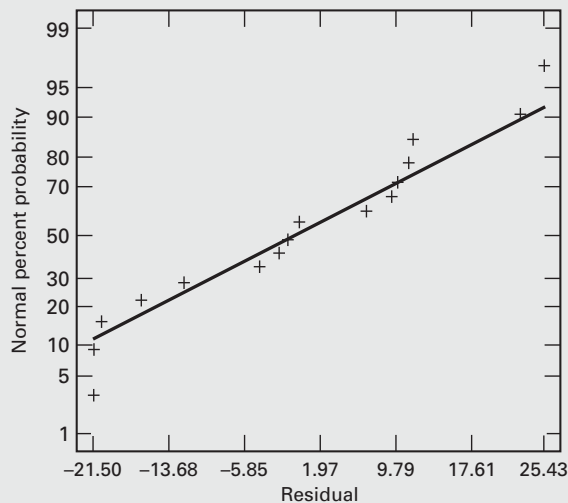
Observation	Temperature ( $x_1$ , °C)	Catalyst Feed Rate ( $x_2$ , lb/h)	Viscosity
1	80	8	2256
2	93	9	2340
3	100	10	2426
4	82	12	2293
5	90	11	2330
6	99	8	2368
7	81	8	2250
8	96	10	2409
9	94	12	2364
10	93	11	2379
11	97	13	2440
12	95	11	2364
13	100	8	2404
14	85	12	2317
15	86	9	2309
16	87	12	2328

The least squares fit, with the regression coefficients reported to two decimal places, is

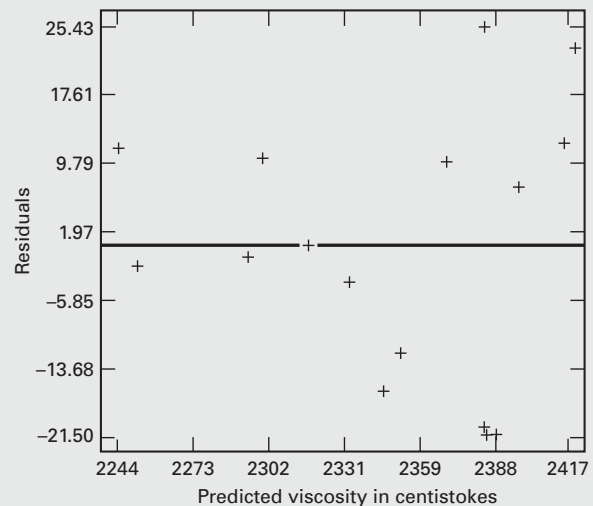
$$\hat{y} = 1566.08 + 7.62x_1 + 8.58x_2$$

The first three columns of Table 10.3 present the actual observations  $y_i$ , the predicted or fitted values  $\hat{y}_i$ , and the residuals. Figure 10.1 is a normal probability plot of the residuals. Plots of the residuals versus the predicted

values  $\hat{y}_i$  and versus the two variables  $x_1$  and  $x_2$  are shown in Figures 10.2, 10.3, and 10.4, respectively. Just as in designed experiments, residual plotting is an integral part of regression model building. These plots indicate that the variance of the observed viscosity tends to increase with the magnitude of viscosity. Figure 10.3 suggests that the variability in viscosity is increasing as temperature increases.



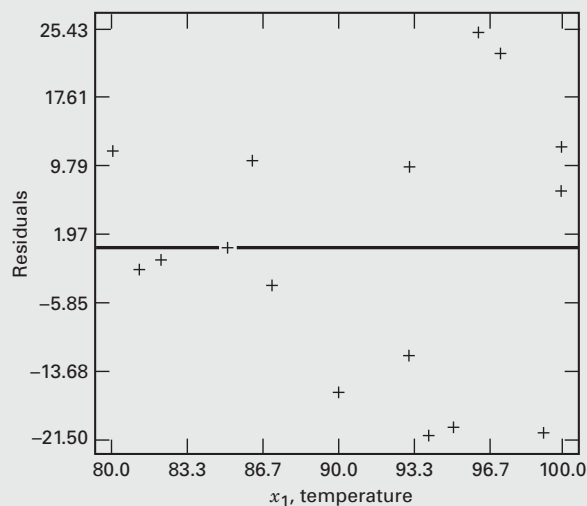
■ **FIGURE 10.1** Normal probability plot of residuals, Example 10.1



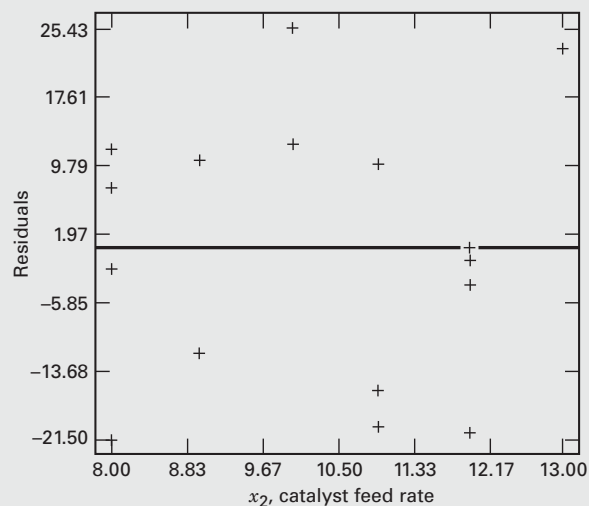
■ **FIGURE 10.2** Plot of residuals versus predicted viscosity, Example 10.1

■ **TABLE 10.3**  
Predicted Values, Residuals, and Other Diagnostics from Example 10.1

Observation $i$	$y_i$	Predicted Value $\hat{y}_i$	Residual $e_i$	$h_{ii}$	Studentized Residual	$D_i$	$R$ -Student
1	2256	2244.5	11.5	0.350	0.87	0.137	0.87
2	2340	2352.1	-12.1	0.102	-0.78	0.023	-0.77
3	2426	2414.1	11.9	0.177	0.80	0.046	0.79
4	2293	2294.0	-1.0	0.251	-0.07	0.001	-0.07
5	2330	2346.4	-16.4	0.077	-1.05	0.030	-1.05
6	2368	2389.3	-21.3	0.265	-1.52	0.277	-1.61
7	2250	2252.1	-2.1	0.319	-0.15	0.004	-0.15
8	2409	2383.6	25.4	0.098	1.64	0.097	1.76
9	2364	2385.5	-21.5	0.142	-1.42	0.111	-1.48
10	2379	2369.3	9.7	0.080	0.62	0.011	0.60
11	2440	2416.9	23.1	0.278	1.66	0.354	1.80
12	2364	2384.5	-20.5	0.096	-1.32	0.062	-1.36
13	2404	2396.9	7.1	0.289	0.52	0.036	0.50
14	2317	2316.9	0.1	0.185	0.01	0.000	<0.01
15	2309	2298.8	10.2	0.134	0.67	0.023	0.66
16	2328	2332.1	-4.1	0.156	-0.28	0.005	-0.27



■ **FIGURE 10.3** Plot of residuals versus  $x_1$  (temperature), Example 10.1



■ **FIGURE 10.4** Plot of residuals versus  $x_2$  (feed rate), Example 10.1



**Using the Computer.** Regression model fitting is almost always done using a statistical software package, such as Minitab or JMP. Table 10.4 shows some of the output obtained when Minitab is used to fit the viscosity regression model in Example 10.1. Many of the quantities in this output should be familiar because they have similar meanings to the quantities in the output displays for computer analysis of data from designed experiments. We have seen many such computer outputs previously in the book. In subsequent sections, we will discuss the analysis of variance and  $t$ -test information in Table 10.4 in detail and will show exactly how these quantities were computed.

**Fitting Regression Models in Designed Experiments.** We have often used a regression model to present the results of a designed experiment in a quantitative form. We now give a complete illustrative example. This is followed by three other brief examples that illustrate other useful applications of regression analysis in designed experiments.

## EXAMPLE 10.2 Regression Analysis of a $2^3$ Factorial Design

A chemical engineer is investigating the yield of a process. Three process variables are of interest: temperature, pressure, and catalyst concentration. Each variable can be run at a low and a high level, and the engineer decides to run a  $2^3$  design with four center points. The design and the resulting yields are shown in Figure 10.5, where we have shown both the natural levels of the design factor and the  $+1, -1$  coded variable notation normally employed in  $2^k$  factorial designs to represent the factor levels.

