7 Multiple Regression: Estimation

7.1 INTRODUCTION

In *multiple regression*, we attempt to predict a *dependent* or *response* variable y on the basis of an assumed linear relationship with several *independent* or *predictor* variables $x_1, x_1, \ldots, x_k$. In addition to constructing a model for prediction, we may wish to assess the extent of the relationship between y and the x variables. For this purpose, we use the multiple correlation coefficient R (Section 7.7).

In this chapter, y is a continuous random variable and the x variables are fixed constants (either discrete or continuous) that are controlled by the experimenter. The case in which the x variables are random variables is covered in Chapter 10. In analysis-of-variance (Chapters 12–15), the x variables are fixed and discrete.

Useful applied expositions of multiple regression for the fixed-x case can be found in Morrison (1983), Myers (1990), Montgomery and Peck (1992), Graybill and Iyer (1994), Mendenhall and Sincich (1996), Ryan (1997), Draper and Smith (1998), and Kutner et al. (2005). Theoretical treatments are given by Searle (1971), Graybill (1976), Guttman (1982), Kshirsagar (1983), Myers and Milton (1991), Jørgensen (1993), Wang and Chow (1994), Christensen (1996), Seber and Lee (2003), and Hocking (1976, 1985, 2003).

7.2 THE MODEL

The multiple linear regression model, as introduced in Section 1.2, can be expressed as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon. \tag{7.1}$$

We discuss estimation of the β parameters when the model is linear in the β's. An example of a model that is linear in the β's but not the x's is the second-order

Linear Models in Statistics, Second Edition, by Alvin C. Rencher and G. Bruce Schaalje

response surface model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \varepsilon. \tag{7.2}$$

To estimate the β's in (7.1), we will use a sample of n observations on y and the associated x variables. The model for the ith observation is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i, \quad i = 1, 2, \ldots, n. \tag{7.3}$$

The assumptions for ε_i or y_i are essentially the same as those for simple linear regression in Section 6.1:

1. $E(\varepsilon_i) = 0$ for $i = 1, 2, \ldots, n$, or, equivalently, $E(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$.
2. $\text{var}(\varepsilon_i) = \sigma^2$ for $i = 1, 2, \ldots, n$, or, equivalently, $\text{var}(y_i) = \sigma^2$.
3. $\text{cov}(\varepsilon_i, \varepsilon_j) = 0$ for all $i \neq j$, or, equivalently, $\text{cov}(y_i, y_j) = 0$.

Assumption 1 states that the model is correct, in other words that all relevant x's are included and the model is indeed linear. Assumption 2 asserts that the variance of y is constant and therefore does not depend on the x's. Assumption 3 states that the y's are uncorrelated with each other, which usually holds in a random sample (the observations would typically be correlated in a time series or when repeated measurements are made on a single plant or animal). Later we will add a normality assumption (Section 7.6), under which the y variable will be independent as well as uncorrelated.

When all three assumptions hold, the least-squares estimators of the β's have some good properties (Section 7.3.2). If one or more assumptions do not hold, the estimators may be poor. Under the normality assumption (Section 7.6), the maximum likelihood estimators have excellent properties.

Any of the three assumptions may fail to hold with real data. Several procedures have been devised for checking the assumptions. These diagnostic techniques are discussed in Chapter 9.

Writing (7.3) for each of the n observations, we have

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_k x_{1k} + \varepsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_k x_{2k} + \varepsilon_2 \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_k x_{nk} + \varepsilon_n. \end{aligned}$$

These n equations can be written in matrix form as

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

or

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{7.4}$$

The preceding three assumptions on ε_i or y_i can be expressed in terms of the model in (7.4):

1. $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ or $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$.
2. $\text{cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$ or $\text{cov}(\mathbf{y}) = \sigma^2\mathbf{I}$.

Note that the assumption $\text{cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$ includes both the previous assumptions $\text{var}(\varepsilon_i) = \sigma^2$ and $\text{cov}(\varepsilon_i, \varepsilon_j) = 0$.

The matrix $\mathbf{X}$ in (7.4) is $n \times (k+1)$. In this chapter we assume that $n > k+1$ and rank $(\mathbf{X}) = k+1$. If $n < k+1$ or if there is a linear relationship among the x's, for example, $x_5 = \sum_{j=1}^{4} x_j/4$, then $\mathbf{X}$ will not have full column rank. If the values of the x_{ij}'s are planned (chosen by the researcher), then the $\mathbf{X}$ matrix essentially contains the experimental design and is sometimes called the *design matrix*.

The β parameters in (7.1) or (7.4) are called *regression coefficients*. To emphasize their collective effect, they are sometimes referred to as *partial regression coefficients*. The word *partial* carries both a mathematical and a statistical meaning. Mathematically, the partial derivative of $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$ with respect to x_1, for example, is β_1. Thus β_1 indicates the change in $E(y)$ with a unit increase in x_1 when $x_2, x_3, \dots, x_k$ are held constant. Statistically, β_1 shows the effect of x_1 on $E(y)$ in the presence of the other x's. This effect would typically be different from the effect of x_1 on $E(y)$ if the other x's were not present in the model. Thus, for example, β_0 and β_1 in

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

will usually be different from β_0^* and β_1^* in

$$y = \beta_0^* + \beta_1^* x_1 + \varepsilon^*.$$

[If x_1 and x_2 are orthogonal, that is, if $\mathbf{x}_1'\mathbf{x}_2 = 0$ or if $(\mathbf{x}_1 - \bar{x}_1\mathbf{j})'(\mathbf{x}_2 - \bar{x}_2\mathbf{j}) = 0$, where $\mathbf{x}_1$ and $\mathbf{x}_2$ are columns in the $\mathbf{X}$ matrix, then $\beta_0 = \beta_0^*$ and $\beta_1 = \beta_1^*$; see Corollary 1 to Theorem 7.9a and Theorem 7.10]. The change in parameters when an x is deleted from the model is illustrated (with estimates) in the following example.

TABLE 7.1 Data for Example 7.2

Observation Number	y	x_1	x_2
1	2	0	2
2	3	2	6
3	2	2	7
4	7	2	5
5	6	4	9
6	8	4	8
7	10	4	7
8	7	6	10
9	8	6	11
10	12	6	9
11	11	8	15
12	14	8	13

Example 7.2. [See Freund and Minton (1979, pp. 36–39)]. Consider the (contrived) data in Table 7.1.

Using (6.5) and (6.6) from Section 6.2 and (7.6) in Section 7.3 (see Example 7.3.1), we obtain prediction equations for y regressed on x_1 alone, on x_2 alone, and on both x_1 and x_2:

$$\hat{y} = 1.86 + 1.30x_1,$$
$$\hat{y} = .86 + .78x_2,$$
$$\hat{y} = 5.37 + 3.01x_1 - 1.29x_2.$$

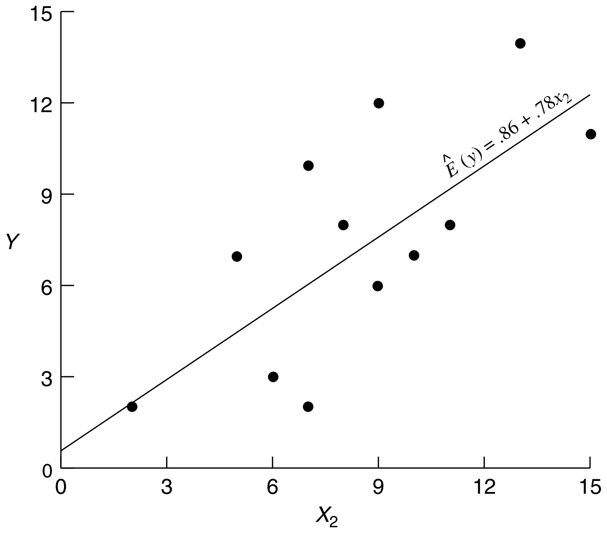

Figure 7.1 Regression of y on x_2 ignoring x_1.

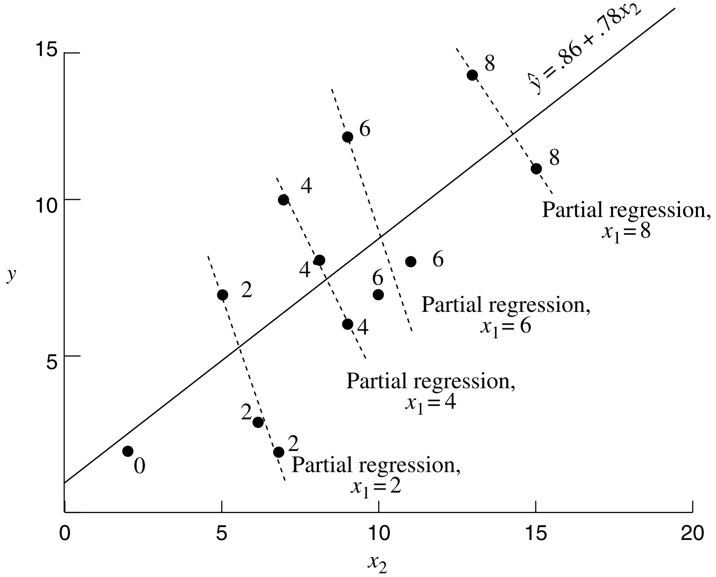

Figure 7.2 Regression of y on x_2 showing the value of x_1 at each point and partial regressions of y on x_2.

As expected, the coefficients change from either of the reduced models to the full model. Note the sign change as the coefficient of x_2 changes from .78 to -1.29.

The values of y and x_2 are plotted in Figure 7.1 along with the prediction equation $\hat{y} = .86 + .78x_2$. The linear trend is clearly evident.

In Figure 7.2 we have the same plot as in Figure 7.1, except that each point is labeled with the value of x_1. Examining values of y and x_2 for a fixed value of x_1 (2, 4, 6, or 8) shows a negative slope for the relationship. These negative relationships are shown as partial regressions of y on x_2 for each value of x_1. The partial regression coefficient $\hat{\beta}_2 = -1.29$ reflects the negative slopes of these four partial regressions.

Further insight into the meaning of the partial regression coefficients is given in Section 7.10. □

7.3 ESTIMATION OF β AND σ^2

7.3.1 Least-Squares Estimator for β

In this section, we discuss the *least-squares approach* to estimation of the β's in the fixed-x model (7.1) or (7.4). No distributional assumptions on y are required to obtain the estimators.

For the parameters $\beta_0, \beta_1, \ldots, \beta_k$, we seek estimators that minimize the sum of squares of deviations of the n observed y's from their predicted values $\hat{y}$. By extension of (6.2), we seek $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ that minimize

$$\begin{aligned}\sum_{i=1}^{n} \hat{\varepsilon}_i^2 &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_k x_{ik})^2. \qquad (7.5)\end{aligned}$$

Note that the predicted value $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik}$ estimates $E(y_i)$, not y_i. A better notation would be $\widehat{E(y_i)}$, but $\hat{y}_i$ is commonly used.

To obtain the least-squares estimators, it is not necessary that the prediction equation $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik}$ be based on $E(y_i)$. It is only necessary to postulate an empirical model that is linear in the $\hat{\beta}$'s, and the least-squares method will find the "best" fit to this model. This was illustrated in Figure 6.2.

To find the values of $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ that minimize (7.5), we could differentiate $\sum_i \hat{\varepsilon}_i^2$ with respect to each $\hat{\beta}_j$ and set the results equal to zero to yield $k + 1$ equations that can be solved simultaneously for the $\hat{\beta}_j$'s. However, the procedure can be carried out in more compact form with matrix notation. The result is given in the following theorem.

Theorem 7.3a. If $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\mathbf{X}$ is $n \times (k+1)$ of rank $k + 1 < n$, then the value of $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)'$ that minimizes (7.5) is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \qquad (7.6)$$

Proof. Using (2.20) and (2.27), we can write (7.5) as

$$\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = \sum_{i=1}^{n} (y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}})^2 = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}), \qquad (7.7)$$

where $\mathbf{x}_i' = (1, x_{i1}, \ldots, x_{ik})$ is the ith row of $\mathbf{X}$. When the product $(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$ in (7.7) is expanded as in (2.17), two of the resulting four terms can be combined to yield

$$\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}.$$

We can find the value of $\hat{\boldsymbol{\beta}}$ that minimizes $\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$ by differentiating $\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$ with respect to $\hat{\boldsymbol{\beta}}$ [using (2.112) and (2.113)] and setting the result equal to zero:

$$\frac{\partial \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{\partial \hat{\boldsymbol{\beta}}} = \mathbf{0} - 2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{0},$$

This gives the *normal equations*

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}. \qquad (7.8)$$

By Theorems 2.4(iii) and 2.6d(i) and Corollary 1 of Theorem 2.6c, if $\mathbf{X}$ is full-rank, $\mathbf{X}'\mathbf{X}$ is nonsingular, and the solution to (7.8) is given by (7.6). □

Since $\hat{\boldsymbol{\beta}}$ in (7.6) minimizes the sum of squares in (7.5), $\hat{\boldsymbol{\beta}}$ is called the *least-squares estimator*. Note that each $\hat{\beta}_j$ in $\hat{\boldsymbol{\beta}}$ is a linear function of $\mathbf{y}$; that is, $\hat{\beta}_j = \mathbf{a}_j'\mathbf{y}$, where $\mathbf{a}_j'$ is the jth row of $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. This usage of the word *linear* in *linear estimator* is different from that in *linear model*, which indicates that the model is linear in the β's.

We now show that $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ minimizes $\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$. Let $\mathbf{b}$ be an alternative estimator that may do better than $\hat{\boldsymbol{\beta}}$ so that $\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$ is

$$\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}).$$

Now adding and subtracting $\mathbf{X}\hat{\boldsymbol{\beta}}$, we obtain

$$= (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\mathbf{b}) \tag{7.9}$$

$$= (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) + (\hat{\boldsymbol{\beta}} - \mathbf{b})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}} - \mathbf{b}) + 2(\hat{\boldsymbol{\beta}} - \mathbf{b})'(\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}). \tag{7.10}$$

The third term on the right side of (7.10) vanishes because of the normal equations $\mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}$ in (7.8). The second term is a positive definite quadratic form (assuming that $\mathbf{X}$ is full-rank; see Theorem 2.6d), and $\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$ is therefore minimized when $\mathbf{b} = \hat{\boldsymbol{\beta}}$.

To examine the structure of $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}'\mathbf{y}$, note that by Theorem 2.2c(i), the $(k+1) \times (k+1)$ matrix $\mathbf{X}'\mathbf{X}$ can be obtained as products of columns of $\mathbf{X}$; similarly, $\mathbf{X}'\mathbf{y}$ contains products of columns of $\mathbf{X}$ and $\mathbf{y}$:

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} n & \sum_i x_{i1} & \sum_i x_{i2} & \cdots & \sum_i x_{ik} \\ \sum_i x_{i1} & \sum_i x_{i1}^2 & \sum_i x_{i1}x_{i2} & \cdots & \sum_i x_{i1}x_{ik} \\ \vdots & \vdots & \vdots & & \vdots \\ \sum x_{ik} & \sum_i x_{i1}x_{ik} & \sum_i x_{i2}x_{ik} & \cdots & \sum_i x_{ik}^2 \end{pmatrix},$$

$$\mathbf{X}'\mathbf{y} = \begin{pmatrix} \sum_i y_i \\ \sum_i x_{i1}y_i \\ \vdots \\ \sum_i x_{ik}y_i \end{pmatrix}.$$

If $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ as in (7.6), then

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{y} - \hat{\mathbf{y}} \tag{7.11}$$

is the vector of *residuals*, $\hat{\varepsilon}_1 = y_1 - \hat{y}_1$, $\hat{\varepsilon}_2 = y_2 - \hat{y}_2, \ldots, \hat{\varepsilon}_n = y_n - \hat{y}_n$. The residual vector $\hat{\boldsymbol{\varepsilon}}$ estimates $\boldsymbol{\varepsilon}$ in the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ and can be used to check the validity of the model and attendant assumptions; see Chapter 9.

Example 7.3.1a. We use the data in Table 7.1 to illustrate computation of $\hat{\boldsymbol{\beta}}$ using (7.6).

$$\mathbf{y} = \begin{pmatrix} 2 \\ 3 \\ 2 \\ 7 \\ 6 \\ 8 \\ 10 \\ 7 \\ 8 \\ 12 \\ 11 \\ 14 \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & 0 & 2 \\ 1 & 2 & 6 \\ 1 & 2 & 7 \\ 1 & 2 & 5 \\ 1 & 4 & 9 \\ 1 & 4 & 8 \\ 1 & 4 & 7 \\ 1 & 6 & 10 \\ 1 & 6 & 11 \\ 1 & 6 & 9 \\ 1 & 8 & 15 \\ 1 & 8 & 13 \end{pmatrix}, \quad \mathbf{X}'\mathbf{X} = \begin{pmatrix} 12 & 52 & 102 \\ 52 & 395 & 536 \\ 102 & 536 & 1004 \end{pmatrix},$$

$$\mathbf{X}'\mathbf{y} = \begin{pmatrix} 90 \\ 482 \\ 872 \end{pmatrix}, \quad (\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} .97476 & .24290 & -.22871 \\ .24290 & .16207 & -.11120 \\ -.22871 & -.11120 & .08360 \end{pmatrix},$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{pmatrix} 5.3754 \\ 3.0118 \\ -1.2855 \end{pmatrix}.$$

□

Example 7.3.1b. Simple linear regression from Chapter 6 can also be expressed in matrix terms:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix},$$

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} n & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{pmatrix}, \quad \mathbf{X}'\mathbf{y} = \begin{pmatrix} \sum_i y_i \\ \sum_i x_i y_i \end{pmatrix},$$

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{n\sum_i x_i^2 - (\sum_i x_i)^2} \begin{pmatrix} \sum_i x_i^2 & -\sum_i x_i \\ -\sum_i x_i & n \end{pmatrix}.$$

Then $\hat{\beta}_0$ and $\hat{\beta}_1$ can be obtained using (7.6), $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$:

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \frac{1}{n\sum_i x_i^2 - (\sum_i x_i)^2} \begin{pmatrix} \left(\sum_i x_i^2\right)\left(\sum y_i\right) - \left(\sum_i x_i\right)\left(\sum_i x_i y_i\right) \\ -\left(\sum_i x_i\right)\left(\sum y_i\right) + n\sum_i x_i y_i \end{pmatrix}. \tag{7.12}$$

The estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ in (7.11) are the same as those in (6.5) and (6.6). □

7.3.2 Properties of the Least-Squares Estimator $\hat{\boldsymbol{\beta}}$

The least-squares estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ in Theorem 7.3a was obtained without using the assumptions $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ and $\text{cov}(\mathbf{y}) = \sigma^2\mathbf{I}$ given in Section 7.2. We merely postulated a model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ as in (7.4) and fitted it. If $E(\mathbf{y}) \neq \mathbf{X}\boldsymbol{\beta}$, the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ could still be fitted to the data, in which case, $\hat{\boldsymbol{\beta}}$ may have poor properties. If $\text{cov}(\mathbf{y}) \neq \sigma^2\mathbf{I}$, there may be additional adverse effects on the estimator $\hat{\boldsymbol{\beta}}$. However, if $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ and $\text{cov}(\mathbf{y}) = \sigma^2\mathbf{I}$ hold, $\hat{\boldsymbol{\beta}}$ has some good properties, as noted in the four theorems in this section. Note that $\hat{\boldsymbol{\beta}}$ is a random vector (from sample to sample). We discuss its mean vector and covariance matrix in this section (with no distributional assumptions on $\mathbf{y}$) and its distribution (assuming that the y variables are normal) in Section 7.6.3. In the following theorems, we assume that $\mathbf{X}$ is fixed (remains constant in repeated sampling) and full rank.

