7.3 ESTIMATION OF $\boldsymbol{\beta}$ AND $\boldsymbol{\sigma^2}$

7.3.1 Least-Squares Estimator for $\boldsymbol{\beta}$

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

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

$$= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_k x_{ik})^2. \tag{7.5}$$

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.

↳ 더 간결하게 나타낼 수 있게함.

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

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

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

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

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:

$u = \alpha'\mathbf{x} = \mathbf{x}'\alpha = (x_1 \cdots x_p)\begin{pmatrix} a_1 \\ \vdots \\ a_p \end{pmatrix}$ 일때,

$\left[\frac{\partial u}{\partial \mathbf{x}} = \frac{\partial(\alpha'\mathbf{x})}{\partial \mathbf{x}} = \frac{\partial(\mathbf{x}'\alpha)}{\partial \mathbf{x}} = \alpha\right]$

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

$u = \mathbf{x}'A\mathbf{x}$, A: 대칭행렬

$\left[\frac{\partial u}{\partial \mathbf{x}} = \frac{\partial(\mathbf{x}'A\mathbf{x})}{\partial \mathbf{x}} = 2A\mathbf{x}\right]$

This gives the *normal equations*

정규 방정식

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}. \tag{7.8}$$

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

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

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.

① **Theorem 7.3b.** If $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$, then $\hat{\boldsymbol{\beta}}$ is an unbiased estimator for $\boldsymbol{\beta}$.

PROOF

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

▨

② **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

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

▨

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

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

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

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

Thus

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

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

We found $\operatorname{var}(\hat{\beta}_0)$ and $\operatorname{var}(\hat{\beta}_1)$ in Section 6.2 but did not obtain $\operatorname{cov}(\hat{\beta}_0, \hat{\beta}_1)$. Note that if $\bar{x} > 0$, then $\operatorname{cov}(\hat{\beta}_0, \hat{\beta}_1)$ 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, $\operatorname{cov}(\hat{\boldsymbol{\beta}})$ is given by

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

The negative value of $\operatorname{cov}(\hat{\beta}_1, \hat{\beta}_2) = -.111$ 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.

③ In addition to $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ and $\operatorname{cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$, 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 $\operatorname{cov}(\mathbf{y}) = \sigma^2\mathbf{I}$, 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

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

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_i'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

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

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, $i = 1, 2, \ldots, n$. 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 $\text{SSE} = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}$. 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:

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

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

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