Welcome to Deep Institute

ISS COACHING

Learn with DEEP Institute

Dear Students,

This Institute is dedicated to cater the needs of students preparing for Indian Statistical Service. We publish videos on Youtube channel for student help.

Guided by - Sudhir Sir 9999001310

Sudhirdse1@gmail.com

www.isscoaching.com

2513, Basement, Hudson Lane Beside HDFC Bank Opp. Laxmi Dairy, GTB Nagar New Delhi: 110009

I.S.S. LINEAR MODEL

GENERALIZED INVERSE

We now consider generalized inverses of those matrices that do not have inverses in the usual sense [see (2.45)]. A solution of a consistent system of equations $\mathbf{Ax} = \mathbf{c}$ can be expressed in terms of a generalized inverse of $\mathbf{A}$.

Definition and Properties

A *generalized inverse* of an $n \times p$ matrix $\mathbf{A}$ is any matrix $\mathbf{A}^-$ that satisfies

$$\mathbf{A}\mathbf{A}^-\mathbf{A} = \mathbf{A}. \tag{2.58}$$

A generalized inverse is not unique except when $\mathbf{A}$ is nonsingular, in which case $\mathbf{A}^- = \mathbf{A}^{-1}$. A generalized inverse is also called a *conditional inverse*.

Every matrix, whether square or rectangular, has a generalized inverse. This holds even for vectors. For example, let

$$\mathbf{x} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}.$$

Then $\mathbf{x}_1^- = (1,0,0,0)$ is a generalized inverse of $\mathbf{x}$ satisfying (2.58). Other examples are $\mathbf{x}_2^- = (0, \frac{1}{2}, 0, 0)$, $\mathbf{x}_3^- = (0, 0, \frac{1}{3}, 0)$, and $\mathbf{x}_4^- = (0, 0, 0, \frac{1}{4})$. For each $\mathbf{x}_i^-$, we have

$$\mathbf{x}\mathbf{x}_i^-\mathbf{x} = \mathbf{x}1 = \mathbf{x}, \quad i = 1, 2, 3, 4.$$

In this illustration, $\mathbf{x}$ is a column vector and $\mathbf{x}_i^-$ is a row vector. This pattern is generalized in the following theorem.

Theorem 2.8a. If $\mathbf{A}$ is $n \times p$, any generalized inverse $\mathbf{A}^-$ is $p \times n$. □

In the following example we give two illustrations of generalized inverses of a singular matrix.

Theorem 2.8b. Suppose $\mathbf{A}$ is $n \times p$ of rank r and that $\mathbf{A}$ is partitioned as

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix},$$

where $\mathbf{A}_{11}$ is $r \times r$ of rank r. Then a generalized inverse of $\mathbf{A}$ is given by

$$\mathbf{A}^- = \begin{pmatrix} \mathbf{A}_{11}^{-1} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{pmatrix},$$

where the three $\mathbf{O}$ matrices are of appropriate sizes so that $\mathbf{A}^-$ is $p \times n$.

Corollary 1. Suppose that $\mathbf{A}$ is $n \times p$ of rank r and that $\mathbf{A}$ is partitioned as in Theorem 2.8b, where $\mathbf{A}_{22}$ is $r \times r$ of rank r. Then a generalized inverse of $\mathbf{A}$ is given by

$$\mathbf{A}^- = \begin{pmatrix} \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{A}_{22}^{-1} \end{pmatrix},$$

where the three $\mathbf{O}$ matrices are of appropriate sizes so that $\mathbf{A}^-$ is $p \times n$. □

Theorem 2.8c. Let $\mathbf{A}$ be $n \times p$ of rank r, let $\mathbf{A}^-$ be any generalized inverse of $\mathbf{A}$, and let $(\mathbf{A}'\mathbf{A})^-$ be any generalized inverse of $\mathbf{A}'\mathbf{A}$. Then

(i) $\operatorname{rank}(\mathbf{A}^-\mathbf{A}) = \operatorname{rank}(\mathbf{A}\mathbf{A}^-) = \operatorname{rank}(\mathbf{A}) = r$.

(ii) $(\mathbf{A}^-)'$ is a generalized inverse of $\mathbf{A}'$; that is, $(\mathbf{A}')^- = (\mathbf{A}^-)'$.

(iii) $\mathbf{A} = \mathbf{A}(\mathbf{A}'\mathbf{A})^-\mathbf{A}'\mathbf{A}$ and $\mathbf{A}' = \mathbf{A}'\mathbf{A}(\mathbf{A}'\mathbf{A})^-\mathbf{A}'$.

(iv) $(\mathbf{A}'\mathbf{A})^-\mathbf{A}'$ is a generalized inverse of $\mathbf{A}$; that is, $\mathbf{A}^- = (\mathbf{A}'\mathbf{A})^-\mathbf{A}'$.

(v) $\mathbf{A}(\mathbf{A}'\mathbf{A})^-\mathbf{A}'$ is symmetric, has rank $= r$, and is invariant to the choice of $(\mathbf{A}'\mathbf{A})^-$; that is, $\mathbf{A}(\mathbf{A}'\mathbf{A})^-\mathbf{A}'$ remains the same, no matter what value of $(\mathbf{A}'\mathbf{A})^-$ is used. □

Theorem 2.8d. If the system of equations $\mathbf{Ax} = \mathbf{c}$ is consistent and if $\mathbf{A}^-$ is any generalized inverse for $\mathbf{A}$, then $\mathbf{x} = \mathbf{A}^-\mathbf{c}$ is a solution.

PROOF. Since $\mathbf{AA}^-\mathbf{A} = \mathbf{A}$, we have

$$\mathbf{AA}^-\mathbf{Ax} = \mathbf{Ax}.$$

Substituting $\mathbf{Ax} = \mathbf{c}$ on both sides, we obtain

$$\mathbf{AA}^-\mathbf{c} = \mathbf{c}.$$

Writing this in the form $\mathbf{A}(\mathbf{A}^-\mathbf{c}) = \mathbf{c}$, we see that $\mathbf{A}^-\mathbf{c}$ is a solution to $\mathbf{Ax} = \mathbf{c}$. □

Different choices of $\mathbf{A}^-$ will result in different solutions for $\mathbf{Ax} = \mathbf{c}$.

Theorem 2.8e. If the system of equations $\mathbf{Ax} = \mathbf{c}$ is consistent, then all possible solutions can be obtained in the following two ways:

(i) Use a specific $\mathbf{A}^-$ in $\mathbf{x} = \mathbf{A}^-\mathbf{c} + (\mathbf{I} - \mathbf{A}^-\mathbf{A})\mathbf{h}$, and use all possible values of the arbitrary vector $\mathbf{h}$.
(ii) Use all possible values of $\mathbf{A}^-$ in $\mathbf{x} = \mathbf{A}^-\mathbf{c}$ if $\mathbf{c} \neq \mathbf{0}$.

Theorem 2.8f. The system of equations $\mathbf{Ax} = \mathbf{c}$ has a solution if and only if for any generalized inverse $\mathbf{A}^-$ of $\mathbf{A}$

$$\mathbf{AA}^-\mathbf{c} = \mathbf{c}.$$

IDEMPOTENT MATRICES

A square matrix $\mathbf{A}$ is said to be *idempotent* if $\mathbf{A}^2 = \mathbf{A}$.

Theorem 2.13a. The only nonsingular idempotent matrix is the identity matrix $\mathbf{I}$.

Theorem 2.13b. If $\mathbf{A}$ is singular, symmetric, and idempotent, then $\mathbf{A}$ is positive semidefinite.

Theorem 2.13c. If $\mathbf{A}$ is an $n \times n$ symmetric idempotent matrix of rank r, then $\mathbf{A}$ has r eigenvalues equal to 1 and $n - r$ eigenvalues equal to 0.

(I.S.S.) Coaching by SUDHIR SIR DEEP INSTITUTE 9560402898

Theorem 2.13d. If $\mathbf{A}$ is symmetric and idempotent of rank r, then $\text{rank}(\mathbf{A}) = \text{tr}(\mathbf{A}) = r$.

Theorem 2.13e. If $\mathbf{A}$ is an $n \times n$ idempotent matrix, $\mathbf{P}$ is an $n \times n$ nonsingular matrix, and $\mathbf{C}$ is an $n \times n$ orthogonal matrix, then

(i) $\mathbf{I} - \mathbf{A}$ is idempotent.
(ii) $\mathbf{A}(\mathbf{I} - \mathbf{A}) = \mathbf{O}$ and $(\mathbf{I} - \mathbf{A})\mathbf{A} = \mathbf{O}$.
(iii) $\mathbf{P}^{-1}\mathbf{A}\mathbf{P}$ is idempotent.
(iv) $\mathbf{C}'\mathbf{A}\mathbf{C}$ is idempotent. (If $\mathbf{A}$ is symmetric, $\mathbf{C}'\mathbf{A}\mathbf{C}$ is a symmetric idempotent matrix.) □

Theorem 2.13f. Let $\mathbf{A}$ be $n \times p$ of rank r, let $\mathbf{A}^-$ be any generalized inverse of $\mathbf{A}$, and let $(\mathbf{A}'\mathbf{A})^-$ be any generalized inverse of $\mathbf{A}'\mathbf{A}$. Then $\mathbf{A}^-\mathbf{A}, \mathbf{A}\mathbf{A}^-$, and $\mathbf{A}(\mathbf{A}'\mathbf{A})^-\mathbf{A}'$ are all idempotent. □

Theorem 2.13g. Suppose that the $n \times n$ symmetric matrix $\mathbf{A}$ can be written as $\mathbf{A} = \sum_{i=1}^{k} \mathbf{A}_i$ for some k, where each $\mathbf{A}_i$ is an $n \times n$ symmetric matrix. Then any two of the following conditions implies the third condition.