Theorem 7.3b. If $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$, then $\hat{\boldsymbol{\beta}}$ is an unbiased estimator for $\boldsymbol{\beta}$.

Proof

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}) &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{y}) \qquad \text{[by (3.38)]} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \\ &= \boldsymbol{\beta}. \end{aligned} \tag{7.13}$$

□

Theorem 7.3c. If $\text{cov}(\mathbf{y}) = \sigma^2\mathbf{I}$, the covariance matrix for $\hat{\boldsymbol{\beta}}$ is given by $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

Proof

$$\begin{aligned} \text{cov}(\hat{\boldsymbol{\beta}}) &= \text{cov}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{cov}(\mathbf{y})[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']' \qquad \text{[by (3.44)]} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sigma^2\mathbf{I})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}. \end{aligned} \tag{7.14}$$

□

Example 7.3.2a. Using the matrix $(\mathbf{X}'\mathbf{X})^{-1}$ for simple linear regression given in Example 7.3.1, we obtain

$$\mathrm{cov}(\hat{\boldsymbol{\beta}}) = \mathrm{cov}\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} \mathrm{var}(\hat{\beta}_0) & \mathrm{cov}(\hat{\beta}_0, \hat{\beta}_1) \\ \mathrm{cov}(\hat{\beta}_0, \hat{\beta}_1) & \mathrm{var}(\hat{\beta}_1) \end{pmatrix} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

$$= \frac{\sigma^2}{n\sum_i x_i^2 - (\sum_i x_i)^2}\begin{pmatrix} \sum_i x_i^2 & -\sum_i x_i \\ -\sum_i x_i & n \end{pmatrix} \tag{7.15}$$

$$= \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}\begin{pmatrix} \sum_i x_i^2/n & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}. \tag{7.16}$$

Thus

$$\mathrm{var}(\hat{\beta}_0) = \frac{\sigma^2 \sum_i x_i^2/n}{\sum_i (x_i - \bar{x})^2}, \quad \mathrm{var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2},$$

$$\mathrm{cov}(\hat{\beta}_0, \hat{\beta}_1) = \frac{-\sigma^2\bar{x}}{\sum_i (x_i - \bar{x})^2}.$$

We found $\mathrm{var}(\hat{\beta}_0)$ and $\mathrm{var}(\hat{\beta}_1)$ in Section 6.2 but did not obtain $\mathrm{cov}(\hat{\beta}_0, \hat{\beta}_1)$. Note that if $\bar{x} > 0$, then $\mathrm{cov}(\hat{\beta}_0, \hat{\beta}_1)$ is negative and the estimated slope and intercept are negatively correlated. In this case, if the estimate of the slope increases from one sample to another, the estimate of the intercept tends to decrease (assuming the x's stay the same). □

Example 7.3.2b. For the data in Table 7.1, $(\mathbf{X}'\mathbf{X})^{-1}$ is as given in Example 7.3.1. Thus, $\mathrm{cov}(\hat{\boldsymbol{\beta}})$ is given by

$$\mathrm{cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2\begin{pmatrix} .975 & .243 & -.229 \\ .243 & .162 & -.111 \\ -.229 & -.111 & .084 \end{pmatrix}.$$

The negative value of $\mathrm{cov}(\hat{\beta}_1, \hat{\beta}_2) = -.111$ indicates that in repeated sampling (using the same 12 values of x_1 and x_2), $\hat{\beta}_1$ and $\hat{\beta}_2$ would tend to move in opposite directions; that is, an increase in one would be accompanied by a decrease in the other. □

In addition to $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ and $\mathrm{cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$, a third important property of $\hat{\boldsymbol{\beta}}$ is that under the standard assumptions, the variance of each $\hat{\beta}_j$ is minimum (see the following theorem).

Theorem 7.3d (Gauss–Markov Theorem). If $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ and $\mathrm{cov}(\mathbf{y}) = \sigma^2\mathbf{I}$, the least-squares estimators $\hat{\beta}_j, j = 0, 1, \ldots, k$, have minimum variance among all linear unbiased estimators.

PROOF. We consider a linear estimator $\mathbf{Ay}$ of $\boldsymbol{\beta}$ and seek the matrix $\mathbf{A}$ for which $\mathbf{Ay}$ is a minimum variance unbiased estimator of $\boldsymbol{\beta}$. In order for $\mathbf{Ay}$ to be an unbiased estimator of $\boldsymbol{\beta}$, we must have $E(\mathbf{Ay}) = \boldsymbol{\beta}$. Using the assumption $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$, this can be expressed as

$$E(\mathbf{Ay}) = \mathbf{A}E(\mathbf{y}) = \mathbf{AX}\boldsymbol{\beta} = \boldsymbol{\beta},$$

which gives the unbiasedness condition

$$\mathbf{AX} = \mathbf{I}$$

since the relationship $\mathbf{AX}\boldsymbol{\beta} = \boldsymbol{\beta}$ must hold for any possible value of $\boldsymbol{\beta}$ [see (2.44)].

The covariance matrix for the estimator $\mathbf{Ay}$ is given by

$$\text{cov}(\mathbf{Ay}) = \mathbf{A}(\sigma^2\mathbf{I})\mathbf{A}' = \sigma^2\mathbf{AA}'.$$

The variances of the $\hat{\beta}_j$'s are on the diagonal of $\sigma^2\mathbf{AA}'$, and we therefore need to choose $\mathbf{A}$ (subject to $\mathbf{AX} = \mathbf{I}$) so that the diagonal elements of $\mathbf{AA}'$ are minimized. To relate $\mathbf{Ay}$ to $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, we add and subtract $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ to obtain

$$\mathbf{AA}' = [\mathbf{A} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'][\mathbf{A} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']'.$$

Expanding this in terms of $\mathbf{A} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, we obtain four terms, two of which vanish because of the restriction $\mathbf{AX} = \mathbf{I}$. The result is

$$\mathbf{AA}' = [\mathbf{A} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'][\mathbf{A} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']' + (\mathbf{X}'\mathbf{X})^{-1}. \tag{7.17}$$

The matrix $[\mathbf{A} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'][\mathbf{A} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']'$ on the right side of (7.17) is positive semidefinite (see Theorem 2.6d), and, by Theorem 2.6a (ii), the diagonal elements are greater than or equal to zero. These diagonal elements can be made equal to zero by choosing $\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. (This value of $\mathbf{A}$ also satisfies the unbiasedness condition $\mathbf{AX} = \mathbf{I}$.) The resulting minimum variance estimator of $\boldsymbol{\beta}$ is

$$\mathbf{Ay} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y},$$

which is equal to the least–squares estimator $\hat{\boldsymbol{\beta}}$. □

The Gauss–Markov theorem is sometimes stated as follows. If $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ and $\text{cov}(\mathbf{y}) = \sigma^2\mathbf{I}$, the least-squares estimators $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ are *best linear unbiased estimators* (BLUE). In this expression, *best* means minimum variance and *linear* indicates that the estimators are linear functions of $\mathbf{y}$.

The remarkable feature of the Gauss–Markov theorem is its distributional generality. The result holds for any distribution of $\mathbf{y}$; normality is not required. The only assumptions used in the proof are $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ and $\text{cov}(\mathbf{y}) = \sigma^2\mathbf{I}$. If these assumptions do not hold, $\hat{\boldsymbol{\beta}}$ may be biased or each $\hat{\beta}_j$ may have a larger variance than that of some other estimator.

The Gauss–Markov theorem is easily extended to a linear combination of the $\hat{\beta}$'s, as follows.

Corollary 1. If $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ and $\text{cov}(\mathbf{y}) = \sigma^2\mathbf{I}$, the best linear unbiased estimator of $\mathbf{a}'\boldsymbol{\beta}$ is $\mathbf{a}'\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is the least–squares estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

Proof. See Problem 7.7. □

Note that Theorem 7.3d is concerned with the form of the estimator $\hat{\boldsymbol{\beta}}$ for a given $\mathbf{X}$ matrix. Once $\mathbf{X}$ is chosen, the variances of the $\hat{\beta}_j$'s are minimized by $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. However, in Theorem 7.3c, we have $\text{cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ and therefore $\text{var}(\hat{\beta}_j)$ and $\text{cov}(\hat{\beta}_i, \hat{\beta}_j)$ depend on the values of the x_j's. Thus the configuration of $\mathbf{X}'\mathbf{X}$ is important in estimation of the β_j's (this was illustrated in Problem 6.4).

In both estimation and testing, there are advantages to choosing the x's (or the centered x's) to be orthogonal so that $\mathbf{X}'\mathbf{X}$ is diagonal. These advantages include minimizing the variances of the $\hat{\beta}_j$'s and maximizing the power of tests about the β_j's (Chapter 8). For clarification, we note that orthogonality is necessary but not sufficient for minimizing variances and maximizing power. For example, if there are two x's, with values to be selected in a rectangular space, the points could be evenly placed on a grid, which would be an orthogonal pattern. However, the optimal orthogonal pattern would be to place one-fourth of the points at each corner of the rectangle.

A fourth property of $\hat{\boldsymbol{\beta}}$ is as follows. The predicted value $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k = \hat{\boldsymbol{\beta}}'\mathbf{x}$ is invariant to simple linear changes of scale on the x's, where $\mathbf{x} = (1, x_1, x_2, \ldots, x_k)'$. Let the rescaled variables be denoted by $z_j = c_j x_j$, $j = 1, 2, \ldots, k$, where the c_j terms are constants. Thus $\mathbf{x}$ is transformed to $\mathbf{z} = (1, c_1 x_1, \ldots, c_k x_k)'$. The following theorem shows that $\hat{y}$ based on $\mathbf{z}$ is the same as $\hat{y}$ based on $\mathbf{x}$.

Theorem 7.3e. If $\mathbf{x} = (1, x_1, \ldots, x_k)'$ and $\mathbf{z} = (1, c_1 x_1, \ldots, c_k x_k)'$, then $\hat{y} = \hat{\boldsymbol{\beta}}'\mathbf{x} = \hat{\boldsymbol{\beta}}_z'\mathbf{z}$, where $\hat{\boldsymbol{\beta}}_z$ is the least squares estimator from the regression of y on $\mathbf{z}$.

Proof. From (2.29), we can rewrite $\mathbf{z}$ as $\mathbf{z} = \mathbf{D}\mathbf{x}$, where $\mathbf{D} = \text{diag}(1, c_1, c_2, \ldots, c_k)$. Then, the $\mathbf{X}$ matrix is transformed to $\mathbf{Z} = \mathbf{X}\mathbf{D}$ [see (2.28)]. We substitute $\mathbf{Z} = \mathbf{X}\mathbf{D}$ in the least-squares estimator $\hat{\boldsymbol{\beta}}_z = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}$ to obtain

$$\begin{aligned} \hat{\boldsymbol{\beta}}_z &= (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y} = [(\mathbf{X}\mathbf{D})'(\mathbf{X}\mathbf{D})]^{-1}(\mathbf{X}\mathbf{D})'\mathbf{y} \\ &= \mathbf{D}^{-1}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \qquad \text{[by (2.49)]} \\ &= \mathbf{D}^{-1}\hat{\boldsymbol{\beta}}, \end{aligned} \tag{7.18}$$

where $\hat{\boldsymbol{\beta}}$ is the usual estimator for y regressed on the x's. Then

$$\hat{\boldsymbol{\beta}}_z'\mathbf{z} = (\mathbf{D}^{-1}\hat{\boldsymbol{\beta}})'\mathbf{D}\mathbf{x} = \hat{\boldsymbol{\beta}}'\mathbf{x}.$$

□

In the following corollary to Theorem 7.3e, the invariance of $\hat{y}$ is extended to any full-rank linear transformation of the x variables.

Corollary 1. The predicted value $\hat{y}$ is invariant to a full-rank linear transformation on the x's.

Proof. We can express a full-rank linear transformation of the x's as

$$\mathbf{Z} = \mathbf{XK} = (\mathbf{j}, \mathbf{X}_1)\begin{pmatrix} 1 & \mathbf{0}' \\ \mathbf{0} & \mathbf{K}_1 \end{pmatrix} = (\mathbf{j} + \mathbf{X}_1\mathbf{0}, \mathbf{j}\mathbf{0}' + \mathbf{X}_1\mathbf{K}_1) = (\mathbf{j}, \mathbf{X}_1\mathbf{K}_1),$$

where $\mathbf{K}_1$ is nonsingular and

$$\mathbf{X}_1 = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix}. \tag{7.19}$$

We partition $\mathbf{X}$ and $\mathbf{K}$ in this way so as to transform only the x's in $\mathbf{X}_1$, leaving the first column of $\mathbf{X}$ unaffected. Now $\hat{\boldsymbol{\beta}}_z$ becomes

$$\hat{\boldsymbol{\beta}}_z = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y} = \mathbf{K}^{-1}\hat{\boldsymbol{\beta}}, \tag{7.20}$$

and we have

$$\hat{y} = \hat{\boldsymbol{\beta}}_z'\mathbf{z} = \hat{\boldsymbol{\beta}}'\mathbf{x}, \tag{7.21}$$

where $\mathbf{z} = \mathbf{K}'\mathbf{x}$. □

In addition to $\hat{y}$, the sample variance s^2 (Section 7.3.3) is also invariant to changes of scale on the x variable (see Problem 7.10). The following are invariant to changes of scale on y as well as on the x's (but not to a joint linear transformation on y and the x's): t statistics (Section 8.5), F statistics (Chapter 8), and R^2 (Sections 7.7 and 10.3).

7.3.3 An Estimator for σ^2

The method of least squares does not yield a function of the y and x values in the sample that we can minimize to obtain an estimator of σ^2. However, we can devise an unbiased estimator for σ^2 based on the least-squares estimator $\hat{\boldsymbol{\beta}}$. By assumption 2 following (7.3), σ^2 is the same for each y_i, $i = 1, 2, \dots, n$. By (3.6), σ^2 is defined by $\sigma^2 = E[y_i - E(y_i)]^2$, and by assumption 1, we obtain

$$E(y_i) = \beta_0 + \beta_i x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} = \mathbf{x}_i'\boldsymbol{\beta},$$

where $\mathbf{x}_i'$ is the ith row of $\mathbf{X}$. Thus σ^2 becomes

$$\sigma^2 = E[y_i - \mathbf{x}_i'\boldsymbol{\beta}]^2.$$

We estimate σ^2 by a corresponding average from the sample

$$s^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}})^2, \tag{7.22}$$

where n is the sample size and k is the number of x's. Note that, by the corollary to Theorem 7.3d, $\mathbf{x}_i'\hat{\boldsymbol{\beta}}$ is the BLUE of $\mathbf{x}_i'\boldsymbol{\beta}$.

Using (7.7), we can write (7.22) as

$$s^2 = \frac{1}{n-k-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \tag{7.23}$$

$$= \frac{\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}}{n-k-1} = \frac{\text{SSE}}{n-k-1}, \tag{7.24}$$

where SSE $= (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}$. With the denominator $n-k-1$, s^2 is an unbiased estimator of σ^2, as shown below.

Theorem 7.3f. If s^2 is defined by (7.22), (7.23), or (7.24) and if $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ and $\text{cov}(\mathbf{y}) = \sigma^2\mathbf{I}$, then

$$E(s^2) = \sigma^2. \tag{7.25}$$

Proof. Using (7.24) and (7.6), we write SSE as a quadratic form:

$$\begin{aligned} \text{SSE} &= \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} = \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= \mathbf{y}'\left[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]\mathbf{y}. \end{aligned} \tag{7.26}$$

By Theorem 5.2a, we have

$$\begin{aligned} E(\text{SSE}) &= \text{tr}\left\{\left[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]\sigma^2\mathbf{I}\right\} \\ &\quad + E(\mathbf{y}')\left[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]E(\mathbf{y}) \\ &= \sigma^2\text{tr}\left[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right] \\ &\quad + \boldsymbol{\beta}'\mathbf{X}'\left[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]\mathbf{X}\boldsymbol{\beta} \\ &= \sigma^2\left\{n - \text{tr}\left[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]\right\} \\ &\quad + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \\ &= \sigma^2\left\{n - \text{tr}[\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}]\right\} \\ &\quad + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \qquad \text{[by (2.87)].} \end{aligned}$$

Since $\mathbf{X}'\mathbf{X}$ is $(k+1) \times (k+1)$, this becomes

$$E(\text{SSE}) = \sigma^2[n - \text{tr}(\mathbf{I}_{k+1})] = \sigma^2(n-k-1).$$

□

Corollary 1. An unbiased estimator of $\text{cov}(\hat{\boldsymbol{\beta}})$ in (7.14) is given by

$$\widehat{\text{cov}}(\hat{\boldsymbol{\beta}}) = s^2(\mathbf{X}'\mathbf{X})^{-1}. \tag{7.27}$$

□

Note the correspondence between $n - (k+1)$ and $\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}$; there are n terms in $\mathbf{y}'\mathbf{y}$ and $k+1$ terms in $\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}$ [see (7.8)]. A corresponding property of the sample is that each additional x (and $\hat{\beta}$) in the model reduces SSE (see Problem 7.13).

