COPYRIGHT

Abraham, B. and Ledolter, J.

Introduction to Regression Modeling

Belmont, CA: Duxbury Press, 2006

4 Multiple Linear Regression Model

4.1 INTRODUCTION

In this chapter we consider the general linear model introduced in Eq. (1.10),

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon \tag{4.1}$$

which links a response variable y to several independent (also called explanatory or predictor) variables $x_1, x_2, \ldots, x_p$. We discuss how to estimate the model parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ and how to test various hypotheses about them. You may find the subsequent discussion interesting from a theoretical standpoint because it uses linear algebra to establish general results. It also maps out an elegant geometric approach to least squares regression. Be prepared for subspaces, basis vectors, and orthogonal projections.

4.1.1 TWO EXAMPLES

In order to motivate the general notation, we start our discussion with two examples: the urea formaldehyde foam insulation (UFFI) example and the gas consumption data of Chapter 1.

UFFI Example

In Example 1.2.5 of Chapter 1, we considered 12 homes without UFFI ($x_1 = 0$) and 12 homes with insulation ($x_1 = 1$). For each home we obtained an air-tightness measure (x_2) and a reading of its ambient formaldehyde concentration (y). The model in Eq. (1.6) relates the ambient formaldehyde concentration (y) of the ith home to its air tightness (x_2) and the presence of UFFI (x_1):

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i, \quad i = 1, 2, \ldots, 24 \tag{4.2}$$

Table 1.2 lists the information on the 12 houses without UFFI ($x_1 = 0$) first; the remaining 12 homes with UFFI ($x_1 = 1$) are listed second. Note that Chapter 1 uses z and x for the predictors x_1 and x_2. The 24 equations resulting from model (4.2),

$$\begin{aligned}
31.33 &= \beta_0 + \beta_1 0 + \beta_2 0 + \epsilon_1 \\
28.57 &= \beta_0 + \beta_1 0 + \beta_2 1 + \epsilon_2 \\
\vdots & \qquad\qquad \vdots \\
56.67 &= \beta_0 + \beta_1 0 + \beta_2 9 + \epsilon_{12} \\
43.58 &= \beta_0 + \beta_1 1 + \beta_2 1 + \epsilon_{13} \\
\vdots & \qquad\qquad \vdots \\
70.34 &= \beta_0 + \beta_1 1 + \beta_2 10 + \epsilon_{24}
\end{aligned}$$

can be written in vector form,

$$\begin{bmatrix} 31.33 \\ 28.57 \\ \vdots \\ 56.67 \\ 43.58 \\ \vdots \\ 70.34 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 9 \\ 1 & 1 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 10 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_{12} \\ \epsilon_{13} \\ \vdots \\ \epsilon_{24} \end{bmatrix}$$

In short,

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{4.3}$$

where

$$\boldsymbol{y} = \begin{bmatrix} 31.33 \\ 28.57 \\ \vdots \\ 70.34 \end{bmatrix}; \quad X = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 9 \\ 1 & 1 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 10 \end{bmatrix}; \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}; \quad \text{and} \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_{24} \end{bmatrix}$$

Gas Consumption Data

In Example 1.2.7 of Chapter 1 we relate the fuel efficiency on each of 38 cars to their weight, engine displacement, and number of cylinders. Consider the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i, \quad i = 1, 2, \ldots, 38 \tag{4.4}$$

where y_i = gas consumption (miles per gallon) for the ith car

x_{i1} = weight of the ith car

x_{i2} = engine displacement for the ith car

x_{i3} = number of cylinders for the ith car

The resulting 38 equations

$$\begin{aligned}
16.9 &= \beta_0 + \beta_1 4.360 + \beta_2 350 + \beta_3 8 + \epsilon_1 \\
15.5 &= \beta_0 + \beta_1 4.054 + \beta_2 351 + \beta_3 8 + \epsilon_2 \\
\vdots \quad & \vdots \qquad\qquad\quad \vdots \\
31.9 &= \beta_0 + \beta_1 1.925 + \beta_2 89 + \beta_3 4 + \epsilon_{38}
\end{aligned}$$

can be written in vector form as

$$\begin{bmatrix} 16.9 \\ 15.5 \\ \vdots \\ 31.9 \end{bmatrix} = \begin{bmatrix} 1 & 4.360 & 350 & 8 \\ 1 & 4.054 & 351 & 8 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1.925 & 89 & 4 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_{38} \end{bmatrix}$$

In short,

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

where

$$\boldsymbol{y} = \begin{bmatrix} 16.9 \\ 15.5 \\ \vdots \\ 31.9 \end{bmatrix}; \quad X = \begin{bmatrix} 1 & 4.360 & 350 & 8 \\ 1 & 4.054 & 351 & 8 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1.925 & 89 & 4 \end{bmatrix}; \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}; \quad \text{and} \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_{38} \end{bmatrix} \tag{4.5}$$

4.1.2 THE GENERAL LINEAR MODEL

These two examples show us how we can write the general linear model (4.1) in vector form. Suppose that we have information on n cases, or subjects $i = 1, 2, \ldots, n$. Let y_i be the observed value on the response variable and let $x_{i1}, x_{i2}, \ldots, x_{ip}$ be the values on the independent or predictor variables of the ith case. The values of the p predictor variables are treated as fixed constants; however, the responses are subject to variability. The model for the response of case i is written as

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i \\ &= \mu_i + \epsilon_i \end{aligned} \tag{4.6}$$

where $\mu_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$ is a deterministic component that is affected by the regressor variables and ϵ_i is a term that captures the effect of all other variables that are not included in the model.

We assume that ϵ_i is a random variable with mean $E(\epsilon_i) = 0$ and variance $V(\epsilon_i) = \sigma^2$, and we suppose that the ϵ_i are normally distributed. Furthermore, we assume that the errors from different cases, $\epsilon_1, \ldots, \epsilon_n$, are independent random variables. These assumptions imply that the responses $y_1, \ldots, y_n$ are independent normal random variables with mean $E(y_i) = \mu_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$ and variance $V(y_i) = \sigma^2$.

We assume that the variance $V(y_i)$ is the same for each case. Note that this is an assumption that needs to be checked because one needs to check all other model assumptions, such as the form of the deterministic relationship and the normal distribution of the errors.

The n equations in (4.6) can be rewritten in vector form,

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 x_{11} + \cdots + \beta_p x_{1p} \\ \vdots \\ \beta_0 + \beta_1 x_{n1} + \cdots + \beta_p x_{np} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

In short,

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{4.7}$$

where

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}; \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}; \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}; \quad \text{and} \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

You should convince yourself that this representation is correct by multiplying out the first few elements of $X\boldsymbol{\beta}$.

The assumptions on the errors in this model can also be written in vector form. We write $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 I)$, a multivariate normal distribution with mean vector $E(\boldsymbol{\epsilon}) = \boldsymbol{0}$ and covariance matrix $V(\boldsymbol{\epsilon}) = \sigma^2 I$. Similarly, we write $\boldsymbol{y} \sim N(X\boldsymbol{\beta}, \sigma^2 I)$, a multivariate normal distribution with mean vector $E(\boldsymbol{y}) = X\boldsymbol{\beta}$ and covariance matrix $V(\boldsymbol{y}) = \sigma^2 I$.

4.2 ESTIMATION OF THE MODEL

We now consider the estimation of the unknown parameters: the $(p+1)$ regression parameters $\boldsymbol{\beta}$, and the variance of the errors σ^2. Since $y_i \sim N(\mu_i, \sigma^2)$ with $\mu_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$ are independent, it is straightforward to write down the joint probability density $p(y_1, \ldots, y_n \mid \boldsymbol{\beta}, \sigma^2)$. Treating this, for given data $\boldsymbol{y}$, as a function of the parameters leads to the likelihood function

$$L(\boldsymbol{\beta}, \sigma^2 \mid y_1, \ldots, y_n) = (1/\sqrt{2\pi}\sigma)^n \exp\left[-\sum_{i=1}^{n} (y_i - \mu_i)^2 / 2\sigma^2 \right] \tag{4.8}$$

Maximizing the likelihood function L with respect to $\boldsymbol{\beta}$ is equivalent to minimizing $S(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \mu_i)^2$ with respect to $\boldsymbol{\beta}$. This is because the exponent in Eq. (4.8) is the only term containing $\boldsymbol{\beta}$. The sum of squares $S(\boldsymbol{\beta})$ can be written in vector notation,

$$S(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{\mu})'(\boldsymbol{y} - \boldsymbol{\mu}) = (\boldsymbol{y} - X\boldsymbol{\beta})'(\boldsymbol{y} - X\boldsymbol{\beta}), \quad \text{since } \boldsymbol{\mu} = X\boldsymbol{\beta} \tag{4.9}$$

The minimization of $S(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ is known as **least squares estimation**, and for normal errors it is equivalent to maximum likelihood estimation. We determine the least squares estimates by obtaining the first derivatives of $S(\boldsymbol{\beta})$ with respect to the parameters $\beta_0, \beta_1, \ldots, \beta_p$, and by setting these $(p+1)$ derivatives equal to zero.

The appendix shows that this leads to the $(p+1)$ equations

$$X'X\hat{\boldsymbol{\beta}} = X'\boldsymbol{y} \tag{4.10}$$

These equations are referred to as the **normal equations**. The matrix X is assumed to have full column rank $p+1$. Hence, the $(p+1) \times (p+1)$ matrix $X'X$ is nonsingular and the solution of Eq. (4.10) is given by

$$\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\boldsymbol{y} \tag{4.11}$$

The estimate $\hat{\boldsymbol{\beta}}$ in Eq. (4.11) minimizes $S(\boldsymbol{\beta})$, and is known as the **least squares estimate** (LSE) of $\boldsymbol{\beta}$.

4.2.1 A GEOMETRIC INTERPRETATION OF LEAST SQUARES

The model in Eq. (4.7) can be written as

$$\begin{aligned} \boldsymbol{y} &= \beta_0\mathbf{1} + \beta_1\boldsymbol{x}_1 + \cdots + \beta_p\boldsymbol{x}_p + \boldsymbol{\epsilon} \\ &= \boldsymbol{\mu} + \boldsymbol{\epsilon} \end{aligned} \tag{4.12}$$

where the $(n \times 1)$ vectors $\boldsymbol{y}$ and $\boldsymbol{\epsilon}$ are as defined before, and the $(n \times 1)$ vectors $\mathbf{1} = (1, 1, \ldots, 1)'$ and $\boldsymbol{x}_j = (x_{1j}, x_{2j}, \ldots, x_{nj})'$, for $j = 1, 2, \ldots, p$, represent the columns of the matrix X. Thus, $X = (\mathbf{1}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$ and $\boldsymbol{\mu} = X\boldsymbol{\beta} = \beta_0\mathbf{1} + \beta_1\boldsymbol{x}_1 + \cdots + \beta_p\boldsymbol{x}_p$.

The representation in Eq. (4.12) shows that the deterministic component $\boldsymbol{\mu}$ is a linear combination of the vectors $\mathbf{1}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$. Let $L(\mathbf{1}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$ be the set of all linear combinations of these vectors. If we assume that these vectors are not linearly dependent, $L(X) = L(\mathbf{1}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$ is a subspace of R^n of dimension $p+1$. Note that the assumption that $\mathbf{1}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ are not linearly dependent is the same as saying that X has rank $p+1$.

We want to explain these concepts slowly because they are essential for understanding the geometric interpretation that follows. First, note that the dimension of the regressor vectors $\mathbf{1}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ is n, the number of cases. When we display the $(p+1)$ regressor vectors, we do that in n-dimensional Euclidean space R^n. The coordinates on each regressor vector correspond to the regressor's values on the n cases. For example, the regressor vector $\boldsymbol{x}$ may represent the air tightness of a home, and the dimension of this vector is 24, if measurements on 24 homes are taken. Note that for models with an intercept, one of the regressor columns is always the vector of ones, $\mathbf{1}$.

Obviously, it is impossible to graph vectors in 24-dimensional space, but you can get a good idea of this by considering lower dimensional situations. Consider the case in which $n = 3$, and use two regressor columns: the unit vector

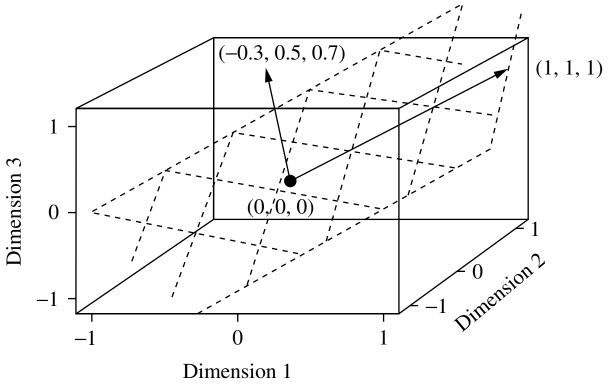


FIGURE 4.1 Two Vectors in Three-Dimensional Space, and the Two-Dimensional Space Spanned by These Two Vectors

$\mathbf{1} = (1, 1, 1)'$ and $\boldsymbol{x} = (-0.3, 0.5, 0.7)'$. These two vectors are graphed in three-dimensional space in Figure 4.1. Any linear combination of these two vectors results in a vector that lies in the two-dimensional space that is spanned by the vectors $\mathbf{1}$ and $\boldsymbol{x}$. We highlight this by shading the plane that contains all linear combinations. We see that $L(\mathbf{1}, \boldsymbol{x})$ is a subspace of R^3, and its dimension is 2.

Observe that we have selected two vectors $\mathbf{1}$ and $\boldsymbol{x}$ that are **not** linearly dependent. This means that one of the two vectors cannot be written as a multiple of the other. This is the case in our example. Note that the matrix

$$X = [\mathbf{1}, \boldsymbol{x}] = \begin{bmatrix} 1 & -0.3 \\ 1 & 0.5 \\ 1 & 0.7 \end{bmatrix}$$

has full column rank, 2.

What would happen if two regressor vectors were linearly dependent; for example, if $\mathbf{1} = (1, 1, 1)'$ and $\boldsymbol{x} = (0.5, 0.5, 0.5)'$? Here, every linear combination of $\mathbf{1}$ and $\boldsymbol{x}$, $\alpha_1 \mathbf{1} + \alpha_2 \boldsymbol{x} = \alpha_1 \mathbf{1} + \alpha_2 (0.5)\mathbf{1} = (\alpha_1 + 0.5\alpha_2)\mathbf{1}$, is a multiple of $\mathbf{1}$. Hence, the set of all linear combinations are points along the unit vector, and $L(\mathbf{1}, \boldsymbol{x})$ defines a subspace of dimension 1. You can also see this from the rank of the matrix X: The rank of

$$X = [\mathbf{1}, \boldsymbol{x}] = \begin{bmatrix} 1 & 0.5 \\ 1 & 0.5 \\ 1 & 0.5 \end{bmatrix}$$

is one; X does not have full column rank.

If we contemplate a model with two regressor columns, $\mathbf{1}$ and $\boldsymbol{x}$, then we suppose that $\mathbf{1}$ and $\boldsymbol{x}$ are not linearly dependent. If they were linearly dependent, we would encounter difficulties because an infinite number of linear combinations could be used to represent each point in the subspace spanned by $\mathbf{1}$ and $\boldsymbol{x}$. You can see this from our example. There is an infinite number of values for α_1 and α_2 that result in a given value $\alpha_1 + 0.5\alpha_2 = c$.

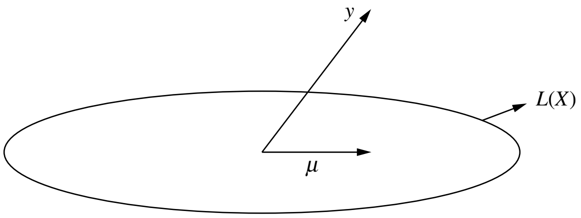


FIGURE 4.2 Geometric Representation of the Response Vector *y* and the Subspace *L*(*X*)

Now we are ready to go to the more general case with a large number of cases, n. Suppose that there are two regressors ($p = 2$) and three regressor columns $\mathbf{1}$, $\boldsymbol{x_1}$, and $\boldsymbol{x_2}$. We assume that these three columns are not linearly dependent and that the matrix $X = [\mathbf{1}, \boldsymbol{x_1}, \boldsymbol{x_2}]$ has full column rank, rank 3. The regressor vectors are elements in R^n, and the set of all linear combinations of $\mathbf{1}, \boldsymbol{x_1}, \boldsymbol{x_2}$, $L(\mathbf{1}, \boldsymbol{x_1}, \boldsymbol{x_2})$, defines a three-dimensional subspace of R^n. If $\mathbf{1}, \boldsymbol{x_1}, \boldsymbol{x_2}$ were linearly dependent, then the subspace would be of lower dimension (either 2 or 1).

Now we consider the case with p regressors shown in Figure 4.2. The oval represents the subspace $L(X)$. The vector $\boldsymbol{\mu} = \beta_0\mathbf{1} + \beta_1\boldsymbol{x_1} + \cdots + \beta_p\boldsymbol{x_p}$ is a linear combination of $\mathbf{1}, \boldsymbol{x_1}, \ldots, \boldsymbol{x_p}$, and is part of the subspace $L(X)$. This picture is simplified as it tries to illustrate a higher dimensional space. You need to use your imagination.

Until now, we have talked about the subspace of R^n that is spanned by the $p + 1$ regressor vectors $\mathbf{1}, \boldsymbol{x_1}, \ldots, \boldsymbol{x_p}$. Next, let us add the ($n \times 1$) response vector $\boldsymbol{y}$ to the picture (see Figure 4.2). The response vector $\boldsymbol{y}$ is not part of the subspace $L(X)$. For a given value of $\boldsymbol{\beta}$, $X\boldsymbol{\beta}$ is a vector in the subspace; $\boldsymbol{y} - X\boldsymbol{\beta}$ is the difference between the response vector $\boldsymbol{y}$ and the vector in the subspace, and $S(\boldsymbol{\beta}) = (\boldsymbol{y} - X\boldsymbol{\beta})'(\boldsymbol{y} - X\boldsymbol{\beta})$ represents the squared length of this difference. Minimizing $S(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ corresponds to finding $\hat{\boldsymbol{\beta}}$ so that $\boldsymbol{y} - X\hat{\boldsymbol{\beta}}$ has minimum length.

In other words, we must find a vector $X\hat{\boldsymbol{\beta}}$ in the subspace $L(X)$ that is "closest" to $\boldsymbol{y}$. The vector in the subspace $L(X)$ that is closest to $\boldsymbol{y}$ is obtained by making the difference $\boldsymbol{y} - X\hat{\boldsymbol{\beta}}$ perpendicular to the subspace $L(X)$; see Figure 4.3. Since $\mathbf{1}, \boldsymbol{x_1}, \ldots, \boldsymbol{x_p}$ are in the subspace, we require that $\boldsymbol{y} - X\hat{\boldsymbol{\beta}}$ is perpendicular to $\mathbf{1}$, $\boldsymbol{x_1}, \ldots,$ and $\boldsymbol{x_p}$.

This implies the equations

$$\begin{aligned} \mathbf{1}'(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}) &= 0 \\ \boldsymbol{x}_1'(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}) &= 0 \\ &\cdots \\ \boldsymbol{x}_p'(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}) &= 0 \end{aligned}$$

Combining these $p + 1$ equations leads to

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

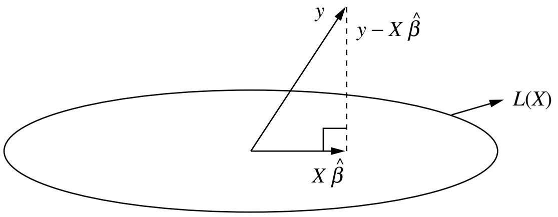


FIGURE 4.3 A Geometric View of Least Squares

and

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

the normal equations in Eq. (4.10) that we previously derived algebraically.

We assume that X has full column rank, $p + 1$. Hence, $X'X$ has rank $(p + 1)$, the inverse $(X'X)^{-1}$ exists, and the least squares estimate is given by $\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\boldsymbol{y}$. Notice that we have obtained the LSE solely through a geometric argument; no algebraic derivation was involved.

The vector of fitted values is given by $\hat{\boldsymbol{\mu}} = X\hat{\boldsymbol{\beta}}$, and the vector of residuals is $\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{\mu}} = \boldsymbol{y} - X\hat{\boldsymbol{\beta}}$. The geometric interpretation of least squares is quite simple. Least squares estimation amounts to finding the vector $\hat{\boldsymbol{\mu}} = X\hat{\boldsymbol{\beta}}$ in the subspace $L(X)$ that is closest to the observation vector $\boldsymbol{y}$. This requires that the difference (i.e., the residual vector) is perpendicular (or othogonal) to the subspace $L(X)$. Hence, the vector of fitted values $\hat{\boldsymbol{\mu}} = X\hat{\boldsymbol{\beta}}$ is the **orthogonal projection** of $\boldsymbol{y}$ onto the subspace $L(X)$. In algebraic terms,

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

where $H = X(X'X)^{-1}X'$ is an $n \times n$ symmetric and idempotent matrix. It is easy to confirm that H is idempotent as

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

The matrix H is an important matrix because it represents the orthogonal projection of $\boldsymbol{y}$ onto $L(X)$. It is referred to as the “hat” matrix.

The vector of residuals $\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{\mu}} = \boldsymbol{y} - X(X'X)^{-1}X'\boldsymbol{y} = (I - H)\boldsymbol{y}$ is also a projection of $\boldsymbol{y}$, this time on the subspace of R^n that is perpendicular to $L(X)$.