(i) $\mathbf{A}$ is idempotent.
(ii) Each of $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_k$ is idempotent.
(iii) $\mathbf{A}_i\mathbf{A}_j = \mathbf{O}$ for $i \neq j$.

□

Theorem 2.13h. If $\mathbf{I} = \sum_{i=1}^{k} \mathbf{A}_i$, where each $n \times n$ matrix $\mathbf{A}_i$ is symmetric of rank r_i, and if $n = \sum_{i=1}^{k} r_i$, then both of the following are true:

(i) Each of $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_k$ is idempotent.
(ii) $\mathbf{A}_i\mathbf{A}_j = \mathbf{O}$ for $i \neq j$.

□

Derivatives of Functions of Vectors and Matrices

Let $u = f(\mathbf{x})$ be a function of the variables $x_1, x_2, \ldots, x_p$ in $\mathbf{x} = (x_1, x_2, \ldots, x_p)'$, and let $\partial u/\partial x_1, \partial u/\partial x_2, \ldots, \partial u/\partial x_p$ be the partial derivatives. We define $\partial u/\partial \mathbf{x}$ as

$$\frac{\partial u}{\partial \mathbf{x}} = \begin{pmatrix} \dfrac{\partial u}{\partial x_1} \\ \dfrac{\partial u}{\partial x_2} \\ \vdots \\ \dfrac{\partial u}{\partial x_p} \end{pmatrix}. \tag{2.111}$$

Two specific functions of interest are $u = \mathbf{a}'\mathbf{x}$ and $u = \mathbf{x}'\mathbf{A}\mathbf{x}$. Their derivatives with respect to $\mathbf{x}$ are given in the following two theorems.

Theorem 2.14a. Let $u = \mathbf{a}'\mathbf{x} = \mathbf{x}'\mathbf{a}$, where $\mathbf{a}' = (a_1, a_2, \ldots, a_p)$ is a vector of constants. Then

$$\frac{\partial u}{\partial \mathbf{x}} = \frac{\partial(\mathbf{a}'\mathbf{x})}{\partial \mathbf{x}} = \frac{\partial(\mathbf{x}'\mathbf{a})}{\partial \mathbf{x}} = \mathbf{a}. \tag{2.112}$$

Theorem 2.14b. Let $u = \mathbf{x}'\mathbf{A}\mathbf{x}$, where $\mathbf{A}$ is a symmetric matrix of constants. Then

$$\frac{\partial u}{\partial \mathbf{x}} = \frac{\partial(\mathbf{x}'\mathbf{A}\mathbf{x})}{\partial \mathbf{x}} = 2\mathbf{A}\mathbf{x}. \tag{2.113}$$

Theorem 2.14c. Let $u = \text{tr}(\mathbf{X}\mathbf{A})$, where $\mathbf{X}$ is a $p \times p$ positive definite matrix and $\mathbf{A}$ is a $p \times p$ matrix of constants. Then

$$\frac{\partial u}{\partial \mathbf{X}} = \frac{\partial[\text{tr}(\mathbf{X}\mathbf{A})]}{\partial \mathbf{X}} = \mathbf{A} + \mathbf{A}' - \text{diag}\,\mathbf{A}. \tag{2.115}$$

Theorem 2.14d. Let $u = \ln|\mathbf{X}|$ where $\mathbf{X}$ is a $p \times p$ positive definite matrix. Then

$$\frac{\partial \ln|\mathbf{X}|}{\partial \mathbf{X}} = 2\mathbf{X}^{-1} - \text{diag}(\mathbf{X}^{-1}). \tag{2.116}$$

Simple Linear Regression

6.1 THE MODEL

By (1.1), the *simple linear regression* model for n observations can be written as

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \ldots, n. \tag{6.1}$$

The designation *simple* indicates that there is only one x to predict the response y, and *linear* means that the model (6.1) is linear in β_0 and β_1. [Actually, it is the assumption $E(y_i) = \beta_0 + \beta_1 x_i$ that is linear; see assumption 1 below.] For example, a model such as $y_i = \beta_0 + \beta_1 x_i^2 + \varepsilon_i$ is linear in β_0 and β_1, whereas the model $y_i = \beta_0 + e^{\beta_1 x_i} + \varepsilon_i$ is not linear.

In this chapter, we assume that y_i and ε_i are random variables and that the values of x_i are known constants, which means that the same values of $x_1, x_2, \ldots, x_n$ would be used in repeated sampling. The case in which the x variables are random variables is treated in Chapter 10.

To complete the model in (6.1), we make the following additional assumptions:

1. $E(\varepsilon_i) = 0$ for all $i = 1, 2, \ldots, n$, or, equivalently, $E(y_i) = \beta_0 + \beta_1 x_i$.
2. $\text{var}(\varepsilon_i) = \sigma^2$ for all $i = 1, 2, \ldots, n$, 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 (6.1) is correct, implying that y_i depends only on x_i and that all other variation in y_i is random. Assumption 2 asserts that the variance of ε or y does not depend on the values of x_i. (Assumption 2 is also known as the assumption of *homoscedasticity*, *homogeneous variance* or *constant variance*.) Under assumption 3, the ε variables (or the y variables) are uncorrelated with each other. In Section 6.3, we will add a normality assumption, and the y (or the ε) variables will thereby be independent as well as uncorrelated. Each assumption has been stated in terms of the ε's or the y's. For example, if $\text{var}(\varepsilon_i) = \sigma^2$, then $\text{var}(y_i) = E[y_i - E(y_i)]^2 = E(y_i - \beta_0 - \beta_1 x_i)^2 = E(\varepsilon_i^2) = \sigma^2$.

6.2 ESTIMATION OF β_0, β_1, AND σ^2

Using a random sample of n observations $y_1, y_2, \ldots, y_n$ and the accompanying fixed values $x_1, x_2, \ldots, x_n$, we can estimate the parameters β_0, β_1, and σ^2. To obtain the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, we use the method of least squares, which does not require any distributional assumptions (for maximum likelihood estimators based on normality, see Section 7.6.2).

In the *least-squares* approach, we seek estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the sum of squares of the deviations $y_i - \hat{y}_i$ of the n observed y_i's from their predicted values $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$:

$$\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2. \tag{6.2}$$

Note that the predicted value $\hat{y}_i$ estimates $E(y_i)$, not y_i; that is, $\hat{\beta}_0 + \hat{\beta}_1 x_i$ estimates $\beta_0 + \beta_1 x_i$, not $\beta_0 + \beta_1 x_i + \varepsilon_i$. A better notation would be $\widehat{E(y_i)}$, but $\hat{y}_i$ is commonly used.

To find the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize $\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$ in (6.2), we differentiate with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$ and set the results equal to 0:

$$\frac{\partial \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{\partial \hat{\beta}_0} = -2\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0, \tag{6.3}$$

$$\frac{\partial \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{\partial \hat{\beta}_1} = -2\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i = 0. \tag{6.4}$$

The solution to (6.3) and (6.4) is given by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \tag{6.5}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}. \tag{6.6}$$

Note that the three assumptions in Section 6.1 were not used in deriving the least-squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ in (6.5) and (6.6). It is not necessary that $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be based on $E(y_i) = \beta_0 + \beta_1 x_i$; that is, $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ can be fit to a set of data for which $E(y_i) \neq \beta_0 + \beta_1 x_i$. This is illustrated in Figure 6.2, where a straight line has been fitted to curved data.

However, if the three assumptions in Section 6.1 hold, then the least-squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased and have minimum variance among all linear unbiased

$$E(\hat{\beta}_1) = \beta_1 \tag{6.7}$$

$$E(\hat{\beta}_0) = \beta_0 \tag{6.8}$$

$$\text{var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{6.9}$$

$$\text{var}(\hat{\beta}_0) = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]. \tag{6.10}$$

Note that in discussing $E(\hat{\beta}_1)$ and $\text{var}(\hat{\beta}_1)$, for example, we are considering random variation of $\hat{\beta}_1$ from sample to sample. It is assumed that the n values x_1, $x_2, \ldots, x_n$ would remain the same in future samples so that $\text{var}(\hat{\beta}_1)$ and $\text{var}(\hat{\beta}_0)$ are constant.

In (6.9), we see that $\text{var}(\hat{\beta}_1)$ is minimized when $\sum_{i=1}^{n}(x_i - \bar{x})^2$ is maximized. If the x_i values have the range $a \le x_i \le b$, then $\sum_{i=1}^{n}(x_i - \bar{x})^2$ is maximized if half the x's are selected equal to a and half equal to b (assuming that n is even; see Problem 6.4). In (6.10), it is clear that $\text{var}(\hat{\beta}_0)$ is minimized when $\bar{x} = 0$.