Since SSE is a quadratic function of $\mathbf{y}$, it is not a best *linear* unbiased estimator. The optimality property of s^2 is given in the following theorem.

Theorem 7.3g. If $E(\boldsymbol{\varepsilon}) = \mathbf{0}$, $\text{cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$, and $E(\varepsilon_i^4) = 3\sigma^4$ for the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, then s^2 in (7.23) or (7.24) is the best (minimum variance) *quadratic* unbiased estimator of σ^2.

Proof. See Graybill (1954), Graybill and Wortham (1956), or Wang and Chow (1994, pp. 161–163). □

Example 7.3.3. For the data in Table 7.1, we have

$$\begin{aligned}
\text{SSE} &= \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}\mathbf{y} \\
&= 840 - (5.3754,\ 3.0118,\ -1.2855)\begin{pmatrix} 90 \\ 482 \\ 872 \end{pmatrix} \\
&= 840 - 814.541 = 25.459, \\
s^2 &= \frac{\text{SSE}}{n-k-1} = \frac{25.459}{12-2-1} = 2.829.
\end{aligned}$$

□

7.4 GEOMETRY OF LEAST SQUARES

In Sections 7.1–7.3 we presented the multiple linear regression model as the matrix equation $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ in (7.4). We defined the principle of least-squares estimation in terms of deviations from the model [see (7.7)], and then used matrix calculus and matrix algebra to derive the estimators of $\boldsymbol{\beta}$ in (7.6) and of σ^2 in (7.23) and (7.24). We now present an alternate but equivalent derivation of these estimators based completely on geometric ideas.

It is important to clarify first what the geometric approach to least squares is *not*. In two dimensions, we illustrated the principle of least squares by creating a two-dimensional scatter plot (Fig. 6.1) of the n points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. We then visualized the least-squares regression line as the best-fitting straight line to the data. This approach can be generalized to present the least-squares estimate in multiple linear regression on the basis of the best-fitting hyperplane in $(k + 1)$-dimensional space to the n points $(x_{11}, x_{12}, \ldots, x_{1k}, y_1), (x_{21}, x_{22}, \ldots, x_{2k}, y_2), \ldots, (x_{n1}, x_{n2}, \ldots, x_{nk}, y_n)$. Although this approach is somewhat useful in visualizing multiple linear regression, the geometric approach to least-squares estimation in multiple linear regression does *not* involve this high-dimensional generalization.

The geometric approach to be discussed below is appealing because of its mathematical elegance. For example, the estimator is derived without the use of matrix calculus. Also, the geometric approach provides deeper insight into statistical inference. Several advanced statistical methods including kernel smoothing (Eubank and Eubank 1999), Fourier analysis (Bloomfield 2000), and wavelet analysis (Ogden 1997) can be understood as generalizations of this geometric approach. The geometric approach to linear models was first proposed by Fisher (Mahalanobis 1964). Christensen (1996) and Jammalamadaka and Sengupta (2003) discuss the linear statistical model almost completely from the geometric perspective.

7.4.1 Parameter Space, Data Space, and Prediction Space

The geometric approach to least squares begins with two high-dimensional spaces, a $(k + 1)$-dimensional space and an n-dimensional space. The unknown parameter vector $\boldsymbol{\beta}$ can be viewed as a single point in $(k + 1)$-dimensional space, with axes corresponding to the $k + 1$ regression coefficients $\beta_0, \beta_1, \beta_0, \ldots, \beta_k$. Hence we call this space the *parameter space* (Fig. 7.3). Similarly, the data vector $\mathbf{y}$ can be viewed as a

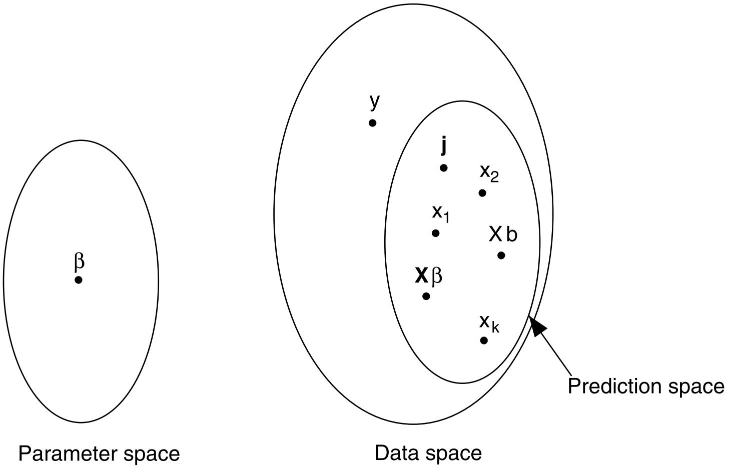

Figure 7.3 Parameter space, data space, and prediction space with representative elements.

single point in n-dimensional space with axes corresponding to the n observations. We call this space the *data space*.

The $\mathbf{X}$ matrix of the multiple regression model (7.4) can be written as a partitioned matrix in terms of its $k+1$ columns as

$$\mathbf{X} = (\mathbf{j}, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_k).$$

The columns of $\mathbf{X}$, including $\mathbf{j}$, are all n-dimensional vectors and are therefore points in the data space. Note that because we assumed that $\mathbf{X}$ is of rank $k+1$, these vectors are linearly independent. The set of all possible linear combinations of the columns of $\mathbf{X}$ (Section 2.3) constitutes a subset of the data space. Elements of this subset can be written as

$$\mathbf{Xb} = b_0\mathbf{j} + b_1\mathbf{x}_1 + b_2\mathbf{x}_2 + \cdots + b_k\mathbf{x}_k, \tag{7.28}$$

where $\mathbf{b}$ is any $k+1$ vector, that is, any vector in the parameter space. This subset actually has the status of a *subspace* because it is closed under addition and scalar multiplication (Harville 1997, pp. 28–29). This subset is said to be the subspace generated or *spanned* by the columns of $\mathbf{X}$, and we will call this subspace the *prediction space*. The columns of $\mathbf{X}$ constitute a *basis set* for the prediction space.

7.4.2 Geometric Interpretation of the Multiple Linear Regression Model

The multiple linear regression model (7.4) states that $\mathbf{y}$ is equal to a vector in the prediction space, $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$, plus a vector of random errors, $\boldsymbol{\varepsilon}$ (Fig. 7.4). The

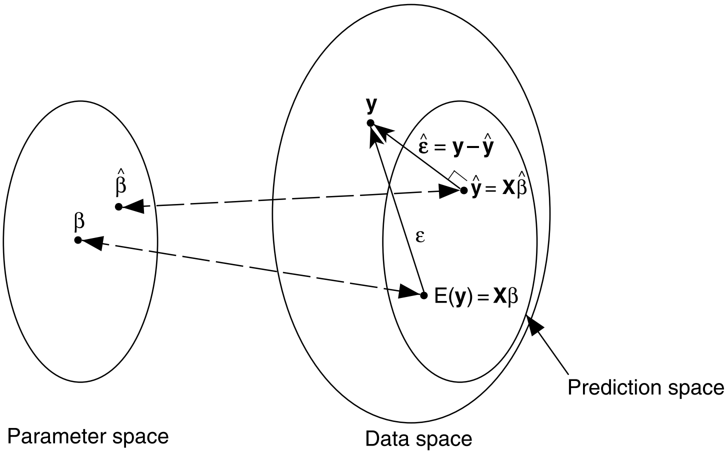
y
$\hat{\varepsilon} = \mathbf{y} - \hat{\mathbf{y}}$
$\hat{\beta}$
$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$
β
ε
$E(\mathbf{y}) = \mathbf{X}\beta$
Prediction space
Parameter space
Data space

Figure 7.4 Geometric relationships of vectors associated with the multiple linear regression model.

problem is that neither $\boldsymbol{\beta}$ nor $\boldsymbol{\varepsilon}$ is known. However, the data vector $\mathbf{y}$, which is not in the prediction space, is known. And it is known that $E(\mathbf{y})$ is in the prediction space.

Multiple linear regression can be understood geometrically as the process of finding a sensible estimate of $E(\mathbf{y})$ in the prediction space and then determining the vector in the parameter space that is associated with this estimate (Fig. 7.4). The estimate of $E(\mathbf{y})$ is denoted as $\hat{\mathbf{y}}$, and the associated vector in the parameter space is denoted as $\hat{\boldsymbol{\beta}}$.

A reasonable geometric idea is to estimate $E(\mathbf{y})$ using the point in the prediction space that is closest to $\mathbf{y}$. It turns out that $\hat{\mathbf{y}}$, the closest point in the prediction space to $\mathbf{y}$, can be found by noting that the difference vector $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \hat{\mathbf{y}}$ must be orthogonal (perpendicular) to the prediction space (Harville 1997, p. 170). Furthermore, because the prediction space is spanned by the columns of $\mathbf{X}$, the point $\hat{\mathbf{y}}$ must be such that $\hat{\boldsymbol{\varepsilon}}$ is orthogonal to the columns of $\mathbf{X}$. Using an extension of (2.80), we therefore seek $\hat{\mathbf{y}}$ such that

$$\mathbf{X}'\hat{\boldsymbol{\varepsilon}} = \mathbf{0}$$

or

$$\mathbf{X}'(\mathbf{y} - \hat{\mathbf{y}}) = \mathbf{X}'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{0}, \tag{7.29}$$

which implies that

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

Thus, using purely geometric ideas, we obtain the normal equations (7.8) and consequently the usual least-squares estimator $\hat{\boldsymbol{\beta}}$ in (7.6). We can then calculate $\hat{\mathbf{y}}$ as $\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}$. Also, $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = (\mathbf{I} - \mathbf{H})\mathbf{y}$ can be taken as an estimate of $\boldsymbol{\varepsilon}$. Since $\hat{\boldsymbol{\varepsilon}}$ is a vector in $(n - k - 1)$-dimensional space, it seems reasonable to estimate σ^2 as the squared length (2.22) of $\hat{\boldsymbol{\varepsilon}}$ divided by $n - k - 1$. In other words, a sensible estimator of σ^2 is $s^2 = \mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y}/(n - k - 1)$, which is equal to (7.25).

7.5 THE MODEL IN CENTERED FORM

The model in (7.3) for each y_i can be written in terms of centered x variables as

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \\ &= \alpha + \beta_1(x_{i1} - \bar{x}_1) + \beta_2(x_{i2} - \bar{x}_2) + \cdots + \beta_k(x_{ik} - \bar{x}_k) + \varepsilon_i, \end{aligned} \tag{7.30}$$

$i = 1, 2, \ldots, n$, where

$$\alpha = \beta_0 + \beta_1\bar{x}_1 + \beta_2\bar{x}_2 + \ldots + \beta_k\bar{x}_k \tag{7.31}$$

and $\bar{x}_j = \sum_{i=1}^{n} x_{ij}/n, j = 1, 2, \ldots, k$. The centered form of the model is useful in expressing certain hypothesis tests (Section 8.1), in a search for influential observations (Section 9.2), and in providing other insights.

In matrix form, the centered model (7.30) for $y_1, y_2, \ldots, y_n$ becomes

$$\mathbf{y} = (\mathbf{j}, \mathbf{X}_c)\begin{pmatrix} \alpha \\ \boldsymbol{\beta}_1 \end{pmatrix} + \boldsymbol{\varepsilon}, \tag{7.32}$$

where $\boldsymbol{\beta}_1 = (\beta_1, \beta_2, \ldots, \beta_k)'$,

$$\mathbf{X}_c = \left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right)\mathbf{X}_1 = \begin{pmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \ldots & x_{1k} - \bar{x}_k \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \ldots & x_{2k} - \bar{x}_k \\ \vdots & \vdots & & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \ldots & x_{nk} - \bar{x}_k \end{pmatrix}, \tag{7.33}$$

and $\mathbf{X}_1$ is as given in (7.19). The matrix $\mathbf{I} - (1/n)\mathbf{J}$ is sometimes called the *centering matrix*.

As in (7.8), the normal equations for the model in (7.32) are

$$(\mathbf{j}, \mathbf{X}_c)'(\mathbf{j}, \mathbf{X}_c)\begin{pmatrix} \hat{\alpha} \\ \hat{\boldsymbol{\beta}}_1 \end{pmatrix} = (\mathbf{j}, \mathbf{X}_c)'\mathbf{y}. \tag{7.34}$$

By (2.35) and (2.39), the product $(\mathbf{j}, \mathbf{X}_c)'(\mathbf{j}, \mathbf{X}_c)$ on the left side of (7.34) becomes

$$(\mathbf{j}, \mathbf{X}_c)'(\mathbf{j}, \mathbf{X}_c) = \begin{pmatrix} \mathbf{j}' \\ \mathbf{X}_c' \end{pmatrix}(\mathbf{j}, \mathbf{X}_c) = \begin{pmatrix} \mathbf{j}'\mathbf{j} & \mathbf{j}'\mathbf{X}_c \\ \mathbf{X}_c'\mathbf{j} & \mathbf{X}_c'\mathbf{X}_c \end{pmatrix}$$

$$= \begin{pmatrix} n & \mathbf{0}' \\ \mathbf{0} & \mathbf{X}_c'\mathbf{X}_c \end{pmatrix}, \tag{7.35}$$

where $\mathbf{j}'\mathbf{X}_c = \mathbf{0}'$ because the columns of $\mathbf{X}_c$ sum to zero (Problem 7.16). The right side of (7.34) can be written as

$$(\mathbf{j}, \mathbf{X}_c)'\mathbf{y} = \begin{pmatrix} \mathbf{j}' \\ \mathbf{X}_c' \end{pmatrix}\mathbf{y} = \begin{pmatrix} n\bar{y} \\ \mathbf{X}_c'\mathbf{y} \end{pmatrix}.$$

The least-squares estimators are then given by

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\boldsymbol{\beta}}_1 \end{pmatrix} = [(\mathbf{j}, \mathbf{X}_c)'(\mathbf{j}, \mathbf{X}_c)]^{-1}(\mathbf{j}, \mathbf{X}_c)'\mathbf{y} = \begin{pmatrix} n & \mathbf{0}' \\ \mathbf{0} & \mathbf{X}_c'\mathbf{X}_c \end{pmatrix}^{-1}\begin{pmatrix} n\bar{y} \\ \mathbf{X}_c'\mathbf{y} \end{pmatrix}$$

$$= \begin{pmatrix} 1/n & \mathbf{0}' \\ \mathbf{0} & (\mathbf{X}_c'\mathbf{X}_c)^{-1} \end{pmatrix}\begin{pmatrix} n\bar{y} \\ \mathbf{X}_c'\mathbf{y} \end{pmatrix} = \begin{pmatrix} \bar{y} \\ (\mathbf{X}_c'\mathbf{X}_c)^{-1}\mathbf{X}_c'\mathbf{y} \end{pmatrix},$$

or

$$\hat{\alpha} = \bar{y}, \tag{7.36}$$

$$\hat{\boldsymbol{\beta}}_1 = (\mathbf{X}_c'\mathbf{X}_c)^{-1}\mathbf{X}_c'\mathbf{y}. \tag{7.37}$$

These estimators are the same as the usual least-squares estimators $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ in (7.6), with the adjustment

$$\hat{\beta}_0 = \hat{\alpha} - \hat{\beta}_1\bar{x}_1 - \hat{\beta}_2\bar{x} - \cdots - \hat{\beta}_k\bar{x}_k = \bar{y} - \hat{\boldsymbol{\beta}}_1'\bar{\mathbf{x}} \tag{7.38}$$

obtained from an estimator of α in (7.31) (see Problem 7.17).

When we express $\hat{\mathbf{y}}$ in centered form

$$\hat{\mathbf{y}} = \hat{\alpha} + \hat{\beta}_1(x_1 - \bar{x}_1) + \cdots + \hat{\beta}_k(x_k - \bar{x}_k),$$

it is clear that the fitted regression plane passes through the point $(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k, \bar{y})$.

Adapting the expression for SSE (7.24) to the centered model with centered $\hat{y}$'s, we obtain

$$\mathrm{SSE} = \sum_{i=1}^{n}(y_i - \bar{y})^2 - \hat{\boldsymbol{\beta}}_1'\mathbf{X}_c'\mathbf{y}, \tag{7.39}$$

which turns out to be equal to $\mathrm{SSE} = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}$ (see Problem 7.19).

We can use (7.36)–(7.38) to express $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_0$ in terms of sample variances and covariances, which will be useful in comparing these estimators with those for the random-x case in Chapter 10. We first define a sample covariance matrix for the x variables and a vector of sample covariances between y and the x's

$$\mathbf{S}_{xx} = \begin{pmatrix} s_1^2 & s_{12} & \cdots & s_{1k} \\ s_{21} & s_2^2 & \cdots & s_{2k} \\ \vdots & \vdots & & \vdots \\ s_{k1} & s_{k2} & \cdots & s_k^2 \end{pmatrix}, \quad \mathbf{s}_{yx} = \begin{pmatrix} s_{y1} \\ s_{y2} \\ \vdots \\ s_{yk} \end{pmatrix}, \tag{7.40}$$

where, s_i^2, s_{ij}, and s_{yi} are analogous to s^2 and s_{xy} defined in (5.6) and (5.15); for example

$$s_2^2 = \frac{\sum_{i=1}^{n}(x_{i2} - \bar{x}_2)^2}{n-1}, \tag{7.41}$$

$$s_{12} = \frac{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)}{n-1}, \tag{7.42}$$

$$s_{y2} = \frac{\sum_{i=1}^{n}(x_{i2} - \bar{x}_2)(y_i - \bar{y})}{n-1}, \tag{7.43}$$

with $\bar{x}_2 = \sum_{i=1}^{n} x_{i2}/n$. However, since the x's are fixed, these sample variances and covariances do not estimate population variances and covariances. If the x's were random variables, as in Chapter 10, the s_i^2, s_{ij}, and s_{yi} values would estimate population parameters.