Suppose that the engineer decides to fit a main effects only model, say

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

For this model the  $\mathbf{X}$  matrix and  $\mathbf{y}$  vector are

$$\mathbf{X} = \begin{bmatrix} 1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \text{ and } \mathbf{y} = \begin{bmatrix} 32 \\ 46 \\ 57 \\ 65 \\ 36 \\ 48 \\ 57 \\ 68 \\ 50 \\ 44 \\ 53 \\ 56 \end{bmatrix}$$

The  $2^3$  is an orthogonal design, and even with the added center runs it is still orthogonal. Therefore

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 12 & 0 & 0 & 0 \\ 0 & 8 & 0 & 0 \\ 0 & 0 & 8 & 0 \\ 0 & 0 & 0 & 8 \end{bmatrix} \text{ and } \mathbf{X}'\mathbf{y} = \begin{bmatrix} 612 \\ 45 \\ 85 \\ 9 \end{bmatrix}$$

Because the design is orthogonal, the  $\mathbf{X}'\mathbf{X}$  matrix is *diagonal*, the required inverse is also diagonal, and the vector of least squares estimates of the regression coefficients is

$$\begin{aligned} \hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} &= \begin{bmatrix} 1/12 & 0 & 0 & 0 \\ 0 & 1/8 & 0 & 0 \\ 0 & 0 & 1/8 & 0 \\ 0 & 0 & 0 & 1/8 \end{bmatrix} \begin{bmatrix} 612 \\ 45 \\ 85 \\ 9 \end{bmatrix} \\ &= \begin{bmatrix} 51.000 \\ 5.625 \\ 10.625 \\ 1.125 \end{bmatrix} \end{aligned}$$

The fitted regression model is

$$\hat{y} = 51.000 + 5.625x_1 + 10.625x_2 + 1.125x_3$$

As we have made use of on many occasions, the regression coefficients are closely connected to the effect estimates that would be obtained from the usual analysis of a  $2^3$  design. For example, the effect of temperature is (refer to Figure 10.5)

$$\begin{aligned} T &= \bar{y}_{T+} - \bar{y}_{T-} \\ &= 56.75 - 45.50 = 11.25 \end{aligned}$$

Notice that the regression coefficient for  $x_1$  is

$$(11.25)/2 = 5.625$$

That is, the regression coefficient is exactly one-half the usual effect estimate. This will always be true for a  $2^k$  design. As noted above, we used this result in Chapters 6 through 8 to produce regression models, fitted values, and residuals for several two-level experiments. This example demonstrates

that the effect estimates from a  $2^k$  design are least squares estimates.

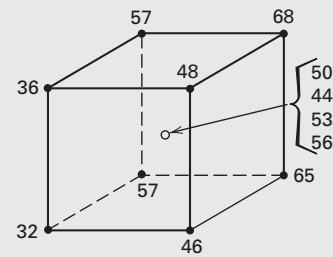
The variance of the regression model parameter are found from the diagonal elements of  $(\mathbf{X}'\mathbf{X})^{-1}$ . That is,

$$V(\hat{\beta}_0) = \frac{\sigma^2}{12}, \quad \text{and} \quad V(\hat{\beta}_i) = \frac{\sigma^2}{8}, i = 1, 2, 3.$$

The **relative variance** are

$$\frac{V(\hat{\beta}_0)}{\sigma^2} = \frac{1}{12} \quad \text{and} \quad \frac{V(\hat{\beta}_i)}{\sigma^2} = \frac{1}{8}, i = 1, 2, 3.$$

Process Variables				Coded Variables			Yield, $y$
Run	Temp (°C)	Pressure (psig)	Conc (g/l)	$x_1$	$x_2$	$x_3$	
1	120	40	15	-1	-1	-1	32
2	160	80	15	1	-1	-1	46
3	120	40	15	-1	1	-1	57
4	160	80	15	1	1	-1	65
5	120	40	30	-1	-1	1	36
6	160	80	30	1	-1	1	48
7	120	40	30	-1	1	1	57
8	160	80	30	1	1	1	68
9	140	60	22.5	0	0	0	50
10	140	60	22.5	0	0	0	44
11	140	60	22.5	0	0	0	53
12	140	60	22.5	0	0	0	56

$$x_1 = \frac{\text{Temp} - 140}{20}, \quad x_2 = \frac{\text{Pressure} - 60}{20}, \quad x_3 = \frac{\text{Conc} - 22.5}{7.5}$$


■ **FIGURE 10.5** Experimental design for Example 10.2

■ **TABLE 10.4**  
Minitab Output for the Viscosity Regression Model, Example 10.1

#### Regression Analysis

The regression equation is

Viscosity = 1566 + 7.62 Temp + 8.58 Feed Rate

Predictor	Coef	Std. Dev.	T	P
Constant	1566.08	61.59	25.43	0.000
Temp	7.6213	0.6184	12.32	0.000
Feed Rat	8.585	2.439	3.52	0.004

S = 16.36      R-Sq = 92.7%      R-Sq (adj) = 91.6%

#### Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	44157	22079	82.50	0.000
Residual Error	13	3479	268		
Total	15	47636			

Source	DF	Seq SS
Temp	1	40841
Feed Rat	1	3316

In Example 10.2, the inverse matrix is easy to obtain because  $\mathbf{X}'\mathbf{X}$  is diagonal. Intuitively, this seems to be advantageous, not only because of the computational simplicity but also because the estimators of all the regression coefficients are uncorrelated; that is,  $\text{Cov}(\hat{\beta}_i, \hat{\beta}_j) = 0$ . If we can choose the levels of the  $x$  variables before the data are collected, we might wish to design the experiment so that a diagonal  $\mathbf{X}'\mathbf{X}$  will result.

In practice, it can be relatively easy to do this. We know that the off-diagonal elements in  $\mathbf{X}'\mathbf{X}$  are the sums of cross products of the columns in  $\mathbf{X}$ . Therefore, we must make the inner product of the columns of  $\mathbf{X}$  equal to zero; that is, these columns must be **orthogonal**. As we have noted before, experimental designs that have this property for fitting a regression model are called **orthogonal designs**. In general, the  $2^k$  factorial design is an orthogonal design for fitting the multiple linear regression model.

Regression methods are extremely useful when something “goes wrong” in a designed experiment. This is illustrated in the next two examples.

### EXAMPLE 10.3 A $2^3$ Factorial Design with a Missing Observation

Consider the  $2^3$  factorial design with four center points from Example 10.2. Suppose that when this experiment was performed, the run with all variables at the high level (run 8 in Figure 10.5) was missing. This can happen for a variety of reasons; the measurement system can produce a faulty reading, the combination of factor levels may prove infeasible, the experimental unit may be damaged, and so forth.

We will fit the main effects model

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

using the 11 remaining observations. The  $\mathbf{X}$  matrix and  $\mathbf{y}$  vector are

$$\mathbf{X} = \begin{bmatrix} 1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{y} = \begin{bmatrix} 32 \\ 46 \\ 57 \\ 65 \\ 36 \\ 48 \\ 57 \\ 50 \\ 44 \\ 53 \\ 56 \end{bmatrix}$$

To estimate the model parameters, we form

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 11 & -1 & -1 & -1 \\ -1 & 7 & -1 & -1 \\ -1 & -1 & 7 & -1 \\ -1 & -1 & -1 & 7 \end{bmatrix} \quad \text{and} \quad \mathbf{X}'\mathbf{y} = \begin{bmatrix} 544 \\ -23 \\ 17 \\ -59 \end{bmatrix}$$

Because there is a missing observation, the design is no longer orthogonal. Now

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= \begin{bmatrix} 9.61538 \times 10^{-2} & 1.92307 \times 10^{-2} \\ 1.92307 \times 10^{-2} & 0.15385 \\ 1.92307 \times 10^{-2} & 2.88462 \times 10^{-2} \\ 1.92307 \times 10^{-2} & 2.88462 \times 10^{-2} \end{bmatrix} \begin{bmatrix} 544 \\ -23 \\ 17 \\ -59 \end{bmatrix} \\ &= \begin{bmatrix} 51.25 \\ 5.75 \\ 10.75 \\ 1.25 \end{bmatrix} \end{aligned}$$

Therefore, the fitted model is

$$\hat{y} = 51.25 + 5.75x_1 + 10.75x_2 + 1.25x_3$$

Compare this model to the one obtained in Example 10.2, where all 12 observations were used. The regression coefficients are very similar. Because the regression coefficients are closely related to the factor effects, our conclusions would not be seriously affected by the missing observation. However, notice that the effect estimates are no longer orthogonal because  $\mathbf{X}'\mathbf{X}$  and its inverse are no longer diagonal. Furthermore the variances of the regression coefficients are larger than they were in the original orthogonal design with no missing data.

**EXAMPLE 10.4** Inaccurate Levels in Design Factors

When running a designed experiment, it is sometimes difficult to reach and hold the precise factor levels required by the design. Small discrepancies are not important, but large ones are potentially of more concern. Regression methods are useful in the analysis of a designed experiment where the experimenter has been unable to obtain the required factor levels.

To illustrate, the experiment presented in Table 10.5 shows a variation of the  $2^3$  design from Example 10.2, where many of the test combinations are not exactly the ones specified in the design. Most of the difficulty seems to have occurred with the temperature variable.

