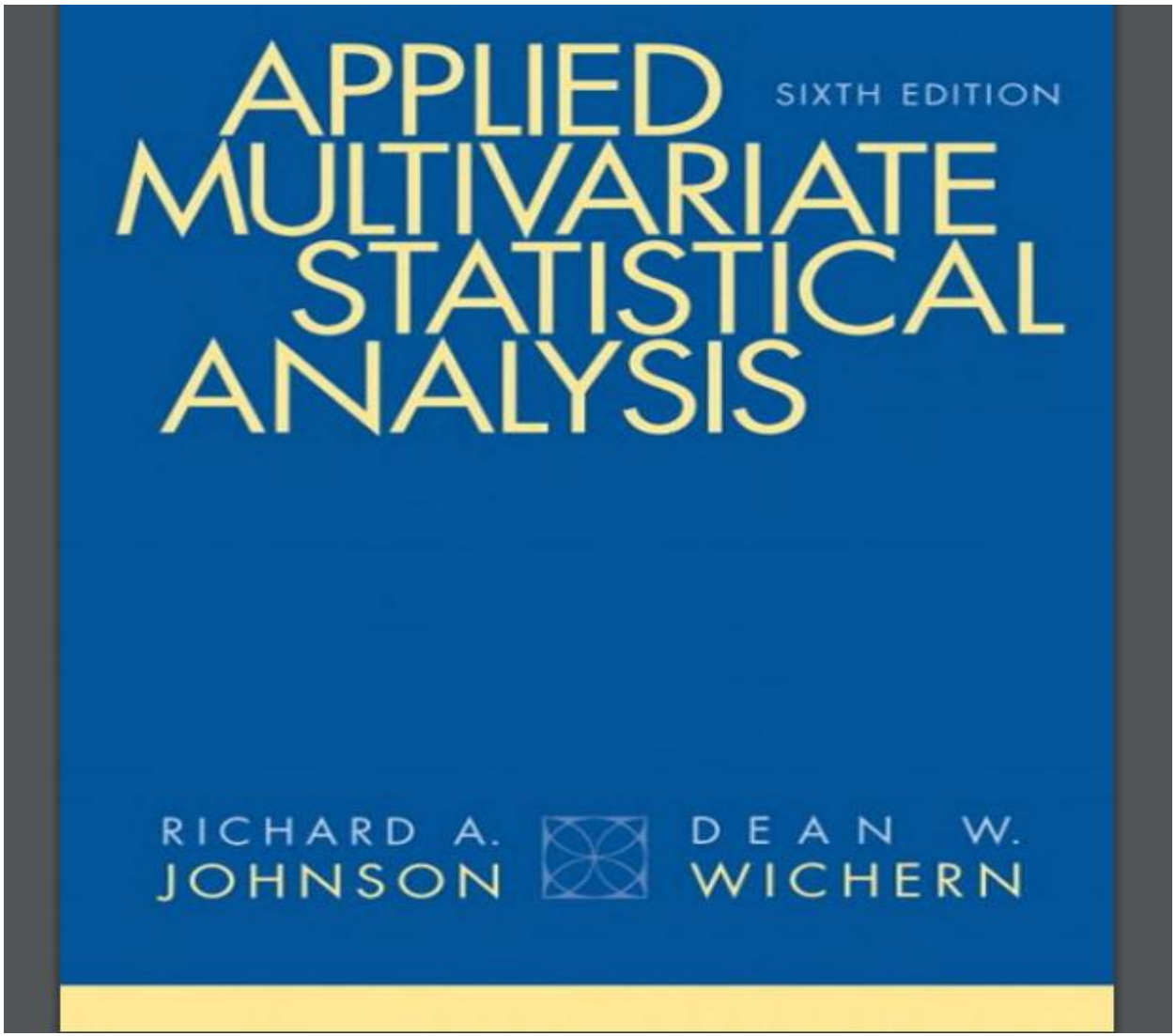
Advanced Multivariate Methods



Multivariate Linear Regression Models

MULTIVARIATE LINEAR REGRESSION MODELS

7.1 Introduction

Regression analysis is the statistical methodology for predicting values of one or more *response* (dependent) variables from a collection of *predictor* (independent) variable values. It can also be used for assessing the effects of the predictor variables on the responses. Unfortunately, the name *regression*, culled from the title of the first paper on the subject by F. Galton [15], in no way reflects either the importance or breadth of application of this methodology.

In this chapter, we first discuss the multiple regression model for the prediction of a *single* response. This model is then generalized to handle the prediction of *several* dependent variables. Our treatment must be somewhat terse, as a vast literature exists on the subject. (If you are interested in pursuing regression analysis, see the following books, in ascending order of difficulty: Abraham and Ledolter [1], Bowerman and O'Connell [6], Neter, Wasserman, Kutner, and Nachtsheim [20], Draper and Smith [13], Cook and Weisberg [11], Seber [23], and Goldberger [16].) Our abbreviated treatment highlights the regression assumptions and their consequences, alternative formulations of the regression model, and the general applicability of regression techniques to seemingly different situations.