To express $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_0$ in terms of $\mathbf{S}_{xx}$ and $\mathbf{s}_{yx}$, we first write $\mathbf{S}_{xx}$ and $\mathbf{s}_{yx}$ in terms of the centered matrix $\mathbf{X}_c$:

$$\mathbf{S}_{xx} = \frac{\mathbf{X}_c'\mathbf{X}_c}{n-1}, \tag{7.44}$$

$$\mathbf{s}_{yx} = \frac{\mathbf{X}_c'\mathbf{y}}{n-1}. \tag{7.45}$$

Note that $\mathbf{X}_c'\mathbf{y}$ in (7.45) contains terms of the form $\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)y_i$ rather than $\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)(y_i-\bar{y})$ as in (7.43). It can readily be shown that $\sum_i(x_{ij}-\bar{x}_j)(y_i-\bar{y}) = \sum_i(x_{ij}-\bar{x}_j)y_i$ (see Problem 6.2).

From (7.37), (7.44), and (7.45), we have

$$\hat{\boldsymbol{\beta}}_1 = (n-1)(\mathbf{X}_c'\mathbf{X}_c)^{-1}\frac{\mathbf{X}_c'\mathbf{y}}{n-1} = \left(\frac{\mathbf{X}_c'\mathbf{X}_c}{n-1}\right)^{-1}\frac{\mathbf{X}_c'\mathbf{y}}{n-1} = \mathbf{S}_{xx}^{-1}\mathbf{s}_{yx}, \tag{7.46}$$

and from (7.38) and (7.46), we obtain

$$\hat{\beta}_0 = \hat{\alpha} - \hat{\boldsymbol{\beta}}_1'\bar{\mathbf{x}} = \bar{y} - \mathbf{s}_{yx}'\mathbf{S}_{xx}^{-1}\bar{\mathbf{x}}. \tag{7.47}$$

Example 7.5. For the data in Table 7.1, we calculate $\hat{\boldsymbol{\beta}}_1$ and $\hat{\beta}_0$ using (7.46) and (7.47).

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_1 &= \mathbf{S}_{xx}^{-1}\mathbf{s}_{yx} = \begin{pmatrix} 6.4242 & 8.5455 \\ 8.5455 & 12.4545 \end{pmatrix}^{-1}\begin{pmatrix} 8.3636 \\ 9.7273 \end{pmatrix} \\
&= \begin{pmatrix} 3.0118 \\ -1.2855 \end{pmatrix}, \\
\hat{\beta}_0 &= \bar{y} - \mathbf{s}_{yx}'\mathbf{S}_{xx}^{-1}\bar{\mathbf{x}} \\
&= 7.5000 - (3.0118,\ -1.2855)\begin{pmatrix} 4.3333 \\ 8.5000 \end{pmatrix} \\
&= 7.500 - 2.1246 = 5.3754.
\end{aligned}$$

These values are the same as those obtained in Example 7.3.1a. □

7.6 NORMAL MODEL

7.6.1 Assumptions

Thus far we have made no normality assumptions about the random variables $y_1, y_2, \ldots, y_n$. To the assumptions in Section 7.2, we now add that

$$\mathbf{y} \text{ is } N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}) \quad \text{or} \quad \boldsymbol{\varepsilon} \text{ is } N_n(\mathbf{0}, \sigma^2\mathbf{I}).$$

Under normality, $\sigma_{ij} = 0$ implies that the y (or ε) variables are independent, as well as uncorrelated.

7.6.2 Maximum Likelihood Estimators for $\boldsymbol{\beta}$ and σ^2

With the normality assumption, we can obtain maximum likelihood estimators. The likelihood function is the joint density of the y's, which we denote by $L(\boldsymbol{\beta}, \sigma^2)$. We seek values of the unknown $\boldsymbol{\beta}$ and σ^2 that maximize $L(\boldsymbol{\beta}, \sigma^2)$ for the given y and x values in the sample.

In the case of the normal density function, it is possible to find maximum likelihood estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ by differentiation. Because the normal density involves a product and an exponential, it is simpler to work with $\ln L(\boldsymbol{\beta}, \sigma^2)$, which achieves its maximum for the same values of $\boldsymbol{\beta}$ and σ^2 as does $L(\boldsymbol{\beta}, \sigma^2)$.

The maximum likelihood estimators for $\boldsymbol{\beta}$ and σ^2 are given in the following theorem.

Theorem 7.6a. If $\mathbf{y}$ is $N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, where $\mathbf{X}$ is $n \times (k+1)$ of rank $k+1 < n$, the maximum likelihood estimators of $\boldsymbol{\beta}$ and σ^2 are

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}, \qquad (7.48)$$

$$\hat{\sigma}^2 = \frac{1}{n}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}). \qquad (7.49)$$

PROOF. We sketch the proof. For the remaining steps, see Problem 7.21. The likelihood function (joint density of $y_1, y_2, \ldots, y_n$) is given by the multivariate normal density (4.9)

$$\begin{aligned} L(\boldsymbol{\beta}, \sigma^2) = f(\mathbf{y}; \boldsymbol{\beta}, \sigma^2) &= \frac{1}{(2\pi)^{n/2}|\sigma^2\mathbf{I}|^{1/2}} e^{-(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\sigma^2\mathbf{I})^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})/2} \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})/2\sigma^2}. \end{aligned} \qquad (7.50)$$

[Since the y_i's are independent, $L(\boldsymbol{\beta}, \sigma^2)$ can also be obtained as $\prod_{i=1}^n f(y_i; \mathbf{x}_i'\boldsymbol{\beta}, \sigma^2)$.] Then $\ln L(\boldsymbol{\beta}, \sigma^2)$ becomes

$$\ln L(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \qquad (7.51)$$

Taking the partial derivatives of $\ln L(\boldsymbol{\beta}, \sigma^2)$ with respect to $\boldsymbol{\beta}$ and σ^2 and setting the results equal to zero will produce (7.48) and (7.49). To verify that $\hat{\boldsymbol{\beta}}$ maximizes (7.50) or (7.51), see (7.10). □

The maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ in (7.48) is the same as the least-squares estimator $\hat{\boldsymbol{\beta}}$ in Theorem 7.3a. The estimator $\hat{\sigma}^2$ in (7.49) is biased since the denominator is n rather than $n - k - 1$. We often use the unbiased estimator s^2 given in (7.23) or (7.24).

7.6.3 Properties of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\sigma}}^2$

We now consider some properties of $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ (or s^2) under the normal model. The distributions of $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are given in the following theorem.

Theorem 7.6b. Suppose that $\mathbf{y}$ is $N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, where $\mathbf{X}$ is $n \times (k+1)$ of rank $k + 1 < n$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)'$. Then the maximum likelihood estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ given in Theorem 7.6a have the following distributional properties:

(i) $\hat{\boldsymbol{\beta}}$ is $N_{k+1}[\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}]$.
(ii) $n\hat{\sigma}^2/\sigma^2$ is $\chi^2(n-k-1)$, or equivalently, $(n-k-1)s^2/\sigma^2$ is $\chi^2(n-k-1)$.
(iii) $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ (or s^2) are independent.

Proof

(i) Since $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ is a linear function of $\mathbf{y}$ of the form $\hat{\boldsymbol{\beta}} = \mathbf{Ay}$, where $\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is a constant matrix, then by Theorem 4.4a(ii), $\hat{\boldsymbol{\beta}}$ is $N_{k+1}[\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}]$.
(ii) The result follows from Corollary 2 to Theorem 5.5.
(iii) The result follows from Corollary 1 to Theorem 5.6a.

□

Another property of $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ under normality is that they are sufficient statistics. Intuitively, a statistic is sufficient for a parameter if the statistic summarizes all the information in the sample about the parameter. Sufficiency of $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ can be established by the Neyman factorization theorem [see Hogg and Craig (1995, p. 318) or Graybill (1976, pp. 69–70)], which states that $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are jointly sufficient for $\boldsymbol{\beta}$ and σ^2 if the density $f(\mathbf{y}; \boldsymbol{\beta}, \sigma^2)$ can be factored as $f(\mathbf{y}; \boldsymbol{\beta}, \sigma^2) = g(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2, \boldsymbol{\beta}, \sigma^2)h(\mathbf{y})$, where $h(\mathbf{y})$ does not depend on $\boldsymbol{\beta}$ or σ^2. The following theorem shows that $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ satisfy this criterion.

Theorem 7.6c. If $\mathbf{y}$ is $N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, then $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are jointly sufficient for $\boldsymbol{\beta}$ and σ^2.

Proof. The density $f(\mathbf{y}; \boldsymbol{\beta}, \sigma^2)$ is given in (7.50). In the exponent, we add and subtract $\mathbf{X}\hat{\boldsymbol{\beta}}$ to obtain

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

Expanding this in terms of $\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ and $\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$, we obtain four terms, two of which vanish because of the normal equations $\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$. The result is

$$\begin{aligned}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) &= (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \qquad (7.52)\\ &= n\hat{\sigma}^2 + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}).\end{aligned}$$

We can now write the density (7.50) as

$$f(\mathbf{y}; \boldsymbol{\beta}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-[n\hat{\sigma}^2 + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})]/2\sigma^2},$$

which is of the form

$$f(\mathbf{y}; \boldsymbol{\beta}, \sigma^2) = g(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2, \boldsymbol{\beta}, \sigma^2)h(\mathbf{y}),$$

where $h(\mathbf{y}) = 1$. Therefore, by the Neyman factorization theorem, $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are jointly sufficient for $\boldsymbol{\beta}$ and σ^2. □

Note that $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are jointly sufficient for $\boldsymbol{\beta}$ and σ^2, not independently sufficient; that is, $f(\mathbf{y}; \boldsymbol{\beta}, \sigma^2)$ does not factor into the form $g_1(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})g_2(\hat{\sigma}^2, \sigma^2)h(\mathbf{y})$. Also note that because $s^2 = n\hat{\sigma}^2/(n - k - 1)$, the proof to Theorem 7.6c can be easily modified to show that $\hat{\boldsymbol{\beta}}$ and s^2 are also jointly sufficient for $\boldsymbol{\beta}$ and σ^2.

Since $\hat{\boldsymbol{\beta}}$ and s^2 are sufficient, no other estimators can improve on the information they extract from the sample to estimate $\boldsymbol{\beta}$ and σ^2. Thus, it is not surprising that $\hat{\boldsymbol{\beta}}$ and s^2 are minimum variance unbiased estimators (each $\hat{\beta}_j$ in $\hat{\boldsymbol{\beta}}$ has minimum variance). This result is given in the following theorem.

Theorem 7.6d. If $\mathbf{y}$ is $N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, then $\hat{\boldsymbol{\beta}}$ and s^2 have minimum variance among all unbiased estimators.

Proof. See Graybill (1976, p. 176) or Christensen (1996, pp. 25–27). □

In Theorem 7.3d, the elements of $\hat{\boldsymbol{\beta}}$ were shown to have minimum variance among all *linear unbiased* estimators. With the normality assumption added in Theorem 7.6d, the elements of $\hat{\boldsymbol{\beta}}$ have minimum variance among all *unbiased* estimators. Similarly, by Theorem 7.3g, s^2 has minimum variance among all *quadratic unbiased* estimators. With the added normality assumption in Theorem 7.6d, s^2 has minimum variance among all *unbiased* estimators.

The following corollary to Theorem 7.6d is analogous to Corollary 1 of Theorem 7.3d.

Corollary 1. If $\mathbf{y}$ is $N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, then the minimum variance unbiased estimator of $\mathbf{a}'\boldsymbol{\beta}$ is $\mathbf{a}'\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is the maximum likelihood estimator given in (7.48). □

7.7 R^2 IN FIXED-x REGRESSION

In (7.39), we have $\text{SSE} = \sum_{i=1}^{n}(y_i - \bar{y})^2 - \hat{\boldsymbol{\beta}}_1'\mathbf{X}_c'\mathbf{y}$. Thus the corrected total sum of squares $\text{SST} = \sum_i (y_i - \bar{y})^2$ can be partitioned as

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \hat{\boldsymbol{\beta}}_1'\mathbf{X}_c'\mathbf{y} + \text{SSE}, \tag{7.53}$$

$$\text{SST} = \text{SSR} + \text{SSE},$$

where $\text{SSR} = \hat{\boldsymbol{\beta}}_1'\mathbf{X}_c'\mathbf{y}$ is the *regression sum of squares.* From (7.37), we obtain $\mathbf{X}_c'\mathbf{y} = \mathbf{X}_c'\mathbf{X}_c\hat{\boldsymbol{\beta}}_1$, and multiplying this by $\hat{\boldsymbol{\beta}}_1'$ gives $\hat{\boldsymbol{\beta}}_1'\mathbf{X}_c'\mathbf{y} = \hat{\boldsymbol{\beta}}_1'\mathbf{X}_c'\mathbf{X}_c\hat{\boldsymbol{\beta}}_1$. Then $\text{SSR} = \hat{\boldsymbol{\beta}}_1'\mathbf{X}_c'\mathbf{y}$ can be written as

$$\text{SSR} = \hat{\boldsymbol{\beta}}_1'\mathbf{X}_c'\mathbf{X}_c\hat{\boldsymbol{\beta}}_1 = (\mathbf{X}_c\hat{\boldsymbol{\beta}}_1)'(\mathbf{X}_c\hat{\boldsymbol{\beta}}_1). \tag{7.54}$$

In this form, it is clear that SSR is due to $\boldsymbol{\beta}_1 = (\beta_1, \beta_2, \ldots, \beta_k)'$.

The proportion of the total sum of squares due to regression is

$$R^2 = \frac{\hat{\boldsymbol{\beta}}_1'\mathbf{X}_c'\mathbf{X}_c\hat{\boldsymbol{\beta}}_1}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\text{SSR}}{\text{SST}}, \tag{7.55}$$

which is known as the *coefficient of determination* or the *squared multiple correlation.* The ratio in (7.55) is a measure of model fit and provides an indication of how well the x's predict y.

The partitioning in (7.53) can be rewritten as the identity

$$\begin{aligned}\sum_{i=1}^{n}(y_i - \bar{y})^2 &= \mathbf{y}'\mathbf{y} - n\bar{y}^2 = (\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - n\bar{y}^2) + (\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}) \\ &= \text{SSR} + \text{SSE},\end{aligned}$$

which leads to an alternative expression for R^2:

$$R^2 = \frac{\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - n\bar{y}^2}{\mathbf{y}'\mathbf{y} - n\bar{y}^2}. \tag{7.56}$$

The positive square root R obtained from (7.55) or (7.56) is called the *multiple correlation coefficient.* If the x variables were random, R would estimate a population multiple correlation (see Section (10.4)).

We list some properties of R^2 and R:

1. The range of R^2 is $0 \leq R^2 \leq 1$. If all the $\hat{\beta}_j$'s were zero, except for $\hat{\beta}_0$, R^2 would be 0. (This event has probability 0 for continuous data.) If all the y values fell on the fitted surface, that is, if $y_i = \hat{y}_i$, $i = 1, 2, \ldots, n$, then R^2 would be 1.

2. $R = r_{y\hat{y}}$; that is, the multiple correlation is equal to the simple correlation [see (6.18)] between the observed y_i's and the fitted $\hat{y}_i$'s.
3. Adding a variable x to the model increases (cannot decrease) the value of R^2.
4. If $\beta_1 = \beta_2 = \cdots = \beta_k = 0$, then

$$E(R^2) = \frac{k}{n-1}. \tag{7.57}$$

   Note that the $\hat{\beta}_j$'s will not be 0 when the β_j's are 0.
5. R^2 cannot be partitioned into k components, each of which is uniquely attributable to an x_j, unless the x's are mutually orthogonal, that is, $\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(x_{im} - \bar{x}_m) = 0$ for $j \neq m$.
6. R^2 is invariant to full-rank linear transformations on the x's and to a scale change on y (but not invariant to a joint linear transformation including y and the x's).

In properties 3 and 4 we see that if k is a relatively large fraction of n, it is possible to have a large value of R^2 that is not meaningful. In this case, x's that do not contribute to predicting y may appear to do so in a particular example, and the estimated regression equation may not be a useful estimator of the population model. To correct for this tendency, an adjusted R^2, denoted by R_a^2, was proposed by Ezekiel (1930). To obtain R_a^2, we first subtract $k/(n-1)$ in (7.57) from R^2 in order to correct for the bias when $\beta_1 = \beta_2 = \ldots = \beta_k = 0$. This correction, however, would make R_a^2 too small when the β's are large, so a further modification is made so that $R_a^2 = 1$ when $R^2 = 1$. Thus R_a^2 is defined as

$$R_a^2 = \frac{(R^2 - \frac{k}{n-1})(n-1)}{n-k-1} = \frac{(n-1)R^2 - k}{n-k-1}. \tag{7.58}$$

Example 7.7. For the data in Table 7.1 in Example 7.2, we obtain R^2 by (7.56) and R_a^2 by (7.58). The values of $\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}$ and $\mathbf{y}'\mathbf{y}$ are given in Example 7.3.3.

$$\begin{aligned} R^2 &= \frac{\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - n\bar{y}^2}{\mathbf{y}'\mathbf{y} - n\bar{y}^2} = \frac{814.5410 - 12(7.5)^2}{840 - 12(7.5)^2} \\ &= \frac{139.5410}{165.0000} = .8457, \\ R_a^2 &= \frac{(n-1)R^2 - k}{n-k-1} = \frac{(11)(.8457) - 2}{9} = .8114. \end{aligned}$$

□

Using (7.44) and (7.46), we can express R^2 in (7.55) in terms of sample variances and covariances:

$$R^2 = \frac{\hat{\boldsymbol{\beta}}_1'\mathbf{X}_c'\mathbf{X}_c\hat{\boldsymbol{\beta}}_1}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\mathbf{s}_{yx}'\mathbf{S}_{xx}^{-1}(n-1)\mathbf{S}_{xx}\mathbf{S}_{xx}^{-1}\mathbf{s}_{yx}}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\mathbf{s}_{yx}'\mathbf{S}_{xx}^{-1}\mathbf{s}_{yx}}{s_y^2}. \tag{7.59}$$

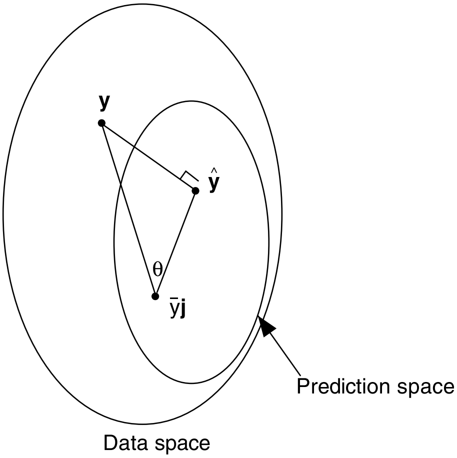

Figure 7.5 Multiple correlation R as cosine of θ, the angle between $\mathbf{y} - \bar{y}\mathbf{j}$ and $\hat{\mathbf{y}} - \bar{y}\mathbf{j}$.