We will fit the main effects model

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \epsilon$$

The  $\mathbf{X}$  matrix and  $\mathbf{y}$  vector are

$$\mathbf{X} = \begin{bmatrix} 1 & -0.75 & -0.95 & -1.133 \\ 1 & 0.90 & -1 & -1 \\ 1 & -0.95 & 1.1 & -1 \\ 1 & 1 & 0 & -1 \\ 1 & -1.10 & -1.05 & 1.4 \\ 1 & 1.15 & -1 & 1 \\ 1 & -0.90 & 1 & 1 \\ 1 & 1.25 & 1.15 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} 32 \\ 46 \\ 57 \\ 65 \\ 36 \\ 48 \\ 57 \\ 68 \\ 50 \\ 44 \\ 53 \\ 56 \end{bmatrix}$$

To estimate the model parameters, we need

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 12 & 0.60 & 0.25 & 0.2670 \\ 0.60 & 8.18 & 0.31 & -0.1403 \\ 0.25 & 0.31 & 8.5375 & -0.3437 \\ 0.2670 & -0.1403 & -0.3437 & 9.2437 \end{bmatrix}$$

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} 612 \\ 77.55 \\ 100.7 \\ 19.144 \end{bmatrix}$$

■ **TABLE 10.5**  
Experimental Design for Example 10.4

Run	Process Variables			Coded Variables			Yield y
	Temp (°C)	Pressure (psig)	Conc (g/l)				
				x <sub>1</sub>	x <sub>2</sub>	x <sub>3</sub>	
1	125	41	14	−0.75	−0.95	−1.133	32
2	158	40	15	0.90	−1	−1	46
3	121	82	15	−0.95	1.1	−1	57
4	160	80	15	1	1	−1	65
5	118	39	33	−1.10	−1.05	1.14	36
6	163	40	30	1.15	−1	1	48
7	122	80	30	−0.90	1	1	57
8	165	83	30	1.25	1.15	1	68
9	140	60	22.5	0	0	0	50
10	140	60	22.5	0	0	0	44
11	140	60	22.5	0	0	0	53
12	140	60	22.5	0	0	0	56

Then

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

$$= \begin{bmatrix} 8.37447 \times 10^{-2} & -6.09871 \times 10^{-3} & -2.33542 \times 10^{-3} & -2.59833 \times 10^{-3} \\ -6.09871 \times 10^{-3} & 0.12289 & -4.20766 \times 10^{-3} & 1.88490 \times 10^{-3} \\ -2.33542 \times 10^{-3} & -4.20766 \times 10^{-3} & 0.11753 & 4.37851 \times 10^{-3} \\ -2.59833 \times 10^{-3} & 1.88490 \times 10^{-3} & 4.37851 \times 10^{-3} & 0.10845 \end{bmatrix} \begin{bmatrix} 612 \\ 77.55 \\ 100.7 \\ 19.144 \end{bmatrix} = \begin{bmatrix} 50.49391 \\ 5.40996 \\ 10.16316 \\ 1.07245 \end{bmatrix}$$

The fitted regression model, with the coefficients reported to two decimal places, is

$$\hat{y} = 50.49 + 5.41x_1 + 10.16x_2 + 1.07x_3$$

Comparing this to the original model in Example 10.2, where the factor levels were exactly those specified by the design,

we note very little difference. The practical interpretation of the results of this experiment would not be seriously affected by the inability of the experimenter to achieve the desired factor levels exactly.

### EXAMPLE 10.5 De-aliasing Interactions in a Fractional Factorial

We observed in Chapter 8 that it is possible to de-alias interactions in a fractional factorial design by a process called fold over. For a resolution III design, a full fold over is constructed by running a second fraction in which the signs are reversed from those in the original fraction. Then the combined design can be used to de-alias all main effects from the two-factor interactions.

A difficulty with a full fold over is that it requires a second group of runs of identical size as the original design. It is usually possible to de-alias certain interactions of interest by augmenting the original design with fewer runs than required in a full fold over. The partial fold-over technique was used to solve this problem. Regression methods are an easy way to see how the partial fold-over technique works and, in some cases, find even more efficient fold-over designs.

To illustrate, suppose that we have run a  $2_{IV}^{4-1}$  design. Table 8.3 shows the principal fraction of this design, in which  $I = ABCD$ . Suppose that after the data from the first eight trials were observed, the largest effects were  $A$ ,  $B$ ,  $C$ ,  $D$  (we ignore the three-factor interactions that are aliased with these main effects) and the  $AB + CD$  alias chain. The other two alias chains can be ignored, but clearly either  $AB$ ,  $CD$ , or both two-factor interactions are large. To find out which interactions are important, we could, of course, run the alternate fraction, which would require another eight trials. Then all 16 runs could be used to estimate the main effects and the two-factor interactions. An alternative

would be to use a partial fold over involving four additional runs.

It is possible to de-alias  $AB$  and  $CD$  in fewer than four additional trials. Suppose that we wish to fit the model

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_4 + \beta_{12}x_1x_2 + \beta_{34}x_3x_4 + \epsilon$$

where  $x_1$ ,  $x_2$ ,  $x_3$ , and  $x_4$  are the coded variables representing  $A$ ,  $B$ ,  $C$ , and  $D$ . Using the design in Table 8.3, the  $\mathbf{X}$  matrix for this model is

$$\mathbf{X} = \begin{bmatrix} & x_1 & x_2 & x_3 & x_4 & x_1x_2 & x_3x_4 \\ 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & 1 & -1 & -1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 & 1 & -1 & -1 \\ 1 & 1 & 1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & 1 & -1 & -1 & -1 \\ 1 & -1 & 1 & 1 & -1 & -1 & -1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

where we have written the variables above the columns to facilitate understanding. Notice that the  $x_1x_2$  column is identical to the  $x_3x_4$  column (as anticipated, because  $AB$  or  $x_1x_2$  is aliased with  $CD$  or  $x_3x_4$ ), implying a linear dependency in the columns of  $\mathbf{X}$ . Therefore, we cannot estimate both  $\beta_{12}$  and  $\beta_{34}$  in the model. However, suppose that we add a single run  $x_1 = -1$ ,  $x_2 = -1$ ,  $x_3 = -1$ , and  $x_4 = 1$

from the alternate fraction to the original eight runs. The  $\mathbf{X}$  matrix for the model now becomes

$$\mathbf{X} = \begin{bmatrix} & x_1 & x_2 & x_3 & x_4 & x_1x_2 & x_3x_4 \\ 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & 1 & -1 & -1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 & 1 & -1 & -1 \\ 1 & 1 & 1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & 1 & -1 & -1 & -1 \\ 1 & -1 & 1 & 1 & -1 & -1 & -1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & -1 & -1 & 1 & 1 & -1 \end{bmatrix}$$

Notice that the columns  $x_1x_2$  and  $x_3x_4$  are now no longer identical, and we can fit the model including both the  $x_1x_2$  ( $AB$ ) and  $x_3x_4$  ( $CD$ ) interactions. The magnitudes of the regression coefficients will give insight regarding which interactions are important.

Although adding a single run will de-alias the  $AB$  and  $CD$  interactions, this approach does have a disadvantage. Suppose that there is a time effect (or a block effect) between the first eight runs and the last run added above. Add a column to the  $\mathbf{X}$  matrix for blocks, and you obtain the following:

$$\mathbf{X} = \begin{bmatrix} & x_1 & x_2 & x_3 & x_4 & x_1x_2 & x_3x_4 & \text{block} \\ 1 & -1 & -1 & -1 & -1 & 1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & 1 & -1 & -1 & -1 \\ 1 & 1 & 1 & -1 & -1 & 1 & 1 & -1 \\ 1 & -1 & -1 & 1 & 1 & 1 & 1 & -1 \\ 1 & 1 & -1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & -1 \\ 1 & -1 & -1 & -1 & 1 & 1 & -1 & 1 \end{bmatrix}$$

We have assumed the block factor was at the low or “ $-$ ” level during the first eight runs, and at the high or “ $+$ ” level during the ninth run. It is easy to see that the sum of the cross products of every column with the block column does not sum to zero, meaning that blocks are no longer orthogonal to treatments, or that the block effect now affects the estimates of the model regression coefficients. To block orthogonally, you must add an even number of runs. For example, the four runs

$x_1$	$x_2$	$x_3$	$x_4$
$-1$	$-1$	$-1$	$1$
$1$	$-1$	$-1$	$-1$
$-1$	$1$	$1$	$1$
$1$	$1$	$1$	$-1$

will de-alias  $AB$  from  $CD$  and allow orthogonal blocking (you can see this by writing out the  $\mathbf{X}$  matrix as we did previously). This is equivalent to a partial fold over, in terms of the number of runs that are required.

In general, it is usually straightforward to examine the  $\mathbf{X}$  matrix for the reduced model obtained from a fractional factorial and determine which runs to augment the original design with to de-alias interactions of potential interest. Furthermore, the impact of specific augmentation strategies can be evaluated using the general results for regression models given later in this chapter. There are also computer-based optimal design methods for constructing designs that can be useful for **design augmentation** to de-alias effects (refer to the supplemental material for Chapter 8).

## 10.4 Hypothesis Testing in Multiple Regression

In multiple linear regression problems, certain tests of hypotheses about the model parameters are helpful in measuring the usefulness of the model. In this section, we describe several important hypothesis-testing procedures. These procedures require that the errors  $\epsilon_i$  in the model be normally and independently distributed with mean zero and variance  $\sigma^2$ , abbreviated  $\epsilon \sim \text{NID}(0, \sigma^2)$ . As a result of this assumption, the observations  $y_i$  are normally and independently distributed with mean  $\beta_0 + \sum_{j=1}^k \beta_j x_{ij}$  and variance  $\sigma^2$ .

### 10.4.1 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  $x_1, x_2, \dots, x_k$ . The appropriate hypotheses are

$$H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0 \quad (10.20)$$

$$H_1: \beta_j \neq 0 \quad \text{for at least one } j$$

Rejection of  $H_0$  in Equation 10.20 implies that at least one of the regressor variables  $x_1, x_2, \dots, x_k$  contributes significantly to the model. The test procedure involves an analysis of variance partitioning of the total sum of squares  $SS_T$  into a sum of squares due to the model (or to regression) and a sum of squares due to residual (or error), say

$$SS_T = SS_R + SS_E \quad (10.21)$$

Now if the null hypothesis  $H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$  is true, then  $SS_R/\sigma^2$  is distributed as  $\chi_k^2$ , where the number of degrees of freedom for  $\chi^2$  is equal to the number of regressor variables in the model. Also, we can show that  $SS_E/\sigma^2$  is distributed as  $\chi_{n-k-1}^2$  and that  $SS_E$  and  $SS_R$  are independent. The test procedure for  $H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$  is to compute

$$F_0 = \frac{SS_R/k}{SS_E/(n-k-1)} = \frac{MS_R}{MS_E} \quad (10.22)$$

and to reject  $H_0$  if  $F_0$  exceeds  $F_{\alpha, k, n-k-1}$ . Alternatively, we could use the  $P$ -value approach to hypothesis testing and, reject  $H_0$  if the  $P$ -value for the statistic  $F_0$  is less than  $\alpha$ . The test is usually summarized in an analysis of variance table such as Table 10.6.