The method of least squares does not yield an estimator of $\text{var}(y_i) = \sigma^2$; minimization of $\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$ yields only $\hat{\beta}_0$ and $\hat{\beta}_1$. To estimate σ^2, we use the definition in (3.6), $\sigma^2 = E[y_i - E(y_i)]^2$. By assumption 2 in Section 6.1, σ^2 is the same for each y_i, $i = 1, 2, \ldots, n$. Using $\hat{y}_i$ as an estimator of $E(y_i)$, we estimate σ^2 by an average from the sample, that is

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} = \frac{\sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{n-2} = \frac{\text{SSE}}{n-2}, \tag{6.11}$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are given by (6.5) and (6.6) and $\text{SSE} = \sum_i (y_i - \hat{y}_i)^2$. The deviation $\hat{\varepsilon}_i = y_i - \hat{y}_i$ is often called the *residual* of y_i, and SSE is called the *residual sum of squares* or *error sum of squares*. With $n-2$ in the denominator, s^2 is an unbiased estimator of σ^2:

$$E(s^2) = \frac{E(\text{SSE})}{n-2} = \frac{(n-2)\sigma^2}{n-2} = \sigma^2. \tag{6.12}$$

Intuitively, we divide by $n-2$ in (6.11) instead of $n-1$ as in $s^2 = \sum_i (y_i - \bar{y})^2/(n-1)$ in (5.6), because $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ has two estimated parameters and should thereby be a better estimator of $E(y_i)$ than $\bar{y}$. Thus we expect $\text{SSE} = \sum_i (y_i - \hat{y}_i)^2$ to be less than $\sum_i (y_i - \bar{y})^2$. In fact, using (6.5) and (6.6), we can write the numerator of (6.11) in the form

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 - \frac{\left[\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})\right]^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \tag{6.13}$$

which shows that $\sum_i (y_i - \hat{y}_i)^2$ is indeed smaller than $\sum_i (y_i - \bar{y})^2$.

6.3 HYPOTHESIS TEST AND CONFIDENCE INTERVAL FOR β_1

Typically, hypotheses about β_1 are of more interest than hypotheses about β_0, since our first priority is to determine whether there is a linear relationship between y and x. (See Problem 6.9 for a test and confidence interval for β_0.) In this section, we consider the hypothesis H_0: $\beta_1 = 0$, which states that there is no linear relationship between y and x in the model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. The hypothesis H_0:$\beta_1 = c$ (for $c \neq 0$) is of less interest.

In order to obtain a test for H_0: $\beta_1 = 0$, we assume that y_i is $N(\beta_0 + \beta_1 x_i, \sigma^2)$. Then $\hat{\beta}_1$ and s^2 have the following properties (these are special cases of results established in Theorem 7.6b in Section 7.6.3):

1. $\hat{\beta}_1$ is $N\left[\beta_1, \sigma^2/\sum_i (x_i - \bar{x})^2\right]$.
2. $(n-2)s^2/\sigma^2$ is $\chi^2(n-2)$.
3. $\hat{\beta}_1$ and s^2 are independent.

From these three properties it follows by (5.29) that

$$t = \frac{\hat{\beta}_1}{s \Big/ \sqrt{\sum_i (x_i - \bar{x})^2}} \tag{6.14}$$

is distributed as $t(n-2, \delta)$, the noncentral t with noncentrality parameter δ. By a comment following (5.29), δ is given by $\delta = E(\hat{\beta}_1)/\sqrt{\text{var}(\hat{\beta}_1)}$ $= \beta_1/[\sigma/\sqrt{\sum_i (x_i - \bar{x})^2}]$. If $\beta_1 = 0$, then by (5.28), t is distributed as $t(n-2)$. For

a two-sided alternative hypothesis $H_1: \beta_1 \neq 0$, we reject $H_0: \beta_1 = 0$ if $|t| \geq t_{\alpha/2,\, n-2}$, where $t_{\alpha/2,\, n-2}$ is the upper $\alpha/2$ percentage point of the central t distribution and α is the desired significance level of the test (probability of rejecting H_0 when it is true). Alternatively, we reject H_0 if $p \leq \alpha$, where p is the p value. For a two-sided test, the p value is defined as twice the probability that $t(n-2)$ exceeds the absolute value of the observed t.

A $100(1-\alpha)\%$ confidence interval for β_1 is given by

$$\hat{\beta}_1 \pm t_{\alpha/2,\, n-2} \frac{s}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}}. \tag{6.15}$$

Confidence intervals are defined and discussed further in Section 8.6. A confidence interval for $E(y)$ and a prediction interval for y are also given in Section 8.6.

6.4 COEFFICIENT OF DETERMINATION

The *coefficient of determination* r^2 is defined as

$$r^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}, \tag{6.16}$$

where $\text{SSR} = \sum_i (\hat{y}_i - \bar{y})^2$ is the regression sum of squares and $\text{SST} = \sum_i (y_i - \bar{y})^2$ is the total sum of squares. The total sum of squares can be partitioned into $\text{SST} = \text{SSR} + \text{SSE}$, that is,

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2. \tag{6.17}$$

Thus r^2 in (6.16) gives the proportion of variation in y that is explained by the model or, equivalently, accounted for by regression on x.

We have labeled (6.16) as r^2 because it is the same as the square of the *sample correlation coefficient* r between y and x

$$r = \frac{s_{xy}}{\sqrt{s_x^2 s_y^2}} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left[\sum_{i=1}^{n} (x_i - \bar{x})^2\right]\left[\sum_{i=1}^{n} (y_i - \bar{y})^2\right]}}, \tag{6.18}$$

where s_{xy} is given by 5.15 (see Problem 6.11). When x is a random variable, r estimates the population correlation in (3.19). The coefficient of determination r^2 is discussed further in Sections 7.7, 10.4, and 10.5.

Multiple Regression: Estimation

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 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.

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

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

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 $x_2, x_3, \ldots, x_k$ 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.

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

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

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

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.

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

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

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

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

Thus

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

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

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

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

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 $\mathbf{z}$.

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

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

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

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

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

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.

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

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

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

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 $\hat{\sigma}^2$

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.

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.

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.

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.

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 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 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 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 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

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \mathbf{y}'\mathbf{y} - n\bar{y}^2 = (\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - n\bar{y}^2) + (\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y})$$

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

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 $0 \le R^2 \le 1$. 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

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

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

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

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

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 $\text{var}(y_i)$. 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 σ^2 when $\text{cov}(\mathbf{y}) = \sigma^2\mathbf{V}$

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).

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

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

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

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.

7.9 MODEL MISSPECIFICATION

In Section 7.8.2, we discussed some consequences of misspecification of cov(**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

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

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 $\text{cov}(\mathbf{y}) = \sigma^2\mathbf{I}$, 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:

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

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

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: $E(\hat{\boldsymbol{\beta}}_1^*) = \boldsymbol{\beta}_1$. □

Multiple Regression: Tests of Hypotheses and Confidence Intervals

In this chapter we consider hypothesis tests and confidence intervals for the parameters $\beta_0, \beta_1, \ldots, \beta_k$ in $\boldsymbol{\beta}$ in the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. We also provide a confidence interval for $\sigma^2 = \text{var}(y_i)$. We will assume throughout the chapter that $\mathbf{y}$ is $N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, where $\mathbf{X}$ is $n \times (k+1)$ of rank $k+1 < n$.

8.1 TEST OF OVERALL REGRESSION

We noted in Section 7.9 that the problems associated with both overfitting and underfitting motivate us to seek an optimal model. Hypothesis testing is a formal tool for, among other things, choosing between a reduced model and an associated full model. The hypothesis H_0, expresses the reduced model in terms of values of a subset of the β_j's in $\boldsymbol{\beta}$. The alternative hypothesis, H_1, is associated with the full model.

To illustrate this tool we begin with a common test, the test of the overall regression hypothesis that none of the x variables predict y. This hypothesis (leading to the reduced model) can be expressed as $H_0 : \boldsymbol{\beta}_1 = \mathbf{0}$, where $\boldsymbol{\beta}_1 = (\beta_1, \beta_2, \ldots, \beta_k)'$. Note that we wish to test $H_0 : \boldsymbol{\beta}_1 = \mathbf{0}$, not $H_0 : \boldsymbol{\beta} = \mathbf{0}$, where

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \boldsymbol{\beta}_1 \end{pmatrix}.$$

Since β_0 is usually not zero, we would rarely be interested in including $\beta_0 = 0$ in the hypothesis. Rejection of H_0: $\boldsymbol{\beta} = \mathbf{0}$ might be due solely to β_0, and we would not learn whether the x variables predict y. For a test of $H_0 : \boldsymbol{\beta} = \mathbf{0}$, see Problem 8.6.