This form of R^2 will facilitate a comparison with R^2 for the random-x case in Section (10.4) [see (10.34)].

Geometrically, R is the cosine of the angle θ between $\mathbf{y}$ and $\hat{\mathbf{y}}$ corrected for their means. The mean of $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$ is $\bar{y}$, the same as the mean of $y_1, y_2, \ldots, y_n$ (see Problem 7.30). Thus the centered forms of $\mathbf{y}$ and $\hat{\mathbf{y}}$ are $\mathbf{y} - \bar{y}\mathbf{j}$ and $\hat{\mathbf{y}} - \bar{y}\mathbf{j}$. The angle between them is illustrated in Figure 7.5. (Note that $\bar{y}\mathbf{j}$ is in the estimation space since it is a multiple of the first column of $\mathbf{X}$.)

To show that $\cos\theta$ is equal to the square root of R^2 as given by (7.56), we use (2.81) for the cosine of the angle between two vectors:

$$\cos\theta = \frac{(\mathbf{y} - \bar{y}\mathbf{j})'(\hat{\mathbf{y}} - \bar{y}\mathbf{j})}{\sqrt{[(\mathbf{y} - \bar{y}\mathbf{j})'(\mathbf{y} - \bar{y}\mathbf{j})][(\hat{\mathbf{y}} - \bar{y}\mathbf{j})'(\hat{\mathbf{y}} - \bar{y}\mathbf{j})]}}. \tag{7.60}$$

To simplify (7.60), we use the identity $\mathbf{y} - \bar{y}\mathbf{j} = (\hat{\mathbf{y}} - \bar{y}\mathbf{j}) + (\mathbf{y} - \hat{\mathbf{y}})$, which can also be seen geometrically in Figure 7.5. The vectors $\hat{\mathbf{y}} - \bar{y}\mathbf{j}$ and $\mathbf{y} - \hat{\mathbf{y}}$ on the right side of this identity are orthogonal since $\hat{\mathbf{y}} - \bar{y}\mathbf{j}$ is in the prediction space. Thus the numerator of (7.60) can be written as

$$\begin{aligned}(\mathbf{y} - \bar{y}\mathbf{j})'(\hat{\mathbf{y}} - \bar{y}\mathbf{j}) &= [(\hat{\mathbf{y}} - \bar{y}\mathbf{j}) + (\mathbf{y} - \hat{\mathbf{y}})]'(\hat{\mathbf{y}} - \bar{y}\mathbf{j}) \\ &= (\hat{\mathbf{y}} - \bar{y}\mathbf{j})'(\hat{\mathbf{y}} - \bar{y}\mathbf{j}) + (\mathbf{y} - \hat{\mathbf{y}})'(\hat{\mathbf{y}} - \bar{y}\mathbf{j}) \\ &= (\hat{\mathbf{y}} - \bar{y}\mathbf{j})'(\hat{\mathbf{y}} - \bar{y}\mathbf{j}) + 0.\end{aligned}$$

Then (7.60) becomes

$$\cos\theta = \frac{\sqrt{(\hat{\mathbf{y}} - \bar{y}\mathbf{j})'(\hat{\mathbf{y}} - \bar{y}\mathbf{j})}}{\sqrt{(\mathbf{y} - \bar{y}\mathbf{j})'(\mathbf{y} - \bar{y}\mathbf{j})}} = R, \tag{7.61}$$

which is easily shown to be the square root of R^2 as given by (7.56). This is equivalent to property 2 following (7.56): $R = r_{y\hat{y}}$.

We can write (7.61) in the form

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\text{SSR}}{\text{SST}},$$

in which $\text{SSR} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ is a sum of squares for the $\hat{y}_i$'s. Then the partitioning $\text{SST} = \text{SSR} + \text{SSE}$ below (7.53) can be written as

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2,$$

which is analogous to (6.17) for simple linear regression.

7.8 GENERALIZED LEAST SQUARES: cov(Y) = $\boldsymbol{\sigma}^2$V

We now consider models in which the y variables are correlated or have differing variances, so that $\text{cov}(\mathbf{y}) \neq \sigma^2\mathbf{I}$. In simple linear regression, larger values of x_i may lead to larger values of var(y_i). In either simple or multiple regression, if $y_1, y_2, \ldots, y_n$ occur at sequential points in time, they are typically correlated. For cases such as these, in which the assumption $\text{cov}(\mathbf{y}) = \sigma^2\mathbf{I}$ is no longer appropriate, we use the model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}, \quad \text{cov}(\mathbf{y}) = \Sigma = \sigma^2\mathbf{V}, \tag{7.62}$$

where $\mathbf{X}$ is full-rank and $\mathbf{V}$ is a known positive definite matrix. The usage $\Sigma = \sigma^2\mathbf{V}$ permits estimation of σ^2 in some convenient contexts (see Examples 7.8.1 and 7.8.2). The $n \times n$ matrix $\mathbf{V}$ has n diagonal elements and $\binom{n}{2}$ elements above (or below) the diagonal. If $\mathbf{V}$ were unknown, these $\binom{n}{2} + n$ distinct elements could not be estimated from a sample of n observations. In certain applications, a simpler structure for $\mathbf{V}$ is assumed that permits estimation. Such structures are illustrated in Examples 7.8.1 and 7.8.2.

7.8.1 Estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}^2$ when cov(y) = $\boldsymbol{\sigma}^2$V

In the following theorem we give estimators of $\boldsymbol{\beta}$ and σ^2 for the model in (7.62).

Theorem 7.8a. Let $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, let $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$, and let $\text{cov}(\mathbf{y}) = \text{cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{V}$, where $\mathbf{X}$ is a full-rank matrix and $\mathbf{V}$ is a known positive definite matrix. For this model, we obtain the following results:

(i) The best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}. \tag{7.63}$$

(ii) The covariance matrix for $\hat{\boldsymbol{\beta}}$ is

$$\text{cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}. \tag{7.64}$$

(iii) An unbiased estimator of σ^2 is

$$s^2 = \frac{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{n-k-1} \tag{7.65}$$

$$= \frac{\mathbf{y}'[\mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}]\mathbf{y}}{n-k-1}, \tag{7.66}$$

where $\hat{\boldsymbol{\beta}}$ is as given by (7.63).

Proof. We prove part (i). For parts (ii) and (iii), see Problems (7.32) and (7.33).

1. Since $\mathbf{V}$ is positive definite, there exists an $n \times n$ nonsingular matrix $\mathbf{P}$ such that $\mathbf{V} = \mathbf{PP}'$ (see Theorem 2.6c). Multiplying $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ by $\mathbf{P}^{-1}$, we obtain $\mathbf{P}^{-1}\mathbf{y} = \mathbf{P}^{-1}\mathbf{X}\boldsymbol{\beta} + \mathbf{P}^{-1}\boldsymbol{\varepsilon}$, for which $E(\mathbf{P}^{-1}\boldsymbol{\varepsilon}) = \mathbf{P}^{-1}E(\boldsymbol{\varepsilon}) = \mathbf{P}^{-1}\mathbf{0} = \mathbf{0}$ and

$$\begin{aligned} \text{cov}(\mathbf{P}^{-1}\boldsymbol{\varepsilon}) &= \mathbf{P}^{-1}\text{cov}(\boldsymbol{\varepsilon})(\mathbf{P}^{-1})' \qquad \text{[by (3.44)]} \\ &= \mathbf{P}^{-1}\sigma^2\mathbf{V}(\mathbf{P}^{-1})' = \sigma^2\mathbf{P}^{-1}\mathbf{PP}'(\mathbf{P}')^{-1} = \sigma^2\mathbf{I}. \end{aligned}$$

Thus the assumptions for Theorem 7.3d are satisfied for the model $\mathbf{P}^{-1}\mathbf{y} = \mathbf{P}^{-1}\mathbf{X}\boldsymbol{\beta} + \mathbf{P}^{-1}\boldsymbol{\varepsilon}$, and the least-squares estimator $\hat{\boldsymbol{\beta}} = [(\mathbf{P}^{-1}\mathbf{X})'(\mathbf{P}^{-1}\mathbf{X})]^{-1}(\mathbf{P}^{-1}\mathbf{X})'\mathbf{P}^{-1}\mathbf{y}$ is BLUE. Using Theorems 2.2b and 2.5b, this can be written as

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= [\mathbf{X}'(\mathbf{P}^{-1})'\mathbf{P}^{-1}\mathbf{X}]^{-1}\mathbf{X}'(\mathbf{P}^{-1})'\mathbf{P}^{-1}\mathbf{y} \\ &= [\mathbf{X}'(\mathbf{P}')^{-1}\mathbf{P}^{-1}\mathbf{X}]^{-1}\mathbf{X}'(\mathbf{P}')^{-1}\mathbf{P}^{-1}\mathbf{y} \qquad \text{[by (2.48)]} \\ &= [\mathbf{X}'(\mathbf{PP}')^{-1}\mathbf{X}]^{-1}\mathbf{X}'(\mathbf{PP}')^{-1}\mathbf{y} \qquad \text{[by (2.49)]} \\ &= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}. \end{aligned}$$

□

Note that since $\mathbf{X}$ is full-rank, $\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}$ is positive definite (see Theorem 2.6b). The estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$ is usually called the *generalized least-squares* estimator. The same estimator is obtained under a normality assumption.

Theorem 7.8b. If $\mathbf{y}$ is $N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{V})$, where $\mathbf{X}$ is full-rank and $\mathbf{V}$ is a known positive definite matrix, where $\mathbf{X}$ is $n \times (k+1)$ of rank $k+1$, then the maximum likelihood estimators for $\boldsymbol{\beta}$ and σ^2 are

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

$$\hat{\sigma}^2 = \frac{1}{n}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}).$$

PROOF. The likelihood function is

$$L(\boldsymbol{\beta}, \sigma^2) = \frac{1}{(2\pi)^{n/2}|\sigma^2\mathbf{V}|^{1/2}} e^{-(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\sigma^2\mathbf{V})^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})/2}.$$

By (2.69), $|\sigma^2\mathbf{V}| = (\sigma^2)^n|\mathbf{V}|$. Hence

$$L(\boldsymbol{\beta}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}|\mathbf{V}|^{1/2}} e^{-(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})/2\sigma^2}.$$

The results can be obtained by differentiation of $\ln L(\boldsymbol{\beta}, \sigma^2)$ with respect to $\boldsymbol{\beta}$ and with respect to σ^2. □

We illustrate an application of generalized least squares.

Example 7.8.1. Consider the centered model in (7.32)

$$\mathbf{y} = (\mathbf{j}, \mathbf{X}_c)\begin{pmatrix} \alpha \\ \boldsymbol{\beta}_1 \end{pmatrix} + \boldsymbol{\varepsilon},$$

with covariance pattern

$$\Sigma = \sigma^2[(1-\rho)\mathbf{I} + \rho\mathbf{J}] = \sigma^2\mathbf{V} \tag{7.67}$$

$$= \sigma^2 \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix},$$

in which all variables have the same variance σ^2 and all pairs of variables have the same correlation ρ. This covariance pattern was introduced in Problem 5.26 and is assumed for certain repeated measures and intraclass correlation designs. See (3.19) for a definition of ρ.

By (7.63), we have

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\alpha} \\ \hat{\boldsymbol{\beta}}_1 \end{pmatrix} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}.$$

For the centered model with $\mathbf{X} = (\mathbf{j}, \mathbf{X}_c)$, the matrix $\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}$ becomes

$$\begin{aligned} \mathbf{X}'\mathbf{V}^{-1}\mathbf{X} &= \begin{pmatrix} \mathbf{j}' \\ \mathbf{X}_c' \end{pmatrix} \mathbf{V}^{-1}(\mathbf{j}, \mathbf{X}_c) \\ &= \begin{pmatrix} \mathbf{j}'\mathbf{V}^{-1}\mathbf{j} & \mathbf{j}'\mathbf{V}^{-1}\mathbf{X}_c \\ \mathbf{X}_c'\mathbf{V}^{-1}\mathbf{j} & \mathbf{X}_c'\mathbf{V}^{-1}\mathbf{X}_c \end{pmatrix}. \end{aligned}$$

The inverse of the $n \times n$ matrix $\mathbf{V} = (1-\rho)\mathbf{I} + \rho\mathbf{J}$ in (7.67) is given by

$$\mathbf{V}^{-1} = a(\mathbf{I} - b\rho\mathbf{J}), \tag{7.68}$$

where $a = 1/(1-\rho)$ and $b = 1/[1+(n-1)\rho]$. Using $\mathbf{V}^{-1}$ in (7.68), $\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}$ becomes

$$\mathbf{X}'\mathbf{V}^{-1}\mathbf{X} = \begin{pmatrix} bn & \mathbf{0}' \\ \mathbf{0} & a\mathbf{X}_c'\mathbf{X}_c \end{pmatrix}. \tag{7.69}$$

Similarly

$$\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} = \begin{pmatrix} bn\bar{y} \\ a\mathbf{X}_c'\mathbf{y} \end{pmatrix}. \tag{7.70}$$

We therefore have

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\boldsymbol{\beta}}_1 \end{pmatrix} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} = \begin{pmatrix} \bar{y} \\ (\mathbf{X}_c'\mathbf{X}_c)^{-1}\mathbf{X}_c'\mathbf{y} \end{pmatrix},$$

which is the same as (7.36) and (7.37). Thus the usual least-squares estimators are BLUE for a covariance structure with equal variances and equal correlations. □

7.8.2 Misspecification of the Error Structure

Suppose that the model is $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with $\text{cov}(\mathbf{y}) = \sigma^2\mathbf{V}$, as in (7.62), and we mistakenly (or deliberately) use the ordinary least-squares estimator $\hat{\boldsymbol{\beta}}^* = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ in (7.6), which we denote here by $\hat{\boldsymbol{\beta}}^*$ to distinguish it from the BLUE estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$ in (7.63). Then the mean vector and covariance matrix for $\hat{\boldsymbol{\beta}}^*$ are

$$E(\hat{\boldsymbol{\beta}}^*) = \boldsymbol{\beta}, \tag{7.71}$$

$$\text{cov}(\hat{\boldsymbol{\beta}}^*) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}. \tag{7.72}$$

Thus the ordinary least-squares estimators are unbiased, but the covariance matrix differs from (7.64). Because of Theorem 7.8a(i), the variances of the $\hat{\beta}_j^*$'s in (7.72) cannot be smaller than the variances in $\text{cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$ in (7.64). This is illustrated in the following example.

Example 7.8.2. Suppose that we have a simple linear regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, where $\text{var}(y_i) = \sigma^2 x_i$ and $\text{cov}(y_i, y_j) = 0$ for $i \neq j$. Thus

$$\text{cov}(\mathbf{y}) = \sigma^2\mathbf{V} = \sigma^2\begin{pmatrix} x_1 & 0 & \dots & 0 \\ 0 & x_2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & x_n \end{pmatrix}.$$

This is an example of *weighted least squares*, which typically refers to the case where $\mathbf{V}$ is diagonal with functions of the x's on the diagonal. In this case

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix},$$

and by (7.63), we have

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$$

$$= \frac{1}{\left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n \frac{1}{x_i}\right) - n^2}\begin{pmatrix} \left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n \frac{y_i}{x_i}\right) - n\sum_{i=1}^n y_i \\ \left(\sum_{i=1}^n y_i\right)\left(\sum_{i=1}^n \frac{1}{x_i}\right) - n\sum_{i=1}^n \frac{y_i}{x_i} \end{pmatrix}. \tag{7.73}$$

The covariance matrix for $\hat{\boldsymbol{\beta}}$ is given by (7.64):

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

$$= \frac{\sigma^2}{\sum_i x_i \sum_i \frac{1}{x_i} - n^2}\begin{pmatrix} \sum_i x_i & -n \\ -n & \sum_i \frac{1}{x_i} \end{pmatrix}. \tag{7.74}$$

If we use the ordinary least-squares estimator $\hat{\boldsymbol{\beta}}^* = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ as given in (6.5) and (6.6) or in (7.12) in Example 7.3.1b, then $\text{cov}(\hat{\boldsymbol{\beta}}^*)$ is given by (7.72); that is,

$$\begin{aligned}
\text{cov}(\hat{\boldsymbol{\beta}}^*) &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2 \begin{pmatrix} n & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_i x_i & \sum_i x_i^2 \\ \sum_i x_i^2 & \sum_i x_i^3 \end{pmatrix} \begin{pmatrix} n & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{pmatrix}^{-1} \\
&= \sigma^2 c \begin{pmatrix} \sum_i x_i^3(\sum_i x_i)^2 - \sum_i x_i(\sum_i x_i^2)^2 & n(\sum_i x_i^2)^2 - n\sum_i x_i \sum_i x_i^3 \\ n(\sum_i x_i^2)^2 - n\sum_i x_i \sum_i x_i^3 & n^2\sum_i x_i^3 - 2n\sum_i x_i \sum_i x_i^2 + (\sum_i x_i)^3 \end{pmatrix},
\end{aligned} \tag{7.75}$$

where $c = 1/\left[n\sum_i x_i^2 - \left(\sum_i x_i\right)^2\right]^2$. The variance of the estimator $\hat{\beta}_1^*$ is given by the lower right diagonal element of (7.75):

$$\text{var}(\hat{\beta}_1^*) = \sigma^2 \frac{n^2\sum_i x_i^3 - 2n\sum_i x_i \sum_i x_i^2 + (\sum_i x_i)^3}{\left[n\sum_i x_i^2 - \left(\sum_i x_i\right)^2\right]^2}, \tag{7.76}$$

and the variance of the estimator $\hat{\beta}_1$ is given by the corresponding element of (7.74):

$$\text{var}(\hat{\beta}_1) = \sigma^2 \frac{\sum_i (1/x_i)}{\sum_i x_i \sum_i (1/x_i) - n^2}. \tag{7.77}$$

Consider the following seven values of x: 1, 2, 3, 4, 5, 6, 7. Using (7.76), we obtain $\text{var}(\hat{\beta}_1^*) = .1429\sigma^2$, and from (7.77), we have $\text{var}(\hat{\beta}_1) = .1099\sigma^2$. Thus for these values of x, the use of ordinary least squares yields a slope estimator with a larger variance, as expected. □

Further consequences of using a wrong model are discussed in the next section.