A computational formula for  $SS_R$  may be found easily. We have derived a computational formula for  $SS_E$  in Equation 10.16—that is,

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

Now, because  $SS_T = \sum_{i=1}^n y_i^2 - (\sum_{i=1}^n y_i)^2/n = \mathbf{y}'\mathbf{y} - (\sum_{i=1}^n y_i)^2/n$ , we may rewrite the foregoing equation 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

$$SS_R = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \frac{\left(\sum_{i=1}^n y_i\right)^2}{n} \quad (10.23)$$

and the error sum of squares is

$$SS_E = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} \quad (10.24)$$

and the total sum of squares is

$$SS_T = \mathbf{y}'\mathbf{y} - \frac{\left(\sum_{i=1}^n y_i\right)^2}{n} \quad (10.25)$$

■ TABLE 10.6

Analysis of Variance for 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	$SS_E$	$n - k - 1$	$MS_E$	
Total	$SS_T$	$n - 1$		

These computations are almost always performed with regression software. For instance, Table 10.4 shows some of the output from Minitab for the viscosity regression model in Example 10.1. The upper portion in this display is the analysis of variance for the model. The test of significance of regression in this example involves the hypotheses

$$\begin{aligned} H_0: \beta_1 = \beta_2 = 0 \\ H_1: \beta_j \neq 0 \quad \text{for at least one } j \end{aligned}$$

The  $P$ -value in Table 10.4 for the  $F$  statistic (Equation 10.22) is very small, so we would conclude that at least one of the two variables—temperature ( $x_1$ ) and feed rate ( $x_2$ )—has a nonzero regression coefficient.

Table 10.4 also reports the coefficient of multiple determination  $R^2$ , where

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T} \quad (10.26)$$

Just as in designed experiments,  $R^2$  is a measure of the amount of reduction in the variability of  $y$  obtained by using the regressor variables  $x_1, x_2, \dots, x_k$  in the model. However, as we have noted previously, a large value of  $R^2$  does not necessarily imply that the regression model is a good one. Adding a variable to the model will always increase  $R^2$ , regardless of whether the additional variable is statistically significant or not. Thus, it is possible for models that have large values of  $R^2$  to yield poor predictions of new observations or estimates of the mean response.

Because  $R^2$  always increases as we add terms to the model, some regression model builders prefer to use an **adjusted  $R^2$  statistic** defined as

$$R_{\text{adj}}^2 = 1 - \frac{SS_E/(n-p)}{SS_T/(n-1)} = 1 - \left( \frac{n-1}{n-p} \right) (1 - R^2) \quad (10.27)$$

In general, the adjusted  $R^2$  statistic will not always increase as variables are added to the model. In fact, if unnecessary terms are added, the value of  $R_{\text{adj}}^2$  will often decrease.

For example, consider the viscosity regression model. The adjusted  $R^2$  for the model is shown in Table 10.4. It is computed as

$$\begin{aligned} R_{\text{adj}}^2 &= 1 - \left( \frac{n-1}{n-p} \right) (1 - R^2) \\ &= 1 - \left( \frac{15}{13} \right) (1 - 0.92697) = 0.915735 \end{aligned}$$

which is very close to the ordinary  $R^2$ . When  $R^2$  and  $R_{\text{adj}}^2$  differ dramatically, there is a good chance that nonsignificant terms have been included in the model.

### 10.4.2 Tests on Individual Regression Coefficients and Groups of Coefficients

We are frequently interested in testing hypotheses on the individual regression coefficients. Such tests would be useful in determining the value of each regressor variable 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 variables already in the model.

Adding a variable to the regression model always causes the sum of squares for regression to increase and the error sum of squares to decrease. We must decide whether the increase in the regression sum of squares is sufficient to warrant using the additional variable in the model. Furthermore, adding an unimportant variable to the model can actually increase the mean square error, thereby decreasing the usefulness of the model.



The hypotheses for testing the significance of any individual regression coefficient, say  $\beta_j$ , are

$$H_0: \beta_j = 0$$

$$H_1: \beta_j \neq 0$$

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

$$t_0 = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} \quad (10.28)$$

where  $C_{jj}$  is the diagonal element of  $(\mathbf{X}'\mathbf{X})^{-1}$  corresponding to  $\hat{\beta}_j$ . The null hypothesis  $H_0: \beta_j = 0$  is rejected if  $|t_0| > t_{\alpha/2, n-k-1}$ . Note that this is really 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.

The denominator of Equation 10.28,  $\sqrt{\hat{\sigma}^2 C_{jj}}$ , is often called the **standard error** of the regression coefficient  $\hat{\beta}_j$ . That is,

$$se(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 C_{jj}} \quad (10.29)$$

Therefore, an equivalent way to write the test statistic in Equation (10.28) is

$$t_0 = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \quad (10.30)$$

Most regression computer programs provide the  $t$ -test for each model parameter. For example, consider Table 10.4, which contains the Minitab output for Example 10.1. The upper portion of this table gives the least squares estimate of each parameter, the standard error, the  $t$  statistic, and the corresponding  $P$ -value. We would conclude that both variables, temperature and feed rate, contribute significantly to the model.

We may also directly examine the contribution to the regression sum of squares for a particular variable, say  $x_j$ , given that other variables  $x_i$  ( $i \neq j$ ) are included in the model. The procedure for doing this is the general regression significance test or, as it is often called, 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:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

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  $x_1, x_2, \dots, x_r$  ( $r < k$ ) contribute significantly to the regression model. Let the vector of regression coefficients be partitioned as follows:

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix}$$

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

$$H_0: \boldsymbol{\beta}_1 = \mathbf{0}$$

$$H_1: \boldsymbol{\beta}_1 \neq \mathbf{0} \quad (10.31)$$

The model may be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon} \quad (10.32)$$

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  $\beta_1$  and  $\beta_2$ ), we know that  $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ . Also, the regression sum of squares for all variables including the intercept is

$$SS_R(\beta) = \hat{\beta}'\mathbf{X}'\mathbf{y} \quad (p \text{ degrees of freedom})$$

and

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

$SS_R(\beta)$  is called the regression sum of squares due to  $\beta$ . To find the contribution of the terms in  $\beta_1$  to the regression, we fit the model assuming the null hypothesis  $H_0: \beta_1 = \mathbf{0}$  to be true. The **reduced model** is found from Equation 10.32 with  $\beta_1 = \mathbf{0}$ :

$$\mathbf{y} = \mathbf{X}_2\beta_2 + \epsilon \quad (10.33)$$

The least squares estimator of  $\beta_2$  is  $\hat{\beta}_2 = (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{y}$ , and

$$SS_R(\beta_2) = \hat{\beta}_2'\mathbf{X}_2'\mathbf{y} \quad (p - r \text{ degrees of freedom}) \quad (10.34)$$

The regression sum of squares due to  $\beta_1$  given that  $\beta_2$  is already in the model is

$$SS_R(\beta_1|\beta_2) = SS_R(\beta) - SS_R(\beta_2) \quad (10.35)$$

This sum of squares has  $r$  degrees of freedom. It is the “extra sum of squares” due to  $\beta_1$ . Note that  $SS_R(\beta_1|\beta_2)$  is the increase in the regression sum of squares due to inclusion of variables  $x_1, x_2, \dots, x_r$  in the model.

Now,  $SS_R(\beta_1|\beta_2)$  is independent of  $MS_E$ , and the null hypothesis  $\beta_1 = \mathbf{0}$  may be tested by the statistic

$$F_0 = \frac{SS_R(\beta_1|\beta_2)/r}{MS_E} \quad (10.36)$$

If  $F_0 > F_{\alpha, r, n-p}$ , we reject  $H_0$ , concluding that at least one of the parameters in  $\beta_1$  is not zero, and, consequently, at least one of the variables  $x_1, x_2, \dots, x_r$  in  $\mathbf{X}_1$  contributes significantly to the regression model. Some authors call the test in Equation 10.36 a **partial  $F$  test**.

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

$$SS_R(\beta_j|\beta_0, \beta_1, \dots, \beta_{j-1}, \beta_{j+1}, \dots, \beta_k)$$

This is the increase in the regression sum of squares due to adding  $x_j$  to a model that already includes  $x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_k$ . Note that the partial  $F$  test on a single variable  $x_j$  is equivalent to the  $t$  test in Equation 10.28. However, the partial  $F$  test is a more general procedure in that we can measure the effect of sets of variables.

## EXAMPLE 10.6

Consider the viscosity data in Example 10.1. Suppose that we wish to investigate the contribution of the variable  $x_2$  (feed rate) to the model. That is, the hypotheses we wish to test are

$$H_0: \beta_2 = 0$$

$$H_1: \beta_2 \neq 0$$

This will require the extra sum of squares due to  $\beta_2$ , or

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

Now from Table 10.4, where we tested for significance of regression, we have

$$SS_R(\beta_1, \beta_2|\beta_0) = 44,157.1$$

which was called the model sum of squares in the table. This sum of squares has two degrees of freedom.