We proceed by proposing a test statistic that is distributed as a central F if H_0 is true and as a noncentral F otherwise. Our approach to obtaining a test statistic is somewhat simplified if we use the centered model (7.32)

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

where $\mathbf{X}_c = [\mathbf{I} - (1/n)\mathbf{J}]\mathbf{X}_1$ is the centered matrix [see (7.33)] and $\mathbf{X}_1$ contains all the columns of $\mathbf{X}$ except the first [see (7.19)]. The corrected total sum of squares $\text{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2$ can be partitioned as

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\hat{\boldsymbol{\beta}}_1'\mathbf{X}_c'\mathbf{y}+\left[\sum_{i=1}^{n}(y_i-\bar{y})^2-\hat{\boldsymbol{\beta}}_1'\mathbf{X}_c'\mathbf{y}\right] \quad \text{[by (7.53)]}$$

$$=\hat{\boldsymbol{\beta}}_1'\mathbf{X}_c'\mathbf{X}_c\hat{\boldsymbol{\beta}}_1+\text{SSE}=\text{SSR}+\text{SSE} \quad \text{[by (7.54)]}, \tag{8.1}$$

where SSE is as given in (7.39). The regression sum of squares $\text{SSR}=\hat{\boldsymbol{\beta}}_1'\mathbf{X}_c'\mathbf{X}_c\hat{\boldsymbol{\beta}}_1$ is clearly due to $\boldsymbol{\beta}_1$.

In order to construct an F test, we first express the sums of squares in (8.1) as quadratic forms in $\mathbf{y}$ so that we can use theorems from Chapter 5 to show that SSR and SSE have chi-square distributions and are independent. Using $\sum_i(y_i-\bar{y})^2=\mathbf{y}'[\mathbf{I}-(1/n)\mathbf{J}]\mathbf{y}$ in (5.2), $\hat{\boldsymbol{\beta}}_1=(\mathbf{X}_c'\mathbf{X}_c)^{-1}\mathbf{X}_c'\mathbf{y}$ in (7.37), and $\text{SSE}=\sum_{i=1}^{n}(y_i-\bar{y})^2-\hat{\boldsymbol{\beta}}_1'\mathbf{X}_c'\mathbf{y}$ in (7.39), we can write (8.1) as

$$\mathbf{y}'\left(\mathbf{I}-\frac{1}{n}\mathbf{J}\right)\mathbf{y}=\text{SSR}+\text{SSE}$$

$$=\mathbf{y}'\mathbf{X}_c(\mathbf{X}_c'\mathbf{X}_c)^{-1}\mathbf{X}_c'\mathbf{y}+\mathbf{y}'\left(\mathbf{I}-\frac{1}{n}\mathbf{J}\right)\mathbf{y}-\mathbf{y}'\mathbf{X}_c(\mathbf{X}_c'\mathbf{X}_c)^{-1}\mathbf{X}_c'\mathbf{y}$$

$$=\mathbf{y}'\mathbf{H}_c\mathbf{y}+\mathbf{y}'\left(\mathbf{I}-\frac{1}{n}\mathbf{J}-\mathbf{H}_c\right)\mathbf{y}, \tag{8.2}$$

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

In the following theorem we establish some properties of the three matrices of the quadratic forms in (8.2).

Theorem 8.1a. The matrices $\mathbf{I}-(1/n)\,\mathbf{J}$, $\mathbf{H}_c=\mathbf{X}_c(\mathbf{X}_c'\mathbf{X}_c)^{-1}\mathbf{X}_c'$, and $\mathbf{I}-(1/n)\,\mathbf{J}-\mathbf{H}_c$ have the following properties:

(i) $\mathbf{H}_c[\mathbf{I}-(1/n)\,\mathbf{J}]=\mathbf{H}_c$. (8.3)

(ii) $\mathbf{H}_c$ is idempotent of rank k.

(iii) $\mathbf{I}-(1/n)\,\mathbf{J}-\mathbf{H}_c$ is idempotent of rank $n-k-1$.

(iv) $\mathbf{H}_c[\mathbf{I}-(1/n)\,\mathbf{J}-\mathbf{H}_c]=\mathbf{O}$. (8.4)

The distributions of SSR/σ^2 and SSE/σ^2 are given in the following theorem.

Theorem 8.1b. If $\mathbf{y}$ is $N_n(\mathbf{X}\boldsymbol{\beta},\,\sigma^2\mathbf{I})$, then $\text{SSR}/\sigma^2=\hat{\boldsymbol{\beta}}_1'\mathbf{X}_c'\mathbf{X}_c\hat{\boldsymbol{\beta}}_1/\sigma^2$ and $\text{SSE}/\sigma^2=\left[\sum_{i=1}^{n}(y_i-\bar{y})^2-\hat{\boldsymbol{\beta}}_1'\mathbf{X}_c'\mathbf{X}_c\hat{\boldsymbol{\beta}}_1\right]/\sigma^2$ have the following distributions:

(i) SSR/σ^2 is $\chi^2(k,\,\lambda_1)$, where $\lambda_1=\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}/2\sigma^2=\boldsymbol{\beta}_1'\mathbf{X}_c'\mathbf{X}_c\boldsymbol{\beta}_1/2\sigma^2$.

(ii) SSE/σ^2 is $\chi^2(n-k-1)$.

Theorem 8.1c. If $\mathbf{y}$ is $N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, then SSR and SSE are independent, where SSR and SSE are defined in (8.1) and (8.2).

Theorem 8.1d. If $\mathbf{y}$ is $N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, the distribution of

$$F = \frac{\text{SSR}/(k\sigma^2)}{\text{SSE}/[(n-k-1)\sigma^2]} = \frac{\text{SSR}/k}{\text{SSE}/(n-k-1)} \tag{8.5}$$

is as follows:

(i) If $H_0: \boldsymbol{\beta}_1 = \mathbf{0}$ is false, then

$$F \text{ is distributed as } F(k, n-k-1, \lambda_1),$$

where $\lambda_1 = \boldsymbol{\beta}_1'\mathbf{X}_c'\mathbf{X}_c\boldsymbol{\beta}_1/2\sigma^2$.

(ii) If $H_0: \boldsymbol{\beta}_1 = \mathbf{0}$ is true, then $\lambda_1 = 0$ and

$$F \text{ is distributed as } F(k, n-k-1).$$

If $H_0: \boldsymbol{\beta}_1 = \mathbf{0}$ is true, both of the expected mean squares in Table 8.1 are equal to σ^2, and we expect F to be near 1. If $\boldsymbol{\beta}_1 \neq \mathbf{0}$, then $E(\text{SSR}/k) > \sigma^2$ since $\mathbf{X}_c'\mathbf{X}_c$ is positive definite, and we expect F to exceed 1. We therefore reject H_0 for large values of F.

ESTIMATION

In this section, we consider various aspects of estimation of $\boldsymbol{\beta}$ in the non-full-rank model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. We do not reparameterize or impose side conditions. These two approaches to estimation are discussed in Sections 12.5 and 12.6, respectively. Normality of $\mathbf{y}$ is not assumed in the present section.

12.2.1 Estimation of β

Consider the model

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

where $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$, $\text{cov}(\mathbf{y}) = \sigma^2\mathbf{I}$, and $\mathbf{X}$ is $n \times p$ of rank $k < p \leq n$. [We will say "$\mathbf{X}$ is $n \times p$ of rank $k < p \leq n$" to indicate that $\mathbf{X}$ is not of full rank; that is, $\text{rank}(\mathbf{X}) < p$ and $\text{rank}(\mathbf{X}) < n$. In some cases, we have $k < n < p$.] In this non-full-rank model, the p parameters in $\boldsymbol{\beta}$ are not unique. We now ascertain whether $\boldsymbol{\beta}$ can be estimated.

Using least-squares, we seek a value of $\hat{\boldsymbol{\beta}}$ that minimizes

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

We can expand $\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$ to obtain

$$\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = \mathbf{y}'\mathbf{y} - 2\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}, \tag{12.10}$$

which can be differentiated with respect to $\hat{\boldsymbol{\beta}}$ and set equal to $\mathbf{0}$ to produce the familiar normal equations

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}. \tag{12.11}$$

Since $\mathbf{X}$ is not full rank, $\mathbf{X}'\mathbf{X}$ has no inverse, and (12.11) does not have a unique solution. However, $\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$ has (an infinite number of) solutions:

Theorem 12.2a. If $\mathbf{X}$ is $n \times p$ of rank $k < p \leq n$, the system of equations $\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$ is consistent.

PROOF. By Theorem 2.8f, the system is consistent if and only if

$$\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{y}, \tag{12.12}$$

where $(\mathbf{X}'\mathbf{X})^-$ is any generalized inverse of $\mathbf{X}'\mathbf{X}$. By Theorem 2.8c(iii), $\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}' = \mathbf{X}'$, and (12.12) therefore holds. (An alternative proof is suggested in Problem 12.3.) □

Since the normal equations $\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$ are consistent, a solution is given by Theorem 2.8d as

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{y}, \tag{12.13}$$

where $(\mathbf{X}'\mathbf{X})^-$ is any generalized inverse of $\mathbf{X}'\mathbf{X}$. For a particular generalized inverse $(\mathbf{X}'\mathbf{X})^-$, the expected value of $\hat{\boldsymbol{\beta}}$ is

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}) &= (\mathbf{X}'\mathbf{X})^-\mathbf{X}'E(\mathbf{y}) \\ &= (\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{X}\boldsymbol{\beta}. \end{aligned} \tag{12.14}$$

Thus, $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $(\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{X}\boldsymbol{\beta}$. Since $(\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{X} \neq \mathbf{I}$, $\hat{\boldsymbol{\beta}}$ is not an unbiased estimator of $\boldsymbol{\beta}$. The expression $(\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{X}\boldsymbol{\beta}$ is not invariant to the choice of $(\mathbf{X}'\mathbf{X})^-$; that is, $E(\hat{\boldsymbol{\beta}})$ is different for each choice of $(\mathbf{X}'\mathbf{X})^-$. [An implication in (12.14) is that having selected a value of $(\mathbf{X}'\mathbf{X})^-$, we would use that same value of $(\mathbf{X}'\mathbf{X})^-$ in repeated sampling.]