The vector of fitted values $\hat{\boldsymbol{\mu}} = X\hat{\boldsymbol{\beta}}$ and the vector of residuals $\boldsymbol{e}$ are orthogonal, which means algebraically that

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

See the normal equations in Eq. (4.10). Hence, least squares decomposes the response vector

$$\boldsymbol{y} = \hat{\boldsymbol{\mu}} + \boldsymbol{e} = X\hat{\boldsymbol{\beta}} + (\boldsymbol{y} - X\hat{\boldsymbol{\beta}})$$

into two orthogonal pieces. The vector of fitted values $X\hat{\boldsymbol{\beta}}$ is in $L(X)$, whereas the vector of residuals $\boldsymbol{y} - X\hat{\boldsymbol{\beta}}$ is in the space orthogonal to $L(X)$.

It may help you to look at this in the very simplest special case in which we have $n = 2$ cases and just a single regressor column, $\mathbf{1} = (1, 1)'$. This represents the "mean" regression model, $y_i = \beta_0 + \epsilon_i$, with $i = 1, 2$. How does this look geometrically? Since the number of cases is 2, we are looking at the two-dimensional Euclidean space. Draw in the unit vector $\mathbf{1} = (1, 1)'$ and the response vector $\boldsymbol{y} = (y_1, y_2)'$. For illustration, take $\boldsymbol{y} = (0, 1)'$. We project $\boldsymbol{y} = (y_1, y_2)' = (0, 1)'$ onto the subspace $L(\mathbf{1})$, which is the 45-degree line in the two-dimensional Euclidean space. The projection leads to the vector of fitted values $\hat{\boldsymbol{\mu}} = 0.5\mathbf{1} = (0.5, 0.5)'$ and the LSE $\hat{\beta}_0 = 0.5$. The estimate is the average of the two observations, 0 and 1. The residual vector $\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{\mu}} = (0 - 0.5, 1 - 0.5)' = (-0.5, 0.5)'$ and the vector of fitted values $\hat{\boldsymbol{\mu}} = (0.5, 0.5)'$ are orthogonal; that is, $\boldsymbol{e}'\hat{\boldsymbol{\mu}} = -(0.5)^2 + (0.5)^2 = 0$.

4.2.2 USEFUL PROPERTIES OF ESTIMATES AND OTHER RELATED VECTORS

Recall our model

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

where X is a fixed (nonrandom) matrix with full rank, and the random error $\boldsymbol{\epsilon}$ follows a distribution with mean $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and covariance matrix $V(\boldsymbol{\epsilon}) = \sigma^2 I$. Usually, we also assume a normal distribution. The model implies that $E(\boldsymbol{y}) = X\boldsymbol{\beta}$ and $V(\boldsymbol{y}) = \sigma^2 I$. The LSE of the parameter vector $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\boldsymbol{y}$. The vector of fitted values is $\hat{\boldsymbol{\mu}} = X\hat{\boldsymbol{\beta}} = H\boldsymbol{y}$ and the residual vector is $\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{\mu}} = (I - H)\boldsymbol{y}$. We now study properties of these vectors and other related quantities, always assuming that the model is true.

i. Estimate $\hat{\boldsymbol{\beta}}$:

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}) &= E(X'X)^{-1}X'\boldsymbol{y} \\ &= (X'X)^{-1}X'E(\boldsymbol{y}) = (X'X)^{-1}X'X\boldsymbol{\beta} = \boldsymbol{\beta} \end{aligned} \tag{4.13}$$

showing that $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$.

$$\begin{aligned} V(\hat{\boldsymbol{\beta}}) &= V[(X'X)^{-1}X'\boldsymbol{y}] \\ &= (X'X)^{-1}X'V(\boldsymbol{y})X(X'X)^{-1} \\ &= (X'X)^{-1}X'(\sigma^2 I)(X'X)^{-1} \\ &= (X'X)^{-1}X'X(X'X)^{-1}\sigma^2 = (X'X)^{-1}\sigma^2 \end{aligned} \tag{4.14}$$

The matrix in Eq. (4.14) contains the variances of the estimates in the diagonal and the covariances in the off-diagonal elements. Let v_{ij} denote the elements of the matrix $(X'X)^{-1}$. Then $V(\hat{\beta}_i) = \sigma^2 v_{ii}$, $\text{Cov}(\hat{\beta}_i, \hat{\beta}_j) = \sigma^2 v_{ij}$, and $\text{Corr}(\hat{\beta}_i, \hat{\beta}_j) = \dfrac{v_{ij}}{(v_{ii}v_{jj})^{1/2}}$.

ii. Linear combination of estimates, $\boldsymbol{a}'\hat{\boldsymbol{\beta}}$:
The linear combination $\boldsymbol{a}'\boldsymbol{\beta}$, where $\boldsymbol{a}$ is a vector of constants of appropriate dimension, can be estimated by $\boldsymbol{a}'\hat{\boldsymbol{\beta}}$. We find

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

and

$$V(\boldsymbol{a}'\hat{\boldsymbol{\beta}}) = \boldsymbol{a}'V(\hat{\boldsymbol{\beta}})\boldsymbol{a} = \boldsymbol{a}'(X'X)^{-1}\boldsymbol{a}\sigma^2 \tag{4.15}$$

iii. Fitted values $\hat{\boldsymbol{\mu}} = X\hat{\boldsymbol{\beta}}$:

$$E(\hat{\boldsymbol{\mu}}) = E(X\hat{\boldsymbol{\beta}}) = XE(\hat{\boldsymbol{\beta}}) = X\boldsymbol{\beta} = \boldsymbol{\mu}$$

and

$$\begin{aligned} V(\hat{\boldsymbol{\mu}}) &= V(X\hat{\boldsymbol{\beta}}) = XV(\hat{\boldsymbol{\beta}})X' = X(X'X)^{-1}X'\sigma^2 \\ &= H\sigma^2 \end{aligned} \tag{4.16}$$

where $H = X(X'X)^{-1}X'$ is the idempotent projection matrix defined earlier.

iv. Residual vector $\boldsymbol{e} = \boldsymbol{y} - X\hat{\boldsymbol{\beta}}$:

$$E(\boldsymbol{e}) = E(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}) = E(\boldsymbol{y}) - XE(\hat{\boldsymbol{\beta}}) = X\boldsymbol{\beta} - X\boldsymbol{\beta} = \mathbf{0} \tag{4.17}$$

$$\begin{aligned} V(\boldsymbol{e}) &= V[(I - H)\boldsymbol{y}] = (I - H)V(\boldsymbol{y})(I - H)' \\ &= (I - H)(I - H)\sigma^2 = (I - H)\sigma^2 \end{aligned} \tag{4.18}$$

v. Statistical independence between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{e}$: We stack the $(p+1)$ vector $\hat{\boldsymbol{\beta}}$ and the $(n \times 1)$ vector of residuals $\boldsymbol{e}$ to obtain the $(p + 1 + n) \times 1$ vector

$$\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ --- \\ \boldsymbol{e} \end{bmatrix} = \begin{bmatrix} A \\ --- \\ I - H \end{bmatrix} \boldsymbol{y} = P\boldsymbol{y}$$

with $A = (X'X)^{-1}X'$ and $H = X(X'X)^{-1}X'$. The stacked vector is a linear transformation of $\boldsymbol{y}$. Our assumption of independent normal random variables for y_i implies a multivariate normal distribution for the vector $\boldsymbol{y}$. Hence, the linear transform $P\boldsymbol{y}$ follows a multivariate normal distribution with mean

$$E\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ --- \\ \boldsymbol{e} \end{bmatrix} = PE(\boldsymbol{y}) = \begin{bmatrix} A \\ --- \\ I - H \end{bmatrix} X\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta} \\ --- \\ \mathbf{0} \end{bmatrix}$$

and covariance matrix

$$\begin{aligned} V\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ --- \\ \boldsymbol{e} \end{bmatrix} &= PV(\boldsymbol{y})P' = \sigma^2 \begin{bmatrix} A \\ --- \\ I - H \end{bmatrix} [A' \mid (I - H)] \\ &= \sigma^2 \begin{bmatrix} AA' & | & A(I - H) \\ --- & & --- \\ (I - H)A' & | & (I - H)(I - H) \end{bmatrix} \\ &= \sigma^2 \begin{bmatrix} (X'X)^{-1} & | & O \\ --- & & --- \\ O' & | & (I - H) \end{bmatrix} \end{aligned}$$

Hence,

$$\begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \boldsymbol{e} \end{pmatrix} \sim N\left\{ \begin{pmatrix} \boldsymbol{\beta} \\ --- \\ \mathbf{0} \end{pmatrix}, \begin{bmatrix} (X'X)^{-1}\sigma^2 & | & O \\ --- & & --- \\ O' & | & (I - H)\sigma^2 \end{bmatrix} \right\} \tag{4.19}$$

Marginal distributions of multivariate normal distributions are themselves normal. Hence, it follows that

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(X'X)^{-1}) \quad \text{and} \quad \boldsymbol{e} \sim N(0, \sigma^2(I-H))$$

Equation (4.19) confirms our earlier results on the means and variances of $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{e}$ in Eqs. (4.13), (4.14) and (4.17), (4.18). Also note that the matrices $(X'X)^{-1}$, H, and $(I-H)$ are very useful quantities, and they will appear repeatedly.

Let Cov$(\boldsymbol{e}, \hat{\boldsymbol{\beta}})$ represents the $n \times (p+1)$ matrix of covariances between the residuals $\hat{\boldsymbol{e}}$ and the parameter estimates $\hat{\boldsymbol{\beta}}$. That is,

$$\text{Cov}(\boldsymbol{e}, \hat{\boldsymbol{\beta}}) = \begin{bmatrix} \text{Cov}(e_1, \hat{\beta}_0) & \text{Cov}(e_1, \hat{\beta}_1) & \cdots & \text{Cov}(e_1, \hat{\beta}_p) \\ \vdots & \vdots & & \vdots \\ \text{Cov}(e_n, \hat{\beta}_0) & \text{Cov}(e_n, \hat{\beta}_1) & \cdots & \text{Cov}(e_n, \hat{\beta}_p) \end{bmatrix} \tag{4.20}$$

Equation (4.19) shows that $\boldsymbol{e}$ and $\hat{\boldsymbol{\beta}}$ are uncorrelated:

$$\text{Cov}(\boldsymbol{e}, \hat{\boldsymbol{\beta}}) = O'$$

which is an $n \times (p+1)$ matrix of zeros. Since they are jointly normal, they are statistically independent and not just uncorrelated. This also implies that $S(\hat{\boldsymbol{\beta}}) = \boldsymbol{e}'\boldsymbol{e}$, which is a function of just $\boldsymbol{e}$, and $\hat{\boldsymbol{\beta}}$ are statistically independent (an alternate proof of this result is given in the appendix). It should be noted that any linear combination $\boldsymbol{a}'\hat{\boldsymbol{\beta}}$ of $\hat{\boldsymbol{\beta}}$ is a linear combination of normal random variables and hence itself normally distributed. That is

$$\boldsymbol{a}'\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{a}'\beta, \sigma^2\boldsymbol{a}'(X'X)^{-1}\boldsymbol{a})$$

vi. $S(\hat{\boldsymbol{\beta}})/\sigma^2 = \boldsymbol{e}'\boldsymbol{e}/\sigma^2 \sim \chi^2_{n-p-1}$, a chi-square distribution with $n-p-1$ degrees of freedom. This result is shown in the appendix. The degrees of freedom are easy to remember: $n-p-1$ is the difference between the number of observations and the number of estimated regression parameters.

The appendix to Chapter 2 mentions that the mean of a chi-square random variable is equal to its degrees of freedom. Hence,

$$E\left(\frac{\boldsymbol{e}'\boldsymbol{e}}{\sigma^2}\right) = \frac{E(\boldsymbol{e}'\boldsymbol{e})}{\sigma^2} = n-p-1$$

and

$$s^2 = \frac{\boldsymbol{e}'\boldsymbol{e}}{n-p-1} = \frac{S(\hat{\boldsymbol{\beta}})}{n-p-1} \tag{4.21}$$

is an unbiased estimate of σ^2. You can see this from $E(s^2) = \dfrac{E(\boldsymbol{e}'\boldsymbol{e})}{n-p-1} = \dfrac{\sigma^2(n-p-1)}{n-p-1} = \sigma^2$.

vii. The residuals $\boldsymbol{e}$ and the fitted values $\hat{\boldsymbol{\mu}}$ are statistically independent. We have already shown that $\boldsymbol{e}$ and $\hat{\boldsymbol{\beta}}$ are independent. X is a fixed (nonrandom) matrix and hence $\boldsymbol{e}$ and $X\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\mu}}$ are separate nonoverlapping functions of independent random variables. Hence, they are independent. This can be proved directly as well; see Exercise 4.18.

viii. **Gauss–Markov Theorem**
Assume that the usual regression assumptions are satisfied and that the $n \times 1$ response vector $\boldsymbol{y}$ has mean $E(\boldsymbol{y}) = \boldsymbol{\mu} = X\boldsymbol{\beta}$ and covariance matrix $V(\boldsymbol{y}) = \sigma^2 I$. The Gauss–Markov theorem says that among all **linear unbiased** estimators, the LSE $\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\boldsymbol{y}$ has the smallest variance. "Smallest" variance means that the covariance matrix of any other linear unbiased estimator exceeds the covariance matrix of $\hat{\boldsymbol{\beta}}$ by a positive semidefinite matrix.

Proof: The LSE $\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\boldsymbol{y}$ is a linear combination of the random response vector $\boldsymbol{y}$. Consider any other linear transformation, for example, $\hat{\boldsymbol{b}} = M^*\boldsymbol{y}$, where M^* is a $(p+1) \times n$ matrix of fixed coefficients. Define $M = M^* - (X'X)^{-1}X'$, and write the new estimator as

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

The requirement of unbiasedness for $\hat{\boldsymbol{b}}$ implies that $MX = O$, a $(p+1) \times (p+1)$ matrix of zeros. With this condition imposed, the covariance matrix for $\hat{\boldsymbol{b}}$ becomes

$$\begin{aligned}V(\hat{\boldsymbol{b}}) &= E[(\hat{\boldsymbol{b}} - \boldsymbol{\beta})(\hat{\boldsymbol{b}} - \boldsymbol{\beta})'] = E\{[M + (X'X)^{-1}X']\boldsymbol{\epsilon}\boldsymbol{\epsilon}'[M + (X'X)^{-1}X']'\}\\ &= [M + (X'X)^{-1}X']E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}')[M + (X'X)^{-1}X']'\\ &= [M + (X'X)^{-1}X']\sigma^2 I[M + (X'X)^{-1}X']'\\ &= \sigma^2[M + (X'X)^{-1}X'][M + (X'X)^{-1}X']'\\ &= \sigma^2[MM' + (X'X)^{-1}] = \sigma^2(X'X)^{-1} + \sigma^2 MM'\\ &= V(\hat{\boldsymbol{\beta}}) + \sigma^2 MM'\end{aligned}$$

Here we have used the fact that $MX = O$, and hence $MX(X'X)^{-1} = O$ and $(X'X)^{-1}X'M' = O$.

This result shows that the variance of the new linear estimator $\hat{\boldsymbol{b}}$ exceeds the variance of the LSE $\hat{\boldsymbol{\beta}}$ by the matrix $\sigma^2 MM'$. However, this matrix is positive semidefinite because for any vector $\boldsymbol{a}$ the quadratic form $\boldsymbol{a}'MM'\boldsymbol{a} = \tilde{\boldsymbol{a}}'\tilde{\boldsymbol{a}} = \sum(\tilde{a}_i)^2 \geq 0$. ■

The Gauss–Markov result also holds when estimating an arbitrary linear combination of the regression parameters. Consider the linear combination $\boldsymbol{a}'\boldsymbol{\beta}$ and the two estimators $\boldsymbol{a}'\hat{\boldsymbol{\beta}}$ and $\boldsymbol{a}'\hat{\boldsymbol{b}}$. The first estimator uses the LSE, whereas the second uses the linear unbiased estimator studied previously. The variances of these estimators are given by

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

and

$$V(\boldsymbol{a}'\hat{\boldsymbol{b}}) = \sigma^2 \boldsymbol{a}'(X'X)^{-1}\boldsymbol{a} + \sigma^2 \boldsymbol{a}'MM'\boldsymbol{a}$$

Since $\boldsymbol{a}'MM'\boldsymbol{a} \geq 0$, the estimator using the LSE has the smaller variance. As a special case, consider the vector $\boldsymbol{a}$ with a one in the ith position and zeros everywhere else. Then the Gauss–Markov result implies that the LSE of the (individual) coefficient β_i has the smallest variance among all other linear unbiased estimators.

Note that it is not necessary to make any assumption about the **form** of the error distribution in order to get the Gauss–Markov property. However, it must be emphasized that the result only proves that the LSE is best within the class of **linear** estimators. For certain nonnormal error distributions, it is possible to find a nonlinear estimator that has smaller variance than the LSE. For normal errors, however, this cannot be done, and in this case the LSE is the best estimator among all estimators—linear as well as nonlinear.

4.2.3 PRELIMINARY DISCUSSION OF RESIDUALS

The residual vector is $\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{\mu}}$. The ith component $e_i = y_i - \hat{\mu}_i$ is the residual associated with the ith individual or case in the experiment. The residual represents the deviation between the response and the fitted value and hence estimates the random component ϵ in the model. Any misspecification or departure from the underlying assumptions in the model will show up as patterns in the residuals. Hence, the analysis of the residuals is an effective way of discovering model inadequacies. Let us examine some important properties of the residuals.

i. The vector of residuals $\boldsymbol{e}$ is orthogonal to $L(X)$, and hence $\boldsymbol{e}'\mathbf{1} = 0$ if β_0 is in the model. This means that $\sum_{i=1}^{n} e_i = 0$ and $\bar{e} = 0$.
ii. $\boldsymbol{e}$ is orthogonal to $\hat{\boldsymbol{\mu}}$.

These two properties are direct consequences of the least squares fitting procedure. They always hold, whether or not the model is adequate.

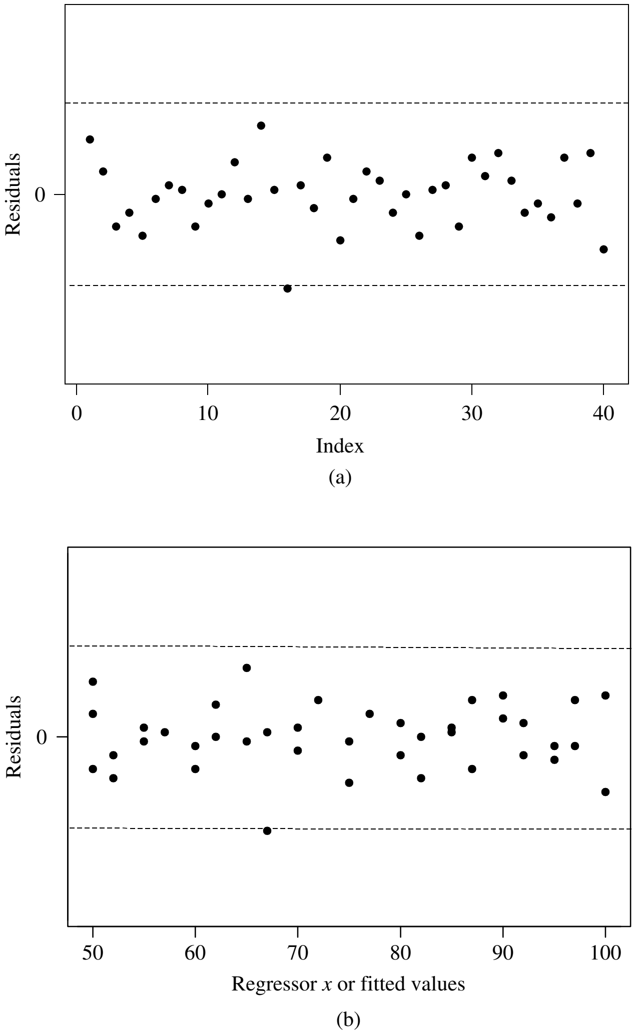
Next, let us summarize the properties of $\boldsymbol{e}$ that only hold if the model is correct. We assume that the funtional form is correct; that is, $E(\boldsymbol{y})$ is in $L(X)$. In addition, we suppose that the errors ϵ are multivariate normal with covariance matrix $\sigma^2 I$.

i. $E(\boldsymbol{e}) = \mathbf{0}$. If $E(\boldsymbol{y})$ is not in the subspace $L(X)$ and the assumed functional form of the model is incorrect, then this property does not hold. We will discuss this more fully in Chapter 6.
ii. $\boldsymbol{e}$ and $\hat{\boldsymbol{\mu}}$ are independent.
iii. $\boldsymbol{e} \sim N(\mathbf{0}, \sigma^2(I - H))$.

If the errors $\boldsymbol{\epsilon}$ do not have a normal distribution with constant variance σ^2, then the residuals $\boldsymbol{e}$ will not satisfy properties (ii) and (iii).

We construct several graphical residual checks that investigate whether the residuals exhibit the properties in (i)–(iii). These graphs can tell us whether the fitted model is an adequate representation. If the model is adequate, we do not expect systematic patterns in the residuals, and hence a plot of the residuals e_i versus the order i should exhibit the noninformative pattern depicted in Figure 4.4(a); that is, the e_i's should fall within an approximate horizontal band around $\bar{e} = 0$. A similar plot should result if e_i is plotted against the values of the jth predictor x_{ij}, $(j = 1, 2, \ldots, p)$. Also, a plot of the residuals e_i against the fitted values $\hat{\mu}_i$ should show no systematic patterns and should look like Figure 4.4(b). Departures from these patterns indicate model inadequacies, and we will discuss those more fully in Chapter 6.