The reduced model is

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

The least squares fit for this model is

$$\hat{y} = 1652.3955 + 7.6397x_1$$

and the regression sum of squares for this model (with one degree of freedom) is

$$SS_R(\beta_1|\beta_0) = 40,840.8$$

Note that  $SS_R(\beta_1|\beta_0)$  is shown at the bottom of the Minitab output in Table 10.4 under the heading “Seq SS.” Therefore,

$$\begin{aligned} SS_R(\beta_2|\beta_0, \beta_1) &= 44,157.1 - 40,840.8 \\ &= 3316.3 \end{aligned}$$

with  $2 - 1 = 1$  degree of freedom. This is the increase in the regression sum of squares that results from adding  $x_2$  to a model already containing  $x_1$ , and it is shown at the bottom of the Minitab output on Table 10.4. To test  $H_0: \beta_2 = 0$ , from the test statistic we obtain

$$F_0 = \frac{SS_R(\beta_2|\beta_0, \beta_1)/1}{MS_E} = \frac{3316.3/1}{267.604} = 12.3926$$

Note that  $MS_E$  from the full model (Table 10.4) is used in the denominator of  $F_0$ . Now, because  $F_{0.05,1,13} = 4.67$ , we would reject  $H_0: \beta_2 = 0$  and conclude that  $x_2$  (feed rate) contributes significantly to the model.

Because this partial  $F$  test involves only a single regressor, it is equivalent to the  $t$ -test because the square of a  $t$  random variable with  $\nu$  degrees of freedom is an  $F$  random variable with 1 and  $\nu$  degrees of freedom. To see this, note from Table 10.4 that the  $t$ -statistic for  $H_0: \beta_2 = 0$  resulted in  $t_0 = 3.5203$  and that  $t_0^2 = (3.5203)^2 = 12.3925 \approx F_0$ .

## 10.5 Confidence Intervals in Multiple Regression

It is often necessary to construct confidence interval estimates for the regression coefficients  $\{\beta_j\}$  and for other quantities of interest from the regression model. The development of a procedure for obtaining these confidence intervals requires that we assume the errors  $\{\epsilon_i\}$  to be normally and independently distributed with mean zero and variance  $\sigma^2$ , the same assumption made in the section on hypothesis testing in Section 10.4.

### 10.5.1 Confidence Intervals on the Individual Regression Coefficients

Because the least squares estimator  $\hat{\beta}$  is a linear combination of the observations, it follows that  $\hat{\beta}$  is normally distributed with mean vector  $\beta$  and covariance matrix  $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ . Then each of the statistics

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} \quad j = 0, 1, \dots, k \quad (10.37)$$

is distributed as  $t$  with  $n - p$  degrees of freedom, where  $C_{jj}$  is the  $(jj)$ th element of the  $(\mathbf{X}'\mathbf{X})^{-1}$  matrix, and  $\hat{\sigma}^2$  is the estimate of the error variance, obtained from Equation 10.17. Therefore, a  $100(1 - \alpha)$  percent confidence interval for the regression coefficient  $\beta_j$ ,  $j = 0, 1, \dots, k$ , is

$$\hat{\beta}_j - t_{\alpha/2, n-p} \sqrt{\hat{\sigma}^2 C_{jj}} \leq \beta_j \leq \hat{\beta}_j + t_{\alpha/2, n-p} \sqrt{\hat{\sigma}^2 C_{jj}} \quad (10.38)$$

Note that this confidence interval could also be written as

$$\hat{\beta}_j - t_{\alpha/2, n-p} se(\hat{\beta}_j) \leq \beta_j \leq \hat{\beta}_j + t_{\alpha/2, n-p} se(\hat{\beta}_j)$$

because  $se(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 C_{jj}}$ .

**EXAMPLE 10.7**

We will construct a 95 percent confidence interval for the parameter  $\beta_1$  in Example 10.1. Now  $\hat{\beta}_1 = 7.62129$ , and because  $\hat{\sigma}^2 = 267.604$  and  $C_{11} = 1.429184 \times 10^{-3}$ , we find that

$$\begin{aligned}\hat{\beta}_1 - t_{0.025,13} \sqrt{\hat{\sigma}^2 C_{11}} &\leq \beta_1 \leq \hat{\beta}_1 + t_{0.025,13} \sqrt{\hat{\sigma}^2 C_{11}} \\ 7.62129 - 2.16 \sqrt{(267.604)(1.429184 \times 10^{-3})} &\leq \beta_1 \\ &\leq 7.62129 + 2.16 \sqrt{(267.604)(1.429184 \times 10^{-3})} \\ 7.62129 - 2.16(0.6184) &\leq \beta_1 \leq 7.62129 + 2.16(0.6184)\end{aligned}$$

and the 95 percent confidence interval on  $\beta_1$  is

$$6.2855 \leq \beta_1 \leq 8.9570$$

**10.5.2 Confidence Interval on the Mean Response**

We may also obtain a confidence interval on the mean response at a particular point, say,  $x_{01}, x_{02}, \dots, x_{0k}$ . We first 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

$$\mu_{y|\mathbf{x}_0} = \beta_0 + \beta_1 x_{01} + \beta_2 x_{02} + \dots + \beta_k x_{0k} = \mathbf{x}_0' \boldsymbol{\beta}$$

The estimated mean response at this point is

$$\hat{y}(\mathbf{x}_0) = \mathbf{x}_0' \hat{\boldsymbol{\beta}} \quad (10.39)$$

This estimator is unbiased because  $E[\hat{y}(\mathbf{x}_0)] = E(\mathbf{x}_0' \hat{\boldsymbol{\beta}}) = \mathbf{x}_0' \boldsymbol{\beta} = \mu_{y|\mathbf{x}_0}$ , and the variance of  $\hat{y}(\mathbf{x}_0)$  is

$$V[\hat{y}(\mathbf{x}_0)] = \sigma^2 \mathbf{x}_0' (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_0 \quad (10.40)$$

Therefore, a  $100(1 - \alpha)$  percent confidence interval on the mean response at the point  $x_{01}, x_{02}, \dots, x_{0k}$  is

$$\hat{y}(\mathbf{x}_0) - t_{\alpha/2, n-p} \sqrt{\hat{\sigma}^2 \mathbf{x}_0' (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_0} \leq \mu_{y|\mathbf{x}_0} \leq \hat{y}(\mathbf{x}_0) + t_{\alpha/2, n-p} \sqrt{\hat{\sigma}^2 \mathbf{x}_0' (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_0} \quad (10.41)$$

**10.6 Prediction of New Response Observations**

A regression model can be used to predict future observations on the response  $y$  corresponding to particular values of the regressor variables, say  $x_{01}, x_{02}, \dots, x_{0k}$ . If  $\mathbf{x}_0' = [1, x_{01}, x_{02}, \dots, x_{0k}]$ , then a point estimate for the future observation  $y_0$  at the point  $x_{01}, x_{02}, \dots, x_{0k}$  is computed from Equation 10.39:

$$\hat{y}(\mathbf{x}_0) = \mathbf{x}_0' \hat{\boldsymbol{\beta}}$$

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

$$\begin{aligned}\hat{y}(\mathbf{x}_0) &= t_{\alpha/2, n-p} \sqrt{\hat{\sigma}^2(1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)} \leq y_0 \\ &\leq \hat{y}(\mathbf{x}_0) + t_{\alpha/2, n-p} \sqrt{\hat{\sigma}^2(1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)}\end{aligned}\quad (10.42)$$

In predicting new observations and in estimating the mean response at a given point  $x_{01}, x_{02}, \dots, x_{0k}$ , 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.

The prediction interval in Equation 10.42 has many useful applications. One of these is in confirmation experiments following a factorial or fractional factorial experiment. In a confirmation experiment, we are usually testing the model developed from the original experiment to determine if our interpretation was correct. Often we will do this by using the model to predict the response at some point of interest in the design space and then comparing the predicted response with an actual observation obtained by conducting another trial at that point. We illustrated this in Chapter 8, using the  $2^{4-1}$  fractional factorial design in Example 8.1. A useful measure of confirmation is to see if the new observation falls inside the prediction interval on the response at that point.

To illustrate, reconsider the situation in Example 8.1. The interpretation of this experiment indicated that three of the four main effects ( $A$ ,  $C$ , and  $D$ ) and two of the two-factor interactions ( $AC$  and  $AD$ ) were important. The point with  $A$ ,  $B$ , and  $D$  at the high level and  $C$  at the low level was considered to be a reasonable confirmation run, and the predicted value of the response at that point was 100.25. If the fractional factorial has been interpreted correctly and the model for the response is valid, we would expect the observed value at this point to fall inside the prediction interval computed from Equation 10.42. This interval is easy to calculate. Since the  $2^{4-1}$  is an orthogonal design, and the model contains six terms (the intercept, the three main effects, and the two two-factor interactions), the  $(\mathbf{X}'\mathbf{X})^{-1}$  matrix has a particularly simple form, namely  $(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{8} \mathbf{I}_6$ . Furthermore, the coordinates of the point of interest are  $x_1 = 1, x_2 = 1, x_3 = -1$ , and  $x_4 = 1$ , but since  $B$  (or  $x_2$ ) isn't in the model and the two interactions  $AC$  and  $AD$  (or  $x_1x_3$  and  $x_1x_4 = 1$ ) are in the model, the coordinates of the point of interest  $\mathbf{x}_0$  are given by  $\mathbf{x}_0' = [1, x_1, x_3, x_4, x_1x_3, x_1x_4] = [1, 1, -1, 1, -1, 1]$ . It is also easy to show that the estimate of  $\sigma^2$  (with two degrees of freedom) for this model is  $\hat{\sigma}^2 = 3.25$ . Therefore, using Equation 10.42, a 95 percent prediction interval on the observation at this point is

$$\begin{aligned}\hat{y}(\mathbf{x}_0) - t_{0.025, 2} \sqrt{\hat{\sigma}^2(1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)} &\leq y_0 \leq \hat{y}(\mathbf{x}_0) + t_{0.025, 2} \sqrt{\hat{\sigma}^2(1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)} \\ 100.25 - 4.30 \sqrt{3.25 \left(1 + \mathbf{x}_0' \frac{1}{8} \mathbf{I}_6 \mathbf{x}_0\right)} &\leq y_0 \leq 100.25 + 4.30 \sqrt{3.25 \left(1 + \mathbf{x}_0' \frac{1}{8} \mathbf{I}_6 \mathbf{x}_0\right)} \\ 100.25 - 4.30 \sqrt{3.25(1 + 0.75)} &\leq y_0 \leq 100.25 + 4.30 \sqrt{3.25(1 + 0.75)} \\ 100.25 - 10.25 &\leq y_0 \leq 100.25 + 10.25 \\ 90 &\leq y_0 \leq 110.50\end{aligned}$$