Thus, $\hat{\boldsymbol{\beta}}$ in (12.13) does not estimate $\boldsymbol{\beta}$. Next, we inquire as to whether there are any linear functions of $\mathbf{y}$ that are unbiased estimators for the elements of $\boldsymbol{\beta}$; that is, whether there exists a $p \times n$ matrix $\mathbf{A}$ such that $E(\mathbf{A}\mathbf{y}) = \boldsymbol{\beta}$. If so, then

$$\boldsymbol{\beta} = E(\mathbf{A}\mathbf{y}) = E[\mathbf{A}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})] = E(\mathbf{A}\mathbf{X}\boldsymbol{\beta}) + \mathbf{A}E(\boldsymbol{\varepsilon}) = \mathbf{A}\mathbf{X}\boldsymbol{\beta}.$$

Since this must hold for all $\boldsymbol{\beta}$, we have $\mathbf{A}\mathbf{X} = \mathbf{I}_p$ [see (2.44)]. But by Theorem 2.4(i), $\text{rank}(\mathbf{A}\mathbf{X}) < p$ since the rank of $\mathbf{X}$ is less than p. Hence $\mathbf{A}\mathbf{X}$ cannot be equal to $\mathbf{I}_p$, and there are no linear functions of the observations that yield unbiased estimators for the elements of $\boldsymbol{\beta}$.

Example 12.2.1. Consider the model $y_{ij} = \mu + \tau_i + \varepsilon_{ij};\ i = 1, 2;\ j = 1, 2, 3$ in (12.2). The matrix $\mathbf{X}$ and the vector $\boldsymbol{\beta}$ are given in (12.3) as

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \end{pmatrix}.$$

By Theorem 2.2c(i), we obtain

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} 6 & 3 & 3 \\ 3 & 3 & 0 \\ 3 & 0 & 3 \end{pmatrix}.$$

By Corollary 1 to Theorem 2.8b, a generalized inverse of $\mathbf{X}'\mathbf{X}$ is given by

$$(\mathbf{X}'\mathbf{X})^- = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \frac{1}{3} & 0 \\ 0 & 0 & \frac{1}{3} \end{pmatrix}.$$

The vector $\mathbf{X}'\mathbf{y}$ is given by

$$\mathbf{X}'\mathbf{y} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \end{pmatrix} = \begin{pmatrix} y_{..} \\ y_{1.} \\ y_{2.} \end{pmatrix},$$

where $y_{..} = \sum_{i=1}^{2} \sum_{j=1}^{3} y_{ij}$ and $y_{i.} = \sum_{j=1}^{3} y_{ij}$. Then

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{y} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \frac{1}{3} & 0 \\ 0 & 0 & \frac{1}{3} \end{pmatrix} \begin{pmatrix} y_{..} \\ y_{1.} \\ y_{2.} \end{pmatrix} = \begin{pmatrix} 0 \\ \bar{y}_{1.} \\ \bar{y}_{2.} \end{pmatrix},$$

where $\bar{y}_{i.} = \sum_{j=1}^{3} y_{ij}/3 = y_{i.}/3$.

To find $E(\hat{\boldsymbol{\beta}})$, we need $E(\bar{y}_{i.})$. Since $E(\boldsymbol{\varepsilon}) = \mathbf{0}$, we have $E(\varepsilon_{ij}) = 0$. Then

$$E(\bar{y}_{i.}) = E\left(\sum_{j=1}^{3} y_{ij}/3\right) = \tfrac{1}{3}\sum_{j=1}^{3} E(y_{ij})$$

$$= \tfrac{1}{3}\sum_{j=1}^{3} E(\mu + \tau_i + \varepsilon_{ij}) = \tfrac{1}{3}(3\mu + 3\tau_i + 0)$$

$$= \mu + \tau_i.$$

Thus

$$E(\hat{\boldsymbol{\beta}}) = \begin{pmatrix} 0 \\ \mu + \tau_1 \\ \mu + \tau_2 \end{pmatrix}.$$

The same result is obtained in (12.14):

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

$$= \begin{pmatrix} 0 & 0 & 0 \\ 0 & \frac{1}{3} & 0 \\ 0 & 0 & \frac{1}{3} \end{pmatrix} \begin{pmatrix} 6 & 3 & 3 \\ 3 & 3 & 0 \\ 3 & 0 & 3 \end{pmatrix} \begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \end{pmatrix}$$

$$= \begin{pmatrix} 0 \\ \mu + \tau_1 \\ \mu + \tau_2 \end{pmatrix}. \qquad \square$$

12.2.2 Estimable Functions of β

Having established that we cannot estimate $\boldsymbol{\beta}$, we next inquire as to whether we can estimate any linear combination of the $\boldsymbol{\beta}$'s, say, $\boldsymbol{\lambda}'\boldsymbol{\beta}$. For example, in Section 12.1.1, we considered the model $y_{ij} = \mu + \tau_i + \varepsilon_{ij}$, $i = 1, 2$, and found that μ, τ_1, and τ_2 in $\boldsymbol{\beta} = (\mu, \tau_1, \tau_2)'$ are not unique but that the linear function $\tau_1 - \tau_2 = (0, 1, -1)\boldsymbol{\beta}$ is unique. In order to show that functions such as $\tau_1 - \tau_2$ can be estimated, we first give a definition of an estimable function $\boldsymbol{\lambda}'\boldsymbol{\beta}$.

A linear function of parameters $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is said to be *estimable* if there exists a linear combination of the observations with an expected value equal to $\boldsymbol{\lambda}'\boldsymbol{\beta}$; that is, $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable if there exists a vector $\mathbf{a}$ such that $E(\mathbf{a}'\mathbf{y}) = \boldsymbol{\lambda}'\boldsymbol{\beta}$.

In the following theorem we consider three methods for determining whether a particular linear function $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable.

Theorem 12.2b. In the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ and $\mathbf{X}$ is $n \times p$ of rank $k < p \leq n$, the linear function $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable if and only if any one of the following equivalent conditions holds:

(i) $\boldsymbol{\lambda}'$ is a linear combination of the rows of $\mathbf{X}$; that is, there exists a vector $\mathbf{a}$ such that

$$\mathbf{a}'\mathbf{X} = \boldsymbol{\lambda}'. \tag{12.15}$$

(ii) $\boldsymbol{\lambda}'$ is a linear combination of the rows of $\mathbf{X}'\mathbf{X}$ or $\boldsymbol{\lambda}$ is a linear combination of the columns of $\mathbf{X}'\mathbf{X}$, that is, there exists a vector $\mathbf{r}$ such that

$$\mathbf{r}'\mathbf{X}'\mathbf{X} = \boldsymbol{\lambda}' \quad \text{or} \quad \mathbf{X}'\mathbf{X}\mathbf{r} = \boldsymbol{\lambda}. \tag{12.16}$$

(iii) $\boldsymbol{\lambda}$ or $\boldsymbol{\lambda}'$ is such that

$$\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\boldsymbol{\lambda} = \boldsymbol{\lambda} \quad \text{or} \quad \boldsymbol{\lambda}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X} = \boldsymbol{\lambda}', \tag{12.17}$$

where $(\mathbf{X}'\mathbf{X})^{-}$ is any (symmetric) generalized inverse of $\mathbf{X}'\mathbf{X}$.

PROOF. For (ii) and (iii), we prove the "if" part. For (i), we prove both "if" and "only if."

(i) If there exists a vector $\mathbf{a}$ such that $\boldsymbol{\lambda}' = \mathbf{a}'\mathbf{X}$, then, using this vector $\mathbf{a}$, we have

$$E(\mathbf{a}'\mathbf{y}) = \mathbf{a}'E(\mathbf{y}) = \mathbf{a}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\lambda}'\boldsymbol{\beta}.$$

Conversely, if $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable, then there exists a vector $\mathbf{a}$ such that $E(\mathbf{a}'\mathbf{y}) = \boldsymbol{\lambda}'\boldsymbol{\beta}$. Thus $\mathbf{a}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\lambda}'\boldsymbol{\beta}$, which implies, among other things, that $\mathbf{a}'\mathbf{X} = \boldsymbol{\lambda}'$.