7.9 MODEL MISSPECIFICATION

In Section 7.8.2, we discussed some consequences of misspecification of $\text{cov}(\mathbf{y})$. We now consider consequences of misspecification of $E(\mathbf{y})$. As a framework for discussion, let the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ be partitioned as

$$\begin{aligned}
\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} &= (\mathbf{X}_1, \mathbf{X}_2)\begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} + \boldsymbol{\varepsilon} \\
&= \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}.
\end{aligned} \tag{7.78}$$

If we leave out $\mathbf{X}_2\boldsymbol{\beta}_2$ when it should be included (i.e., when $\boldsymbol{\beta}_2 \neq \mathbf{0}$), we are *underfitting*. If we include $\mathbf{X}_2\boldsymbol{\beta}_2$ when it should be excluded (i.e., when $\boldsymbol{\beta}_2 = \mathbf{0}$), we are *overfitting*. We discuss the effect of underfitting or overfitting on the bias and the variance of the $\hat{\beta}_j$, $\hat{y}$, and s^2 values.

We first consider estimation of $\boldsymbol{\beta}_1$ when underfitting. We write the *reduced* model as

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1^* + \boldsymbol{\varepsilon}^*, \tag{7.79}$$

using $\boldsymbol{\beta}_1^*$ to emphasize that these parameters (and their estimates $\hat{\boldsymbol{\beta}}_1^*$) will be different from $\boldsymbol{\beta}_1$ (and $\hat{\boldsymbol{\beta}}_1$) in the *full* model (7.78) (unless the x's are orthogonal; see Corollary 1 to Theorem 7.9a and Theorem 7.10). This was illustrated in Example 7.2. In the following theorem, we discuss the bias in the estimator $\hat{\boldsymbol{\beta}}_1^*$ obtained from (7.79) and give the covariance matrix for $\hat{\boldsymbol{\beta}}_1^*$.

Theorem 7.9a. If we fit the model $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1^* + \boldsymbol{\varepsilon}^*$ when the correct model is $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$ with $\mathrm{cov}(\mathbf{y}) = \sigma^2\mathbf{I}$, then the mean vector and covariance matrix for the least-squares estimator $\hat{\boldsymbol{\beta}}_1^* = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}$ are as follows:

$$\text{(i) } E(\hat{\boldsymbol{\beta}}_1^*) = \boldsymbol{\beta}_1 + \mathbf{A}\boldsymbol{\beta}_2, \text{ where } \mathbf{A} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2, \tag{7.80}$$

$$\text{(ii) } \mathrm{cov}(\hat{\boldsymbol{\beta}}_1^*) = \sigma^2(\mathbf{X}_1'\mathbf{X}_1)^{-1}. \tag{7.81}$$

PROOF

$$\begin{aligned}\text{(i) } E(\hat{\boldsymbol{\beta}}_1^*) &= E[(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}] = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'E(\mathbf{y})\\ &= (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'(\mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2)\\ &= \boldsymbol{\beta}_1 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\boldsymbol{\beta}_2.\\ \text{(ii) } \mathrm{cov}(\hat{\boldsymbol{\beta}}_1^*) &= \mathrm{cov}[(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}]\\ &= (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'(\sigma^2\mathbf{I})\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1} \qquad \text{[by (3.44)]}\\ &= \sigma^2(\mathbf{X}_1'\mathbf{X}_1)^{-1}.\end{aligned}$$

□

Thus, when underfitting, $\hat{\boldsymbol{\beta}}_1^*$ is biased by an amount that depends on the values of the x's in both $\mathbf{X}_1$ and $\mathbf{X}_2$. The matrix $\mathbf{A} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2$ in (7.81) is called the *alias* matrix.

Corollary 1. If $\mathbf{X}_1'\mathbf{X}_2 = \mathbf{O}$, that is, if the columns of $\mathbf{X}_1$ are orthogonal to the columns of $\mathbf{X}_2$, then $\hat{\boldsymbol{\beta}}_1^*$ is unbiased: $E(\hat{\boldsymbol{\beta}}_1^*) = \boldsymbol{\beta}_1$. □

In the next three theorems, we discuss the effect of underfitting or overfitting on $\hat{y}$, s^2, and the variances of the $\hat{\beta}_j$'s. In some of the proofs we follow Hocking (1996, pp. 245–247).

Let $\mathbf{x}_0 = (1, x_{01}, x_{02}, \ldots, x_{0k})'$ be a particular value of $\mathbf{x}$ for which we desire to estimate $E(y_0) = \mathbf{x}_0'\boldsymbol{\beta}$. If we partition $\mathbf{x}_0'$ into $(\mathbf{x}_{01}', \mathbf{x}_{02}')$ corresponding to the partitioning $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ and $\boldsymbol{\beta}' = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')$, then we can use either $\hat{y}_0 = \mathbf{x}_0'\hat{\boldsymbol{\beta}}$ or $\hat{y}_{01} = \mathbf{x}_{01}'\hat{\boldsymbol{\beta}}_1^*$ to estimate $\mathbf{x}_0'\boldsymbol{\beta}$. In the following theorem, we consider the mean of $\hat{y}_{01}$.

Theorem 7.9b. Let $\hat{y}_{01} = \mathbf{x}_{01}'\hat{\boldsymbol{\beta}}_1^*$, where $\hat{\boldsymbol{\beta}}_1^*1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}$. Then, if $\boldsymbol{\beta}_2 \neq \mathbf{0}$, we obtain

$$E(\mathbf{x}_{01}'\hat{\boldsymbol{\beta}}_1^*) = \mathbf{x}_{01}'(\boldsymbol{\beta}_1 + \mathbf{A}\boldsymbol{\beta}_2), \tag{7.82}$$

$$= \mathbf{x}_0'\boldsymbol{\beta} - (\mathbf{x}_{02} - \mathbf{A}'\mathbf{x}_{01})'\boldsymbol{\beta}_2 \neq \mathbf{x}_0'\boldsymbol{\beta}. \tag{7.83}$$

Proof. See Problem 7.43. □

In Theorem 7.9b, we see that, when underfitting, $\mathbf{x}_{01}'\hat{\boldsymbol{\beta}}_1^*$ is biased for estimating $\mathbf{x}_0'\boldsymbol{\beta}$. [When overfitting, $\mathbf{x}_0'\hat{\boldsymbol{\beta}}$ is unbiased since $E(\mathbf{x}_0'\hat{\boldsymbol{\beta}}) = \mathbf{x}_0'\boldsymbol{\beta} = \mathbf{x}_{01}'\boldsymbol{\beta}_1 + \mathbf{x}_{02}'\boldsymbol{\beta}_2$, which is equal to $\mathbf{x}_{01}'\boldsymbol{\beta}_1$ if $\boldsymbol{\beta}_2 = \mathbf{0}$.]

In the next theorem, we compare the variances of $\hat{\beta}_j^*$ and $\hat{\beta}_j$, where $\hat{\beta}_j^*$ is from $\hat{\boldsymbol{\beta}}_1^*$ and $\hat{\beta}_j$ is from $\hat{\boldsymbol{\beta}}_1$. We also compare the variances of $\mathbf{x}_{01}'\hat{\boldsymbol{\beta}}_1^*$ and $\mathbf{x}_0'\hat{\boldsymbol{\beta}}$.

Theorem 7.9c. Let $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ from the full model be partitioned as $\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{pmatrix}$, and let $\hat{\boldsymbol{\beta}}_1^* = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}$ be the estimator from the reduced model. Then

(i) $\text{cov}(\hat{\boldsymbol{\beta}}_1) - \text{cov}(\hat{\boldsymbol{\beta}}_1^*) = \sigma^2\mathbf{A}\mathbf{B}^{-1}\mathbf{A}'$, which is a positive definite matrix, where $\mathbf{A} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2$ and $\mathbf{B} = \mathbf{X}_2'\mathbf{X}_2 - \mathbf{X}_2'\mathbf{X}_1\mathbf{A}$. Thus $\text{var}(\hat{\beta}_j) > \text{var}(\hat{\beta}_j^*)$.

(ii) $\text{var}(\mathbf{x}_0'\hat{\boldsymbol{\beta}}) \geq \text{var}(\mathbf{x}_{01}'\hat{\boldsymbol{\beta}}_1^*)$.

Proof

(i) Using $\mathbf{X}'\mathbf{X}$ partitioned to conform to $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$, we have

$$\text{cov}(\hat{\boldsymbol{\beta}}) = \text{cov}\begin{pmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{pmatrix} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2\begin{pmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 \end{pmatrix}^{-1}$$

$$= \sigma^2\begin{pmatrix} \mathbf{G}_{11} & \mathbf{G}_{12} \\ \mathbf{G}_{21} & \mathbf{G}_{22} \end{pmatrix}^{-1} = \sigma^2\begin{pmatrix} \mathbf{G}^{11} & \mathbf{G}^{12} \\ \mathbf{G}^{21} & \mathbf{G}^{22} \end{pmatrix},$$

where $\mathbf{G}_{ij} = \mathbf{X}_i'\mathbf{X}_j$ and $\mathbf{G}^{ij}$ is the corresponding block of the partitioned inverse matrix $(\mathbf{X}'\mathbf{X})^{-1}$. Thus $\text{cov}(\hat{\boldsymbol{\beta}}_1) = \sigma^2\mathbf{G}^{11}$. By (2.50), $\mathbf{G}^{11} = \mathbf{G}_{11}^{-1} + \mathbf{G}_{11}^{-1}\mathbf{G}_{12}\mathbf{B}^{-1}\mathbf{G}_{21}\mathbf{G}_{11}^{-1}$, where $\mathbf{B} = \mathbf{G}_{22} - \mathbf{G}_{21}\mathbf{G}_{11}^{-1}\mathbf{G}_{12}$. By (7.81), $\text{cov}(\hat{\boldsymbol{\beta}}_1^*) = \sigma^2(\mathbf{X}_1'\mathbf{X}_1)^{-1} = \sigma^2\mathbf{G}_{11}^{-1}$. Hence

$$\begin{aligned}\text{cov}(\hat{\boldsymbol{\beta}}_1) - \text{cov}(\hat{\boldsymbol{\beta}}_1^*) &= \sigma^2(\mathbf{G}^{11} - \mathbf{G}_{11}^{-1}) \\ &= \sigma^2(\mathbf{G}_{11}^{-1} + \mathbf{G}_{11}^{-1}\mathbf{G}_{12}\mathbf{B}^{-1}\mathbf{G}_{21}\mathbf{G}_{11}^{-1} - \mathbf{G}_{11}^{-1}) \\ &= \sigma^2\mathbf{A}\mathbf{B}^{-1}\mathbf{A}'.\end{aligned}$$

(ii)
$$\begin{aligned}\text{var}(\mathbf{x}_0'\hat{\boldsymbol{\beta}}) &= \sigma^2\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0\\ &= \sigma^2(\mathbf{x}_{01}', \mathbf{x}_{02}')\begin{pmatrix}\mathbf{G}^{11} & \mathbf{G}^{12}\\ \mathbf{G}^{21} & \mathbf{G}^{22}\end{pmatrix}\begin{pmatrix}\mathbf{x}_{01}\\ \mathbf{x}_{02}\end{pmatrix}\\ &= \sigma^2(\mathbf{x}_{01}'\mathbf{G}^{11}\mathbf{x}_{01} + \mathbf{x}_{01}'\mathbf{G}^{12}\mathbf{x}_{02} + \mathbf{x}_{02}'\mathbf{G}^{21}\mathbf{x}_{01} + \mathbf{x}_{02}'\mathbf{G}^{22}\mathbf{x}_{02}).\end{aligned}$$

Using (2.50), it can be shown that

$$\text{var}(\mathbf{x}_0'\hat{\boldsymbol{\beta}}) - \text{var}(\mathbf{x}_{01}'\hat{\boldsymbol{\beta}}_1^*) = \sigma^2(\mathbf{x}_{02} - \mathbf{A}'\mathbf{x}_{01})'\mathbf{G}^{22}(\mathbf{x}_{02} - \mathbf{A}'\mathbf{x}_{01}) \geq 0$$

because $\mathbf{G}^{22}$ is positive definite. □

By Theorem 7.9c(i), var($\hat{\beta}_j$) in the full model is greater than var($\hat{\beta}_j^*$) in the reduced model. Thus underfitting reduces the variance of the $\hat{\beta}_j$'s but introduces bias. On the other hand, overfitting increases the variance of the $\hat{\beta}_j$'s. In Theorem 7.9c (ii), var($\hat{y}_0$) based on the full model is greater than var($\hat{y}_{01}$) based on the reduced model. Again, underfitting reduces the variance of the estimate of $E(y_0)$ but introduces bias. Overfitting increases the variance of the estimate of $E(y_0)$.

We now consider s^2 for the full model and for the reduced model. For the full model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$, the sample variance s^2 is given by (7.23) as

$$s^2 = \frac{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{n-k-1}.$$

In Theorem 7.3f, we have $E(s^2) = \sigma^2$. The expected value of s^2 for the reduced model is given in the following theorem.

Theorem 7.9d. If $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ is the correct model, then for the reduced model $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1^* + \boldsymbol{\varepsilon}^*$ (underfitting), where $\mathbf{X}_1$ is $n \times (p+1)$ with $p < k$, the variance estimator

$$s_1^2 = \frac{(\mathbf{y} - \mathbf{X}_1\hat{\boldsymbol{\beta}}_1^*)'(\mathbf{y} - \mathbf{X}_1\hat{\boldsymbol{\beta}}_1^*)}{n-p-1} \tag{7.84}$$

has expected value

$$E(s_1^2) = \sigma^2 + \frac{\boldsymbol{\beta}_2'\mathbf{X}_2'[\mathbf{I} - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1']\mathbf{X}_2\boldsymbol{\beta}_2}{n-p-1}. \tag{7.85}$$

Proof. We write the numerator of (7.84) as

$$\begin{aligned}\text{SSE}_1 &= \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}_1^*\mathbf{X}_1'\mathbf{y} = \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}\\ &= \mathbf{y}'[\mathbf{I} - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1']\mathbf{y}.\end{aligned}$$

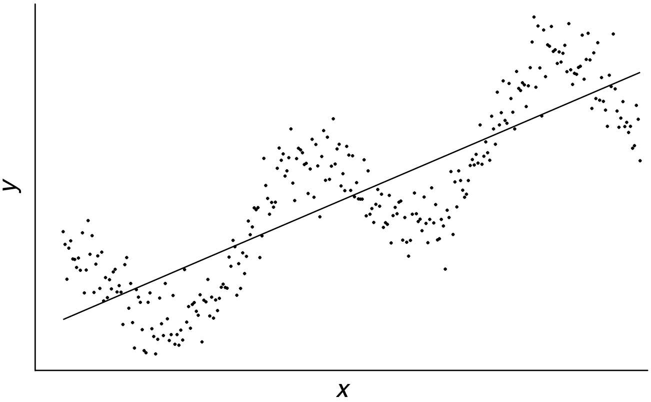

Figure 7.6 Straight-line fit to a curved pattern of points.

Since $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ by assumption, we have, by Theorem 5.2a,

$$\begin{aligned} E(\text{SSE}_1) &= \text{tr}\{[\mathbf{I} - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1']\sigma^2\mathbf{I}\} + \boldsymbol{\beta}'\mathbf{X}'[\mathbf{I} - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1']\mathbf{X}\boldsymbol{\beta} \\ &= (n-p-1)\sigma^2 + \boldsymbol{\beta}_2'\mathbf{X}_2'[\mathbf{I} - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1']\mathbf{X}_2\boldsymbol{\beta}_2 \end{aligned}$$

(see Problem 7.45). □

Since the quadratic form in (7.85) is positive semidefinite, s^2 is biased upward when underfitting (see Fig. 7.6). We can also examine (7.85) from the perspective of overfitting, in which case $\boldsymbol{\beta}_2 = \mathbf{0}$ and s^2 is unbiased.

To summarize the results in this section, underfitting leads to biased $\hat{\beta}_j$'s, biased $\hat{y}$'s, and biased s^2. Overfitting increases the variances of the $\hat{\beta}_j$'s and of the $\hat{y}$'s. We are thus compelled to seek an appropriate balance between a biased model and one with large variances. This is the task of the model builder and serves as motivation for seeking an optimum subset of x's.