FIGURE 4.4
Adequate Residual Plots



UFFI Example Continued

Consider Example 1.2.5 in Chapter 1 and the data in Table 1.2, where we relate the formaldehyde concentrations y_i to the presence or absence of UFFI ($x_{i1} = 1$ or $x_{i1} = 0$) and the airtightness (TIGHT, x_{i2}) of the home. The model in Eq. (4.2) specifies $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$, with the usual assumptions on ϵ_i. The $\boldsymbol{y}$ vector and the X matrix are given in Eq. (4.3). One can compute the LSE $\hat{\boldsymbol{\beta}}$ and s^2, the unbiased estimator of $V(\epsilon_i) = \sigma^2$, from Eqs. (4.11) and (4.21).

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

and

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

These computations are usually performed by a statistical software package (such as S-Plus, R, Minitab, SAS, SPSS, or even EXCEL). Computational details (commands and outputs from the well-known software S-Plus) are shown in Table 4.1.

TABLE 4.1 S-PLUS INPUT AND OUTPUT

```
> ch2o<-matrix(scan('uffi.dat',multi.line = T),byrow = T,ncol = 3,nrow = 24)
> uffi<-ch2o[,1]
> tight<-ch2o[,2]
> form<-ch2o[,3]
> ch2fit<-lm(form ~ uffi+tight)
> summary(ch2fit)
Call: lm(formula = form ~ uffi + tight)
Residuals:
Min              1Q              Median          3Q            Max
−9.546           −3.131          −0.1389         3.578         8.362
Coefficients:
                    Value        Std. Error      t value       Pr(> |t|)
(Intercept)       31.3734          2.4607        12.7500        0.0000
       uffi        9.3120          2.1325         4.3666        0.0003
      tight        2.8545          0.3764         7.5843        0.0000
Residual standard error: 5.223 on 21 degrees of freedom
Multiple R-Squared: 0.7827
F-statistic: 37.82 on 2 and 21 degrees of freedom, the p-value is 1.095e-07
Correlation of Coefficients:
                 (Intercept)          uffi
uffi              −0.4449
tight             −0.7903           0.0147
> X11()
> par(mfrow = c(2,2))
> obsno< − 1:24
> plot(obsno,ch2fit$res,xlab = 'Observation Number',
+ylab = 'Residuals',main = 'Residuals vs Obs. No.')
> plot(ch2fit$ fit,ch2fit $ res,xlab = 'Fitted Values',
+ylab = 'Residuals',main = 'Residuals vs Fitted Values')
>plot(tight,ch2fit $ res, xlab = 'Airtightness',ylab = 'Residuals',
+ main = 'Residuals Vs Airtightness')
```

You may want to consider other packages and convince yourself that the output from other programs will be similar. For the time being, we ignore much of the output and concentrate on the vector of LSEs $\hat{\boldsymbol{\beta}}' = (31.37, 9.31, 2.85)$, the sum of squared errors $S(\hat{\boldsymbol{\beta}}) = 572.72$, and the estimate of σ^2, $s^2 = 572.72/21 = 27.27$. The square root of s^2 is listed in the output. S-Plus calls it the residual standard error.

In addition, the software can calculate and store the vector of fitted values and the vector of residuals. This is useful for generating residual plots that help us check the model assumptions. Figure 4.5(a) shows plots of the residuals e_i against the order i, Figure 4.5(b) residuals e_i against fitted values $\hat{\mu}_i$, and Figure 4.5(c) residuals e_i against the explanatory variable airtightness. These plots do not show any systematic patterns in the residuals. Hence, we conclude, at least for now, that the model assumptions are reasonable. We will revisit this topic in a later chapter.

The estimate $\hat{\beta}_1 = 9.31$ implies that, on average, there is a difference of 9.31 parts per billion (ppb) in the ambient formaldehyde concentration in two homes having identical airtightness but different insulations—one with UFFI present and the other without it.

4.3 STATISTICAL INFERENCE

For the following discussion we assume that the errors in Eq. (4.6) are normally distributed. We discuss how to construct confidence intervals and how to test hypotheses.

4.3.1 CONFIDENCE INTERVALS AND TESTS OF HYPOTHESES FOR A SINGLE PARAMETER

Usually, one is interested in making inferences about a single regression parameter β_i or about a single linear combination of the coefficients $\theta = \boldsymbol{a}'\boldsymbol{\beta}$. We have studied the distribution of $\hat{\boldsymbol{\beta}}$ previously and have found that

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(X'X)^{-1})$$

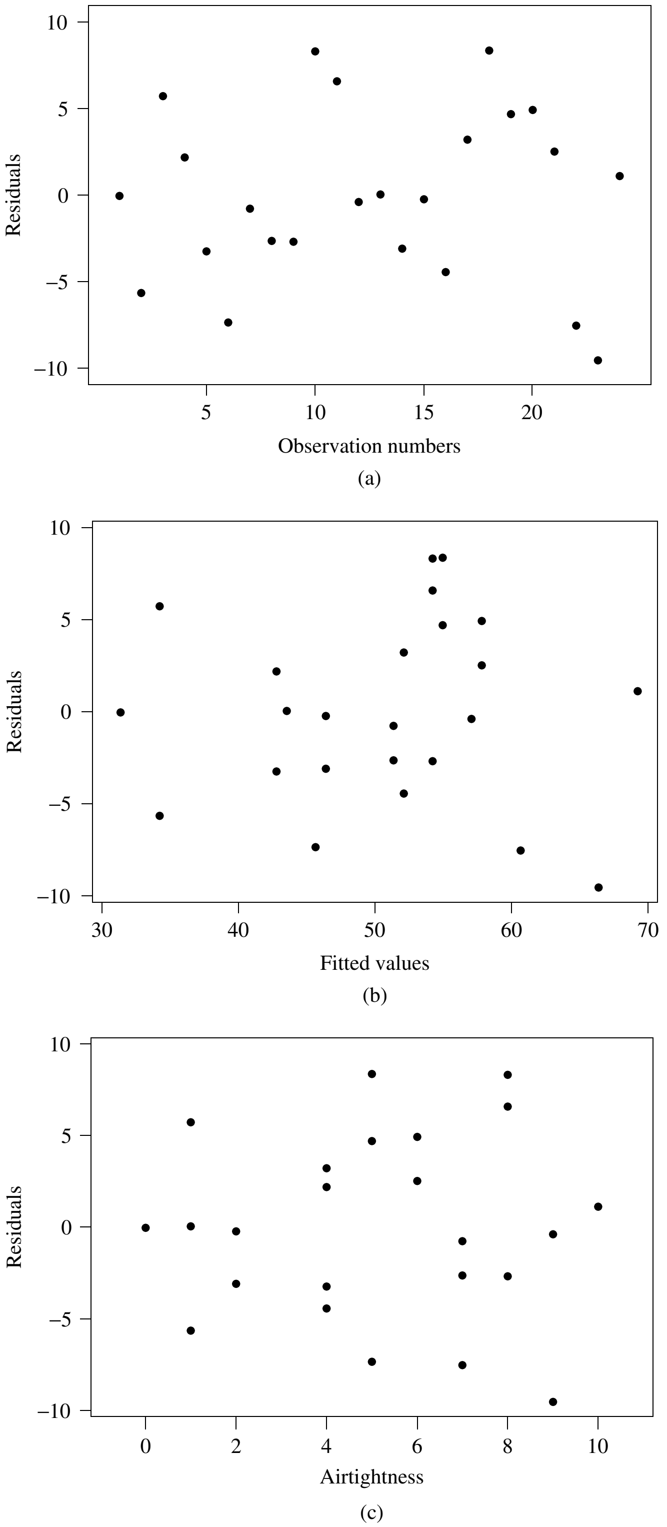
Suppose we are interested in making inferences about one of these coefficients, β_i. The estimate of β_i is given by $\hat{\beta}_i$, and its variance is given by $\sigma^2 v_{ii}$, where v_{ii} is the corresponding diagonal element in the matrix $(X'X)^{-1}$. The sampling distribution of $\hat{\beta}_i$ is

$$\hat{\beta}_i \sim N(\beta_i, \sigma^2 v_{ii}) \tag{4.22}$$

The variance of the errors, σ^2, is unknown and must be estimated. As estimate we use the unbiased estimator of σ^2,

$$s^2 = \frac{1}{n-p-1} S(\hat{\boldsymbol{\beta}}) \tag{4.23}$$

FIGURE 4.5
Residual Plots: UFFI Data



(a)

(b)

(c)

We know that

i. $(\hat{\beta}_i - \beta_i)/\sigma\sqrt{v_{ii}} \sim N(0, 1)$. This follows from Eq. (4.22).

ii. $\dfrac{(n-p-1)s^2}{\sigma^2} \sim \chi^2_{n-p-1}$.

iii. s^2 and $\hat{\boldsymbol{\beta}}$ are independent .

The results in (ii) and (iii) were shown in Section 4.2.2 and are also shown in the appendix. It follows from properties of the t distribution (see appendix to Chapter 2) that

$$T = \frac{\hat{\beta}_i - \beta_i}{s\sqrt{v_{ii}}} = \frac{(\hat{\beta}_i - \beta_i)/\sigma\sqrt{v_{ii}}}{\sqrt{\frac{(n-p-1)s^2}{\sigma^2}\Big/ n-p-1}} \sim t(n-p-1) \tag{4.24}$$

This quantity is used to construct confidence intervals and to test hypotheses about β_i. The ratio T is easy to remember:

$$T = \frac{\text{estimate} - \text{parameter}}{\text{s.e.(estimate)}}$$

relates the difference between the estimate and the true value to the standard error of the estimate, s.e.$(\hat{\beta}_i) = s\sqrt{v_{ii}}$. The standard error estimates the overall variability of the estimate in repeated random samples.

UFFI Example Continued

In this example, $\hat{\boldsymbol{\beta}}' = (31.37, 9.31, 2.85)$, $s^2 = 27.2$, and

$$(X'X)^{-1} = \begin{bmatrix} 0.2219 & -0.0856 & -0.0268 \\ -0.0856 & 0.1667 & 0.0004 \\ -0.0268 & 0.0004 & 0.0052 \end{bmatrix}$$

Let us study β_1, the effect of formaldehyde insulation on the ambient formaldehyde concentration. Is there a difference in the average concentration between homes of equal airtightness but different UFFI insulation? If insulation does not matter, then $\beta_1 = 0$. To answer this question, we test the hypothesis $\beta_1 = 0$. We know that $\hat{\beta}_1 = 9.31$, $V(\hat{\beta}_1) = 0.1667\sigma^2$, and s.e.$(\hat{\beta}_1) = s\sqrt{0.1667} = 5.22\sqrt{0.1667} = 2.13$. The t statistic for the coefficient $\hat{\beta}_1$ is

$$t_0(\hat{\beta}_1) = (\hat{\beta}_1 - 0)/\text{s.e.}(\hat{\beta}_1) = 9.31/2.13 = 4.37$$

The subscript zero indicates that we test the hypothesis $\beta_1 = 0$; the argument $\hat{\beta}_1$ in parentheses indicates that the statistic refers to the estimate $\hat{\beta}_1$. If there is no danger of confusion, we just write $t(\hat{\beta}_1)$. Since there are 24 observations and three parameters in the model, the residual sum of squares has 21 degrees of freedom. The probability value of this test statistic for a two-sided alternative $(\beta_1 \neq 0)$ is given by

$$P(|T| > 4.37) = 2P(T > 4.37) \approx 0.0003$$

Here, we use the t distribution with 21 degrees of freedom. The probability is very small—smaller than any reasonable significance level. Thus, there is very strong evidence that β_1 differs from 0. There is very strong evidence that homes with UFFI insulation have higher formaldehyde concentration levels.

A 95% confidence interval for β_1 is given by

$$\hat{\beta}_1 \pm t(0.975; 21)\text{s.e.}(\hat{\beta}_1),$$
$$9.31 \pm (2.08)(2.13), \quad 9.31 \pm 4.43, \quad \text{or} \quad (4.88, 13.74)$$

Note that $t(0.975; 21) = 2.08$ and $t(0.025; 21) = -2.08$ are the 97.5th and the 2.5th percentiles of the t distribution with 21 degrees of freedom, respectively. We are 95% confident that the interval (4.88, 13.74) covers the true, but unknown, difference in the average ambient formaldehyde concentration of homes with and without UFFI insulation.

One can repeat these calculations for the other parameter β_2, which represents the effect of airtightness on the average ambient formaldehyde concentration. The relevant diagonal element of $(X'X)^{-1}$ is 0.0052, and the standard error is

$$\text{s.e.}(\hat{\beta}_2) = s\sqrt{0.0052} = (5.22)\sqrt{0.0052} = 0.37$$

The t ratio, $t(\hat{\beta}_2) = (2.85 - 0)/0.37 = 7.58$ is very large, and the probability of obtaining such an extreme value from a t distribution with 21 degrees of freedom is negligible; the probability value for a two-sided alternative, $2P(T > 7.58)$, is essentially zero. Hence, there is little doubt that airtightness of a home increases the formaldehyde concentration in the home. A 99% confidence interval for β_2 is given by

$$2.85 \pm t(0.995; 21)(0.37)$$
$$2.85 \pm (2.83)(0.37), \quad \text{or from } 1.80 \text{ to } 3.90$$

We could repeat the calculations for the intercept β_0, which mathematically is the average concentration for homes without UFFI and with airtightness zero. Here (and also in many other applications) the intercept does not have much physical meaning, and we skip the calculation.

Note that estimates, standard errors, t ratios, and probability values for the coefficients are standard output of all statistical software packages.

Linear Combination of Coefficients

Suppose that we are interested in a linear combination of the regression coefficients. For instance, suppose we are interested in estimating the average formaldehyde concentration in homes with UFFI and with airtightness 5. That is,we are interested in

$$\theta = \beta_0 + \beta_1 + 5\beta_2 = \boldsymbol{a}'\boldsymbol{\beta}$$

where $\boldsymbol{a}' = (1,\ 1,\ 5)$ is a vector of known coefficients. The estimate of θ is given by

$$\hat{\theta} = \boldsymbol{a}'\hat{\boldsymbol{\beta}} = (1,\ 1,\ 5)\begin{bmatrix} 31.37 \\ 9.31 \\ 2.85 \end{bmatrix} = 54.96$$

Before we can construct a confidence interval for θ, we need to study the sampling distribution of $\hat{\theta}$. From properties of linear combinations of normal random variables, we know that

$$\hat{\theta} = \boldsymbol{a}'\hat{\boldsymbol{\beta}} \sim N(\theta, \sigma^2 \boldsymbol{a}'(X'X)^{-1}\boldsymbol{a})$$

Replacing σ^2 by the estimate s^2, and after going through similar steps as those in Eq. (4.24), we find that

$$T = \frac{\hat{\theta} - \theta}{s\sqrt{\boldsymbol{a}'(X'X)^{-1}\boldsymbol{a}}} \sim t(21) \tag{4.25}$$

Hence, a 95% confidence interval is given by

$$\hat{\theta} \pm t(0.975;\, 21)s\sqrt{\boldsymbol{a}'(X'X)^{-1}\boldsymbol{a}} \tag{4.26}$$

With $s = 5.22$ and

$$\boldsymbol{a}'(X'X)^{-1}\boldsymbol{a} = (1,\ 1,\ 5)(X'X)^{-1}\begin{pmatrix} 1 \\ 1 \\ 5 \end{pmatrix} = 0.0833$$

the 95% confidence interval for θ is

$$54.96 \pm 2.08(5.22\sqrt{0.0833}),$$
$$54.96 \pm (2.08)(1.51), \quad 54.96 \pm 3.13, \quad \text{or } (51.83, 58.09)$$

We are 95% confident that the interval (51.83, 58.09) will cover the true average concentration of homes with UFFI and airtightness 5. Most statistics software packages allow you to ask for this information.

Gas Consumption Example

In this example, we are interested in predicting the gas consumption of an automobile from its size and engine characteristics. The data are given in Table 1.4. There are $n = 38$ cars and measurements on fuel efficiency in miles per gallon (y), weight (x_1), engine displacement (x_2), number of cylinders (x_3), horsepower (x_4), acceleration (x_5), and engine type (x_6). Part of the data and variables x_1, x_2, and x_3 were discussed in Chapter 1 and also at the beginning of this chapter. Now let us consider all six regressor variables. Initial exploration with the data indicates that it is preferable to consider $z = 100/y$, the gas consumption per 100 traveled miles, as the response. A thorough discussion of this point will be given in Section 6.5 when we discuss transformations. In the following, we consider

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \epsilon \tag{4.27}$$

TABLE 4.2 SAS OUTPUT OF THE FUEL CONSUMPTION EXAMPLE

```
The SAS System
Model: MODEL1
Dependent Variable: Z
Analysis of Variance
                             Sum of      Mean
Source            DF         Squares     Square        F Value      Prob > F
Model              6         46.41156    7.73526       79.015       0.0001
Error             31          3.03477    0.09790
C Total           37         49.44632
   Root MSE  0.31288    R-square 0.9386
   Dep Mean  4.33061    Adj R-sq 0.9267
    C.V.     7.22491
Parameter Estimates
                            Parameter    Standard      T for H0:
Variable          DF        Estimate     Error         Parameter = 0   Prob > |T|
INTERCEP           1       −2.599749     0.66312133       −3.920       0.0005
X1                 1        0.787706     0.45173293        1.744       0.0911
X2                 1       −0.004892     0.00269495       −1.815       0.0792
X3                 1        0.444251     0.12263114        3.623       0.0010
X4                 1        0.023605     0.00673885        3.503       0.0014
X5                 1        0.068804     0.04419393        1.557       0.1297
X6                 1       −0.959720     0.26667148       −3.599       0.0011
```

Least squares estimates of the parameters, their standard errors, t ratios, and probability values are given in Table 4.2, the output from SAS, another popular software package. Different software packages will use slightly different formats, but all of them will supply most of the information in Table 4.2.

Furthermore, in this example, $s^2 = 0.0979$, and the inverse of the matrix $X'X$, is given by

$$(X'X)^{-1}$$

$$= \begin{bmatrix} 4.4918 & 0.6045 & 0.0019 & -0.1734 & -0.0210 & -0.2361 & 0.1872 \\ & 2.0845 & -0.0092 & -0.2052 & -0.0239 & -0.1081 & 0.7099 \\ & & 0.0001 & -0.0005 & 0.0001 & 0.0005 & -0.0030 \\ & & & 0.1536 & 0.0009 & -0.0011 & -0.2001 \\ & \textit{Symmetric} & & & 0.0005 & 0.0017 & -0.0073 \\ & & & & & 0.0200 & -0.0051 \\ & & & & & & 0.7264 \end{bmatrix}$$

Consider the parameter β_5, which measures the effect of x_5 (acceleration) on the average fuel consumption. The estimate is $\hat{\beta}_5 = 0.0688$. From the relevant diagonal element in $(X'X)^{-1}$, we find that $V(\hat{\beta}_5) = \sigma^2(0.0200)$, and s.e.$(\hat{\beta}_5) = \sqrt{0.0979}\,\sqrt{0.0200} = 0.0442$.

For a test of the hypothesis $\beta_5 = 0$, we consider the t statistic

$$t(\hat{\beta}_5) = \hat{\beta}_5/\text{s.e.}(\hat{\beta}_5) = 0.0688/0.0442 = 1.56$$

and its corresponding p value

$$P(|T| > 1.56) = 2P(T > 1.56) = 0.1297 \tag{4.28}$$

Note that the appropriate degrees of freedom are $n - 7 = 38 - 7 = 31$.

The probability value indicates that at the 5% significance level one cannot reject the hypothesis that $\beta_5 = 0$, given that the other variables x_1, x_2, x_3, x_4, x_6 have been included in the model. The t ratio $t(\hat{\beta}_5)$ assesses the potential effect of x_5, having adjusted the analysis for all other variables in the model. The result implies that on top of x_1, x_2, x_3, x_4, x_6 in the model, x_5 is not an important predictor of gas consumption. On the other hand, the probability value for $\hat{\beta}_6$ indicates that there is evidence for rejecting the hypothesis that $\beta_6 = 0$. It implies that x_6 is important in predicting gas consumption, even if x_1, x_2, x_3, x_4, x_5 are already in the model.

We need to remember that any inference procedure depends on the validity of the assumptions that we make about the errors ϵ. We must always check the residuals for any violations of the assumptions. The residual plots in Figure 4.6 [(a) residuals against observation number, (b) residuals against fitted values, (c) residuals against weight, and (d) residuals against displacement] indicate no systematic unusual patterns, and we conclude that the model assumptions are justified.