(ii) If there exists a solution $\mathbf{r}$ for $\mathbf{X}'\mathbf{X}\mathbf{r} = \boldsymbol{\lambda}$, then, by defining $\mathbf{a} = \mathbf{X}\mathbf{r}$, we obtain

$$\begin{aligned} E(\mathbf{a}'\mathbf{y}) &= E(\mathbf{r}'\mathbf{X}'\mathbf{y}) = \mathbf{r}'\mathbf{X}'E(\mathbf{y}) \\ &= \mathbf{r}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\lambda}'\boldsymbol{\beta}. \end{aligned}$$

(iii) If $\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\boldsymbol{\lambda} = \boldsymbol{\lambda}$, then $(\mathbf{X}'\mathbf{X})^{-}\boldsymbol{\lambda}$ is a solution to $\mathbf{X}'\mathbf{X}\mathbf{r} = \boldsymbol{\lambda}$ in part(ii). (For proof of the converse, see Problem 12.4.) □

We illustrate the use of Theorem 12.2b in the following example.

Example 12.2.2a. For the model $y_{ij} = \mu + \tau_i + \varepsilon_{ij}$; $i = 1, 2$; $j = 1, 2, 3$ in Example 12.2.1, the matrix $\mathbf{X}$ and the vector $\boldsymbol{\beta}$ are given as

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \end{pmatrix}.$$

We noted in Section 12.1.1 that $\tau_1 - \tau_2$ is unique. We now show that $\tau_1 - \tau_2 = (0, 1, -1)\boldsymbol{\beta} = \boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable, using all three conditions of Theorem 12.2b.

(i) To find a vector $\mathbf{a}$ such that $\mathbf{a}'\mathbf{X} = \boldsymbol{\lambda}' = (0, 1, -1)$, consider $\mathbf{a}' = (0, 0, 1, -1, 0, 0)$, which gives

$$\begin{aligned} \mathbf{a}'\mathbf{X} &= (0, 0, 1, -1, 0, 0)\mathbf{X} = (1, 1, 0) - (1, 0, 1) \\ &= (0, 1, -1) = \boldsymbol{\lambda}'. \end{aligned}$$

There are many other choices for $\mathbf{a}$, of course, that will yield $\mathbf{a}'\mathbf{X} = \boldsymbol{\lambda}'$, for example $\mathbf{a}' = (1, 0, 0, 0, 0, -1)$ or $\mathbf{a}' = (2, -1, 0, 0, 1, -2)$. Note that we can likewise obtain $\boldsymbol{\lambda}'\boldsymbol{\beta}$ from $E(\mathbf{y})$:

$$\boldsymbol{\lambda}'\boldsymbol{\beta} = \mathbf{a}'\mathbf{X}\boldsymbol{\beta} = \mathbf{a}'E(\mathbf{y}) = (0, 0, 1, -1, 0, 0)E(\mathbf{y})$$

$$= (0, 0, 1, -1, 0, 0) \begin{pmatrix} E(y_{11}) \\ E(y_{12}) \\ E(y_{13}) \\ E(y_{21}) \\ E(y_{22}) \\ E(y_{23}) \end{pmatrix}$$

$$= E(y_{13}) - E(y_{21}) = \mu + \tau_1 - (\mu + \tau_2) = \tau_1 - \tau_2.$$

(ii) The matrix $\mathbf{X}'\mathbf{X}$ is given in Example 12.2.1 as

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} 6 & 3 & 3 \\ 3 & 3 & 0 \\ 3 & 0 & 3 \end{pmatrix}.$$

To find a vector $\mathbf{r}$ such that $\mathbf{X}'\mathbf{X}\mathbf{r} = \boldsymbol{\lambda} = (0, 1, -1)'$, consider $\mathbf{r} = (0, \frac{1}{3}, -\frac{1}{3})'$, which gives

$$\mathbf{X}'\mathbf{X}\mathbf{r} = \begin{pmatrix} 6 & 3 & 3 \\ 3 & 3 & 0 \\ 3 & 0 & 3 \end{pmatrix} \begin{pmatrix} 0 \\ \frac{1}{3} \\ -\frac{1}{3} \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix} = \boldsymbol{\lambda}.$$

There are other possible values of $\mathbf{r}$, of course, such as $\mathbf{r} = (-\frac{1}{3}, \frac{2}{3}, 0)'$.

(iii) Using the generalized inverse $(\mathbf{X}'\mathbf{X})^- = \text{diag}(0, \frac{1}{3}, \frac{1}{3})$ given in Example 12.2.1, the product $\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^-$ becomes

$$\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^- = \begin{pmatrix} 0 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Then, for $\boldsymbol{\lambda} = (0, 1, -1)'$, we see that $\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^-\boldsymbol{\lambda} = \boldsymbol{\lambda}$ in (12.17) holds:

$$\begin{pmatrix} 0 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix}. \qquad \square$$

A set of functions $\boldsymbol{\lambda}_1'\boldsymbol{\beta}, \boldsymbol{\lambda}_2'\boldsymbol{\beta}, \ldots, \boldsymbol{\lambda}_m'\boldsymbol{\beta}$ is said to be linearly independent if the coefficient vectors $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \ldots, \boldsymbol{\lambda}_m$ are linearly independent [see (2.40)]. The number of linearly independent estimable functions is given in the next theorem.

Theorem 12.2c. In the non-full-rank model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, the number of linearly independent estimable functions of $\boldsymbol{\beta}$ is the rank of $\mathbf{X}$.

From Theorem 12.2b(i), we see that $\mathbf{x}_i'\boldsymbol{\beta}$ is estimable for $i = 1, 2, \ldots, n$, where $\mathbf{x}_i'$ is the ith row of $\mathbf{X}$. Thus every row (element) of $\mathbf{X}\boldsymbol{\beta}$ is estimable, and $\mathbf{X}\boldsymbol{\beta}$ itself can be said to be estimable. Likewise, from Theorem 12.2b(ii), every row (element) of $\mathbf{X}'\mathbf{X}\boldsymbol{\beta}$ is estimable, and $\mathbf{X}'\mathbf{X}\boldsymbol{\beta}$ is therefore estimable. Conversely, all estimable functions can be obtained from $\mathbf{X}\boldsymbol{\beta}$ or $\mathbf{X}'\mathbf{X}\boldsymbol{\beta}$:

12.3 ESTIMATORS

12.3.1 Estimators of $\boldsymbol{\lambda}'\boldsymbol{\beta}$

From Theorem 12.2b(i) and (ii) we have the estimators $\mathbf{a}'\mathbf{y}$ and $\mathbf{r}'\mathbf{X}'\mathbf{y}$ for $\boldsymbol{\lambda}'\boldsymbol{\beta}$, where $\mathbf{a}'$ and $\mathbf{r}'$ satisfy $\boldsymbol{\lambda}' = \mathbf{a}'\mathbf{X}$ and $\boldsymbol{\lambda}' = \mathbf{r}'\mathbf{X}'\mathbf{X}$, respectively. A third estimator of $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is $\boldsymbol{\lambda}'\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is a solution of $\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$. In the following theorem, we discuss some properties of $\mathbf{r}'\mathbf{X}'\mathbf{y}$ and $\boldsymbol{\lambda}'\hat{\boldsymbol{\beta}}$. We do not discuss the estimator $\mathbf{a}'\mathbf{y}$ because it is not guaranteed to have minimum variance (see Theorem 12.3d).

Theorem 12.3a. Let $\boldsymbol{\lambda}'\boldsymbol{\beta}$ be an estimable function of $\boldsymbol{\beta}$ in the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ and $\mathbf{X}$ is $n \times p$ of rank $k < p \leq n$. Let $\hat{\boldsymbol{\beta}}$ be any solution to the normal equations $\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$, and let $\mathbf{r}$ be any solution to $\mathbf{X}'\mathbf{X}\mathbf{r} = \boldsymbol{\lambda}$. Then the two estimators $\boldsymbol{\lambda}'\hat{\boldsymbol{\beta}}$ and $\mathbf{r}'\mathbf{X}'\mathbf{y}$ have the following properties:

(i) $E(\boldsymbol{\lambda}'\hat{\boldsymbol{\beta}}) = E(\mathbf{r}'\mathbf{X}'\mathbf{y}) = \boldsymbol{\lambda}'\boldsymbol{\beta}$.
(ii) $\boldsymbol{\lambda}'\hat{\boldsymbol{\beta}}$ is equal to $\mathbf{r}'\mathbf{X}'\mathbf{y}$ for any $\hat{\boldsymbol{\beta}}$ or any $\mathbf{r}$.
(iii) $\boldsymbol{\lambda}'\hat{\boldsymbol{\beta}}$ and $\mathbf{r}'\mathbf{X}'\mathbf{y}$ are invariant to the choice of $\hat{\boldsymbol{\beta}}$ or $\mathbf{r}$.