Example 7.9a. Suppose that the model $y_i = \beta_0^* + \beta_1^* x_i + \varepsilon_i^*$ has been fitted when the true model is $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$. (This situation is similar to that illustrated in Figure 6.2.) In this case, $\hat{\beta}_0^*$, $\hat{\beta}_1^*$, and s_1^2 would be biased by an amount dependent on the choice of the x_i's [see (7.80) and (7.86)]. The error term $\hat{\varepsilon}_i^*$ in the misspecified model $y_i = \beta_0^* + \beta_1^* x_i + \varepsilon_i^*$ does not have a mean of 0:

$$\begin{aligned} E(\varepsilon_i^*) &= E(y_i - \beta_0^* - \beta_1^* x_i) \\ &= E(y_i) - \beta_0^* - \beta_1^* x_i \\ &= \beta_0 + \beta_1 x_i + \beta_2 x_i^2 - \beta_0^* - \beta_1^* x_i \\ &= \beta_0 - \beta_0^* + (\beta_1 - \beta_1^*)x_i + \beta_2 x_i^2. \end{aligned}$$

□

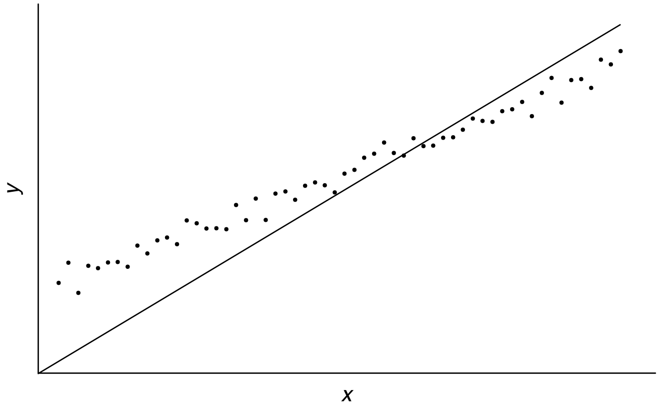

Figure 7.7 No-intercept model fit to data from an intercept model.

Example 7.9b. Suppose that the true model is $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ and we fit the model $y_i = \beta_1^* x_i + \varepsilon_i^*$, as illustrated in Figure 7.7.

For the model $y_i = \beta_1^* x_i + \varepsilon_i^*$, the least-squares estimator is

$$\hat{\beta}_1^* = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2} \tag{7.86}$$

(see Problem 7.46). Then, under the full model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, we have

$$\begin{aligned} E(\hat{\beta}_1^*) &= \frac{1}{\sum_i x_i^2} \sum_i x_i E(y_i) \\ &= \frac{1}{\sum_i x_i^2} \sum_i x_i (\beta_0 + \beta_1 x_i) \\ &= \frac{1}{\sum_i x_i^2} \left(\beta_0 \sum_i x_i + \beta_1 \sum_i x_i^2 \right) \\ &= \beta_0 \frac{\sum_i x_i}{\sum_i x_i^2} + \beta_1. \end{aligned} \tag{7.87}$$

Thus $\hat{\beta}_1^*$ is biased by an amount that depends on β_0 and the values of the x's. □

7.10 ORTHOGONALIZATION

In Section 7.9, we discussed estimation of $\boldsymbol{\beta}_1^*$ in the model $\mathbf{y} = \mathbf{X}_1 \boldsymbol{\beta}_1^* + \boldsymbol{\varepsilon}^*$ when the true model is $\mathbf{y} = \mathbf{X}_1 \boldsymbol{\beta}_1 + \mathbf{X}_2 \boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$. By Theorem 7.9a, $E(\hat{\boldsymbol{\beta}}_1^*) = \boldsymbol{\beta}_1 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\boldsymbol{\beta}_2$, so that estimation of $\boldsymbol{\beta}_1$ is affected by the presence of $\mathbf{X}_2$, unless $\mathbf{X}_1'\mathbf{X}_2 = \mathbf{O}$, in which case, $E(\hat{\boldsymbol{\beta}}_1^*) = \boldsymbol{\beta}_1$. In the following theorem, we show that if $\mathbf{X}_1'\mathbf{X}_2 = \mathbf{O}$, the estimators of $\boldsymbol{\beta}_1^*$ and $\boldsymbol{\beta}_1$ not only have the same expected value, but are exactly the same.

Theorem 7.10. If $\mathbf{X}_1'\mathbf{X}_2 = \mathbf{O}$, then the estimator of $\boldsymbol{\beta}_1$ in the full model $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$ is the same as the estimator of $\boldsymbol{\beta}_1^*$ in the reduced model $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1^* + \boldsymbol{\varepsilon}^*$.

PROOF. The least-squares estimator of $\boldsymbol{\beta}_1^*$ is $\hat{\boldsymbol{\beta}}_1^* = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}$. For the estimator of $\boldsymbol{\beta}_1$ in the full model, we partition $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ to obtain

$$\begin{pmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{X}_1'\mathbf{y} \\ \mathbf{X}_2'\mathbf{y} \end{pmatrix}.$$

Using the notation in the proof of Theorem 7.9c, this becomes

$$\begin{aligned} \begin{pmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{pmatrix} &= \begin{pmatrix} \mathbf{G}_{11} & \mathbf{G}_{12} \\ \mathbf{G}_{21} & \mathbf{G}_{22} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{X}_1'\mathbf{y} \\ \mathbf{X}_2'\mathbf{y} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{G}^{11} & \mathbf{G}^{12} \\ \mathbf{G}^{21} & \mathbf{G}^{22} \end{pmatrix} \begin{pmatrix} \mathbf{X}_1'\mathbf{y} \\ \mathbf{X}_2'\mathbf{y} \end{pmatrix}. \end{aligned}$$

By (2.50), we obtain

$$\begin{aligned} \hat{\boldsymbol{\beta}}_1 &= \mathbf{G}^{11}\mathbf{X}_1'\mathbf{y} + \mathbf{G}^{12}\mathbf{X}_2'\mathbf{y} \\ &= (\mathbf{G}_{11}^{-1} + \mathbf{G}_{11}^{-1}\mathbf{G}_{12}\mathbf{B}^{-1}\mathbf{G}_{21}\mathbf{G}_{11}^{-1})\mathbf{X}_1'\mathbf{y} - \mathbf{G}_{11}^{-1}\mathbf{G}_{12}\mathbf{B}^{-1}\mathbf{X}_2'\mathbf{y}, \end{aligned}$$

where $\mathbf{B} = \mathbf{G}_{22} - \mathbf{G}_{21}\mathbf{G}_{11}^{-1}\mathbf{G}_{12}$. If $\mathbf{G}_{12} = \mathbf{X}_1'\mathbf{X}_2 = \mathbf{O}$, then $\hat{\boldsymbol{\beta}}_1$ reduces to

$$\hat{\boldsymbol{\beta}}_1 = \mathbf{G}_{11}^{-1}\mathbf{X}_1'\mathbf{y} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y},$$

which is the same as $\hat{\boldsymbol{\beta}}_1^*$. □

Note that Theorem 7.10 will also hold if $\mathbf{X}_1$ and $\mathbf{X}_2$ are "essentially orthogonal," that is, if the centered columns of $\mathbf{X}_1$ are orthogonal to the centered columns of $\mathbf{X}_2$.

In Theorem 7.9a, we discussed estimation of $\boldsymbol{\beta}_1^*$ in the presence of $\boldsymbol{\beta}_2$ when $\mathbf{X}_1'\mathbf{X}_2 \neq \mathbf{O}$. We now consider a process of orthogonalization to give additional insights into the meaning of partial regression coefficients.

In Example 7.2, we illustrated the change in the estimate of a regression coefficient when another x was added to the model. We now use the same data to further examine this change.The prediction equation obtained in Example 7.2 was

$$\hat{y} = 5.3754 + 3.0118x_1 - 1.2855x_2, \qquad (7.88)$$

and the negative partial regressions of y on x_2 were shown in Figure 7.2. By means of orthogonalization, we can give additional meaning to the term $-1.2855x_2$. In order to add x_2 to the prediction equation containing only x_1, we need to determine how much variation in y is due to x_2 after the effect of x_1 has been accounted for, and we must also correct for the relationship between x_1 and x_2. Our approach is to consider the relationship between the residual variation after regressing y on x_1 and the residual variation after regressing x_2 on x_1. We follow a three-step process.

1. Regress y on x_1, and calculate residuals [see (7.11)]. The prediction equation is

$$\hat{y} = 1.8585 + 1.3019x_1, \tag{7.89}$$

   and the residuals $y_i - \hat{y}_i(x_1)$ are given in Table 7.2, where $\hat{y}_i(x_1)$ indicates that $\hat{y}$ is based on a regression of y on x_1 as in (7.89).
2. Regress x_2 on x_1 and calculate residuals. The prediction equation is

$$\hat{x}_2 = 2.7358 + 1.3302x_1, \tag{7.90}$$

   and the residuals $x_{2i} - \hat{x}_{2i}(x_1)$ are given in Table 7.2, where $\hat{x}_{2i}(x_1)$ indicates that x_2 has been regressed on x_1 as in (7.90).
3. Now regress $y - \hat{y}(x_1)$ on $x_2 - \hat{x}_2(x_1)$, which gives

$$\widehat{y - \hat{y}} = -1.2855(x_2 - \hat{x}_2). \tag{7.91}$$

There is no intercept in (7.91) because both sets of residuals have a mean of 0.

TABLE 7.2 Data from Table 7.1 and Residuals

y	x_1	x_2	$y - \hat{y}(x_1)$	$x_2 - \hat{x}_2(x_1)$
2	0	2	0.1415	−0.7358
3	2	6	−1.4623	0.6038
2	2	7	−2.4623	1.6038
7	2	5	2.5377	−0.3962
6	4	9	−1.0660	0.9434
8	4	8	0.9340	−0.0566
10	4	7	2.9340	−1.0566
7	6	10	−2.6698	−0.7170
8	6	11	−1.6698	0.2830
12	6	9	2.3302	−1.7170
11	8	15	−1.2736	1.6226
14	8	13	1.7264	−0.3774

In (7.91), we obtain a clearer insight into the meaning of the partial regression coefficient -1.2855 in (7.88). We are using the "unexplained" portion of x_2 (after x_1 is accounted for) to predict the "unexplained" portion of y (after x_1 is accounted for).

Since $x_2 - \hat{x}_2(x_1)$ is orthogonal to x_1 [see Section 7.4.2, in particular (7.29)], fitting $y - \hat{y}(x_1)$ to $x_2 - \hat{x}_2(x_1)$ yields the same coefficient, -1.2855, as when fitting y to x_1 and x_2 together. Thus -1.2855 represents the additional effect of x_2 beyond the effect of x_1 and also after taking into account the overlap between x_1 and x_2 in their effect on y. The orthogonality of x_1 and $x_2 - \hat{x}_2(x_1)$ makes this simplified breakdown of effects possible.

We can substitute $\hat{y}(x_1)$ and $\hat{x}_2(x_1)$ in (7.91) to obtain

$$\widehat{y - \hat{y}} = \hat{y}(x_1, x_2) - \hat{y}(x_1) = -1.2855[x_2 - \hat{x}_2(x_1)],$$

or

$$\hat{y} - (1.8585 + 1.3019x_1) = -1.2855[x_2 - (2.7358 + 1.3302x_1)], \tag{7.92}$$

which reduces to

$$\hat{y} = 5.3754 + 3.0118x_1 - 1.2855x_2, \tag{7.93}$$

the same as (7.88). If we regress y (rather than $y - \hat{y}$) on $x_2 - \hat{x}_2(x_1)$, we will still obtain $-1.2855x_2$, but we will not have $5.3754 + 3.0118x_1$.

The correlation between the residuals $y - \hat{y}(x_1)$ and $x_2 - \hat{x}_2(x_1)$ is the same as the (sample) partial correlation of y and x_2 with x_1 held fixed:

$$r_{y2\cdot 1} = r_{y-\hat{y},\, x_2-\hat{x}_2}. \tag{7.94}$$

This is discussed further in Section 10.8.

We now consider the general case with full model

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$$

and reduced model

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1^* + \boldsymbol{\varepsilon}^*.$$

We use an orthogonalization approach to obtain an estimator of $\boldsymbol{\beta}_2$, following the same three steps as in the illustration with x_1 and x_2 above:

1. Regress $\mathbf{y}$ on $\mathbf{X}_1$ and calculate residuals $\mathbf{y} - \hat{\mathbf{y}}(\mathbf{X}_1)$, where $\hat{\mathbf{y}}(\mathbf{X}_1) = \mathbf{X}_1\hat{\boldsymbol{\beta}}_1^* = \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}$ [see (7.11)].
2. Regress the columns of $\mathbf{X}_2$ on $\mathbf{X}_1$ and obtain residuals $\mathbf{X}_{2\cdot 1} = \mathbf{X}_2 - \hat{\mathbf{X}}_2(\mathbf{X}_1)$. If $\mathbf{X}_2$ is written in terms of its columns as $\mathbf{X}_2 = (\mathbf{x}_{21}, \ldots, \mathbf{x}_{2j}, \ldots, \mathbf{x}_{2p})$, then the

regression coefficient vector for $\mathbf{x}_{2j}$ on $\mathbf{X}_1$ is $\mathbf{b}_j = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{x}_{2j}$, and $\hat{\mathbf{x}}_{2j} = \mathbf{X}_1\mathbf{b}_j = \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{x}_{2j}$. For all columns of $\mathbf{X}_2$, this becomes $\hat{\mathbf{X}}_2(\mathbf{X}_1) = \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2 = \mathbf{X}_1\mathbf{A}$, where $\mathbf{A} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2$ is the alias matrix defined in (7.80). Note that $\mathbf{X}_{2\cdot 1} = \mathbf{X}_2 - \hat{\mathbf{X}}_2(\mathbf{X}_1)$ is orthogonal to $\mathbf{X}_1$:

$$\mathbf{X}_1'\mathbf{X}_{2\cdot 1} = \mathbf{O}. \tag{7.95}$$

Using the alias matrix $\mathbf{A}$, the residual matrix can be expressed as

$$\mathbf{X}_{2\cdot 1} = \mathbf{X}_2 - \hat{\mathbf{X}}_2(\mathbf{X}_1) \tag{7.96}$$

$$= \mathbf{X}_2 - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2 = \mathbf{X}_2 - \mathbf{X}_1\mathbf{A}. \tag{7.97}$$

3. Regress $\mathbf{y} - \hat{\mathbf{y}}(\mathbf{X}_1)$ on $\mathbf{X}_{2\cdot 1} = \mathbf{X}_2 - \hat{\mathbf{X}}_2(\mathbf{X}_1)$. Since $\mathbf{X}_{2\cdot 1}$ is orthogonal to $\mathbf{X}_1$, we obtain the same $\hat{\boldsymbol{\beta}}_2$ as in the full model $\hat{\mathbf{y}} = \mathbf{X}_1\hat{\boldsymbol{\beta}}_1 + \mathbf{X}_2\hat{\boldsymbol{\beta}}_2$. Adapting the notation of (7.91) and (7.92), this can be expressed as

$$\hat{\mathbf{y}}(\mathbf{X}_1, \mathbf{X}_2) - \hat{\mathbf{y}}(\mathbf{X}_1) = \mathbf{X}_{2\cdot 1}\hat{\boldsymbol{\beta}}_2. \tag{7.98}$$

If we substitute $\hat{\mathbf{y}}(\mathbf{X}_1) = \mathbf{X}_1\hat{\boldsymbol{\beta}}_1^*$ and $\mathbf{X}_{2\cdot 1} = \mathbf{X}_2 - \mathbf{X}_1\mathbf{A}$ into (7.98) and use $\hat{\boldsymbol{\beta}}_1^* = \hat{\boldsymbol{\beta}}_1 + \mathbf{A}\hat{\boldsymbol{\beta}}_2$ from (7.80), we obtain

$$\begin{aligned}\hat{\mathbf{y}}(\mathbf{X}_1, \mathbf{X}_2) &= \mathbf{X}_1\hat{\boldsymbol{\beta}}_1^* + (\mathbf{X}_2 - \mathbf{X}_1\mathbf{A})\hat{\boldsymbol{\beta}}_2 \\ &= \mathbf{X}_1(\hat{\boldsymbol{\beta}}_1 + \mathbf{A}\hat{\boldsymbol{\beta}}_2) + (\mathbf{X}_2 - \mathbf{X}_1\mathbf{A})\hat{\boldsymbol{\beta}}_2 \\ &= \mathbf{X}_1\hat{\boldsymbol{\beta}}_1 + \mathbf{X}_2\hat{\boldsymbol{\beta}}_2\end{aligned}$$

which is analogous to (7.93). This confirms that the orthogonality of $\mathbf{X}_1$ and $\mathbf{X}_{2\cdot 1}$ leads to the estimator $\hat{\boldsymbol{\beta}}_2$ in (7.98). For a formal proof, see Problem 7.50.

PROBLEMS

7.1 Show that $\sum_{i=1}^{n}(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}})^2 = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$, thus verifying (7.7).

7.2 Show that (7.10) follows from (7.9). Why is $\mathbf{X}'\mathbf{X}$ positive definite, as noted below (7.10)?

7.3 Show that $\hat{\beta}_0$ and $\hat{\beta}_1$ in (7.12) in Example 7.3.1 are the same as in (6.5) and (6.6).

7.4 Obtain $\text{cov}(\hat{\boldsymbol{\beta}})$ in (7.16) from (7.15).

7.5 Show that $\text{var}(\hat{\beta}_0) = \sigma^2(\sum_i x_i^2/n)/\sum_i (x_i - \bar{x})^2$ in (7.16) in Example 7.3.2a is the same as $\text{var}(\hat{\beta}_0)$ in (6.10).

7.6 Show that $\mathbf{AA}'$ can be expressed as $\mathbf{AA}' = [\mathbf{A} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'][\mathbf{A} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']' + (\mathbf{X}'\mathbf{X})^{-1}$ as in (7.17) in Theorem 7.3d.

7.7 Prove Corollary 1 to Theorem 7.3d in the following two ways:

(**a**) Use an approach similar to the proof of Theorem 7.3d.
(**b**) Use the method of Lagrange multipliers (Section 2.14.3).