Therefore, we would expect the confirmation run with  $A$ ,  $B$ , and  $D$  at the high level and  $C$  at the low level to result in an observation on the filtration rate response that falls between 90 and 110.50. The actual observation was 104. The successful confirmation run provides some assurance that the fractional factorial was interpreted correctly.

## 10.7 Regression Model Diagnostics

As we emphasized in designed experiments, **model adequacy checking** is an important part of the data analysis procedure. This is equally important in building regression models, and as we illustrated in Example 10.1, the **residual plots** that we used with designed experiments should always be examined for a regression model. In general, it is always necessary to (1) examine the fitted model to ensure that it provides an adequate approximation to the true system and (2) verify that none of the least squares regression assumptions are violated. The regression model will probably give poor or misleading results unless it is an adequate fit.

In addition to residual plots, other model diagnostics are frequently useful in regression. This section briefly summarizes some of these procedures. For more complete presentations, see Montgomery, Peck, and Vining (2006) and Myers (1990).

### 10.7.1 Scaled Residuals and PRESS

**Standardized and Studentized Residuals.** Many model builders prefer to work with **scaled residuals** in contrast to the ordinary least squares residuals. These scaled residuals often convey more information than do the ordinary residuals.

One type of scaled residual is the **standardized residual**:

$$d_i = \frac{e_i}{\hat{\sigma}} \quad i = 1, 2, \dots, n \quad (10.43)$$

where we generally use  $\hat{\sigma} = \sqrt{MS_E}$  in the computation. These standardized residuals have mean zero and approximately unit variance; consequently, they are useful in looking for **outliers**. Most of the standardized residuals should lie in the interval  $-3 \leq d_i \leq 3$ , and any observation with a standardized residual outside of this interval is potentially unusual with respect to its observed response. These outliers should be carefully examined because they may represent something as simple as a data-recording error or something of more serious concern, such as a region of the regressor variable space where the fitted model is a poor approximation to the true response surface.

The standardizing process in Equation 10.43 scales the residuals by dividing them by their approximate average standard deviation. In some data sets, residuals may have standard deviations that differ greatly. We now present a scaling that takes this into account.

The vector of fitted values  $\hat{y}_i$  corresponding to the observed values  $y_i$  is

$$\begin{aligned} \hat{\mathbf{y}} &= \mathbf{X}\hat{\boldsymbol{\beta}} \\ &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= \mathbf{H}\mathbf{y} \end{aligned} \quad (10.44)$$

The  $n \times n$  matrix  $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$  is usually called the “hat” matrix because it maps the vector of observed values into a vector of fitted values. The hat matrix and its properties play a central role in regression analysis.

The residuals from the fitted model may be conveniently written in matrix notation as

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$$

and it turns out that the covariance matrix of the residuals is

$$\text{Cov}(\mathbf{e}) = \sigma^2 (\mathbf{I} - \mathbf{H}) \quad (10.45)$$

The matrix  $\mathbf{I} - \mathbf{H}$  is generally not diagonal, so the residuals have different variances and they are correlated.

Thus, the variance of the  $i$ th residual is

$$V(e_i) = \sigma^2(1 - h_{ii}) \quad (10.46)$$

where  $h_{ii}$  is the  $i$ th diagonal element of  $\mathbf{H}$ . Because  $0 \leq h_{ii} \leq 1$ , using the residual mean square  $MS_E$  to estimate the variance of the residuals actually overestimates  $V(e_i)$ . Furthermore, because  $h_{ii}$  is a measure of the location of the  $i$ th point in  $x$ -space, the variance of  $e_i$  depends on where the point  $x_i$  lies. Generally, residuals near the center of the  $x$  space have larger variance than do residuals at more remote locations. Violations of model assumptions are more likely at remote points, and these violations may be hard to detect from inspection of  $e_i$  (or  $d_i$ ) because their residuals will usually be smaller.

We recommend taking this inequality of variance into account when scaling the residuals. We suggest plotting the **studentized residuals**:

$$r_i = \frac{e_i}{\sqrt{\hat{\sigma}^2(1 - h_{ii})}} \quad i = 1, 2, \dots, n \quad (10.47)$$

with  $\hat{\sigma}^2 = MS_E$  instead of  $e_i$  (or  $d_i$ ). The studentized residuals have constant variance  $V(r_i) = 1$  regardless of the location of  $\mathbf{x}_i$  when the form of the model is correct. In many situations the variance of the residuals stabilizes, particularly for large data sets. In these cases, there may be little difference between the standardized and studentized residuals. Thus standardized and studentized residuals often convey equivalent information. However, because any point with a large residual and a large  $h_{ii}$  is potentially highly influential on the least squares fit, examination of the studentized residuals is generally recommended. Table 10.3 displays the hat diagonals  $h_{ii}$  and the studentized residuals for the viscosity regression model in Example 10.1.

**PRESS Residuals.** The prediction error sum of squares (PRESS) provides a useful residual scaling. To calculate PRESS, we select an observation—for example,  $i$ . We fit the regression model to the remaining  $n - 1$  observations and use this equation to predict the withheld observation  $y_i$ . Denoting this predicted value  $\hat{y}_{(i)}$ , we may find the prediction error for point  $i$  as  $e_{(i)} = y_i - \hat{y}_{(i)}$ . The prediction error is often called the  $i$ th PRESS residual. This procedure is repeated for each observation  $i = 1, 2, \dots, n$ , producing a set of  $n$  PRESS residuals  $e_{(1)}, e_{(2)}, \dots, e_{(n)}$ . Then the PRESS statistic is defined as the sum of squares of the  $n$  PRESS residuals as in

$$\text{PRESS} = \sum_{i=1}^n e_{(i)}^2 = \sum_{i=1}^n [y_i - \hat{y}_{(i)}]^2 \quad (10.48)$$

Thus PRESS uses each possible subset of  $n - 1$  observations as an estimation data set, and every observation in turn is used to form a prediction data set.

It would initially seem that calculating PRESS requires fitting  $n$  different regressions. However, it is possible to calculate PRESS from the results of a single least squares fit to all  $n$  observations. It turns out that the  $i$ th PRESS residual is

$$e_{(i)} = \frac{e_i}{1 - h_{ii}} \quad (10.49)$$

Thus because PRESS is just the sum of the squares of the PRESS residuals, a simple computing formula is

$$\text{PRESS} = \sum_{i=1}^n \left( \frac{e_i}{1 - h_{ii}} \right)^2 \quad (10.50)$$

From Equation 10.49, it is easy to see that the PRESS residual is just the ordinary residual weighted according to the diagonal elements of the hat matrix  $h_{ii}$ . Data points for which  $h_{ii}$  are large will have large PRESS residuals. These observations will generally be **high influence**



points. Generally, a large difference between the ordinary residual and the PRESS residuals will indicate a point where the model fits the data well, but a model built without that point predicts poorly. In the next section we will discuss some other measures of influence.

Finally, we note that PRESS can be used to compute an approximate  $R^2$  for prediction, say

$$R_{\text{Prediction}}^2 = 1 - \frac{\text{PRESS}}{SS_T} \quad (10.51)$$

This statistic gives some indication of the predictive capability of the regression model. For the viscosity regression model from Example 10.1, we can compute the PRESS residuals using the ordinary residuals and the values of  $h_{ii}$  found in Table 10.3. The corresponding value of the PRESS statistic is  $\text{PRESS} = 5207.7$ . Then

$$\begin{aligned} R_{\text{Prediction}}^2 &= 1 - \frac{\text{PRESS}}{SS_T} \\ &= 1 - \frac{5207.7}{47,635.9} = 0.8907 \end{aligned}$$

Therefore, we could expect this model to “explain” about 89 percent of the variability in predicting new observations, as compared to the approximately 93 percent of the variability in the original data explained by the least squares fit. The overall predictive capability of the model based on this criterion seems very satisfactory.