PROOF

(i) By (12.14)

$$E(\boldsymbol{\lambda}'\hat{\boldsymbol{\beta}}) = \boldsymbol{\lambda}'E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\lambda}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X}\boldsymbol{\beta}.$$

By Theorem 12.2b(iii), $\boldsymbol{\lambda}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X} = \boldsymbol{\lambda}'$, and $E(\boldsymbol{\lambda}'\hat{\boldsymbol{\beta}})$ becomes

$$E(\boldsymbol{\lambda}'\hat{\boldsymbol{\beta}}) = \boldsymbol{\lambda}'\boldsymbol{\beta}.$$

By Theorem 12.2b(ii)

$$E(\mathbf{r}'\mathbf{X}'\mathbf{y}) = \mathbf{r}'\mathbf{X}'E(\mathbf{y}) = \mathbf{r}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\lambda}'\boldsymbol{\beta}.$$

(ii) By Theorem 12.2b(ii), if $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is estimable, $\boldsymbol{\lambda}' = \mathbf{r}'\mathbf{X}'\mathbf{X}$ for some $\mathbf{r}$. Multiplying the normal equations $\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$ by $\mathbf{r}'$ gives

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

Since $\mathbf{r}'\mathbf{X}'\mathbf{X} = \boldsymbol{\lambda}'$, we have

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

(iii) To show that $\mathbf{r}'\mathbf{X}'\mathbf{y}$ is invariant to the choice of $\mathbf{r}$, let $\mathbf{r}_1$ and $\mathbf{r}_2$ be such that $\mathbf{X}'\mathbf{X}\mathbf{r}_1 = \mathbf{X}'\mathbf{X}\mathbf{r}_2 = \boldsymbol{\lambda}$. Then

$$\mathbf{r}_1'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{r}_1'\mathbf{X}'\mathbf{y} \quad \text{and} \quad \mathbf{r}_2'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{r}_2'\mathbf{X}'\mathbf{y}.$$

Since $\mathbf{r}_1'\mathbf{X}'\mathbf{X} = \mathbf{r}_2'\mathbf{X}'\mathbf{X}$, we have $\mathbf{r}_1'\mathbf{X}'\mathbf{y} = \mathbf{r}_2'\mathbf{X}'\mathbf{y}$. It is clear that each is equal to $\boldsymbol{\lambda}'\hat{\boldsymbol{\beta}}$. (For a direct proof that $\boldsymbol{\lambda}'\hat{\boldsymbol{\beta}}$ is invariant to the choice of $\hat{\boldsymbol{\beta}}$, see Problem 12.6.) □

We illustrate the estimators $\mathbf{r}'\mathbf{X}'\mathbf{y}$ and $\boldsymbol{\lambda}\hat{\boldsymbol{\beta}}$ in the following example.

Example 12.3.1. The linear function $\boldsymbol{\lambda}'\boldsymbol{\beta} = \tau_1 - \tau_2$ was shown to be estimable in Example 12.2.2a. To estimate $\tau_1 - \tau_2$ with $\mathbf{r}'\mathbf{X}'\mathbf{y}$, we use $\mathbf{r}' = (0, \frac{1}{3}, -\frac{1}{3})$ from Example 12.2.2a to obtain

$$\mathbf{r}'\mathbf{X}'\mathbf{y} = \left(0, \tfrac{1}{3}, -\tfrac{1}{3}\right)\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}\begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \end{pmatrix}$$

$$= \left(0, \tfrac{1}{3}, -\tfrac{1}{3}\right)\begin{pmatrix} y_{..} \\ y_{1.} \\ y_{2.} \end{pmatrix} = \frac{y_{1.}}{3} - \frac{y_{2.}}{3} = \bar{y}_{1.} - \bar{y}_{2.},$$

where $y_{..} = \sum_{i=1}^{2}\sum_{j=1}^{3} y_{ij}$, $y_{i.} = \sum_{j=1}^{3} y_{ij}$, and $\bar{y}_{i.} = y_{i.}/3 = \sum_{j=1}^{3} y_{ij}/3$.

To obtain the same result using $\boldsymbol{\lambda}'\hat{\boldsymbol{\beta}}$, we first find a solution to the normal equations $\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$

$$\begin{pmatrix} 6 & 3 & 3 \\ 3 & 3 & 0 \\ 3 & 0 & 3 \end{pmatrix}\begin{pmatrix} \hat{\mu} \\ \hat{\tau}_1 \\ \hat{\tau}_2 \end{pmatrix} = \begin{pmatrix} y_{..} \\ y_{1.} \\ y_{2.} \end{pmatrix}$$

or

$$\begin{aligned} 6\hat{\mu} + 3\hat{\tau}_1 + 3\hat{\tau}_2 &= y_{..} \\ 3\hat{\mu} + 3\hat{\tau}_1 &= y_{1.} \\ 3\hat{\mu} + 3\hat{\tau}_2 &= y_{2.}. \end{aligned}$$

The first equation is redundant since it is the sum of the second and third equations. We can take $\hat{\mu}$ to be an arbitrary constant and obtain

$$\hat{\tau}_1 = \tfrac{1}{3}y_{1.} - \hat{\mu} = \bar{y}_{1.} - \hat{\mu}, \quad \hat{\tau}_2 = \tfrac{1}{3}y_{2.} - \hat{\mu} = \bar{y}_{2.} - \hat{\mu}.$$

Thus

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\mu} \\ \hat{\tau}_1 \\ \hat{\tau}_2 \end{pmatrix} = \begin{pmatrix} 0 \\ \bar{y}_{1.} \\ \bar{y}_{2.} \end{pmatrix} + \hat{\mu}\begin{pmatrix} 1 \\ -1 \\ -1 \end{pmatrix}.$$

To estimate $\tau_1 - \tau_2 = (0, 1, -1)\boldsymbol{\beta} = \boldsymbol{\lambda}'\boldsymbol{\beta}$, we can set $\hat{\mu} = 0$ to obtain $\hat{\boldsymbol{\beta}} = (0, \bar{y}_{1.}, \bar{y}_{2.})'$ and $\boldsymbol{\lambda}'\hat{\boldsymbol{\beta}} = \bar{y}_{1.} - \bar{y}_{2.}$. If we leave $\hat{\mu}$ arbitrary, we likewise obtain

$$\boldsymbol{\lambda}'\hat{\boldsymbol{\beta}} = (0, 1, -1)\begin{pmatrix} \hat{\mu} \\ \bar{y}_{1.} - \hat{\mu} \\ \bar{y}_{2.} - \hat{\mu} \end{pmatrix}$$

$$= \bar{y}_{1.} - \hat{\mu} - (\bar{y}_{2.} - \hat{\mu}) = \bar{y}_{1.} - \bar{y}_{2.}. \qquad \square$$

Since $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{y}$ is not unique for the non-full-rank model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with $\text{cov}(\mathbf{y}) = \sigma^2\mathbf{I}$, it does not have a unique covariance matrix. However, for a particular (symmetric) generalized inverse $(\mathbf{X}'\mathbf{X})^-$, we can use Theorem 3.6d(i) to obtain the following covariance matrix:

$$\begin{aligned} \text{cov}(\hat{\boldsymbol{\beta}}) &= \text{cov}[(\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{y}] \\ &= (\mathbf{X}'\mathbf{X})^-\mathbf{X}'(\sigma^2\mathbf{I})\mathbf{X}[(\mathbf{X}'\mathbf{X})^-]' \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^-\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^-. \end{aligned} \tag{12.18}$$

The expression in (12.18) is not invariant to the choice of $(\mathbf{X}'\mathbf{X})^-$.

The variance of $\boldsymbol{\lambda}'\hat{\boldsymbol{\beta}}$ or of $\mathbf{r}'\mathbf{X}'\mathbf{y}$ is given in the following theorem.

Theorem 12.3b. Let $\boldsymbol{\lambda}'\boldsymbol{\beta}$ be an estimable function in the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\mathbf{X}$ is $n \times p$ of rank $k < p \leq n$ and $\text{cov}(\mathbf{y}) = \sigma^2\mathbf{I}$. Let $\mathbf{r}$ be any solution to $\mathbf{X}'\mathbf{X}\mathbf{r} = \boldsymbol{\lambda}$, and let $\hat{\boldsymbol{\beta}}$ be any solution to $\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$. Then the variance of $\boldsymbol{\lambda}'\hat{\boldsymbol{\beta}}$ or $\mathbf{r}'\mathbf{X}'\mathbf{y}$ has the following properties:

(i) $\text{var}(\mathbf{r}'\mathbf{X}'\mathbf{y}) = \sigma^2\mathbf{r}'\mathbf{X}'\mathbf{X}\mathbf{r} = \sigma^2\mathbf{r}'\boldsymbol{\lambda}$.

(ii) $\text{var}(\boldsymbol{\lambda}'\hat{\boldsymbol{\beta}}) = \sigma^2\boldsymbol{\lambda}'(\mathbf{X}'\mathbf{X})^-\boldsymbol{\lambda}$.

(iii) $\text{var}(\boldsymbol{\lambda}'\hat{\boldsymbol{\beta}})$ is unique, that is, invariant to the choice of $\mathbf{r}$ or $(\mathbf{X}'\mathbf{X})^-$.