4.3.2 PREDICTION OF A NEW OBSERVATION

Suppose that we are interested in predicting the response of a new case, for example, the formaldehyde concentration in a new home with UFFI insulation and airtightness 5. Let y_p represent this unknown concentration,

$$y_p = \beta_0 + \beta_1 1 + \beta_2 5 + \epsilon_p = \mu_p + \epsilon_p$$

In other words, $y_p \sim N(\mu_p, \sigma^2)$. The mean $\mu_p = (1, 1, 5)\boldsymbol{\beta} = \boldsymbol{a}'\boldsymbol{\beta}$ depends on the specified (fixed) levels of the regressor variables and the parameter $\boldsymbol{\beta}$. If the parameter $\boldsymbol{\beta}$ were known exactly, then we could use μ_p as our prediction. Any other choice would have a larger expected squared error. You can see this by using any other prediction f and considering the expected squared future error,

$$\begin{aligned} E(y_p - f)^2 &= E[(y_p - \mu_p) + (\mu_p - f)]^2 \\ &= E(y_p - \mu_p)^2 + (\mu_p - f)^2 + (\mu_p - f)E(y_p - \mu_p) \\ &= \sigma^2 + (\mu_p - f)^2 \geq \sigma^2 \end{aligned}$$

However, since the parameter $\boldsymbol{\beta}$ and μ_p are unknown, we need to replace them with their LSEs. Our point prediction for the response at the new case is given by $\hat{\mu}_p = (1, 1, 5)\hat{\boldsymbol{\beta}} = 54.96$. We had calculated this earlier, and had denoted it by $\hat{\theta}$. To assess the precision of this prediction, we need to take account of two sources of variability. First, we have only an estimate $\hat{\mu}_p$ of μ_p, and there is uncertainty from the estimation. Second, there is variability of a single observation y_p around its mean μ_p. Consider the prediction error,

$$y_p - \hat{\mu}_p = \mu_p - \hat{\mu}_p + \epsilon_p$$

FIGURE 4.6
Residual Plots: Gas Consumption Data

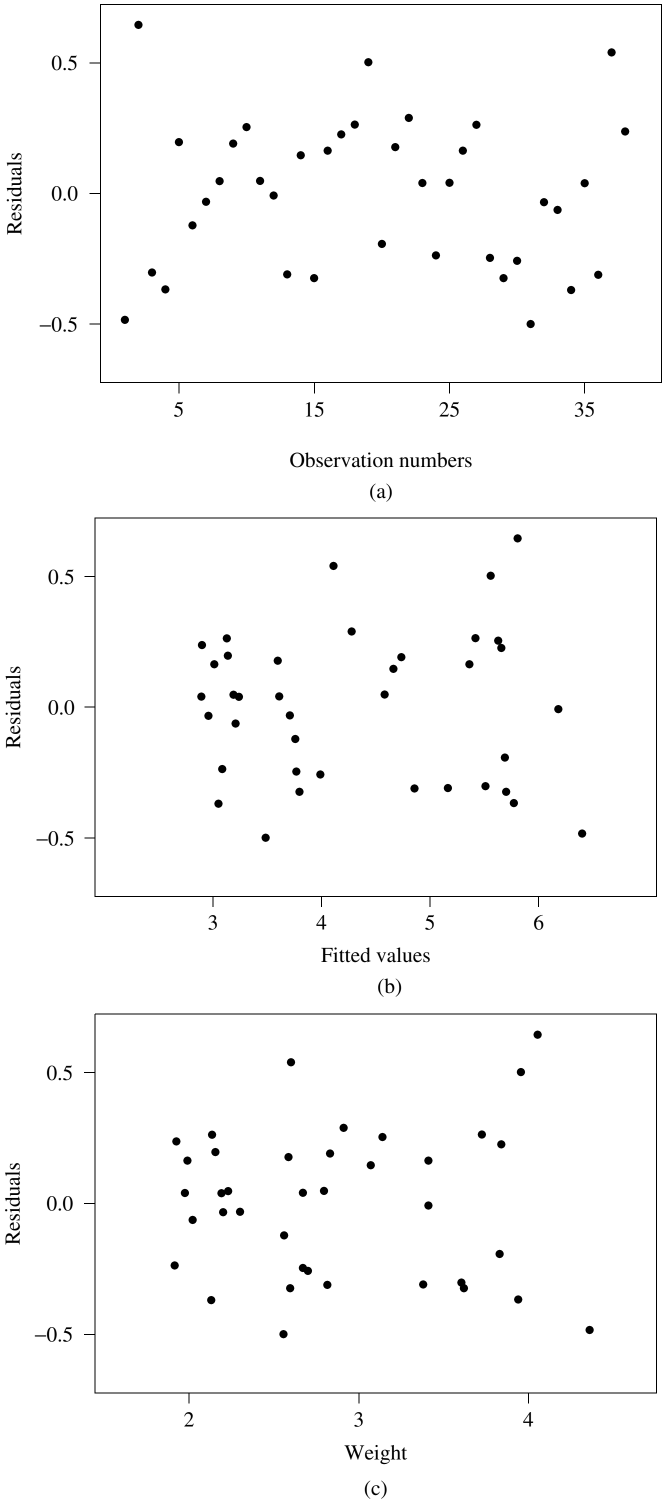
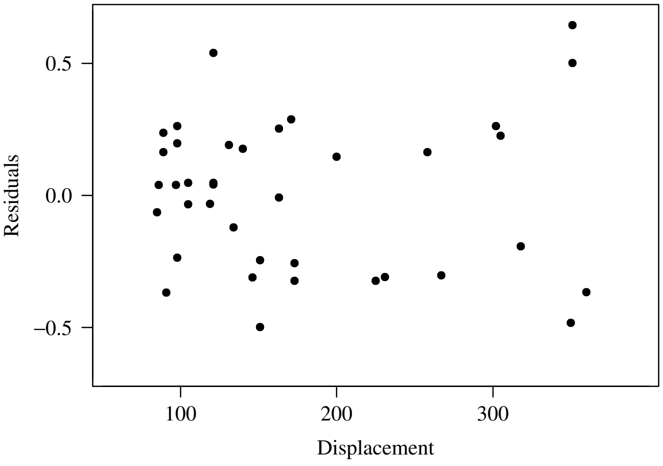


FIGURE 4.6 (Continued)



(d)

Since we consider a new case, and since the error for a new observation is independent of the errors in the observations that we used for estimating $\boldsymbol{\beta}$, the two errors, $(\mu_p - \hat{\mu}_p)$ and ϵ_p, are independent. Hence, the variance is given by

$$\begin{aligned} V(y_p - \hat{\mu}_p) = V(\hat{\mu}_p) + V(\epsilon_p) &= \sigma^2 \boldsymbol{a}'(X'X)^{-1}\boldsymbol{a} + \sigma^2 \\ &= (1 + \boldsymbol{a}'(X'X)^{-1}\boldsymbol{a})\sigma^2 \end{aligned}$$

where in our special case $\boldsymbol{a}' = (1,\ 1,\ 5)$. Hence,

$$y_p - \hat{\mu}_p \sim N(0, \sigma^2(1 + \boldsymbol{a}'(X'X)^{-1}\boldsymbol{a}))$$

and

$$T = \frac{y_p - \hat{\mu}_p}{s\sqrt{1 + \boldsymbol{a}'(X'X)^{-1}\boldsymbol{a}}} \sim t(n - p - 1)$$

The denominator in this ratio is the standard error of the prediction error

$$\text{s.e.}(y_p - \hat{\mu}_p) = s\sqrt{1 + \boldsymbol{a}'(X'X)^{-1}\boldsymbol{a}}$$

Here we have used the same argument as in the derivation of the t ratios for individual coefficients; see Eq. (4.24).

This result implies that

$$P\left(-t\left(1 - \frac{\alpha}{2}; n - p - 1\right) \le \frac{y_p - \hat{\mu}_p}{\text{s.e.}(y_p - \hat{\mu}_p)} \le t\left(1 - \frac{\alpha}{2}; n - p - 1\right)\right) = 1 - \alpha$$

$$\begin{aligned} P\Bigg(\hat{\mu}_p - t\left(1 - \frac{\alpha}{2}; n - p - 1\right)&\text{s.e.}(y_p - \hat{\mu}_p) \\ < y_p < \hat{\mu}_p + t\left(1 - \frac{\alpha}{2}; n - p - 1\right)&\text{s.e.}(y_p - \hat{\mu}_p)\Bigg) = 1 - \alpha \end{aligned}$$

Hence, a $100(1-\alpha)\%$ prediction interval for y_p is given as

$$\hat{\mu}_p \pm t\left(1-\frac{\alpha}{2}; n-p-1\right) s\sqrt{1+\boldsymbol{a}'(X'X)^{-1}\boldsymbol{a}} \qquad (4.29)$$

For our example with $n-p-1=24-3=21$, $\boldsymbol{a}'=(1,1,5)$, and the estimates in Table 4.1, we obtain the 95% prediction interval

$$54.96 \pm (2.08)(5.22)\sqrt{1+0.0833}$$

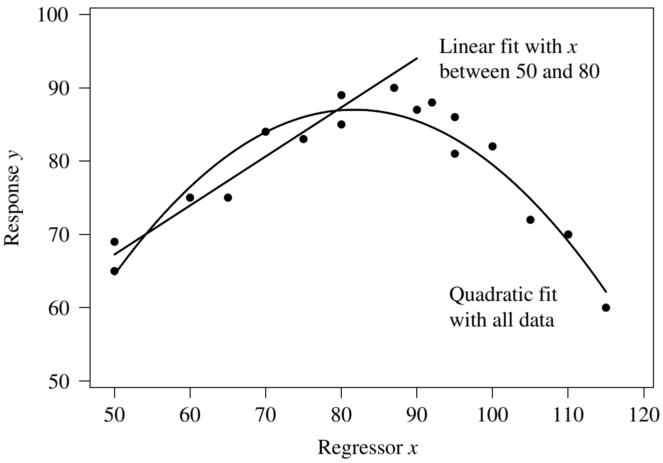
$$54.96 \pm (2.08)(5.43),\ 54.96 \pm 11.30, \quad \text{or} \quad (43.66, 66.26)$$

Our best prediction is 54.96; our uncertainty for the new value ranges from 43.7 to 66.3. Note that the prediction interval is much wider than the confidence interval for the mean response $\mu_p(=\theta)$. This is because a prediction interval is concerned with a single new observation and not the average for a fixed setting on the regressor variables.

A Caution on Predictions One should be cautious when predicting values of y for sets of x's that are very different from those that are used to fit the model. Extrapolation beyond the experimental region may lead to unreasonable results because the model is descriptive of the relationship between y and $x_1, \ldots, x_p$ only in the region of the observed x's. We are always unsure about the form of the model in a region of the x's for which we have no data. For illustration, consider Figure 4.7, which displays the relationship between a dependent variable y and a single regressor variable x. For values of x in the range from 50 to 120, we entertain a quadratic model and the fitted curve is shown in the figure. Now suppose that we had x only over the range from 50 to 80. Then a straight line model will fit the data quite well. However, Figure 4.7 shows that the prediction from the linear model of y for $x=120$ would be very misleading.

Good predictions require a valid regression model—that is, a model in which the predictor variables are significant. A model in which the influence of regressor varibles is established poorly will not do much for prediction.

FIGURE 4.7 Predicting Outside the Study Region



4.4 THE ADDITIONAL SUM OF SQUARES PRINCIPLE

4.4.1 INTRODUCTION

In this section, we describe a procedure for testing simultaneous statements about several parameters. For illustration, consider the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon \tag{4.30}$$

Suppose that previous studies suggest

$$\beta_1 = 2\beta_2 \quad \text{and} \quad \beta_3 = 0$$

How can we simulteneously test these two restrictions? The restrictions specify values for two linear combinations of the parameters. Our restrictions can be restated in vector form as

$$\begin{bmatrix} 0 & 1 & -2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

or as $A\boldsymbol{\beta} = \mathbf{0}$, where the 2 × 4 matrix $A = \begin{bmatrix} 0 & 1 & -2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$, $\boldsymbol{\beta}' = (\beta_0, \beta_1, \beta_2, \beta_3)$, and $\mathbf{0} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$.

Note that the two linear restrictions are **not** linearly dependent, and hence the matrix A has rank 2. A situation in which this would not be the case is $\beta_1 = 2\beta_2$ and $4\beta_1 = 8\beta_2$. In this case, the second condition is superfluous and can be ignored. The rank of the implied A matrix would be one.

Under the null hypothesis $H_0 : A\boldsymbol{\beta} = \mathbf{0}$, the **full model** in Eq. (4.30) simplifies to

$$y = \beta_0 + \beta_2(2x_1 + x_2) + \epsilon \tag{4.31}$$

We call this the **restricted model** because its form is constrained by H_0. For a test of $A\boldsymbol{\beta} = \mathbf{0}$ we compare two models: the full model in Eq. (4.30) and the restricted model in Eq. (4.31). We illustrate the general approach with the following two examples.

Gas Consumption Example Continued

Previously, we considered the model

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \epsilon \tag{4.32}$$

The estimates, t ratios, and probability values of the estimates were listed in Table 4.2.

Consider a test of the hypothesis that the last three regressor variables can be omitted from the model. The hypothesis

$$\beta_4 = \beta_5 = \beta_6 = 0$$

can be written in matrix form

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

where the 3 × 7 matrix

$$A = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}; \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_6 \end{bmatrix}; \quad \text{and} \quad \mathbf{0} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Under H_0, the model reduces to the restricted model

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon \qquad (4.33)$$

Note that we have reduced the number of model parameters from seven to four. The matrix A has rank 3, and our hypothesis involves three independent restrictions. Failing to reject the null hypothesis implies that the associated variables x_4, x_5, x_6 are not important, given that the rest of the variables are already in the model. On the other hand, a rejection of the null hypothesis indicates that at least one of the variables x_4, x_5, x_6 is important, in addition to the regressor variables x_1, x_2, and x_3.

OC Example

Our model in Chapter 1 relates the HDL (cholesterol) at the end of the study (y) to the initial HDL (z) and indicators for five different drug regimes,

$$y = \alpha z + \beta_1 x_1 + \cdots + \beta_5 x_5 + \epsilon \qquad (4.34)$$

Here $x_1, \ldots, x_5$ are indicator variables denoting the five oral contraceptive groups. The most interesting question is whether there are differences among the five groups. In terms of our parameters, we ask whether there is any evidence that $\beta_1, \ldots, \beta_5$ differ. To examine this question, we consider the hypothesis

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5$$

The model (4.34), written in vector notation, is

$$\mathbf{y} = \alpha \mathbf{z} + \beta_1 \mathbf{x}_1 + \cdots + \beta_5 \mathbf{x}_5 + \boldsymbol{\epsilon} \qquad (4.35)$$

Under the null hypothesis that the five β's are equal,

$$\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \gamma \qquad (4.36)$$

the model becomes

$$\begin{aligned} \mathbf{y} &= \alpha \mathbf{z} + \gamma \mathbf{x}_1 + \gamma \mathbf{x}_2 + \gamma \mathbf{x}_3 + \gamma \mathbf{x}_4 + \gamma \mathbf{x}_5 + \boldsymbol{\epsilon} \\ &= \gamma \mathbf{1} + \alpha \mathbf{z} + \boldsymbol{\epsilon} \end{aligned} \qquad (4.37)$$

since the indicator structure of $\mathbf{x}_i$ implies that $\mathbf{x_1} + \mathbf{x_2} + \cdots + \mathbf{x_5} = \mathbf{1}$, a $(n \times 1)$ vector of ones. The full model contains six coefficients and the restricted model only two. Hence, we reduced the dimension of the model from six parameters $(\alpha, \beta_1, \ldots, \beta_5)$ in Eq. (4.35) to just two parameters (α, γ) in Eq. (4.37).

Geometrically, one can visualize the restriction as follows. In the original model, $E(\boldsymbol{y})$ is an element of $L(\boldsymbol{z}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_5)$, a six-dimensional subspace of R^{50}. Under the null hypothesis, $E(\boldsymbol{y})$ is an element of $L(\boldsymbol{z}, \mathbf{1})$, a two-dimensional subspace of the subspace $L(\boldsymbol{z}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_5)$.

The restrictions in Eq. (4.36) are equivalent to

$$\begin{aligned} \beta_1 - \beta_2 &= 0 \\ \beta_1 - \beta_3 &= 0 \\ \beta_1 - \beta_4 &= 0 \\ \beta_1 - \beta_5 &= 0 \end{aligned} \tag{4.38}$$

and can also be written as

$$A\boldsymbol{\beta} = \mathbf{0} \tag{4.39}$$

where

$$A = \begin{bmatrix} 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 \end{bmatrix}; \quad \boldsymbol{\beta}' = [\alpha, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5]$$

The rank of the matrix A is 4; we are testing four linearly independent restrictions among the six parameters. Note that there are many other ways of parameterizing the restrictions. One could write $\beta_1 = \beta_2$, $\beta_2 = \beta_3$, $\beta_3 = \beta_4$, $\beta_4 = \beta_5$, and select

$$A = \begin{bmatrix} 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix}$$

which is another matrix of rank 4. It turns out that the particular parameterization does not matter. Testing the hypothesis in Eq. (4.39) is usually referred to as testing a general linear hypothesis since the test involves linear functions of the parameters.

4.4.2 TEST OF A SET OF LINEAR HYPOTHESES

Suppose we have the model

$$\boldsymbol{y} = \beta_0 \mathbf{1} + \beta_1 \boldsymbol{x}_1 + \cdots + \beta_p \boldsymbol{x}_p + \boldsymbol{\epsilon} \tag{4.40}$$

and want to test the hypothesis that a certain set of linear combinations of $\beta_0, \beta_1, \ldots, \beta_p$ are zero. That is,

$$A\boldsymbol{\beta} = \mathbf{0} \tag{4.41}$$

where A is an $l \times (p+1)$ matrix of rank l.

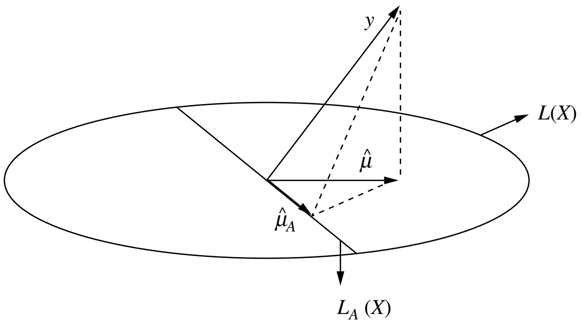


FIGURE 4.8 Geometric Representation of $L(X)$ and $L_A(X)$

Our test procedure relies on the **additional sum of squares** principle. First, we look at the problem geometrically. As usual, we write the model (4.40) in vector notation,

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

The model component $\boldsymbol{\mu} = X\boldsymbol{\beta}$ is in the $p + 1$ dimensional subspace $L(X)$ that is spanned by the regressor vectors in $X = [\mathbf{1}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p]$.
Let

$$L_A(X) = \{\beta_0\mathbf{1} + \beta_1\boldsymbol{x}_1 + \cdots + \beta_p\boldsymbol{x}_p \mid A\boldsymbol{\beta} = \mathbf{0}\}$$

be the subspace spanned by all linear combinations of the regressor vectors $\mathbf{1}, \boldsymbol{x}_{\mathbf{1}}, \ldots, \boldsymbol{x}_{\boldsymbol{p}}$ with coefficients satisfying the restriction $A\boldsymbol{\beta} = \mathbf{0}$. $L_A(X)$ is a subspace of $L(X)$, with dimension $p + 1 - l$. This is easy to see because every linear combination in $L_A(X)$ is also an element in $L(X)$ (see Figure 4.8). If $E(\boldsymbol{y})$ is an element of $L_A(X)$, then the hypothesis $A\boldsymbol{\beta} = \mathbf{0}$ is true exactly.

Let $\hat{\boldsymbol{\mu}}$ be the orthogonal projection of $\boldsymbol{y}$ onto $L(X)$, and let $\hat{\boldsymbol{\mu}}_A$ be the orthogonal projection of $\boldsymbol{y}$ onto the subspace $L_A(X)$. If the null hypothesis is true, then $\hat{\boldsymbol{\mu}}$ should be close to $L_A(X)$, and the difference $\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_A$ and its squared length $(\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_A)'(\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_A) = \|\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_A\|^2$ should be small. We would be surprised if this quantity was exactly 0 because there is random variation in the model. The formal procedure takes this variability into account. In the results that follow, we calculate the distribution of the random variable $\|\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_A\|^2$ under the hypothesis $A\boldsymbol{\beta} = \mathbf{0}$.

Once again, we use a technique similar to the one we used in showing that $\hat{\boldsymbol{\beta}}$ and the residual sum of squares $S(\hat{\boldsymbol{\beta}})$ are independent.

Theorem Suppose that $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 I)$. Let $\hat{\boldsymbol{\mu}}$ be the orthogonal projection of $\boldsymbol{y}$ onto $L(\boldsymbol{X})$ and $\hat{\boldsymbol{\mu}}_A$ the orthogonal projection of $\boldsymbol{y}$ onto $L_A(X)$. Assume that $A\boldsymbol{\beta} = \mathbf{0}$. We can show that (i) $\|\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_A\|^2/\sigma^2 \sim \chi^2_l$, and (ii) $\|\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_A\|^2$ is independent of $S(\hat{\boldsymbol{\beta}}) = (\boldsymbol{y} - \hat{\boldsymbol{\mu}})'(\boldsymbol{y} - \hat{\boldsymbol{\mu}})$.

Proof: See the appendix. ■

The Theorem implies the following

Corollary: Under the hypothesis $A\boldsymbol{\beta} = \mathbf{0}$, the ratio

$$F = \frac{\|\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_A\|^2 / l}{S(\hat{\boldsymbol{\beta}})/(n - p - 1)} \tag{4.42}$$

has an F distribution with l and $n - p - 1$ degrees of freedom.