***R-Student.*** The studentized residual  $r_i$  discussed above is often considered an outlier diagnostic. It is customary to use  $MS_E$  as an estimate of  $\sigma^2$  in computing  $r_i$ . This is referred to as internal scaling of the residual because  $MS_E$  is an internally generated estimate of  $\sigma^2$  obtained from fitting the model to all  $n$  observations. Another approach would be to use an estimate of  $\sigma^2$  based on a data set with the  $i$ th observation removed. We denote the estimate of  $\sigma^2$  so obtained by  $S_{(i)}^2$ . We can show that

$$S_{(i)}^2 = \frac{(n-p)MS_E - e_i^2/(1-h_{ii})}{n-p-1} \quad (10.52)$$

The estimate of  $\sigma^2$  in Equation 10.52 is used instead of  $MS_E$  to produce an externally studentized residual, usually called *R-student*, given by

$$t_i = \frac{e_i}{\sqrt{S_{(i)}^2(1-h_{ii})}} \quad i = 1, 2, \dots, n \quad (10.53)$$

In many situations,  $t_i$  will differ little from the studentized residual  $r_i$ . However, if the  $i$ th observation is influential, then  $S_{(i)}^2$  can differ significantly from  $MS_E$ , and thus the *R-student* will be more sensitive to this point. Furthermore, under the standard assumptions,  $t_i$  has a  $t_{n-p-1}$  distribution. Thus *R-student* offers a more formal procedure for outlier detection via hypothesis testing. Table 10.3 displays the values of *R-student* for the viscosity regression model in Example 10.1. None of those values are unusually large.

## 10.7.2 Influence Diagnostics

We occasionally find that a small subset of the data exerts a disproportionate influence on the fitted regression model. That is, parameter estimates or predictions may depend more on the influential subset than on the majority of the data. We would like to locate these influential points and assess their impact on the model. If these influential points are “bad” values, they should be eliminated. On the contrary, there may be nothing wrong with these points. But if they control key model properties, we would like to know it because it could



affect the use of the model. In this section we describe and illustrate some useful measures of influence.

**Leverage Points.** The disposition of points in  $x$  space is important in determining model properties. In particular, remote observations potentially have disproportionate leverage on the parameter estimates, predicted values, and the usual summary statistics.

The hat matrix  $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$  is very useful in identifying influential observations. As noted earlier,  $\mathbf{H}$  determines the variances and covariances of  $\hat{\mathbf{y}}$  and  $\mathbf{e}$  because  $V(\hat{\mathbf{y}}) = \sigma^2\mathbf{H}$  and  $V(\mathbf{e}) = \sigma^2(\mathbf{I} - \mathbf{H})$ . The elements  $h_{ij}$  of  $\mathbf{H}$  may be interpreted as the amount of leverage exerted by  $y_j$  on  $\hat{y}_i$ . Thus, inspection of the elements of  $\mathbf{H}$  can reveal points that are potentially influential by virtue of their location in  $x$  space. Attention is usually focused on the diagonal elements  $h_{ii}$ . Because  $\sum_{i=1}^n h_{ii} = \text{rank}(\mathbf{H}) = \text{rank}(\mathbf{X}) = p$ , the average size of the diagonal element of the  $\mathbf{H}$  matrix is  $p/n$ . As a rough guideline, then, if a diagonal element  $h_{ii}$  is greater than  $2p/n$ , observation  $i$  is a high-leverage point. To apply this to the viscosity model in Example 10.1, note that  $2p/n = 2(3)/16 = 0.375$ . Table 10.3 gives the hat diagonals  $h_{ii}$  for the first-order model; because none of the  $h_{ii}$  exceeds 0.375, we would conclude that there are no leverage points in these data.

**Influence on Regression Coefficients.** The hat diagonals will identify points that are potentially influential due to their location in  $x$  space. It is desirable to consider both the location of the point and the response variable in measuring influence. Cook (1977, 1979) has suggested using a measure of the squared distance between the least squares estimate based on all  $n$  points  $\hat{\boldsymbol{\beta}}$  and the estimate obtained by deleting the  $i$  point, say  $\hat{\boldsymbol{\beta}}_{(i)}$ . This distance measure can be expressed as

$$D_i = \frac{(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})' \mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})}{pMS_E} \quad i = 1, 2, \dots, n \quad (10.54)$$

A reasonable cutoff for  $D_i$  is unity. That is, we usually consider observations for which  $D_i > 1$  to be influential.

The  $D_i$  statistic is actually calculated from

$$D_i = \frac{r_i^2}{p} \frac{V[\hat{y}(x_i)]}{V(e_i)} = \frac{r_i^2}{p} \frac{h_{ii}}{(1 - h_{ii})} \quad i = 1, 2, \dots, n \quad (10.55)$$

Note that, apart from the constant  $p$ ,  $D_i$  is the product of the square of the  $i$ th studentized residual and  $h_{ii}/(1 - h_{ii})$ . This ratio can be shown to be the distance from the vector  $\mathbf{x}_i$  to the centroid of the remaining data. Thus,  $D_i$  is made up of a component that reflects how well the model fits the  $i$ th observation  $y_i$  and a component that measures how far that point is from the rest of the data. Either component (or both) may contribute to a large value of  $D_i$ .

Table 10.3 presents the values of  $D_i$  for the regression model fit to the viscosity data in Example 10.1. None of these values of  $D_i$  exceeds 1, so there is no strong evidence of influential observations in these data.

## 10.8 Testing for Lack of Fit

In Section 6.8 we showed how adding center points to a  $2^k$  factorial design allows the experimenter to obtain an estimate of pure experimental error. This allows the partitioning of the residual sum of squares  $SS_E$  into two components; that is

$$SS_E = SS_{PE} + SS_{LOF}$$

where  $SS_{PE}$  is the sum of squares due to pure error and  $SS_{LOF}$  is the sum of squares due to lack of fit.

We may give a general development of this partitioning in the context of a regression model. Suppose that we have  $n_i$  observations on the response at the  $i$ th level of the regressors  $\mathbf{x}_i$ ,  $i = 1, 2, \dots, m$ . Let  $y_{ij}$  denote the  $j$ th observation on the response at  $\mathbf{x}_i$ ,  $i = 1, 2, \dots, m$  and  $j = 1, 2, \dots, n_i$ . There are  $n = \sum_{i=1}^m n_i$  total observations. We may write the  $(ij)$ th residual as

$$y_{ij} - \hat{y}_i = (y_{ij} - \bar{y}_i) + (\bar{y}_i - \hat{y}_i) \quad (10.56)$$

where  $\bar{y}_i$  is the average of the  $n_i$  observations at  $\mathbf{x}_i$ . Squaring both sides of Equation 10.56 and summing over  $i$  and  $j$  yields

$$\sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \hat{y}_i)^2 = \sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^m n_i (\bar{y}_i - \hat{y}_i)^2 \quad (10.57)$$

The left-hand side of Equation 10.57 is the usual residual sum of squares. The two components on the right-hand side measure pure error and lack of fit. We see that the pure error sum of squares

$$SS_{PE} = \sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 \quad (10.58)$$

is obtained by computing the corrected sum of squares of the repeat observations at each level of  $\mathbf{x}$  and then pooling over the  $m$  levels of  $\mathbf{x}$ . If the assumption of constant variance is satisfied, this is a **model-independent** measure of pure error because only the variability of the  $y$ 's at each  $\mathbf{x}_i$  level is used to compute  $SS_{PE}$ . Because there are  $n_i - 1$  degrees of freedom for pure error at each level  $\mathbf{x}_i$ , the total number of degrees of freedom associated with the pure error sum of squares is

$$\sum_{i=1}^m (n_i - 1) = n - m \quad (10.59)$$

The sum of squares for lack of fit

$$SS_{LOF} = \sum_{i=1}^m n_i (\bar{y}_i - \hat{y}_i)^2 \quad (10.60)$$

is a weighted sum of squared deviations between the mean response  $\bar{y}_i$  at each  $\mathbf{x}_i$  level and the corresponding fitted value. If the fitted values  $\hat{y}_i$  are close to the corresponding average responses  $\bar{y}_i$ , then there is a strong indication that the regression function is linear. If the  $\hat{y}_i$  deviate greatly from the  $\bar{y}_i$ , then it is likely that the regression function is not linear. There are  $m - p$  degrees of freedom associated with  $SS_{LOF}$  because there are  $m$  levels of  $\mathbf{x}$ , and  $p$  degrees of freedom are lost because  $p$  parameters must be estimated for the model. Computationally we usually obtain  $SS_{LOF}$  by subtracting  $SS_{PE}$  from  $SS_E$ .

The test statistic for lack of fit is

$$F_0 = \frac{SS_{LOF}/(m - p)}{SS_{PE}/(n - m)} = \frac{MS_{LOF}}{MS_{PE}} \quad (10.61)$$

The expected value of  $MS_{PE}$  is  $\sigma^2$ , and the expected value of  $MS_{LOF}$  is

$$E(MS_{LOF}) = \sigma^2 + \frac{\sum_{i=1}^m n_i \left[ E(y_i) - \beta_0 - \sum_{j=1}^k \beta_j x_{ij} \right]^2}{m - 2} \quad (10.62)$$

If the true regression function is linear, then  $E(y_i) = \beta_0 + \sum_{j=1}^k \beta_j x_{ij}$ , and the second term of Equation 10.62 is zero, resulting in  $E(MS_{LOF}) = \sigma^2$ . However, if the true regression function is not linear, then  $E(y_i) \neq \beta_0 + \sum_{j=1}^k \beta_j x_{ij}$ , and  $E(MS_{LOF}) > \sigma^2$ . Furthermore, if the true regression function is linear, then the statistic  $F_0$  follows the  $F_{m-p, n-m}$  distribution. Therefore, to test

for lack of fit, we would compute the test statistic  $F_0$  and conclude that the regression function is not linear if  $F_0 > F_{\alpha, m-p, n-m}$ .