7.8 Show that if the x's are rescaled as $z_j = c_j x_j$, $j = 1, 2, \ldots, k$, then $\hat{\boldsymbol{\beta}}_z = \mathbf{D}^{-1}\hat{\boldsymbol{\beta}}$, as in (7.18) in the proof of the Theorem 7.3e.

7.9 Verify (7.20) and (7.21) in the proof of Corollary 1 to Theorem 7.3e.

7.10 Show that s^2 is invariant to changes of scale on the x's, as noted following Corollary 1 to Theorem 7.3e.

7.11 Show that $(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}$ as in (7.24).

7.12 Show that $E(\text{SSE}) = \sigma^2(n - k - 1)$, as in Theorem 7.3f, using the following approach. Show that $\text{SSE} = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}$. Show that $E(\mathbf{y}'\mathbf{y}) = n\sigma^2 + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}$ and that $E(\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}) = (k+1)\sigma^2 + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}$.

7.13 Prove that an additional x reduces SSE, as noted following Theorem 7.3f.

7.14 Show that the noncentered model preceding (7.30) can be written in the centered form in (7.30), with α defined as in (7.31).

7.15 Show that $\mathbf{X}_c = [\mathbf{I} - (1/n)\mathbf{J}]\mathbf{X}_1$ as in (7.33), where $\mathbf{X}_1$ is as given in (7.19).

7.16 Show that $\mathbf{j}'\mathbf{X}_c = \mathbf{0}'$, as in (7.35), where $\mathbf{X}_c$ is the centered $\mathbf{X}$ matrix defined in (7.33).

7.17 Show that the estimators $\hat{\alpha} = \bar{y}$ and $\hat{\boldsymbol{\beta}}_1 = (\mathbf{X}_c'\mathbf{X}_c)^{-1}\mathbf{X}_c'\mathbf{y}$ in (7.36) and (7.37) are the same as $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ in (7.6). Use the following two methods:

(**a**) Work with the normal equations in both cases.
(**b**) Use the inverse of $\mathbf{X}'\mathbf{X}$ in partitioned form: $(\mathbf{X}'\mathbf{X})^{-1} = [(\mathbf{j}, \mathbf{X}_1)'(\mathbf{j}, \mathbf{X}_1)]^{-1}$.

7.18 Show that the fitted regression plane $\hat{y} = \hat{\alpha} + \hat{\beta}_1(x_1 - \bar{x}_1) + \cdots + \hat{\beta}_k(x_k - \bar{x}_k)$ passes through the point $(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k, \bar{y})$, as noted below (7.38).

7.19 Show that $\text{SSE} = \sum_i (y_i - \bar{y})^2 - \hat{\boldsymbol{\beta}}_1'\mathbf{X}_c'\mathbf{y}$ in (7.39) is the same as $\text{SSE} = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}$ in (7.24).

7.20 (**a**) Show that $\mathbf{S}_{xx} = \mathbf{X}_c'\mathbf{X}_c/(n-1)$ as in (7.44).

(**b**) Show that $\mathbf{s}_{yx} = \mathbf{X}_c'\mathbf{y}/(n-1)$ as in (7.45).

7.21 (**a**) Show that if $\mathbf{y}$ is $N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, the likelihood function is

$$L(\boldsymbol{\beta}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})/2\sigma^2},$$

as in (7.50) in the proof of Theorem 7.6a.

(**b**) Differentiate $\ln L(\boldsymbol{\beta}, \sigma^2)$ in (7.51) with respect to $\boldsymbol{\beta}$ to obtain $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ in (7.48).

(**c**) Differentiate $\ln L(\boldsymbol{\beta}, \sigma^2)$ with respect to σ^2 to obtain $\hat{\sigma}^2 = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})/n$ as in (7.49).

7.22 Prove parts (ii) and (iii) of Theorem 7.6b.

7.23 Show that $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ as in (7.52) in the proof of Theorem 7.6c.

7.24 Explain why $f(\mathbf{y}; \boldsymbol{\beta}, \sigma^2)$ does not factor into $g_1(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})g_2(\hat{\sigma}^2, \sigma^2)h(\mathbf{y})$, as noted following Theorem 7.6c.

7.25 Verify the equivalence of (7.55) and (7.56); that is, show that $\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - n\bar{y}^2 = \hat{\boldsymbol{\beta}}_1'\mathbf{X}_c'\mathbf{X}_c\hat{\boldsymbol{\beta}}_1$.

7.26 Verify the comments in property 1 in Section 7.7, namely, that if $\hat{\beta}_1 = \hat{\beta}_2 = \cdots = \hat{\beta}_k = 0$, then $R^2 = 0$, and if $y_i = \hat{y}_i$, $i = 1, 2, \ldots, n$, then $R^2 = 1$.

7.27 Show that adding an x to the model increases (cannot decrease) the value of R^2, as in property 3 in Section 7.7.

7.28 (**a**) Verify that R^2 is invariant to full-rank linear transformations on the x's as in property 6 in Section 7.7.

(**b**) Show that R^2 is invariant to a scale change $\mathbf{z} = c\mathbf{y}$ on $\mathbf{y}$.

7.29 (**a**) Show that R^2 in (7.55) can be written in the form $R^2 = 1 - \text{SSE}/\sum_i (y_i - \bar{y})^2$.

(**b**) Replace SSE and $\sum_i (y_i - \bar{y})^2$ in part (a) by variance estimators $\text{SSE}/(n-k-1)$ and $\sum_i (y_i - \bar{y})^2/(n-1)$ and show that the result is the same as R_a^2 in (7.56).

7.30 Show that $\sum_{i=1}^n \hat{y}_i/n = \sum_{i=1}^n y_i/n$, as noted following (7.59) in Section 7.7.

7.31 Show that $\cos\theta = R$ as in (7.61), where R^2 is as given by (7.56).

7.32 (**a**) Show that $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$, where $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$ as in (7.63).

(**b**) Show that $\text{cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$ as in (7.64).

7.33 (**a**) Show that the two forms of s^2 in (7.65) and (7.66) are equal.

(**b**) Show that $E(s^2) = \sigma^2$, where s^2 is as given by (7.66).

7.34 Complete the steps in the proof of Theorem 7.8b.

7.35 Show that for $\mathbf{V} = (1-\rho)\mathbf{I} + \rho\mathbf{J}$ in (7.67), the inverse is given by $\mathbf{V}^{-1} = a(\mathbf{I} - b\rho\mathbf{J})$ as in (7.68), where $a = 1/(1-\rho)$ and $b = 1/[1+(n-1)\rho]$.

7.36 (**a**) Show that $\mathbf{X}'\mathbf{V}^{-1}\mathbf{X} = \begin{pmatrix} bn & \mathbf{0}' \\ \mathbf{0} & a\mathbf{X}_c'\mathbf{X}_c \end{pmatrix}$ as in (7.69).

(**b**) Show that $\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} = \begin{pmatrix} bn\bar{y} \\ a\mathbf{X}_c'\mathbf{y} \end{pmatrix}$ as in (7.70).

7.37 Show that $\text{cov}(\hat{\boldsymbol{\beta}}^*) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$ as in (7.72), where $\hat{\boldsymbol{\beta}}^* = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ and $\text{cov}(\mathbf{y}) = \sigma^2\mathbf{V}$.

7.38 (**a**) Show that the weighted least-squares estimator $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)'$ for the model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ with $\text{var}(y_i) = \sigma^2 x_i$ has the form given in (7.73).

(**b**) Verify the expression for $\text{cov}(\hat{\boldsymbol{\beta}})$ in (7.74).

7.39 Obtain the expression for $\text{cov}(\hat{\boldsymbol{\beta}}^*)$ in (7.75).

7.40 As an alternative derivation of $\text{var}(\hat{\beta}_1^*)$ in (7.76), use the following two steps to find $\text{var}(\hat{\beta}_1^*)$ using $\hat{\beta}_1^* = \sum_i (x_i - \bar{x})y_i / \sum_i (x_i - \bar{x})^2$ from the answer to Problem 6.2:

(**a**) Using $\text{var}(y_i) = \sigma^2 x_i$, show that $\text{var}(\hat{\beta}_1^*) = \sigma^2 \sum_i (x_i - \bar{x})^2 x_i / \left[\sum_i (x_i - \bar{x})^2\right]^2$.

(**b**) Show that this expression for $\text{var}(\hat{\beta}_1^*)$ is equal to that in (7.76).

7.41 Using $x = 2, 3, 5, 7, 8, 10$, compare $\text{var}(\hat{\beta}_1^*)$ in (7.76) with $\text{var}(\hat{\beta}_1)$ in (7.77).

7.42 Provide an alternative proof of $\text{cov}(\hat{\boldsymbol{\beta}}_1^*) = \sigma^2(\mathbf{X}_1'\mathbf{X}_1)^{-1}$ in (7.81) using the definition in (3.24), $\text{cov}(\hat{\boldsymbol{\beta}}_1^*) = E\{[\hat{\boldsymbol{\beta}}_1^* - E(\hat{\boldsymbol{\beta}}_1^*)][\hat{\boldsymbol{\beta}}_1^* - E(\hat{\boldsymbol{\beta}}_1^*)]'\}$.

7.43 Prove Theorem 7.9b.

7.44 Provide the missing steps in the proof of Theorem 7.9c(ii).

7.45 Show that $\mathbf{x}_{01}\hat{\boldsymbol{\beta}}_1^*$ is biased for estimating $\mathbf{x}_{01}\boldsymbol{\beta}_1$ if $\boldsymbol{\beta}_2 \neq \mathbf{0}$ and $\mathbf{X}_1'\mathbf{X}_2 \neq \mathbf{O}$.

7.46 Show that $\text{var}(\mathbf{x}_{01}\hat{\boldsymbol{\beta}}_1) \geq \text{var}(\mathbf{x}_{01}\hat{\boldsymbol{\beta}}_1^*)$.

7.47 Complete the steps in the proof of Theorem 7.9d.

7.48 Show that for the no-intercept model $y_i = \beta_1^* x_i + \varepsilon_i^*$, the least-squares estimator is $\hat{\beta}_1^* = \sum_i x_i y_i / \sum_i x_i^2$ as in (7.86).

7.49 Obtain $E(\hat{\beta}_1^*) = \beta_0 \sum_i x_i / \sum_i x_i^2 + \beta_1$ in (7.87) using (7.80), $E(\hat{\boldsymbol{\beta}}_1^*) = \boldsymbol{\beta}_1 + \mathbf{A}\boldsymbol{\beta}_2$.

7.50 Suppose that we use the model $y_i = \beta_0^* + \beta_1^* x_i + \varepsilon_i^*$ when the true model is $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \varepsilon_i$.

(**a**) Using (7.80), find $E(\hat{\beta}_0^*)$ and $E(\hat{\beta}_1^*)$ if observations are taken at $x = -3, \ -2, \ -1, 0, 1, 2, 3$.

(**b**) Using (7.85), find $E(s_1^2)$ for the same values of x.

7.51 Show that $\mathbf{X}_{2\cdot 1} = \mathbf{X}_2 - \hat{\mathbf{X}}_2(\mathbf{X}_1)$ is orthogonal to $\mathbf{X}_1$, that is, $\mathbf{X}_1'\mathbf{X}_{2\cdot 1} = \mathbf{O}$, as in (7.95).

7.52 Show that $\hat{\boldsymbol{\beta}}_2$ in (7.98) is the same as in the full fitted model $\hat{\mathbf{y}} = \mathbf{X}_1\hat{\boldsymbol{\beta}}_1 + \mathbf{X}_2\hat{\boldsymbol{\beta}}_2$.

7.53 When gasoline is pumped into the tank of a car, vapors are vented into the atmosphere. An experiment was conducted to determine whether y, the amount of vapor, can be predicted using the following four variables based on initial conditions of the tank and the dispensed gasoline:

$$x_1 = \text{tank temperature } (^\circ\text{F})$$
$$x_2 = \text{gasoline temperature } (^\circ\text{F})$$
$$x_3 = \text{vapor pressure in tank (psi)}$$
$$x_4 = \text{vapor pressure of gasoline (psi)}$$

The data are given in Table 7.3 (Weisberg 1985, p. 138).

(**a**) Find $\hat{\boldsymbol{\beta}}$ and s^2.

(**b**) Find an estimate of $\text{cov}(\hat{\boldsymbol{\beta}})$.

(**c**) Find $\hat{\boldsymbol{\beta}}_1$ and $\hat{\beta}_0$ using $\mathbf{S}_{xx}$ and $\mathbf{s}_{yx}$ as in (7.46) and (7.47).

(**d**) Find R^2 and R_a^2.

7.54 In an effort to obtain maximum yield in a chemical reaction, the values of the following variables were chosen by the experimenter:

$$x_1 = \text{temperature } (^\circ\text{C})$$
$$x_2 = \text{concentration of a reagent } (\%)$$
$$x_3 = \text{time of reaction (hours)}$$

Two different response variables were observed:

$$y_1 = \text{percent of unchanged starting material}$$
$$y_2 = \text{percent converted to the desired product}$$

TABLE 7.3 Gas Vapor Data

y	x_1	x_2	x_3	x_4	y	x_1	x_2	x_3	x_4
29	33	53	3.32	3.42	40	90	64	7.32	6.70
24	31	36	3.10	3.26	46	90	60	7.32	7.20
26	33	51	3.18	3.18	55	92	92	7.45	7.45
22	37	51	3.39	3.08	52	91	92	7.27	7.26
27	36	54	3.20	3.41	29	61	62	3.91	4.08
21	35	35	3.03	3.03	22	59	42	3.75	3.45
33	59	56	4.78	4.57	31	88	65	6.48	5.80
34	60	60	4.72	4.72	45	91	89	6.70	6.60
32	59	60	4.60	4.41	37	63	62	4.30	4.30
34	60	60	4.53	4.53	37	60	61	4.02	4.10
20	34	35	2.90	2.95	33	60	62	4.02	3.89
36	60	59	4.40	4.36	27	59	62	3.98	4.02
34	60	62	4.31	4.42	34	59	62	4.39	4.53
23	60	36	4.27	3.94	19	37	35	2.75	2.64
24	62	38	4.41	3.49	16	35	35	2.59	2.59
32	62	61	4.39	4.39	22	37	37	2.73	2.59

The data are listed in Table 7.4 (Box and Youle 1955, Andrews and Herzberg 1985, p. 188). Carry out the following for y_1:

(**a**) Find $\hat{\boldsymbol{\beta}}$ and s^2.

(**b**) Find an estimate of cov($\hat{\boldsymbol{\beta}}$).

TABLE 7.4 Chemical Reaction Data

y_1	y_2	x_1	x_2	x_3
41.5	45.9	162	23	3
33.8	53.3	162	23	8
27.7	57.5	162	30	5
21.7	58.8	162	30	8
19.9	60.6	172	25	5
15.0	58.0	172	25	8
12.2	58.6	172	30	5
4.3	52.4	172	30	8
19.3	56.9	167	27.5	6.5
6.4	55.4	177	27.5	6.5
37.6	46.9	157	27.5	6.5
18.0	57.3	167	32.5	6.5
26.3	55.0	167	22.5	6.5
9.9	58.9	167	27.5	9.5
25.0	50.3	167	27.5	3.5
14.1	61.1	177	20	6.5
15.2	62.9	177	20	6.5
15.9	60.0	160	34	7.5
19.6	60.6	160	34	7.5

TABLE 7.5 Land Rent Data

y	x_1	x_2	x_3	y	x_1	x_2	x_3
18.38	15.50	17.25	.24	8.50	9.00	8.89	.08
20.00	22.29	18.51	.20	36.50	20.64	23.81	.24
11.50	12.36	11.13	.12	60.00	81.40	4.54	.05
25.00	31.84	5.54	.12	16.25	18.92	29.62	.72
52.50	83.90	5.44	.04	50.00	50.32	21.36	.19
82.50	72.25	20.37	.05	11.50	21.33	1.53	.10
25.00	27.14	31.20	.27	35.00	46.85	5.42	.08
30.67	40.41	4.29	.10	75.00	65.94	22.10	.09
12.00	12.42	8.69	.41	31.56	38.68	14.55	.17
61.25	69.42	6.63	.04	48.50	51.19	7.59	.13
60.00	48.46	27.40	.12	77.50	59.42	49.86	.13
57.50	69.00	31.23	.08	21.67	24.64	11.46	.21
31.00	26.09	28.50	.21	19.75	26.94	2.48	.10
60.00	62.83	29.98	.17	56.00	46.20	31.62	.26
72.50	77.06	13.59	.05	25.00	26.86	53.73	.43
60.33	58.83	45.46	.16	40.00	20.00	40.18	.56
49.75	59.48	35.90	.32	56.67	62.52	15.89	.05

(**c**) Find R^2 and R_a^2.

(**d**) In order to find the maximum yield for y_1, a second-order model is of interest. Find $\hat{\boldsymbol{\beta}}$ and s^2 for the model $y_1 = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1^2 + \beta_5 x_2^2 + \beta_6 x_3^2 + \beta_7 x_1 x_2 + \beta_8 x_1 x_3 + \beta_9 x_2 x_3 + \varepsilon$.

(**e**) Find R^2 and R_a^2 for the second-order model.

7.55 The following variables were recorded for several counties in Minnesota in 1977:

$$y = \text{average rent paid per acre of land with alfalfa}$$
$$x_1 = \text{average rent paid per acre for all land}$$
$$x_2 = \text{average number of dairy cows per square mile}$$
$$x_3 = \text{proportion of farmland in pasture}$$

The data for 34 counties are given in Table 7.5 (Weisberg 1985, p. 162). Can rent for alfalfa land be predicted from the other three variables?

(**a**) Find $\hat{\boldsymbol{\beta}}$ and s^2.

(**b**) Find $\hat{\boldsymbol{\beta}}_1$ and $\hat{\beta}_0$ using $\mathbf{S}_{xx}$ and $\mathbf{s}_{yx}$ as in (7.46) and (7.47).

(**c**) Find R^2 and R_a^2.