Proof: The Theorem states that

$$\|\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_A\|^2 / \sigma^2 \sim \chi_l^2$$

and that it is independent of $S(\hat{\boldsymbol{\beta}})$. Earlier, we showed that

$$S(\hat{\boldsymbol{\beta}})/\sigma^2 \sim \chi_{n-p-1}^2$$

In the appendix to Chapter 2, we stated that the ratio of two independent normalized chi-square distributions leads to an F distribution. Hence,

$$\frac{\|\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_A\|^2 / \sigma^2 l}{S(\hat{\boldsymbol{\beta}})/\sigma^2(n - p - 1)} = F \sim F(l, n - p - 1)$$

The quantity F in Eq. (4.42) helps us test the hypothesis $A\boldsymbol{\beta} = \mathbf{0}$. Large values of F provide evidence against the hypothesis. ■

Comments

It is easy to see why F in Eq. (4.42) is a sensible test statistic. The denominator in F is the unbiased estimator of σ^2, that we have used previously. It gives an estimate of σ^2, irrespective of whether or not H_0 is true. If $A\boldsymbol{\beta} = \mathbf{0}$, the numerator has distribution $\sigma^2 \chi_l^2 / l$ and expected value σ^2. The numerator of F also estimates σ^2, but only if $A\boldsymbol{\beta} = \mathbf{0}$. If the hypothesis is not true, we expect $\hat{\boldsymbol{\mu}}$ to differ substantially from $\hat{\boldsymbol{\mu}}_A$, and as a consequence $E(\|\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_A\|^2 / l)$ will exceed σ^2. Hence, the ratio F in Eq. (4.42) will tend to be larger than 1 if the hypothesis is false.

The term $\|\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_A\|^2$ can be interpreted as the "**additional sum of squares**," hence the title of this section. We can see this from the geometry. Look at the diagram in Figure 4.9, now slightly relabeled. Consider the right-angled triangle with end points ABC. Denote the squared distances between the points by AB^2, AC^2, and BC^2. You notice that

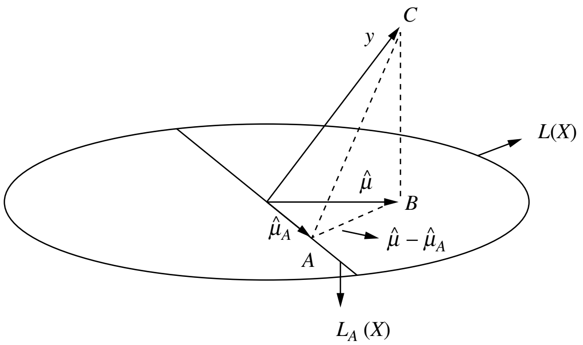
$$BC^2 = S(\hat{\boldsymbol{\beta}}),\ AB^2 = \|\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_A\|^2,\ AC^2 = S(\hat{\boldsymbol{\beta}}_A)$$

where $S(\hat{\boldsymbol{\beta}}_A)$ is the minimum value of $S(\boldsymbol{\beta})$ when $\boldsymbol{\beta}$ is restricted so that $A\boldsymbol{\beta} = \mathbf{0}$. Pythagoras theorem tells us that

$$\|\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_A\|^2 = AC^2 - BC^2 = S(\hat{\boldsymbol{\beta}}_A) - S(\hat{\boldsymbol{\beta}}) \tag{4.43}$$

Hence, the numerator of our test statistic in Eq. (4.42) is the difference of two error sum of squares: $S(\hat{\boldsymbol{\beta}}) = (\boldsymbol{y} - X\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - X\hat{\boldsymbol{\beta}})$ is the error sum of squares in

FIGURE 4.9 Geometric Interpretation of Additional Sum of Squares



the full model and $S(\hat{\boldsymbol{\beta}}_A) = (\boldsymbol{y} - X\hat{\boldsymbol{\beta}}_A)'(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}_A)$ is the error sum of squares in the restricted model—that is, the model under $A\boldsymbol{\beta} = \mathbf{0}$. This cannot be smaller than $S(\hat{\boldsymbol{\beta}})$ because we have restricted the minimization. The difference $S(\hat{\boldsymbol{\beta}}_A) - S(\hat{\boldsymbol{\beta}}) = \|\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_A\|^2$ is the **additional or extra sum of squares** that our restricted model has failed to pick up. We can also think of it as the extra sum of squares that is picked up when omitting the constraints $A\boldsymbol{\beta} = \mathbf{0}$.

Note

i. Here we have considered the hypothesis $A\boldsymbol{\beta} = \mathbf{0}$. The vector $\mathbf{0}$ on the right-hand side can be replaced by any known vector, for example, $\boldsymbol{\delta}$. The results will remain the same.

   Consider the following illustration. Take the full model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$ and the restriction $2\beta_1 + 3\beta_2 = 5$. We can write this restriction as $\beta_2 = (5/3) - (2/3)\beta_1$ and obtain the restricted model as

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + [(5/3) - (2/3)\beta_1]x_2 + \epsilon \\ &= (5/3)x_2 + \beta_0 + \beta_1[x_1 - (2/3)x_2] + \epsilon \end{aligned}$$

   The restricted estimates $\boldsymbol{\beta}_A$ can be obtained by regressing the transformed response $y - (5/3)x_2$ on the new regressor $x_1 - (2/3)x_2$. From the estimates $\hat{\beta}_{0,A}$ and $\hat{\beta}_{1,A}$ we can obtain $\hat{\beta}_{2,A} = (5/3) - (2/3)\hat{\beta}_{1,A}$. From these estimates we can obtain the residual vector $y - [\hat{\beta}_{0,A} + \hat{\beta}_{1,A}x_1 + \hat{\beta}_{2,A}x_2]$ and compute $S(\hat{\boldsymbol{\beta}}_A)$.

ii. The test for the hypothesis $A\boldsymbol{\beta} = \boldsymbol{\delta}$ can be implemented in a slightly different way as well. Consider the statistic

$$F^* = \frac{(A\hat{\boldsymbol{\beta}} - \boldsymbol{\delta})'[A'(X'X)^{-1}A]^{-1}(A\hat{\boldsymbol{\beta}} - \boldsymbol{\delta})/l}{S(\hat{\boldsymbol{\beta}})/(n - p - 1)}$$

   where l is the rank of the constraint matrix A as defined earlier. It can be shown that F^* has an F distribution with l and $(n - p - 1)$ degrees of freedom, and that the test based on F^* is identical to our earlier approach in Eq. (4.42). Some software packages provide the F^* statistic and its associated probability value automatically if you supply the A matrix and the $\boldsymbol{\delta}$ vector. However, we prefer our approach in Eq. (4.42) because we believe it to be more intuitive.

Gas Consumption Example Continued

Let us return to the restriction that we had specified,

$$H_0 : \beta_4 = \beta_5 = \beta_6 = 0$$

Not rejecting this hypothesis implies that the variables x_4, x_5, and x_6 are not important in predicting the fuel consumption z, given that the variables x_1, x_2, and x_3 are already in the model. On the other hand, rejecting the null hypothesis means that one or more of the regressor variables x_4, x_5, or x_6 contribute explanatory power beyond that provided by the variables x_1, x_2, and x_3.

The restricted model under the null hypothesis is

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon \tag{4.44}$$

The LSEs can be obtained, and it turns out that the residual sum of squares from this restricted model is $S(\hat{\boldsymbol{\beta}}_A) = 4.8036$.

Previously, we obtained the residual sum of squares, $S(\hat{\boldsymbol{\beta}}) = 3.0348$, with $n - p - 1 = 38 - 7 = 31$ degrees of freedom. Hence, the additional sum of squares is $S(\hat{\boldsymbol{\beta}}_A) - S(\hat{\boldsymbol{\beta}}) = 4.8036 - 3.0348 = 1.7688$. The full model reduces the error sum of squares by 1.7688; in other words, the constraints in the parameters have cost us an extra sum of squares of 1.7688. This sum of squares has $l = 3$ degrees of freedom since we have constrained three parameters, or equivalently since there are three independent rows in A.

Thus,

$$F = \frac{\text{additional sum of squares}/3}{S(\hat{\boldsymbol{\beta}})/31} = \frac{1.7688/3}{3.0348/31} = 6.02$$

The sampling distribution of the F ratio under the null hypothesis is F with 3 and 31 degrees of freedom. The probability value is given by

$$P(F(3, 31) > 6.02) \simeq 0.01$$

The probability value expresses the likelihood of obtaining the observed F ratio 6.02 under the null hypothesis. It is small, which makes the null hypothesis unlikely. Hence, one can rule out the null hypothesis and reject $\beta_4 = \beta_5 = \beta_6 = 0$. This implies that at least one of the variables x_4, x_5, and x_6 is important, even if variables x_1, x_2, and x_3 are already in the model.

Note that the t tests on individual parameters that we discussed earlier can also be carried out within the additional sums of squares framework. For example, consider the test $\beta_5 = 0$. This test can be formulated as testing the hypothesis

$$H_0 : A\boldsymbol{\beta} = 0$$

where $A = (0, 0, 0, 0, 0, 1, 0)$ is a row vector, and $\boldsymbol{\beta}$ is the (7×1) vector of parameters in the full model. Under this hypothesis, the restricted model becomes

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_6 x_6 + \epsilon \tag{4.45}$$

It is the model without x_5. Estimates in the restricted model, $\hat{\boldsymbol{\beta}}_A$, can be obtained. The residual sum of squares from this restricted model is $S(\hat{\boldsymbol{\beta}}_A) = 3.2720$. Comparing this with the residual sum of squares from the original (full) model, $S(\hat{\boldsymbol{\beta}}) = 3.0348$, gives us the additional sum of squares $S(\hat{\boldsymbol{\beta}}_A) - S(\hat{\boldsymbol{\beta}}) = 3.2720 - 3.0348 = 0.2372$. It has $l = 1$ degree of freedom. Hence,

$$F = \frac{\text{additional sum of squares}/1}{S(\hat{\boldsymbol{\beta}})/(n-p-1)} = \frac{0.2372}{3.0348/(38-7)} = 2.42 \qquad (4.46)$$

with probability value

$$P(F(1,31) > 2.42) = 0.13$$

Note that the computed F in Eq. (4.46) is the square of the t ratio for $\hat{\beta}_5$ in the full model: $2.42 = (1.56)^2$. This equality can be shown in general. Furthermore, note that the previous probability value is exactly the same as the one we found for the t statistic for testing $\beta_5 = 0$. Hence, the conclusions from both tests, the F test for one extra parameter and the t test for an individual coefficient, are identical. We know in general that the square of a t distributed random variable follows a certain F distribution ($t_{df}^2 = F(1, df)$; see Chapter 2, Exercise 2.2). Such F tests for one extra parameter are referred to as **partial *F* tests**, and the **additional sum of squares** is referred to as the **partial sum of squares** due to this extra variable (in our case, x_5), given that all other variables are already in the model.

OC Example Continued

The full model is

$$\boldsymbol{y} = \alpha\boldsymbol{z} + \beta_1\boldsymbol{x}_1 + \cdots + \beta_5\boldsymbol{x}_5 + \boldsymbol{\epsilon} \qquad (4.47)$$

A test of the hypothesis that there are no differences among the five drugs restricts the parameters as follows:

$$\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \gamma \text{ (say)},$$

or

$$A\boldsymbol{\beta} = \boldsymbol{0}$$

where A and $\boldsymbol{\beta}$ are given in Eq. (4.39). The restricted model

$$\boldsymbol{y} = \gamma\boldsymbol{1} + \alpha\boldsymbol{z} + \boldsymbol{\epsilon} \qquad (4.48)$$

can be estimated, and the residual sum of squares can be calculated. We find $S(\hat{\boldsymbol{\beta}}_{\boldsymbol{A}}) = 2,932.0$. The residual sum of squares of the full model in Eq. (4.47) is given by $S(\hat{\boldsymbol{\beta}}) = 2,505.0$. Hence, the additional sum of squares is

$$\|\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_A\|^2 = S(\hat{\boldsymbol{\beta}}_A) - S(\hat{\boldsymbol{\beta}}) = 2,932.0 - 2,505.0 = 427.0$$

We have placed $l = 4$ linear restrictions on the six coefficients $\alpha, \beta_1, \ldots, \beta_5$, and there are $n - p - 1 = 50 - 6 = 44$ degrees of freedom for the error sum of squares in the full model. Hence,

$$F = \frac{427/4}{2,505/44} = 1.87$$

We use the $F(4, 44)$ distribution to obtain the probability value,

$$p \text{ value} = P(F(4, 44) > 1.87) \simeq 0.13$$

Since this is quite large (certainly larger than the commonly used significance level 0.05), we believe in our null hypothesis. Hence, we find no real evidence that the drugs differ in their effect on the final HDLC.

Next, suppose we are also interested in knowing whether or not the drugs have an effect at all. This hypothesis specifies that $\beta_1 = \cdots = \beta_5 = 0$. Under this hypothesis, the model (4.47) becomes

$$\boldsymbol{y} = \alpha \boldsymbol{z} + \boldsymbol{\epsilon} \tag{4.49}$$

The LSE of α can be found $\left(\text{it is } \hat{\alpha}_A = \dfrac{\sum z_i y_i}{\sum z_i^2}\right)$, leading to the residual sum of squares $S(\hat{\boldsymbol{\beta}}_A) = \sum (y_i - \hat{\alpha}_A z_i)^2 = 3,410.68$. Hence, the additional sum of squares is

$$\|\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_A\|^2 = S(\hat{\boldsymbol{\beta}}_A) - S(\hat{\boldsymbol{\beta}}) = 3,410.68 - 2,505.0 = 905.68$$

Since there are five restrictions, 5 degrees of freedom are associated with this extra sum of squares. The test statistic for the previous hypothesis is

$$F = \frac{905.68/5}{2505/44} = 3.18$$

The probability value is given by

$$P(F(5, 44) > 3.18) \simeq 0.02$$

The probability value is small—smaller than the usual significance level 0.05. $F = 3.18$ is an extreme value under the null hypothesis. We can reject H_0 and conclude that there is evidence that the drugs affect the final HDLC.

4.4.3 JOINT CONFIDENCE REGIONS FOR SEVERAL PARAMETERS

In the UFFI example, we constructed confidence intervals for individual parameters. For instance, the 95% confidence interval for β_1 has the form

$$P(L_1 \le \beta_1 \le U_1) = 0.95$$

where $L_1 = \hat{\beta}_1 - t(0.975; n - p - 1)\text{s.e.}(\hat{\beta}_1)$, $U_1 = \hat{\beta}_1 + t(0.975; n - p - 1)$ s.e.$(\hat{\beta}_1)$, and $t(0.975; n - p - 1)$ is the 97.5% percentile of a t distribution with degrees of freedom $n - p - 1$ (see Section 4.3.1). A 95% confidence interval for β_2 has a similar form with lower and upper limits L_2 and U_2. For our data, $L_1 = 4.88$, $U_1 = 13.74$ and $L_2 = 2.08$, $U_2 = 3.62$.

In some contexts, it may be necessary to make joint confidence statements about β_1 and β_2. For example, we may want to construct a confidence region CR such that $P((\beta_1, \beta_2) \text{ is in CR}) = 0.95$. It is known that

$$\begin{aligned} P(L_1 \le \beta_1 \le U_1, L_2 \le \beta_2 \le U_2) &\le P(L_1 \le \beta_1 \le U_1) P(L_2 \le \beta_2 \le U_2) \\ &= 0.95^2 = 0.9025 \end{aligned}$$

The confidence level associated with the rectangular region obtained by taking the two marginal intervals as shown above is less than 0.95.

There is a procedure, however, to construct a joint confidence region for a set of parameters that has the required coverage. In the linear model with parameter vector $\boldsymbol{\beta}$ it can be shown that

$$\frac{(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'X'X(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{(p+1)s^2} \sim F(p+1, n-p-1)$$

where $F(p+1, n-p-1)$ denotes an F distribution with degrees of freedom $(p+1)$ and $(n-p-1)$. This result can be shown as follows.

For the linear model, $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with the standard assumptions, $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(X'X)^{-1})$, and hence $(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \sim N(\mathbf{0}, \sigma^2(X'X)^{-1})$. The results on the distribution of quadratic forms in Section 3.4 imply that

$$(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'[\sigma^2(X'X)^{-1}]^{-1}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) = (\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'X'X(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})/\sigma^2 \sim \chi^2_{p+1}$$

We also know from previous results in this chapter that $(n-p-1)s^2/\sigma^2 \sim \chi^2_{n-p-1}$, and that $\hat{\boldsymbol{\beta}}$ and s^2 are statistically independent. The ratio of two independent chi-square random variables, standardized by their degrees of freedom, has an F distribution; see the appendix in Chapter 2. Hence,

$$\frac{(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'X'X(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})/(p+1)\sigma^2}{(n-p-1)s^2/(n-p-1)\sigma^2} = \frac{(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'X'X(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{(p+1)s^2} \sim F(p+1, n-p-1)$$

This result implies that a $100(1-\alpha)\%$ joint confidence region for all parameters in $\boldsymbol{\beta}$ is given by

$$\frac{(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'X'X(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{(p+1)s^2} \le F(1-\alpha; p+1, n-p-1)$$

where $F(1-\alpha; p+1, n-p-1)$ is the $100(1-\alpha)$ percentile of an F distribution with degrees of freedom $(p+1)$ and $(n-p-1)$. Algebraically, and this is somewhat cumbersome, one needs to find the values $\boldsymbol{\beta}$ such that the previous equality is satisfied. Choosing submatrices of $(X'X)^{-1}$ appropriately, confidence regions for subsets of parameters in $\boldsymbol{\beta}$ can also be obtained. For example, a joint confidence region for β_1, β_2 uses a submatrix of $(X'X)^{-1}$ that corresponds to these coefficients (see Exercise 4.24).

For a pair of parameters, the joint confidence region is an ellipse on a two-dimensional plot. For more than two parameters it is an ellipsoid. Since joint confidence regions are rarely used in practice, we will not pursue the topic further.

4.5 THE ANALYSIS OF VARIANCE AND THE COEFFICIENT OF DETERMINATION, R^2

Let us consider the general linear model

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon \tag{4.50}$$

and the hypothesis

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

Under this hypothesis, the model reduces to

$$y = \beta_0 + \epsilon \tag{4.51}$$

This model implies that the response y has a mean $E(y) = \beta_0$ that is not affected by any of the explanatory variables. The hypothesis expresses the fact that $x_1, \ldots, x_p$ do not influence the response.

We can use the **additional sum of squares principle** to test this hypothesis. The residual sum of squares of the full model in (Eq. 4.50) is given by $S(\hat{\boldsymbol{\beta}})$. In the restricted model (Eq. 4.51), the estimate of β_0 is given by $\bar{y}$ and the "residual sum of squares" by

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{4.52}$$

This is called the **total sum of squares**, corrected for the mean. It is a measure of how the observations fluctuate around their mean. The **additional sum of squares** from the p regressor variables is given by $\sum (y_i - \bar{y})^2 - S(\hat{\boldsymbol{\beta}})$. It has $l = p$ degrees of freedom since our null hypothesis specifies p independent constraints. This quantity is usually called the **regression**, or **the model sum of squares**. It tells us how much variability is explained by the full model, over and above the simple mean model. It can be shown that the regression sum of squares (SSR) is given by

$$\begin{aligned}\text{SSR} &= \sum (y_i - \bar{y})^2 - S(\hat{\boldsymbol{\beta}}) = \mathbf{y}'\mathbf{y} - n\bar{y}^2 - (\mathbf{y} - X\hat{\boldsymbol{\beta}})'(\mathbf{y} - X\hat{\boldsymbol{\beta}}) \\ &= \hat{\boldsymbol{\beta}}'X'\mathbf{y} - n\bar{y}^2 = \hat{\boldsymbol{\beta}}'X'X\hat{\boldsymbol{\beta}} - n\bar{y}^2 \end{aligned} \tag{4.53}$$

The three sums of squares—the regression sum of squares, the residual sum of squares, and the total sum of squares—are usually displayed in a table called the analysis of variance (ANOVA) table (Table 4.3).

The degrees of freedom column in the table contains the relevant degrees of freedom. The regression sum of squares, SSR, has p degrees of freedom, because there are p regressor variables that make up the model. The error sum of squares has $n - p - 1$ degrees of freedom; the number of observations n minus the number of parameters in the model, $p + 1$. The total sum of squares has

TABLE 4.3 ANALYSIS OF VARIANCE (ANOVA) TABLE

Source	df	Sum of Squares	Mean Squares	F
Model (Regression)	p	$\text{SSR} = \hat{\boldsymbol{\beta}}'X'\mathbf{y} - n\bar{y}^2$	$\text{MSR} = \text{SSR}/p$	$\dfrac{\text{MSR}}{\text{MSE}}$
Residual (Error)	$n - p - 1$	$\text{SSE} = S(\hat{\boldsymbol{\beta}}) = (\mathbf{y} - X\hat{\boldsymbol{\beta}})'(\mathbf{y} - X\hat{\boldsymbol{\beta}})$	$\text{MSE} = \dfrac{\text{SSE}}{(n - p - 1)}$	
Corrected total	$n - 1$	$\text{SST} = \sum (y_i - \bar{y})^2$		

$n-1$ degrees of freedom, because there are n deviations from the mean, but the sum of these deviations is zero.

The fourth column contains the **mean squares**, the sums of squares divided by their respective degrees of freedom. The mean square error, $\text{MSE} = \text{SSE}/(n-p-1) = S(\hat{\beta})/(n-p-1)$, was seen earlier. It is the unbiased estimator of σ^2. The fifth column contains the F ratio for testing the hypothesis $H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$,

$$F = \frac{\text{additional sum of squares}/p}{S(\hat{\boldsymbol{\beta}})/(n-p-1)} = \frac{\text{SSR}/p}{\text{SSE}/(n-p-1)} \tag{4.54}$$

Observe that, by construction, the regression sum of squares and the error sum of squares must add up to the total sum of squares. Hence, the ANOVA table partitions the variability (the total sum of squares) into two interpretable components: a sum of squares that is explained by the model and a sum of squares that has been left unexplained.