Theorem 12.3c. If $\boldsymbol{\lambda}_1'\boldsymbol{\beta}$ and $\boldsymbol{\lambda}_2'\boldsymbol{\beta}$ are two estimable functions in the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\mathbf{X}$ is $n \times p$ of rank $k < p \leq n$ and $\text{cov}(\mathbf{y}) = \sigma^2\mathbf{I}$, the covariance of their estimators is given by

$$\text{cov}(\boldsymbol{\lambda}_1'\hat{\boldsymbol{\beta}}, \boldsymbol{\lambda}_2'\hat{\boldsymbol{\beta}}) = \sigma^2\mathbf{r}_1'\boldsymbol{\lambda}_2 = \sigma^2\boldsymbol{\lambda}_1'\mathbf{r}_2 = \sigma^2\boldsymbol{\lambda}_1'(\mathbf{X}'\mathbf{X})^-\boldsymbol{\lambda}_2,$$

where $\mathbf{X}'\mathbf{X}\mathbf{r}_1 = \boldsymbol{\lambda}_1$ and $\mathbf{X}'\mathbf{X}\mathbf{r}_2 = \boldsymbol{\lambda}_2$.

Theorem 12.3d. If $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is an estimable function in the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\mathbf{X}$ is $n \times p$ of rank $k < p \leq n$, then the estimators $\boldsymbol{\lambda}'\hat{\boldsymbol{\beta}}$ and $\mathbf{r}'\mathbf{X}'\mathbf{y}$ are BLUE.

12.3.2 Estimation of σ^2

By analogy with (7.23), we define

$$\text{SSE} = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}). \tag{12.19}$$

where $\hat{\boldsymbol{\beta}}$ is any solution to the normal equations $\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$. Two alternative expressions for SSE are

$$\text{SSE} = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}, \tag{12.20}$$

$$\text{SSE} = \mathbf{y}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}']\mathbf{y}. \tag{12.21}$$

For an estimator of σ^2, we define

$$s^2 = \frac{\text{SSE}}{n - k}, \tag{12.22}$$

where n is the number of rows of $\mathbf{X}$ and $k = \text{rank}(\mathbf{X})$.

Two properties of s^2 are given in the following theorem.

Theorem 12.3e. For s^2 defined in (12.22) for the non-full-rank model, we have the following properties:

(i) $E(s^2) = \sigma^2$.

(ii) s^2 is invariant to the choice of $\hat{\boldsymbol{\beta}}$ or to the choice of generalized inverse $(\mathbf{X}'\mathbf{X})^{-}$.

12.3.3 Normal Model

For the non-full-rank model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, we now assume that

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

With the normality assumption we can obtain maximum likelihood estimators.

Theorem 12.3f. If $\mathbf{y}$ is $N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, where $\mathbf{X}$ is $n \times p$ of rank $k < p \leq n$, then the maximum likelihood estimators for $\boldsymbol{\beta}$ and σ^2 are given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y}, \tag{12.23}$$

$$\hat{\sigma}^2 = \frac{1}{n}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}). \tag{12.24}$$

Theorem 12.3g. If $\mathbf{y}$ is $N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, where $\mathbf{X}$ is $n \times p$ of rank $k < p \leq n$, then th maximum likelihood estimators $\hat{\boldsymbol{\beta}}$ and s^2 (corrected for bias) have the followin properties:

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

Theorem 12.3h. If $\mathbf{y}$ is $N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, where $\mathbf{X}$ is $n \times p$ of rank $k < p \leq n$, and if $\boldsymbol{\lambda}'\boldsymbol{\beta}$ is an estimable function, then $\boldsymbol{\lambda}'\hat{\boldsymbol{\beta}}$ has minimum variance among all unbiased estimators. □

Theorem 12.7b. If $\mathbf{y}$ is $N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, where $\mathbf{X}$ is $n \times p$ of rank $k < p \leq n$, if $\mathbf{C}$ is $m \times p$ of rank $m \leq k$ such that $\mathbf{C}\boldsymbol{\beta}$ is a set of m linearly independent estimable functions, and if $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y}$, then

(i) $\mathbf{C}(\mathbf{X}'\mathbf{X})^{-}\mathbf{C}'$ is nonsingular.
(ii) $\mathbf{C}\hat{\boldsymbol{\beta}}$ is $N_m[\mathbf{C}\boldsymbol{\beta}, \sigma^2\mathbf{C}(\mathbf{X}'\mathbf{X})^{-}\mathbf{C}']$.
(iii) $\mathrm{SSH}/\sigma^2 = (\mathbf{C}\hat{\boldsymbol{\beta}})'[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-}\mathbf{C}']^{-1}\mathbf{C}\boldsymbol{\beta}/\sigma^2$ is $\chi^2(m, \lambda)$, where $\lambda = (\mathbf{C}\boldsymbol{\beta})'[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-}\mathbf{C}']^{-1}\mathbf{C}\boldsymbol{\beta}/2\sigma^2$.
(iv) $\mathrm{SSE}/\sigma^2 = \mathbf{y}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}']\mathbf{y}/\sigma^2$ is $\chi^2(n-k)$.
(v) SSH and SSE are independent.

Theorem 12.7c. Let $\mathbf{y}$ be $N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, where $\mathbf{X}$ is $n \times p$ of rank $k < p \leq n$, and let $\mathbf{C}$, $\mathbf{C}\boldsymbol{\beta}$, and $\hat{\boldsymbol{\beta}}$ be defined as in Theorem 12.7b. Then, if $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$ is true, the statistic

$$F = \frac{\text{SSH}/m}{\text{SSE}/(n-k)} = \frac{(\mathbf{C}\hat{\boldsymbol{\beta}})'[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-}\mathbf{C}']^{-1}\mathbf{C}\hat{\boldsymbol{\beta}}/m}{\text{SSE}/(n-k)} \tag{12.46}$$

is distributed as $F(m, n-k)$.

(I.S.S.) Coaching by SUDHIR SIR
DEEP INSTITUTE 9560402898

ISS COACHING

DEEP INSTITUTE

INDIAN STATISTICAL SERVICE (I.S.S.)

Best I.S.S. Coaching by SUDHIR SIR

(2023 SELECTION)

PRAKHAR GUPTA
[5th RANK]

SWATI GUPTA
[9th RANK]

SIMRAN
[19th RANK]

LOKESH KUMAR
[32th RANK]

www.isscoaching.com | 956 040 2898 | www.deepinstitute.co.in

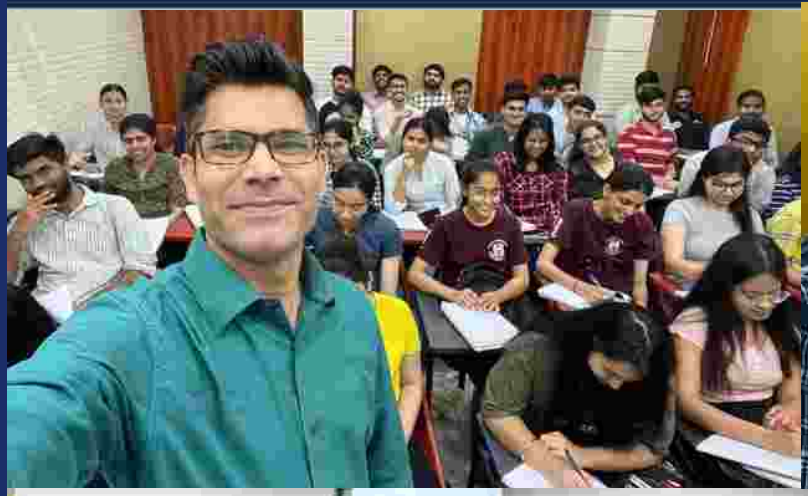
Guided by - Sudhir Sir 9999001310

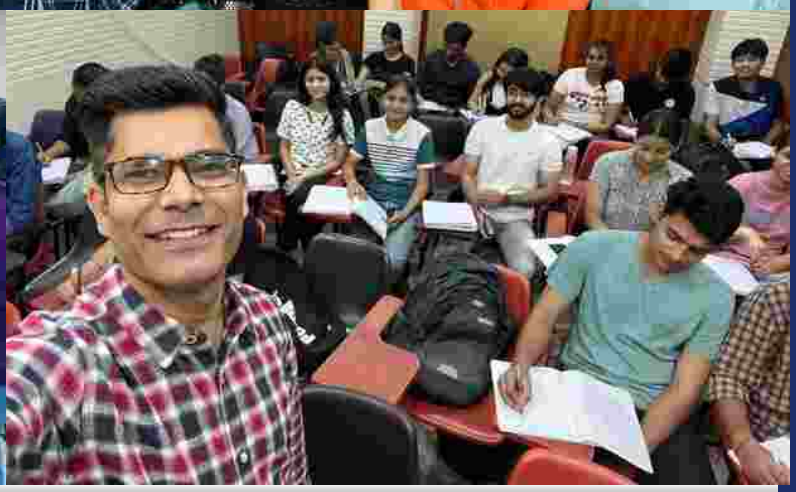
Sudhirdse1@gmail.com

www.isscoaching.com

2513, Basement, Hudson Lane Beside HDFC Bank Opp.
Laxmi Dairy, GTB Nagar New Delhi: 110009

ISS COACHING



Guided by - Sudhir Sir 9999001310

Sudhirdse1@gmail.com

www.isscoaching.com

2513, Basement, Hudson Lane Beside HDFC Bank Opp.
Laxmi Dairy, GTB Nagar New Delhi: 110009