This test procedure may be easily incorporated into the analysis of variance. If we conclude that the regression function is not linear, then the tentative model must be abandoned and attempts made to find a more appropriate equation. Alternatively, if  $F_0$  does not exceed  $F_{\alpha, m-p, n-m}$ , there is no strong evidence of lack of fit and  $MS_{PE}$  and  $MS_{LOF}$  are often combined to estimate  $\sigma^2$ . Example 6.7 is a very complete illustration of this procedure, where the replicate runs are center points in a  $2^4$  factorial design.

## 10.9 Problems



**10.1.** The tensile strength of a paper product is related to the amount of hardwood in the pulp. Ten samples are produced in the pilot plant, and the data obtained are shown in the following table.

Strength	Percent Hardwood	Strength	Percent Hardwood
160	10	181	20
171	15	188	25
175	15	193	25
182	20	195	28
184	20	200	30

- Fit a linear regression model relating strength to percent hardwood.
- Test the model in part (a) for significance of regression.
- Find a 95 percent confidence interval on the parameter  $\beta_1$ .

**10.2.** A plant distills liquid air to produce oxygen, nitrogen, and argon. The percentage of impurity in the oxygen is thought to be linearly related to the amount of impurities in the air as measured by the “pollution count” in parts per million (ppm). A sample of plant operating data is shown below:

Purity (%)	93.3	92.0	92.4	91.7	94.0	94.6	93.6	
Pollution count (ppm)	1.10	1.45	1.36	1.59	1.08	0.75	1.20	
Purity (%)	93.1	93.2	92.9	92.2	91.3	90.1	91.6	91.9
Pollution count (ppm)	0.99	0.83	1.22	1.47	1.81	2.03	1.75	1.68

- Fit a linear regression model to the data.
- Test for significance of regression.
- Find a 95 percent confidence interval on  $\beta_1$ .

**10.3.** Plot the residuals from Problem 10.1 and comment on model adequacy.

**10.4.** Plot the residuals from Problem 10.2 and comment on model adequacy.

**10.5.** Using the results of Problem 10.1, test the regression model for lack of fit.

**10.6.** 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:

$y$	$x_1$	$x_2$
193	1.6	851
230	15.5	816
172	22.0	1058
91	43.0	1201
113	33.0	1357
125	40.0	1115

- Fit a multiple linear regression model to the data.
- Test for significance of regression.
- Compute  $t$  statistics for each model parameter. What conclusions can you draw?


**10.7.** The brake horsepower developed by an automobile engine on a dynamometer is thought to be a function of the engine speed in revolutions per minute (rpm), the road octane number of the fuel, and the engine compression. An experiment is run in the laboratory and the data that follow are collected:

Brake Horsepower	rpm	Road Octane Number	Compression
225	2000	90	100
212	1800	94	95
229	2400	88	110
222	1900	91	96
219	1600	86	100
278	2500	96	110

246	3000	94	98
237	3200	90	100
233	2800	88	105
224	3400	86	97
223	1800	90	100
230	2500	89	104

- (a) Fit a multiple regression model to these data.  
 (b) Test for significance of regression. What conclusions can you draw?  
 (c) Based on  $t$ -tests, do you need all three regressor variables in the model?

**10.8.** Analyze the residuals from the regression model in Problem 10.7. Comment on model adequacy.

 **10.9.** The yield of a chemical process is related to the concentration of the reactant and the operating temperature. An experiment has been conducted with the following results.

Yield	Concentration	Temperature
81	1.00	150
89	1.00	180
83	2.00	150
91	2.00	180
79	1.00	150
87	1.00	180
84	2.00	150
90	2.00	180

- (a) Suppose we wish to fit a main effects model to this data. Set up the  $\mathbf{X}'\mathbf{X}$  matrix using the data exactly as it appears in the table.  
 (b) Is the matrix you obtained in part (a) diagonal? Discuss your response.  
 (c) Suppose we write our model in terms of the “usual” coded variables

$$x_1 = \frac{\text{Conc} - 1.5}{0.5} \quad x_2 = \frac{\text{Temp} - 165}{15}$$

Set up the  $\mathbf{X}'\mathbf{X}$  matrix for the model in terms of these coded variables. Is this matrix diagonal? Discuss your response.

- (d) Define a new set of coded variables

$$x_1 = \frac{\text{Conc} - 1.0}{1.0} \quad x_2 = \frac{\text{Temp} - 150}{30}$$

Set up the  $\mathbf{X}'\mathbf{X}$  matrix for the model in terms of this set of coded variables. Is this matrix diagonal? Discuss your response.

- (e) Summarize what you have learned from this problem about coding the variables.

**10.10.** Consider the  $2^4$  factorial experiment in Example 6.2. Suppose that the last observation is missing. Reanalyze the data and draw conclusions. How do these conclusions compare with those from the original example?

**10.11.** Consider the  $2^4$  factorial experiment in Example 6.2. Suppose that the last two observations are missing. Reanalyze the data and draw conclusions. How do these conclusions compare with those from the original example?

**10.12.** Given the following data, fit the second-order polynomial regression model

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

y	$x_1$	$x_2$
26	1.0	1.0
24	1.0	1.0
175	1.5	4.0
160	1.5	4.0
163	1.5	4.0
55	0.5	2.0
62	1.5	2.0
100	0.5	3.0
26	1.0	1.5
30	0.5	1.5
70	1.0	2.5
71	0.5	2.5

After you have fit the model, test for significance of regression.

**10.13.**

- (a) Consider the quadratic regression model from Problem 10.12. Compute  $t$  statistics for each model parameter and comment on the conclusions that follow from these quantities.  
 (b) Use the extra sum of squares method to evaluate the value of the quadratic terms  $x_1^2$ ,  $x_2^2$ , and  $x_1 x_2$  to the model.

**10.14. Relationship between analysis of variance and regression.** Any analysis of variance model can be expressed in terms of the general linear model  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ , where the  $\mathbf{X}$  matrix consists of 0s and 1s. Show that the single-factor model  $y_{ij} = \mu + \tau_i + \epsilon_{ij}$ ,  $i = 1, 2, 3$ ,  $j = 1, 2, 3, 4$  can be written in general linear model form. Then,

- (a) Write the normal equations  $(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$  and compare them with the normal equations found for this model in Chapter 3.  
 (b) Find the rank of  $\mathbf{X}'\mathbf{X}$ . Can  $(\mathbf{X}'\mathbf{X})^{-1}$  be obtained?  
 (c) Suppose the first normal equation is deleted and the restriction  $\sum_{i=1}^3 n\hat{\tau}_i = 0$  is added. Can the resulting system of equations be solved? If so, find the solution. Find the regression sum of squares  $\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}$ , and compare it to the treatment sum of squares in the single-factor model.

**10.15.** Suppose that we are fitting a straight line and we desire to make the variance of  $\hat{\beta}_1$  as small as possible. Restricting ourselves to an even number of experimental points, where should we place these points so as to minimize  $V(\hat{\beta}_1)$ ? [Note: Use the design called for in this exercise with great caution because, even though it minimizes  $V(\hat{\beta}_1)$ , it has some undesirable properties; for example, see Myers, Montgomery and Anderson-Cook (2009). Only if you are *very sure* the true functional relationship is linear should you consider using this design.]

**10.16. Weighted least squares.** Suppose that we are fitting the straight line  $y = \beta_0 + \beta_1 x + \epsilon$ , but the variance of the  $y$ 's now depends on the level of  $x$ ; that is,

$$V(y|x_i) = \sigma_i^2 = \frac{\sigma^2}{w_i} \quad i = 1, 2, \dots, n$$

where the  $w_i$  are known constants, often called weights. Show that if we choose estimates of the regression coefficients to minimize the weighted sum of squared errors given by  $\sum_{i=1}^n w_i(y_i - \beta_0 - \beta_1 x_i)^2$ , the resulting least squares normal equations are

$$\begin{aligned} \hat{\beta}_0 \sum_{i=1}^n w_i + \hat{\beta}_1 \sum_{i=1}^n w_i x_i &= \sum_{i=1}^n w_i y_i \\ \hat{\beta}_0 \sum_{i=1}^n w_i x_i + \hat{\beta}_1 \sum_{i=1}^n w_i x_i^2 &= \sum_{i=1}^n w_i x_i y_i \end{aligned}$$

**10.17.** Consider the  $2_{IV}^{4-1}$  design discussed in Example 10.5.

- (a) Suppose you elect to augment the design with the single run selected in that example. Find the variances and covariances of the regression coefficients in the model (ignoring blocks):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_{12} x_1 x_2 + \beta_{34} x_3 x_4 + \epsilon$$

- (b) Are there any other runs in the alternate fraction that would de-alias  $AB$  from  $CD$ ?  
 (c) Suppose you augment the design with the four runs suggested in Example 10.5. Find the variances and covariances of the regression coefficients (ignoring blocks) for the model in part (a).  
 (d) Considering parts (a) and (c), which augmentation strategy would you prefer, and why?

**10.18.** Consider a  $2_{III}^{7-4}$  design. Suppose after running the experiment, the largest observed effects are  $A + BD$ ,  $B + AD$ , and  $D + AB$ . You wish to augment the original design with a group of four runs to de-alias these effects.

- (a) Which four runs would you make?  
 (b) Find the variances and covariances of the regression coefficients in the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4 + \beta_{12} x_1 x_2 + \beta_{14} x_1 x_4 + \beta_{24} x_2 x_4 + \epsilon.$$

- (c) Is it possible to de-alias these effects with fewer than four additional runs?