Gas Consumption Example Continued

The basic model is

$$z = 100/y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \epsilon \tag{4.55}$$

The hypothesis

$$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$$

can be written as $A\boldsymbol{\beta} = \mathbf{0}$, where A is the 6×7 matrix of rank 6,

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$\boldsymbol{\beta}$ is the (7×1) vector of parameters, and $\mathbf{0}$ a (6×1) vector of zeros. Failure to reject this hypothesis implies that none of the variables are important in predicting the gas consumption of the vehicle. Rejection of the hypothesis implies that at least one of the variables is important in predicting gas consumption. Under the null hypothesis, the model reduces to

$$y = \beta_0 + \epsilon \tag{4.56}$$

Estimation of the full model in Eq. (4.55) gives $\text{SSE} = S(\hat{\boldsymbol{\beta}}) = 3.0348$. The total sum of squares (corrected for the mean) is easy to calculate; $\text{SST} = \sum (y_i - \bar{y})^2 = \sum y_i^2 - n\bar{y}^2 = 49.4463$. Hence, by subtraction, we find the regression sum of squares, $\text{SSR} = \text{SST} - \text{SSE} = 46.4115$. These are the sum of squares entries in Table 4.4.

The degrees of freedom are 6 (as there are six regressor variables), 31 (because we estimate seven coefficients from $n = 38$ cases), and 37 ($= n - 1$). The F ratio,

TABLE 4.4 ANOVA TABLE FOR GAS CONSUMPTION DATA

Source	df	Sum of Squares	Mean Squares	F	Prob > F
Model (Regression)	6	46.4115	7.7352	79.015	0.0001
Residual (Error)	31	3.0348	0.0979		
Corrected total	37	49.4463			

Note that this table was part of the SAS output; see Section 4.3.1.

$F = 79.015$, is large; its probability value 0.0001 is tiny. It indicates that there is strong evidence against the claim that none of the regressor variables have an influence ($\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$). In other words, we cannot discard x_1, x_2, x_3, x_4, x_5, and x_6 simultaneously; at least one of the variables is important in predicting z.

The F test in the ANOVA table is also known as a test for the overall significance of the regression. If we reject H_0, some regression relations exist. Which ones, we do not know at this point.

4.5.1 COEFFICIENT OF DETERMINATION, R^2

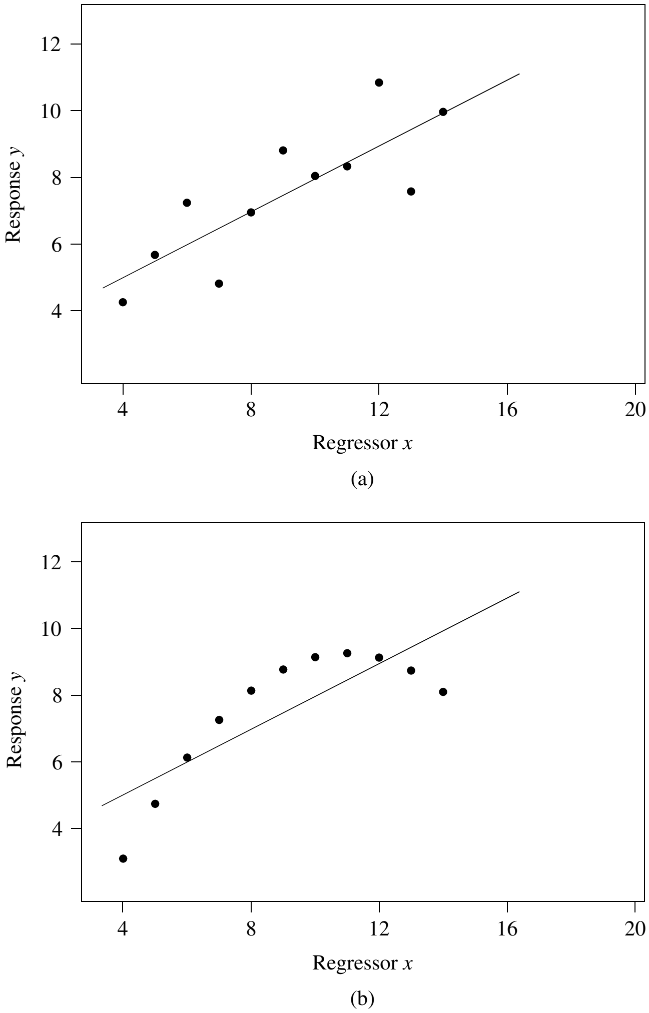
The ANOVA table partitions the total response variation into two components: SST = SSR + SSE, the variation that is explained by the regression model (SSR), and the variation that is left unexplained (SSE). The coefficient of determination R^2 is defined as the proportion of the total response variation that is explained by the model,

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\text{SST} - \text{SSE}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}} \tag{4.57}$$

For the gas consumption example and model (4.55), the total sum of squares is SST = 49.4463, the residual sum of squares (the unexplained response variation) is SSE = 3.0348, and the response variation that is explained by the model (regression sum of squares) is SSR = 49.4463 − 3.0348 = 46.4115. Hence, $R^2 = 46.4115/49.4463 = 0.9386$. This means that 94% of the variation in the response is explained by the linear model with the regressor variables $x_1, x_2, x_3, x_4, x_5, x_6$.

R^2 is a useful summary measure. It provides an overall measure of how well the model fits. It can also give feedback on the importance of adding a variable to (or deleting a variable from) a model. For instance, if we delete x_5, x_6 from the model in Eq. (4.55), the regression sum of squares reduces to 44.9505 and $R^2 = 0.9091$. This is slightly smaller, but it appears that a model without x_5 and x_6 is not much worse than the full model. This casts doubt on the inclusion of these two variables in the model. Note that adding a variable to a model increases the regression sum of squares, and hence the R^2. (In the worst case, it can stay the same.) R^2 can be made 1 by adding increasingly more explanatory variables. If we fit a model with $(n - 1)$ explanatory variables to n cases, the fit is perfect; the residual sum of squares will be zero, and $R^2 = 1$. One certainly does not want to do this because one would "overfit" the data, trying to find an explanation for every random perturbation. Hence, the use of large numbers of explanatory variables, especially when n is small, is not a good idea.

FIGURE 4.10
Different Data Plots Yielding Identical R^2



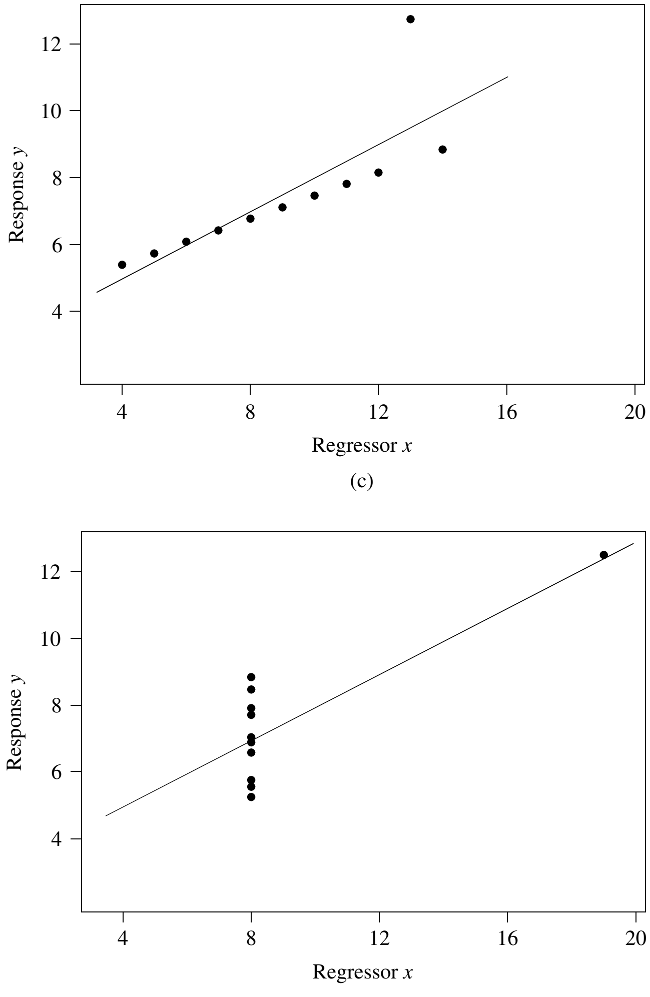
R^2 is just one summary measure of the regression fit. It alone does not tell us whether the fitted model is appropriate. Look at the data that are listed in Exercise 2.3 and plotted in Figure 4.10. It turns out (you should check this) that all four data sets lead to the same least squares estimates, the same ANOVA table, and identical R^2. However, there is only one situation (case a) in which one would say that a simple linear regression describes the data. One needs to be careful when interpreting the R^2.

4.6 GENERALIZED LEAST SQUARES

4.6.1 INTRODUCTION

The standard regression $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ assumes that the vector of errors $\boldsymbol{\epsilon}$ has zero mean and covariance matrix $V(\boldsymbol{\epsilon}) = \sigma^2 I$, which implies that all errors have the

FIGURE 4.10
(Continued)



(c)

(d)

same precision and that they are uncorrelated. In some situations, these assumptions will not be reasonable.

In certain applications, some errors have more variability than others. Consider the situation in which the response for the ith case is obtained as an average of several measurements and the number of measurements that go into that average changes from case to case. In this situation, $V(y_i) = V(\epsilon_i) = \sigma^2/n_i$, where n_i represents the number of measurements in the average y_i. The assumption of equal variance is clearly violated.

Consider the case in which responses are taken over time. For example, consider modeling the relationship between the sales of your product, its price, as well as the prices of major competitors, and the amount your company spends on advertisement. Suppose that monthly observations (e.g., the past 5 years) are available to estimate the coefficients in the regression model. You expect that a regression of sales on price and advertising will be useful, and that these regressor variables will "explain" sales. However, even after controlling for prices and advertising, deviations of the sales from their implied expected levels tend to exhibit "runs." If sales in a certain month are unusually high, then there is a good chance that they will also be high in adjacent months. This is because the economic "driving forces" (which are not in your model) are persistent, moving only slowly over time; if the economy is poor today, then it tends to be poor also in preceeding and following months. You could try to specify an additional variable, "state of the economy," and use this as an additional regressor variable. This may help, but most likely there will be some other unknown and slowly changing variables that affect your sales, and measurement errors from different periods will tend to be correlated. We refer to this as **autocorrelation** or **serial correlation** because the errors are correlated among themselves at different lags. Very often, the amount of (auto)correlation diminishes with the lag. For example, adjacent observations are the most strongly correlated, whereas errors several steps apart are less correlated. Usually, one assumes a certain structure for the auto- (or serial) correlation. Often, one assumes that $\text{Cov}(\epsilon_i, \epsilon_{i-1}) = \text{Cov}(\epsilon_i, \epsilon_{i+1}) = \sigma^2\phi$, $\text{Cov}(\epsilon_i, \epsilon_{i-2}) = \text{Cov}(\epsilon_i, \epsilon_{i+2}) = \sigma^2\phi^2, \ldots, \text{Cov}(\epsilon_i, \epsilon_{i-k}) = \text{Cov}(\epsilon_i, \epsilon_{i+k}) = \sigma^2\phi^k$, for lags $k = 1, 2, \ldots$. In this case, the $n \times n$ covariance matrix of the errors is given by

$$V(\epsilon) = \sigma^2 \begin{bmatrix} 1 & \phi & \phi^2 & \ldots & \phi^{n-1} \\ \phi & 1 & \phi & \ldots & \phi^{n-2} \\ \phi^2 & \phi & 1 & \ldots & \phi^{n-3} \\ \phi^3 & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \phi^{n-1} & \phi^{n-2} & \ldots & \ldots & 1 \end{bmatrix}$$

The model that corresponds to this covariance matrix is known as the **first-order autoregressive model**. It is a particularly simple and useful parameterization because it requires only one additional parameter, the autoregressive parameter ϕ. However, many other models are available for representing autocorrelations among observations. In Chapter 10 on regression time series models, we discuss these models in detail.

A third example in which independence of the errors is violated arises when spatial observations are involved. Consider measurements on a certain groundwater pollutant that are taken at the same time but at different locations. In this situation, it is likely that errors for measurements taken in close spatial proximity are correlated. Many different models have been developed to characterize **spatial correlation**, and most express the spatial correlation as a function of the (Euclidean) distance between the measurement locations. A common assumption is that the correlation decreases with distance among measurement sites.

Expand on this example, and consider the situation when spatial observations are involved but when observations are also taken at several time periods. Here, one faces the situation in which observations (or errors) exhibit a spatial as well as a temporal correlation structure. A common approach is to model the covariance matrix of the errors with several (hopefully few) additional parameters that characterize the spatial and temporal correlations and then estimate the parameters in the regression model $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ under this more general error model.

4.6.2 GENERALIZED LEAST SQUARES ESTIMATION

Assume that the vector of errors $\boldsymbol{\epsilon}$ in the regression model $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ has zero mean and general covariance matrix $V(\boldsymbol{\epsilon}) = E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') = \sigma^2 V$. Now V is no longer the identity matrix. The proportionality coefficient, σ^2, is unknown, but we assume—at least initially—that all elements in the matrix V are known.

We will try to find a linear transformation, $L\boldsymbol{\epsilon}$ of $\boldsymbol{\epsilon}$, which satisfies the assumptions of the standard model. The matrix V is symmetric and positive definite, and we can apply our results in Chapter 3 on the spectral decomposition of a symmetric matrix. We can write the matrix as $V = P\Lambda P' = P\Lambda^{1/2}\Lambda^{1/2}P'$, where the matrix Λ is diagonal. Its elements $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m > 0$ are the eigenvalues of the positive definite matrix V; the column vectors of the matrix P are the corresponding normalized eigenvectors. Note that $V^{-1} = (P\Lambda^{1/2}\Lambda^{1/2}P')^{-1} = P\Lambda^{-1/2}\Lambda^{-1/2}P' = L'L$.

Premultiplying the regression model by the matrix $L = \Lambda^{-1/2}P'$ results in the model

$$L\boldsymbol{y} = LX\boldsymbol{\beta} + L\boldsymbol{\epsilon} = LX\boldsymbol{\beta} + \tilde{\boldsymbol{\epsilon}}$$

where the vector $L\boldsymbol{y}$ represents the transformed response, and the columns in the matrix LX represent the transformed regressor variables. Then $E(\tilde{\boldsymbol{\epsilon}}) = E(L\boldsymbol{\epsilon}) = LE(\boldsymbol{\epsilon}) = \mathbf{0}$ and

$$V(\tilde{\boldsymbol{\epsilon}}) = LV(\boldsymbol{\epsilon})L' = LVL'\sigma^2 = \Lambda^{-1/2}P'P\Lambda^{1/2}\Lambda^{1/2}P'P\Lambda^{-1/2}\sigma^2 = I\sigma^2$$

The new disturbance vector $\tilde{\boldsymbol{\epsilon}}$ satisfies the standard regression assumptions. According to the Gauss–Markov theorem, least squares—applied to the transformed variables—will yield the best linear unbiased estimator of $\boldsymbol{\beta}$. Replacing $\boldsymbol{y}$ and X in the standard least squares estimator in Eq. (4.11) by $L\boldsymbol{y}$ and LX, respectively, leads to the **generalized least squares (GLS) estimator**

$$\hat{\boldsymbol{\beta}}^{\text{GLS}} = (X'L'LX)^{-1}X'L'L\boldsymbol{y} = (X'V^{-1}X)^{-1}X'V^{-1}\boldsymbol{y} \tag{4.58}$$

and its covariance matrix

$$\begin{aligned} V(\hat{\boldsymbol{\beta}}^{\text{GLS}}) &= (X'L'LX)^{-1}X'L'V(L\boldsymbol{y})LX(X'L'LX)^{-1} \\ &= \sigma^2(X'L'LX)^{-1}X'L'LX(X'L'LX)^{-1} \\ &= \sigma^2(X'L'LX)^{-1} = \sigma^2(X'V^{-1}X)^{-1} \end{aligned} \tag{4.59}$$

The GLS estimator can be used to compute the error sum of squares in the transformed model,

$$\begin{aligned} S(\hat{\boldsymbol{\beta}}^{\text{GLS}}) &= (L\boldsymbol{y} - LX\hat{\boldsymbol{\beta}}^{\text{GLS}})'(L\boldsymbol{y} - LX\hat{\boldsymbol{\beta}}^{\text{GLS}}) \\ &= (\boldsymbol{y} - X\hat{\boldsymbol{\beta}}^{\text{GLS}})'L'L(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}^{\text{GLS}}) \\ &= (\boldsymbol{y} - X\hat{\boldsymbol{\beta}}^{\text{GLS}})'V^{-1}(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}^{\text{GLS}}) \end{aligned} \tag{4.60}$$

The model in the transformed variables satisfies the standard regression assumptions. Hence, $S(\hat{\boldsymbol{\beta}}^{\text{GLS}})/\sigma^2$ follows a chi-square distribution with $n-(p+1)$ degrees of freedom, where n represents the number of cases and $p+1$ the number of regression coefficients. Hence,

$$s^2_{\text{GLS}} = S(\hat{\boldsymbol{\beta}}^{\text{GLS}})/(n-p-1) \tag{4.61}$$

is an unbiased estimator of σ^2. This can be used in Eq. (4.59) to obtain an estimate of $V(\hat{\boldsymbol{\beta}}^{\text{GLS}})$.

What are the properties of the standard least squares estimator $\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\boldsymbol{y}$ that has been derived under the wrong assumption of independent and equally precise errors? It also is unbiased, but it is no longer “best” among all linear unbiased estimators. The Gauss–Markov result has already shown us that it is the GLS estimator $\hat{\boldsymbol{\beta}}^{\text{GLS}}$ that has the smallest covariance matrix. The covariance matrix of the standard least squares estimator

$$\begin{aligned} V(\hat{\boldsymbol{\beta}}) &= V[(X'X)^{-1}X'\boldsymbol{y}] = (X'X)^{-1}X'V(\boldsymbol{y})X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}X'VX(X'X)^{-1} \end{aligned}$$

exceeds the covariance matrix in Eq. (4.59) by a positive semidefinite matrix.

Remark

So far, our analysis has assumed that all elements in the matrix V are specified. For this reason, we call the estimator in Eq. (4.58) the **feasible** generalized least squares estimator. In the first example of our introduction, the precision $V(y_i) = V(\epsilon_i) = \sigma^2/n_i$ depends on the **known** number of measurements that go into the observation y_i. Here, V is specified, and the generalized least squares estimator can be calculated. In the second illustration, the matrix V contains the autoregressive parameter ϕ. In practice, this parameter is unknown, and one must estimate the regression coefficients $\boldsymbol{\beta}$ and ϕ jointly. This issue will be addressed in Chapter 10, when we discuss regression models with time series errors.

4.6.3 WEIGHTED LEAST SQUARES

Weighted least squares is a special case of generalized least squares. The weighted least squares estimator minimizes the weighted error sum of squares

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} w_i(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2$$

where $w_i > 0$ are known specified weights. This criterion is equivalent to the one for generalized least squares, with V^{-1} a diagonal matrix having diagonal elements w_i.

Equations (4.58) and (4.59) imply that the weighted least squares (WLS) estimator is given by

$$\hat{\boldsymbol{\beta}}^{\text{WLS}} = \left[\sum_{i=1}^{n} w_i \boldsymbol{x}_i \boldsymbol{x}_i'\right]^{-1} \left[\sum_{i=1}^{n} w_i \boldsymbol{x}_i y_i\right] \tag{4.62}$$

with variance

$$V(\hat{\boldsymbol{\beta}}^{\text{WLS}}) = \sigma^2 \left[\sum_{i=1}^{n} w_i \boldsymbol{x}_i \boldsymbol{x}_i'\right]^{-1} \tag{4.63}$$

APPENDIX: PROOFS OF RESULTS

1. MINIMIZATION OF $S(\beta)$ IN EQ. (4.9)

We wish to minimize

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \mu_i)^2 = \sum_{i=1}^{n} \epsilon_i^2$$

subject to the restriction that $\boldsymbol{\mu} = X\boldsymbol{\beta}$. The elements of the vector $\boldsymbol{\mu}$ are given by

$$\mu_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} = \beta_0 + \sum_{j=1}^{p} x_{ij} \beta_j$$

for $i = 1, 2, \ldots, n$. The partial derivatives of $S(\boldsymbol{\beta})$ with respect to the parameters $\beta_0, \beta_1, \ldots, \beta_p$ are

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \beta_0} = 2 \sum_{i=1}^{n} \epsilon_i \frac{\partial \epsilon_i}{\partial \beta_0} = -2 \sum_{i=1}^{n} \epsilon_i$$

and

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \beta_j} = 2 \sum_{i=1}^{n} \epsilon_i \frac{\partial \epsilon_i}{\partial \beta_j} = -2 \sum_{i=1}^{n} x_{ij} \epsilon_i, \quad j = 1, 2, \ldots, p$$

At the minimum of $S(\boldsymbol{\beta})$ these derivatives are zero. Hence,

$$\sum_{i=1}^{n} \epsilon_i = \sum (y_i - \mu_i) = 0$$

$$\sum_{i=1}^{n} x_{ij} (y_i - \mu_i) = 0, \quad j = 1, 2, \ldots, p$$

In vector form,

$$\mathbf{1}'(\boldsymbol{y} - \boldsymbol{\mu}) = 0$$
$$\boldsymbol{x}_j'(\boldsymbol{y} - \boldsymbol{\mu}) = 0, \quad j = 1, 2, \ldots, p$$

where the $n \times 1$ vector $\mathbf{1}$, a vector of ones, and $\boldsymbol{x_j}$, the vector with elements $x_{1j}, x_{2j}, \dots, x_{nj}$, are columns of the matrix $X = [\mathbf{1}, \boldsymbol{x_1}, \dots, \boldsymbol{x_p}]$. Combining these $p+1$ equations, we obtain

$$X'(\boldsymbol{y} - X\boldsymbol{\beta}) = \mathbf{0}$$

or

$$X'X\boldsymbol{\beta} = X'\boldsymbol{y}$$

Let $\hat{\boldsymbol{\beta}}$ denote a solution of this equation. Solving the normal equations $(X'X)\hat{\boldsymbol{\beta}} = X'\boldsymbol{y}$ leads to $\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\boldsymbol{y}$; the inverse $(X'X)^{-1}$ exists since we assume that X has full column rank. In order to prove that $\hat{\boldsymbol{\beta}}$ actually minimizes $S(\boldsymbol{\beta})$ we show that any other estimate will lead to a larger value of $S(\boldsymbol{\beta})$:

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

since the normal equations imply that the cross-product term $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'X'(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}) = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(X'\boldsymbol{y} - X'X\hat{\boldsymbol{\beta}}) = 0$. Thus,

$$S(\boldsymbol{\beta}) = S(\hat{\boldsymbol{\beta}}) + \boldsymbol{c}'\boldsymbol{c}$$

where $\boldsymbol{c} = X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$. Since $\boldsymbol{c}'\boldsymbol{c} = \sum_{i=1}^{n} c_i^2 \geq 0$, $S(\boldsymbol{\beta}) \geq S(\hat{\boldsymbol{\beta}})$; the equality is true if and only if $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$.

2. ANOTHER PROOF OF THE UNBIASEDNESS OF s^2 AS AN ESTIMATE OF σ^2: $E\left(\sum_{i=1}^{n} e_i^2\right) = (n - p - 1)\sigma^2$

Consider

$$\begin{aligned} E\left(\sum_{i=1}^{n} e_i^2\right) &= E(\boldsymbol{e}'\boldsymbol{e}) = E[\boldsymbol{y}'(I - H)(I - H)\boldsymbol{y}] && \text{since } \boldsymbol{e} = (I - H)\boldsymbol{y} \\ &= E[\boldsymbol{y}'(I - H)\boldsymbol{y}] && \text{since } (I - H) \text{ is idempotent} \\ &= E[\operatorname{tr}(\boldsymbol{y}'(I - H)\boldsymbol{y})] && \text{since } \boldsymbol{y}'(I - H)\boldsymbol{y} \text{ is a scalar} \\ &= E[\operatorname{tr}(I - H)\boldsymbol{y}\boldsymbol{y}'] && \text{since } \operatorname{tr} AB = \operatorname{tr} BA \\ &= \operatorname{tr}[(I - H)E(\boldsymbol{y}\boldsymbol{y}')] \end{aligned}$$

Now

$$\begin{aligned} E(\boldsymbol{y}\boldsymbol{y}') &= E[(X\boldsymbol{\beta} + \boldsymbol{\epsilon})(X\boldsymbol{\beta} + \boldsymbol{\epsilon})'] \\ &= X\boldsymbol{\beta}\boldsymbol{\beta}'X' + E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') = X\boldsymbol{\beta}\boldsymbol{\beta}'X' + \sigma^2 I \end{aligned}$$

Here we have used the fact that $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $V(\boldsymbol{\epsilon}) = E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') = \sigma^2 I$. Hence,

$$\begin{aligned} E\left(\sum_{i=1}^{n} e_i^2\right) &= \operatorname{tr}[(I - H)(\sigma^2 I + X\boldsymbol{\beta}\boldsymbol{\beta}'X')] \\ &= \operatorname{tr}[I - H]\sigma^2, \text{ since } \operatorname{tr}(A + B) = \operatorname{tr}(A) + \operatorname{tr}(B), \text{ and } (I - H)X = O \\ &= \sigma^2[n - \operatorname{tr} X(X'X)^{-1}X'] \\ &= \sigma^2[n - (p + 1)] = (n - p - 1)\sigma^2 \end{aligned}$$

since $\mathrm{tr}[X(X'X)^{-1}X'] = \mathrm{tr}[(X'X)^{-1}X'X] = \mathrm{tr}(I_{p+1})$, where I_{p+1} is the $(p+1) \times (p+1)$ identity matrix.

3. DIRECT PROOF THAT $\hat{\beta}$ AND $S(\hat{\beta})$ ARE STATISTICALLY INDEPENDENT AND THAT $S(\hat{\beta})/\sigma^2$ FOLLOWS A χ^2_{n-p-1} DISTRIBUTION

The set of all linear functions of $L(X) = L(\mathbf{1}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$ forms a $(p+1)$-dimensional subspace of R^n. We can always find $p+1$ orthonormal vectors $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_{p+1}$ (i.e., $\boldsymbol{c}_i'\boldsymbol{c}_i = 1, \boldsymbol{c}_i'\boldsymbol{c}_j = 0, i \neq j$) that form a basis of $L(X)$. These orthonormal vectors are linearly related to the regressor columns. The Gram–Schmidt orthogonalization procedure (see Chapter 3) shows us how to obtain these vectors.

$L(X)$ is a subset of R^n. Hence, we need to add $(n-p-1)$ additional orthonormal vectors $\boldsymbol{c}_{p+2}, \ldots, \boldsymbol{c}_n$ such that $(\boldsymbol{c}_1, \ldots, \boldsymbol{c}_n)$ forms an orthonormal basis of the larger space R^n. You can visualize the construction as follows:

$$\underbrace{\underbrace{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_{p+1}}_{L(X)}, \boldsymbol{c}_{p+2}, \ldots, \boldsymbol{c}_n}_{R^n}$$

The vectors in the matrix

$$P = (\boldsymbol{c}_1, \ldots, \boldsymbol{c}_n) = (P_1, P_2)$$

where $P_1 = (\boldsymbol{c}_1, \ldots, \boldsymbol{c}_{p+1})$ and $P_2 = (\boldsymbol{c}_{p+2}, \ldots, \boldsymbol{c}_n)$ are $n \times (p+1)$ and $n \times (n-p-1)$ matrices, provide an orthonormal basis. By construction, P is an orthogonal matrix. That is, $P'P = PP' = I$.

Our model specifies $\boldsymbol{y} \sim N(\boldsymbol{\mu}, \sigma^2 I)$, where $\boldsymbol{\mu} = X\boldsymbol{\beta}$ is in $L(X)$. Consider the orthogonal transformation

$$\boldsymbol{z} = P'\boldsymbol{y} = \begin{pmatrix} P_1' \\ P_2' \end{pmatrix} \boldsymbol{y}$$

Then $\boldsymbol{z} \sim N(P'\boldsymbol{\mu}, \sigma^2 I)$ since $P'P = I$. This says that the z_i's are independent and have the same variance σ^2. Furthermore,

$$P'\boldsymbol{\mu} = \begin{pmatrix} P_1'\boldsymbol{\mu} \\ P_2'\boldsymbol{\mu} \end{pmatrix} = \begin{pmatrix} P_1'\boldsymbol{\mu} \\ \mathbf{0} \end{pmatrix}$$

since $P_2'\boldsymbol{\mu} = \mathbf{0}$. This is because $\boldsymbol{\mu}$ is in $L(X)$ and the columns of P_2 are perpendicular to $L(X)$.

Turning the transformation around results in

$$\begin{aligned} \boldsymbol{y} = (P')^{-1}\boldsymbol{z} = P\boldsymbol{z} &= \sum_{i=1}^{n} \boldsymbol{c}_i z_i \\ &= \sum_{i=1}^{p+1} \boldsymbol{c}_i z_i + \sum_{i=p+2}^{n} \boldsymbol{c}_i z_i \\ &= \hat{\boldsymbol{\mu}} + (\boldsymbol{y} - \hat{\boldsymbol{\mu}}) \end{aligned}$$

Here we have used the fact that P is orthogonal, and $P^{-1} = P'$. Now,

$$\begin{aligned} S(\hat{\boldsymbol{\beta}}) = (\boldsymbol{y} - X\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}) = \|\boldsymbol{y} - \hat{\boldsymbol{\mu}}\|^2 &= \left(\sum_{i=p+2}^{n} \boldsymbol{c}_i z_i\right)' \left(\sum_{i=p+2}^{n} \boldsymbol{c}_i z_i\right) \\ &= \sum_{i=p+2}^{n} \sum_{j=p+2}^{n} z_i z_j \boldsymbol{c}_i' \boldsymbol{c}_j \\ &= \sum_{i=p+2}^{n} z_i^2 \text{ since } \boldsymbol{c}_i'\boldsymbol{c}_i = 1 \text{ and } \boldsymbol{c}_i'\boldsymbol{c}_j = 0, i \neq j \end{aligned}$$

Since $z_{p+2}, \ldots, z_n$ are i.i.d. $N(0, \sigma^2)$, it follows that

$$\frac{S(\hat{\boldsymbol{\beta}})}{\sigma^2} = \sum_{i=p+2}^{n} z_i^2/\sigma^2$$

is the sum of $(n - p - 1)$ independent χ_1^2 random variables. It has a χ_{n-p-1}^2 distribution. Furthermore, this is independent of $z_1, \ldots, z_{p+1}$. Now

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (X'X)^{-1}X'\boldsymbol{y} = (X'X)^{-1}X'P\boldsymbol{z} \\ &= (X'X)^{-1}X'(P_1, P_2)\boldsymbol{z} \end{aligned}$$

However, $X'(P_1, P_2) = (X'P_1, O)$ since the columns of P_2 are perpendicular to $L(X)$ while rows of X' are in $L(X)$. Hence, $\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'P_1\boldsymbol{z}_{(1)}$, where $\boldsymbol{z}_{(1)} = (z_1, \ldots, z_{p+1})'$. The least squares estimator $\hat{\boldsymbol{\beta}}$ depends on $z_1, \ldots, z_{p+1}$, whereas $S(\hat{\boldsymbol{\beta}})$ depends on $z_{p+2}, \ldots, z_n$. Thus, $\hat{\boldsymbol{\beta}}$ is independent of $S(\hat{\boldsymbol{\beta}})$.

4. PROOF OF THEOREM

Proof: $L_A(X)$ is a subset of $L(X)$, and $L(X)$ is a subset of R^n. $L(X)$ is of dimension $p + 1$. $L_A(X)$ imposes l independent restrictions on the subset $L(X)$. Hence, the dimension of $L_A(X)$ is $p + 1 - l$. Choose an orthonormal basis $(\boldsymbol{c}_1, \ldots, \boldsymbol{c}_{p+1-l})$ for $L_A(X)$ and extend it successively to form orthonormal bases for $L(X)$ and R^n. Visualize the process as follows:

$$\underbrace{\underbrace{\underbrace{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_{p+1-l}}_{L_A(X)}, \boldsymbol{c}_{p+2-l}, \ldots, \boldsymbol{c}_{p+1}}_{L(X)}, \boldsymbol{c}_{p+2}, \ldots, \boldsymbol{c}_n}_{R^n}$$

The vectors are collected in the $n \times n$ matrix P,

$$P = (\boldsymbol{c}_1, \ldots, \boldsymbol{c}_n) = (P_1, P_2, P_3)$$

where $P_1 = (\boldsymbol{c}_1, \ldots, \boldsymbol{c}_{p+1-\ell})$, $P_2 = (c_{p+2-\ell}, \ldots, \boldsymbol{c}_{p+1})$, and $P_3 = (\boldsymbol{c}_{p+2}, \ldots, \boldsymbol{c}_n)$ are $n \times (p + 1 - l)$, $n \times l$, and $n \times (n - p - 1)$ matrices. The matrix P is orthogonal: $PP' = P'P = I$.

Consider the orthogonal transformation $\boldsymbol{z} = P'\boldsymbol{y}$ and its inverse

$$\begin{aligned}\boldsymbol{y} = P\boldsymbol{z} &= \sum_{i=1}^{n} \boldsymbol{c}_i z_i \\ &= \sum_{i=1}^{p+1-l} \boldsymbol{c}_i z_i + \sum_{i=p+2-l}^{p+1} \boldsymbol{c}_i z_i + \sum_{i=p+2}^{n} \boldsymbol{c}_i z_i \\ &= \hat{\boldsymbol{\mu}}_A + (\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_A) + (\boldsymbol{y} - \hat{\boldsymbol{\mu}})\end{aligned}$$

where $\hat{\boldsymbol{\mu}}_A$ is the projection of $\boldsymbol{y}$ on $L_A(X)$, and $\hat{\boldsymbol{\mu}}$ is the projection of $\boldsymbol{y}$ on $L(X)$; $\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_A$ is in $L(X)$ and perpendicular to $L_A(X)$; $\|\boldsymbol{y} - \hat{\boldsymbol{\mu}}\|^2 = \sum_{i=p+2}^{n} z_i^2 = S(\hat{\boldsymbol{\beta}})$ and $\|\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_A\|^2 = \sum_{i=p+2-l}^{p+1} z_i^2$.

Since $\boldsymbol{y} \sim N(\boldsymbol{\mu}, \sigma^2 I)$, it follows that $\boldsymbol{z} = P'\boldsymbol{y} \sim N(P'\boldsymbol{\mu}, \sigma^2 I)$. Under the null hypothesis $A\boldsymbol{\beta} = \boldsymbol{0}$, the mean vector

$$P'\boldsymbol{\mu} = \begin{bmatrix} P_1'\boldsymbol{\mu} \\ P_2'\boldsymbol{\mu} \\ P_3'\boldsymbol{\mu} \end{bmatrix} = \begin{bmatrix} P_1'\boldsymbol{\mu} \\ \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}$$

This is because under the null hypothesis $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}_A$ is in $L_A(X)$ and the columns of P_2 are perpendicular to $L_A(X)$. In addition, $P_3'\boldsymbol{\mu} = \boldsymbol{0}$ since the columns of P_3 are perpendicular to $L(X)$. Hence,

i. $z_1, z_2, \ldots, z_n$ are independent normal random variables with variance σ^2.
ii. $z_{p+2-l}, \ldots, z_{p+1}$ have zero means under the null hypothesis $A\boldsymbol{\beta} = \boldsymbol{0}$.
iii. $z_{p+2}, \ldots, z_n$ have zero means under the original model, even if the null hypothesis is false.

Thus,

i. $\|\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_A\|^2 / \sigma^2 = \sum_{p+2-l}^{p+1} z_i^2$ is the sum of l independent χ_1^2 random variables. It has a χ_l^2 distribution.
ii. $S(\hat{\boldsymbol{\beta}})$ is a function of $z_{p+2}, \ldots, z_n$, whereas $\|\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_A\|^2$ is a function of $z_{p+2-l}, \ldots, z_{p+1}$. Furthermore, $z_1, z_2, \ldots, z_n$ are independent. This shows that $S(\hat{\boldsymbol{\beta}})$ and $\|\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_A\|^2$ are independent.

■

EXERCISES

4.1. Consider the regression on time, $y_t = \beta_0 + \beta_1 t + \epsilon_t$, with $t = 1, 2, \ldots, n$. Here, the regressor vector is $\boldsymbol{x}' = (1, 2, \ldots, n)$. Take $n = 10$. Write down the matrices $X'X$, $(X'X)^{-1}$, $V(\hat{\boldsymbol{\beta}})$, and the variances of $\hat{\beta}_0$ and $\hat{\beta}_1$.

4.2. For the regression model $y_t = \beta_0 + \epsilon_t$ with $n = 2$ and $\boldsymbol{y}' = (2, 4)$, draw the data in two-dimensional space. Identify the orthogonal projection of $\boldsymbol{y}$ onto $L(X) = L(\boldsymbol{1})$. Explain geometrically $\hat{\beta}_0$, $\hat{\boldsymbol{\mu}}$, and $\boldsymbol{e}$.

4.3. Consider the regression model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, $i = 1, 2, 3$. With

$$\boldsymbol{x} = \begin{bmatrix} 1 \\ 3 \\ 2 \end{bmatrix} \quad \boldsymbol{y} = \begin{bmatrix} 2.2 \\ 3.9 \\ 3.1 \end{bmatrix}$$

draw the data in three-dimensional space and identify the orthogonal projection of $\boldsymbol{y}$ onto $L(X) = L(\mathbf{1}, \boldsymbol{x})$. Explain geometrically $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\mu}}$, and $\boldsymbol{e}$.

4.4. Consider the regression model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, $i = 1, 2, 3$. With

$$\boldsymbol{x} = \begin{bmatrix} 1 \\ 3 \\ 2 \end{bmatrix} \quad \boldsymbol{y} = \begin{bmatrix} 2 \\ 4 \\ 6 \end{bmatrix}$$

draw the data in three-dimensional space and identify the orthogonal projection of $\boldsymbol{y}$ onto $L(X) = L(\mathbf{1}, \boldsymbol{x})$. Explain geometrically $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\mu}}$, and $\boldsymbol{e}$.

4.5. After fitting the regression model,

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

on 15 cases, it is found that the mean square error $s^2 = 3$ and

$$(X'X)^{-1} = \begin{bmatrix} 0.5 & 0.3 & 0.2 & 0.6 \\ 0.3 & 6.0 & 0.5 & 0.4 \\ 0.2 & 0.5 & 0.2 & 0.7 \\ 0.6 & 0.4 & 0.7 & 3.0 \end{bmatrix}$$

Find

a. The estimate of $V(\hat{\beta}_1)$.

b. The estimate of $\mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_3)$.

c. The estimate of $\mathrm{Corr}(\hat{\beta}_1, \hat{\beta}_3)$.

d. The estimate of $V(\hat{\beta}_1 - \hat{\beta}_3)$.

4.6. When fitting the model

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

to a set of $n = 15$ cases, we obtained the least squares estimates $\hat{\beta}_0 = 10$, $\hat{\beta}_1 = 12$, $\hat{\beta}_2 = 15$, and $s^2 = 2$. It is also known that

$$(X'X)^{-1} = \begin{bmatrix} 1 & 0.25 & 0.25 \\ 0.25 & 0.5 & -0.25 \\ 0.25 & -0.25 & 2 \end{bmatrix}$$

a. Estimate $V(\hat{\beta}_2)$.

b. Test the hypothesis that $\beta_2 = 0$.

c. Estimate the covariance between $\hat{\beta}_1$ and $\hat{\beta}_2$.

d. Test the hypothesis that $\beta_1 = \beta_2$, using both the t ratio and the 95% confidence interval.

e. The corrected total sum of squares, SST = 120. Construct the ANOVA table and test the hypothesis that $\beta_1 = \beta_2 = 0$. Obtain the percentage of variation in y that is explained by the model.

4.7. Consider a multiple regression model of the price of houses (y) on three explanatory variables: taxes paid (x_1), number of bathrooms (x_2), and square feet (x_3). The incomplete (Minitab) output from a regression on $n = 28$ houses is given as follows:

The regression equation is price $= -10.7 +$ 0.190 taxes + 81.9 baths + 0.101 sqft

Predictor	Coef	SE Coef	t	p
Constant	−10.65	24.02		
taxes	0.18966	0.05623		
baths	81.87	47.82		
sqft	0.10063	0.03125		

Analysis of variance

Source	DF	SS	MS	F	p
Regression	3	504541			
Residual Error					
Total	27	541119			

a. Calculate the coefficient of determination R^2.

b. Test the null hypothesis that all three regression coefficients are zero (H_0: $\beta_1 = \beta_2 = \beta_3 = 0$). Use significance level 0.05.

c. Obtain a 95% confidence interval of the regression coefficient for "taxes." Can you simplify the model by dropping "taxes"? Obtain a 95% confidence interval of the regression coefficient for "baths." Can you simplify the model by dropping "baths"?

4.8. Continuation of Exercise 4.7. The incomplete (Minitab) output from a multiple regression of the price of houses on the two explanatory variables, taxes paid and square feet, is given as follows:

The regression equation is price = 4.9 + 0.242 taxes + 0.134 sqft

Predictor	Coef	SE Coef	t	p
Constant	4.89	23.08		
taxes	0.24237	0.04884		
sqft	0.13397	0.02537		

Analysis of variance

Source	DF	SS	MS	F	p
Regression	2	500074	250037		
Residual Error					
Total		541119			

a. Calculate the coefficient of determination R^2.

b. Test the null hypothesis that both regression coefficients are zero (H_0: $\beta_1 = \beta_2 = 0$). Use significance level 0.05.

c. Test whether you can omit the variable “taxes” from the regression model. Use significance level 0.05.

d. Comment on the fact that the regression coefficients for taxes and square feet are different than those shown in Exercise 4.7.

4.9. Fitting the regression $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$ on $n = 30$ cases leads to the following results:

$$X'X = \begin{bmatrix} 30 & 2{,}108 & 5{,}414 \\ 2{,}108 & 152{,}422 & 376{,}562 \\ 5{,}414 & 376{,}562 & 1{,}015{,}780 \end{bmatrix}$$

$$X'\mathbf{y} = \begin{bmatrix} 5{,}263 \\ 346{,}867 \\ 921{,}939 \end{bmatrix} \quad \text{and} \quad \mathbf{y}'\mathbf{y} = 1{,}148{,}317$$

a. Use computer software to find $(X'X)^{-1}$. Obtain the least squares estimates and their standard errors.

b. Compute the t statistics to test the simple hypotheses that each regression coefficient is zero.

c. Determine the coefficient of variation R^2. (The complete data are given in the file **abrasion**.)

4.10. The following matrices were computed for a certain regression problem:

$$X'X = \begin{bmatrix} 15 & 3{,}626 & 44{,}428 \\ 3{,}626 & 1{,}067{,}614 & 11{,}419{,}181 \\ 44{,}428 & 11{,}419{,}181 & 139{,}063{,}428 \end{bmatrix},$$

$$X'\mathbf{y} = \begin{bmatrix} 2{,}259 \\ 647{,}107 \\ 7{,}096{,}619 \end{bmatrix}$$

$$(X'X)^{-1} = \begin{bmatrix} 1.2463484 & 2.1296642 \times 10^{-4} & -4.1567125 \times 10^{-4} \\ & 7.7329030 \times 10^{-6} & -7.0302518 \times 10^{-7} \\ & & 1.9771851 \times 10^{-7} \end{bmatrix},$$

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 3.452613 \\ 0.496005 \\ 0.009191 \end{bmatrix}$$

$$\mathbf{y}'\mathbf{y} = 394{,}107$$

a. Write down the estimated regression equation. Obtain the standard errors of the regression coefficients.

b. Compute the t statistics to test the simple hypotheses that each regression coefficient is equal to zero. Carry out these tests. State your conclusions.

4.11. A study was conducted to investigate the determinants of survival size of nonprofit U.S. hospitals. Survival size, y, was defined to be the largest U.S. hospital (in terms of the number of beds) exhibiting growth in market share. For the investigation, 10 states were selected at random, and the survival size for nonprofit hospitals in each of the selected states was determined for two time periods t: 1981–1982 and 1984–1985.

Furthermore, the following characteristics were collected on each selected state for each of the two time periods:

x_1 = Percentage of beds that are in for-profit hospitals.

x_2 = Number of people enrolled in health maintenance organizations as a fraction of the number of people covered by hospital insurance.

$x_3 =$ State population in thousands.

$x_4 =$ Percentage of state that is urban.

The data are given in the file **hospital.**

a. Fit the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$$

b. The influence of the percentage of beds in for-profit hospitals was of particular interest to the investigators. What does the analysis tell us?

c. What further investigation might you do with this data set. Give reasons?

d. Rather than selecting 10 states at random, how else might you collect the data on survival size? Would your approach be an improvement over the random selection?

4.12. The amount of water used by the production facilities of a plant varies. Observations on water usage and other, possibily related, variables were collected for 17 months. The data are given in the file **water**. The explanatory variables are

TEMP = average monthly temperature(°F)

PROD = amount of production

DAYS = number of operating days in the month

PAYR = number of people on the monthly plant payroll

HOUR = number of hours shut down for maintenance

The response variable is USAGE = monthly water usage (gallons/100).

a. Fit the model containing all five independent variables,

$$y = \beta_0 + \beta_1 \text{TEMP} + \beta_2 \text{PROD} + \beta_3 \text{DAYS} + \beta_4 \text{PAYR} + \beta_5 \text{HOUR} + \epsilon$$

Plot residuals against fitted values and residuals against the case index, and comment about model adequacy.

b. Test the hypothesis that $\beta_1 = \beta_3 = \beta_5 = 0$.

c. Which model or set of models would you suggest for predictive purposes? Briefly justify.

d. Which independent variable seems to be the most important one in determining the amount of water used?

e. Write a **nontechnical** paragraph that summarizes your conclusions about plant water usage that is supported by the data.

4.13. Data on last year's sales (y, in 100,000s of dollars) in 15 sales districts are given in the file **sales**. This file also contains promotional expenditures (x_1, in thousands of dollars), the number of active accounts (x_2), the number of competing brands (x_3), and the district potential (x_4, coded) for each of the districts.

a. A model with all four regressors is proposed:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

Interpret the parameters β_0, β_1, and β_4.

b. Fit the proposed model in (a) and calculate estimates of β_i, $i = 0, 1, \ldots, 4$, and σ^2.

c. Test the following hypotheses:

(i) $\beta_4 = 0$; (ii) $\beta_3 = \beta_4 = 0$;
(iii) $\beta_2 = \beta_3$; (iv) $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

d. Consider the reduced (restricted) model with $\beta_4 = 0$. Estimate its coefficients and give an expression for the expected sales.

e. Using the model in (d), obtain a prediction for the sales in a district where $x_1 = 3.0$, $x_2 = 45$, and $x_3 = 10$. Obtain the corresponding 95% prediction interval.

4.14. The survival rate (in percentage) of bull semen after storage is measured at various combinations of concentrations of three materials (additives) that are thought to increase the chance of survival. The data listed below are given in the file **bsemen**.

% Survival (y)	% Weight 1 (x_1)	% Weight 2 (x_2)	% Weight 3 (x_3)
25.5	1.74	5.30	10.80
31.2	6.32	5.42	9.40
25.9	6.22	8.41	7.20
38.4	10.52	4.63	8.50
18.4	1.19	11.60	9.40
26.7	1.22	5.85	9.90

% Survival (y)	% Weight 1 (x_1)	% Weight 2 (x_2)	% Weight 3 (x_3)
26.4	4.10	6.62	8.00
25.9	6.32	8.72	9.10
32.0	4.08	4.42	8.70
25.2	4.15	7.60	9.20
39.7	10.15	4.83	9.40
35.9	1.72	3.12	7.60
26.5	1.70	5.30	8.20

Assume the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$.

a. Compute $X'X$, $(X'X)^{-1}$, and $X'\boldsymbol{y}$.

b. Plot the response y versus each predictor variable. Comment on these plots.

c. Obtain the least squares estimates of $\boldsymbol{\beta}$ and give the fitted equation.

d. Construct a 90% confidence interval for

  i. the predicted mean value of y when $x_1 = 3$, $x_2 = 8$, and $x_3 = 9$;

  ii. the predicted individual value of y when $x_1 = 3$, $x_2 = 8$, and $x_3 = 9$.

e. Construct the ANOVA table and test for a significant linear relationship between y and the three predictor variables.

4.15. An experiment was conducted to study the toxic action of a certain chemical on silkworm larvae. The relationship of $\log_{10}$(survival time) to $\log_{10}$(dose) and $\log_{10}$(larvae weight) was investigated. The data, obtained by feeding each larvae a precisely measured dose of the chemical in an aqueous solution and recording the survival time until death, are given in the following table. The data are stored in the file **silkw**.

$\log_{10}$ Survival Time (y)	$\log_{10}$ Dose (x_1)	$\log_{10}$ Weight (x_2)
2.836	0.150	0.425
2.966	0.214	0.439
2.687	0.487	0.301
2.679	0.509	0.325
2.827	0.570	0.371
2.442	0.590	0.093
2.421	0.640	0.140
2.602	0.781	0.406
2.556	0.739	0.364
2.441	0.832	0.156
2.420	0.865	0.247
2.439	0.904	0.278
2.385	0.942	0.141
2.452	1.090	0.289
2.351	1.194	0.193

Assume the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$.

a. Plot the response y versus each predictor variable. Comment on these plots.

b. Obtain the least squares estimates for $\boldsymbol{\beta}$ and give the fitted equation.

c. Construct the ANOVA table and test for a significant linear relationship between y and the two predictor variables.

d. Which independent variable do you consider to be the better predictor of log(survival time)? What are your reasons?

e. Of the models involving one or both of the independent variables, which do you prefer, and why?

4.16. You are given the following matrices computed for a regression analysis:

$$X'X = \begin{bmatrix} 9 & 136 & 269 & 260 \\ 136 & 2{,}114 & 4{,}176 & 3{,}583 \\ 269 & 4{,}176 & 8{,}257 & 7{,}104 \\ 260 & 3{,}583 & 7{,}104 & 12{,}276 \end{bmatrix}$$

$$X'\boldsymbol{y} = \begin{bmatrix} 45 \\ 648 \\ 1{,}283 \\ 1{,}821 \end{bmatrix}$$

$$(X'X)^{-1} = \begin{bmatrix} 9.610 & 0.008 & -0.279 & -0.044 \\ 0.008 & 0.509 & -0.258 & 0.001 \\ -0.279 & -0.258 & 0.139 & 0.001 \\ -0.044 & 0.001 & 0.001 & 0.0003 \end{bmatrix}$$

$$\hat{\boldsymbol{\beta}} = (X'X)^{-1}(X'\boldsymbol{y}) = \begin{bmatrix} -1.163461 \\ 0.135270 \\ 0.019950 \\ 0.121954 \end{bmatrix}$$

$\boldsymbol{y}'\boldsymbol{y} = 285$

a. Use these results to construct the analysis of variance table.

b. Give the computed regression equation and the standard errors of the regression coefficients.

c. Compare each estimated regression coefficient to its standard error and use the t test to test the simple hypotheses that each individual regression coefficient is equal to zero. State your conclusions about β_1, β_2, and β_3.

4.17. Consider the following two models:

$$\text{Model A}: \quad y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$
$$\text{Model B}: \quad y_i = \beta_1 x_i + \epsilon_i$$

Suppose that model A is fitted to 22 data points (x_i, y_i) with the following results:

$$\hat{\boldsymbol{\beta}}' = (\hat{\beta}_0, \hat{\beta}_1) = (4.0, -4.5), \quad V(\hat{\beta}_0) = 4.0,$$
$$V(\hat{\beta}_1) = 9.0, \quad \text{and} \quad \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = 0.0$$

a. Construct individual 95% confidence intervals for β_0 and for β_1. What conclusions can you draw?

b. Construct a joint 95% confidence region for (β_0, β_1). Draw this confidence region on the plane of possible values for (β_0, β_1). On the basis of this region, what conclusions can you draw about the relative merits of models A and B?

c. Do the results of (a) and (b) conflict? Carefully explain your reasoning.

4.18. Consider the model

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 I)$$

Let $\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\boldsymbol{y}$, $\hat{\boldsymbol{\mu}} = H\boldsymbol{y}$, and $\boldsymbol{e} = (I - H)\boldsymbol{y}$, where $H = X(X'X)^{-1}X'$. Show that $\hat{\boldsymbol{\mu}}$ and $\boldsymbol{e}$ are statistically independent.

4.19. Consider a regression through the origin,

$$y_i = \beta x_i + \epsilon_i, \quad \text{with } E(\epsilon_i) = 0,$$
$$V(\epsilon_i) = \sigma^2 x_i^2, \quad i = 1, 2, \ldots, 12$$

a. Derive the generalized least squares estmate of β in Eq. (4.58) and obtain its variance. Note that the covariance matrix V and its inverse V^{-1} are diagonal matrices. The generalized least squares estimate minimizes a weighted sum of squares with weights given by the diagonal elements in V^{-1}. Hence, one refers to it as the **weighted least squares** estimate.

b. Suppose that $z_i = y_i/x_i$ and $\sum_{i=1}^{12} z_i = 30$. Find the numerical value for the weighted least squares estimate in (a) and express its variance as a function of σ^2.

4.20. Consider a regression through the origin,

$$y_i = \beta x_i + \epsilon_i, \quad \text{with } E(\epsilon_i) = 0,$$
$$V(\epsilon_i) = \sigma^2 x_i, \quad x_i > 0, \quad i = 1, 2, \ldots, 10$$

a. Derive the generalized (weighted) least squares estmator of β and obtain its variance.

b. Assume that the experimenter recorded only the sample means $\bar{x} = 15$ and $\bar{y} = 30$. If possible, obtain a numerical value for the weighted least squares estimate in (a) and express its variance as a function of σ^2.

4.21. The data are taken from Davies, O. L., and Goldsmith, P. L. (Eds.). *Statistical Methods in Research and Production* (4th ed.). Edinburgh, UK: Oliver & Boyd, 1972. The data are given in the file **abrasion**.

The hardness and the tensile strength of rubber affect its resistance to abrasion. Thirty samples of rubber are tested for hardness (in degrees Shore; the larger the number, the harder the rubber) and tensile strength (in kilograms per square centimeter). Each sample was subjected to steady abrasion for a certain fixed period of time, and the loss of rubber (in grams per hour of testing) was measured.

Develop a model that relates the abrasion loss to hardness and tensile strength.

Construct scatter plots of abrasion loss against hardness and tensile strength. Fit appropriate regression models, obtain and interpret the estimates of the coefficients, calculate the ANOVA table, and discuss the adequacy of the model fit. Use your model(s) to obtain a 95% confidence interval for the mean abrasion loss for rubber with hardness 70 and tensile strength 200.

$y =$ Abrasion Loss (g/hr)	$x_1 =$ Hardness (degree Shore)	$x_2 =$ Tensile Strength (kg/cm^2)
372	45	162
206	55	233
175	61	232
154	66	231
136	71	231
112	71	237
55	81	224
45	86	219
221	53	203
166	60	189
164	64	210
113	68	210
82	79	196
32	81	180
228	56	200
196	68	173
128	75	188
97	83	161
64	88	119
249	59	161
219	71	151
186	80	165
155	82	151
114	89	128
341	51	161
340	59	146
283	65	148
267	74	144
215	81	134
148	86	127

4.22. The data are taken from Joglekar, G., Schuenemeyer, J. H., and LaRiccia, V. Lack-of-fit testing when replicates are not available. *American Statistician,* 43, 135–143, 1989. The data are given in the file **woodstrength**.

The tensile strength of Kraft paper (in pounds per square inch) is measured against the percentage of hardwood in the batch of pulp from which the paper was produced. Data for 19 observations are given here.

Develop a model that relates tensile strength to the percentage of hardwood in the paper. Construct scatter plots of tensile strength against the percentage of hardwood.

a. Fit a linear model and comment on your findings.

b. Consider a model that also includes the square of the percentage of hardwood. Fit the quadratic model, obtain and interpret the estimates of the coefficients, calculate the ANOVA table, and discuss the adequacy of the model fit. Add the fitted line to your scatter plot. Discuss whether the quadratic component is needed. Use your model to obtain a 95% confidence interval for the mean tensile strength of paper with 6% hardwood content. How is this interval different from a corresponding prediction interval? Discuss whether it is reasonable to obtain a confidence interval for the mean tensile strength of paper with 20% hardwood content.

$x =$ Hardwood Concentration	$y =$ Tensile Strength
1.0	6.3
1.5	11.1
2.0	20.0
3.0	24.0
4.0	26.1
4.5	30.0
5.0	33.8
5.5	34.0
6.0	38.1
6.5	39.9
7.0	42.0
8.0	46.1
9.0	53.1
10.0	52.0

$x =$ Hardwood Concentration	$y =$ Tensile Strength
11.0	52.5
12.0	48.0
13.0	42.8
14.0	27.8
15.0	21.9

4.23. The data are taken from Humphreys, R. M. Studies of luminous stars in nearby galaxies. I. Supergiants and O stars in the Milky Way. *Astrophysics Journal, Supplementary Series,* 38, 309–350, 1978. The data are given in the file **lightintensity**.

Light intensity and surface temperature were determined for 47 stars taken from the Hertzsprung–Russel diagram of Star Cluster CYG OB1. The objective is to find a relationship between light intensity and surface temperature.

Construct a scatter plot of light intensity against surface temperature. Fit a quadratic regression model, obtain and interpret the estimates of the coefficients, calculate the ANOVA table, and discuss the adequacy of the model fit. Add the fitted line to your scatter plot.

What other interpretations of the scatter plot are possible? For example, could it be that four stars are different in the sense that they do not follow the linear pattern established by the other stars? What questions would you ask the astrophysicist?

Index	$x =$ Log Surface Temp	$y =$ Log Light Intensity
1	4.37	5.23
2	4.56	5.74
3	4.26	4.93
4	4.56	5.74
5	4.30	5.19
6	4.46	5.46
7	3.84	4.65
8	4.57	5.27
9	4.26	5.57
10	4.37	5.12
11	3.49	5.73
12	4.43	5.45
13	4.48	5.42
14	4.01	4.05
15	4.29	4.26
16	4.42	4.58
17	4.23	3.94
18	4.42	4.18
19	4.23	4.18
20	3.49	5.89
21	4.29	4.38
22	4.29	4.22
23	4.42	4.42
24	4.49	4.85
25	4.38	5.02
26	4.42	4.66
27	4.29	4.66
28	4.38	4.90
29	4.22	4.39
30	3.48	6.05
31	4.38	4.42
32	4.56	5.10
33	4.45	5.22
34	3.49	6.29
35	4.23	4.34
36	4.62	5.62
37	4.53	5.10
38	4.45	5.22
39	4.53	5.18
40	4.43	5.57
41	4.38	4.62
42	4.45	5.06
43	4.50	5.34
44	4.45	5.34
45	4.55	5.54
46	4.45	4.98
47	4.42	4.50

4.24. Consider the UFFI data set in Table 1.2 ($n = 24$ observations). Estimate the model with three regression coefficients, $y = \beta_0 + \beta_1 x_1(\text{UFFI}) + \beta_2 x_2(\text{TIGHT}) + \varepsilon$.

a. Use the statistical software of your choice and confirm the regression results in Table 4.1.

b. Determine the 3×3 matrix $X'X$ and its inverse $(X'X)^{-1}$. Determine the standard errors of the three estimates and the pairwise correlations among the estimates (there are three correlations).

c. Determine a 95% confidence region (ellipse) for the two slopes $\boldsymbol{\beta} = (\beta_1, \beta_2)'$. We know that the marginal distribution of $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2)'$ is a bivariate normal distribution with covariance matrix $\sigma^2 A^{-1}$, where A^{-1} is the appropriate 2×2 submatrix of $(X'X)^{-1}$ found in (b). Hence, the contours of the confidence ellipse can be traced out by solving $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})' A (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = 2s^2 F(0.95; 2, n-3)$. Here, $F(0.95; 2, n-3 = 21)$ is the 95th percentile of the F distribution, and s^2 is the mean square error.

4.25. Confidence intervals for regression coefficients and the mean response and prediction intervals for future observations in Section 4.3 make use of the t distribution. The t distribution as the resulting sampling distribution of the coefficient estimates in Eq. (4.24) depends critically on the model assumptions, in particular the assumption that the independent errors are normally distributed. The distribution in Eq. (4.24) is not a t distribution and it is no longer known if the distribution of the errors is nonnormal.

Bootstrapping (or resampling) methods are commonly used to overcome problems of unknown sampling distributions. The bootstrap, originally proposed by Efron (1979), approximates the unknown theoretical sampling distribution of the coefficient estimates by an empirical distribution that is obtained through a resampling process.

Several versions of the bootstrap are proposed for the regression situation, and the references listed at the end of this exercise will give you more details. Here, we discuss the "bootstrap in pairs" method, which resamples directly from the original data $(y_i, \boldsymbol{x}_i)$, $i = 1, 2, \ldots, n$. This method repeats the following steps B times. Sample with replacement n pairs from the original n observations $(y_i, \boldsymbol{x}_i)$. From these n sampled pairs, calculate the least squares estimates and denote the jth coefficient estimate by $\hat{\beta}_j^{*(b)}$. The superscript asterisk denotes the fact that the estimate is obtained from data generated by the bootstrap procedure, the superscript b denotes the bth replication, and the subscript j refers to a particular scalar coefficient. The B independent replications supply the empirical bootstrap distribution function.

Percentile bootstrap intervals are proposed as confidence intervals for the regression coefficients. One approach determines the $100(\alpha/2)$ and $100(1 - (\alpha/2))$ percentiles of the empirical bootstrap distribution function, $\hat{\beta}_j^*(\alpha/2)$ and $\hat{\beta}_j^*(1 - (\alpha/2))$, and computes a $100(1-\alpha)\%$ bootstrap confidence interval for the parameter β_j as

$$\hat{\beta}_j^*(\alpha/2), \quad \hat{\beta}_j^*(1 - (\alpha/2))$$

Here, we have given the very simplest bootstrap method for the regression situation. Modifications that improve on this simple procedure have been proposed and are discussed in the references. The modifications involve sampling residuals (compared to the resampling of cases discussed here) and refinements for improving the coverage properties of percentile bootstrap intervals [one modification calculates the lower and upper limits as $\hat{\beta}_j - [\hat{\beta}_j^*(1 - (\alpha/2)) - \hat{\beta}_j]$ and $\hat{\beta}_j - [\hat{\beta}_j^*(\alpha/2) - \hat{\beta}_j]$, where $\hat{\beta}_j$ is the estimate from the original sample].

a. Select one or more of the listed references and write a brief summary that explains the bootstrap methods in regression and discusses their importance.

b. Consider the simple linear regression model. Use the fuel efficiency data in Table 1.3 and regress fuel efficiency (gallons per 100 traveled miles) on the weight of the car. Obtain a 95% bootstrap confidence interval for the slope. Use $B =$ 1,000 and 2,000 replications. Relate the results to the standard confidence interval based on the t distribution.

Literature on the Bootstrap and Its Applications to Regression

Davison, A. C., and Hinkley, D. V. *Bootstrap Methods and Their Applications*. New York: Cambridge University Press, 1997.

Efron, B. Bootstrap methods: Another look at the jackknife. *Annals of Statistics,* 7, 1–26, 1979.

Efron, B., and Tibshirani, R. J. *An Introduction to the Bootstrap*. New York: Chapman & Hall, 1993.

Horowitz, J. L. The Bootstrap. In *Handbook of Econometrics* (Vol. 6). Amsterdam: North Holland, 1999.