Classical Linear Regression Model

7.2 The Classical Linear Regression Model

Let $z_1, z_2, \ldots, z_r$ be r predictor variables thought to be related to a response variable Y. For example, with $r = 4$, we might have

Y = current market value of home

and

z_1 = square feet of living area
z_2 = location (indicator for zone of city)
z_3 = appraised value last year
z_4 = quality of construction (price per square foot)

The classical linear regression model states that Y is composed of a mean, which depends in a continuous manner on the z_i's, and a random error ε, which accounts for measurement error and the effects of other variables not explicitly considered in the model. The values of the predictor variables recorded from the experiment or set by the investigator are treated as *fixed*. The error (and hence the response) is viewed as a random variable whose behavior is characterized by a set of distributional assumptions.

Format Linear Regression

Specifically, the linear regression model with a single response takes the form

$$Y = \beta_0 + \beta_1 z_1 + \cdots + \beta_r z_r + \varepsilon$$

$$[\text{Response}] = [\text{mean (depending on } z_1, z_2, \ldots, z_r)] + [\text{error}]$$

The term "linear" refers to the fact that the mean is a linear function of the unknown parameters $\beta_0, \beta_1, \ldots, \beta_r$. The predictor variables may or may not enter the model as first-order terms.

With n independent observations on Y and the associated values of z_i, the complete model becomes

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 z_{11} + \beta_2 z_{12} + \cdots + \beta_r z_{1r} + \varepsilon_1 \\ Y_2 &= \beta_0 + \beta_1 z_{21} + \beta_2 z_{22} + \cdots + \beta_r z_{2r} + \varepsilon_2 \\ &\vdots \qquad\qquad\qquad \vdots \\ Y_n &= \beta_0 + \beta_1 z_{n1} + \beta_2 z_{n2} + \cdots + \beta_r z_{nr} + \varepsilon_n \end{aligned} \tag{7-1}$$

where the error terms are assumed to have the following properties:

1. $E(\varepsilon_j) = 0$;
2. $\text{Var}(\varepsilon_j) = \sigma^2$ (constant); and (7-2)
3. $\text{Cov}(\varepsilon_j, \varepsilon_k) = 0, j \neq k$.

Error Terms and Specified Properties

$$Y_n = \beta_0 + \beta_1 z_{n1} + \beta_2 z_{n2} + \cdots + \beta_r z_{nr} + \varepsilon_n$$

where the error terms are assumed to have the following properties:

1. $E(\varepsilon_j) = 0$;
2. $\text{Var}(\varepsilon_j) = \sigma^2$ (constant); and (7-2)
3. $\text{Cov}(\varepsilon_j, \varepsilon_k) = 0, j \neq k$.

In matrix notation, (7-1) becomes

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & z_{11} & z_{12} & \cdots & z_{1r} \\ 1 & z_{21} & z_{22} & \cdots & z_{2r} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & z_{n1} & z_{n2} & \cdots & z_{nr} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_r \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

or

$$\underset{(n\times 1)}{\mathbf{Y}} = \underset{(n\times(r+1))}{\mathbf{Z}} \underset{((r+1)\times 1)}{\boldsymbol{\beta}} + \underset{(n\times 1)}{\boldsymbol{\varepsilon}}$$

and the specifications in (7-2) become

1. $E(\boldsymbol{\varepsilon}) = \mathbf{0}$; and
2. $\text{Cov}(\boldsymbol{\varepsilon}) = E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2\mathbf{I}$.

Classical Linear Regression Model

Note that a one in the first column of the *design matrix* $\mathbf{Z}$ is the multiplier of the constant term β_0. It is customary to introduce the artificial variable $z_{j0} = 1$, so that

$$\beta_0 + \beta_1 z_{j1} + \cdots + \beta_r z_{jr} = \beta_0 z_{j0} + \beta_1 z_{j1} + \cdots + \beta_r z_{jr}$$

Each column of $\mathbf{Z}$ consists of the n values of the corresponding predictor variable, while the jth row of $\mathbf{Z}$ contains the values for all predictor variables on the jth trial.

Classical Linear Regression Model

$$\underset{(n\times 1)}{\mathbf{Y}} = \underset{(n\times(r+1))}{\mathbf{Z}}\ \underset{((r+1)\times 1)}{\boldsymbol{\beta}} + \underset{(n\times 1)}{\boldsymbol{\varepsilon}},$$

$$E(\boldsymbol{\varepsilon}) = \underset{(n\times 1)}{\mathbf{0}} \text{ and } \operatorname{Cov}(\boldsymbol{\varepsilon}) = \underset{(n\times n)}{\sigma^2 \mathbf{I}}, \tag{7-3}$$

where $\boldsymbol{\beta}$ and σ^2 are unknown parameters and the design matrix $\mathbf{Z}$ has jth row $[z_{j0}, z_{j1}, \ldots, z_{jr}]$.

Although the error-term assumptions in (7-2) are very modest, we shall later need to add the assumption of joint normality for making confidence statements and testing hypotheses.

We now provide some examples of the linear regression model.

Example: Fitting Straight Line Regression Model

Example 7.1 (Fitting a straight-line regression model) Determine the linear regression model for fitting a straight line

$$\text{Mean response} = E(Y) = \beta_0 + \beta_1 z_1$$

to the data

z_1	0	1	2	3	4
y	1	4	3	8	9

Before the responses $\mathbf{Y}' = [Y_1, Y_2, \ldots, Y_5]$ are observed, the errors $\boldsymbol{\varepsilon}' = [\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_5]$ are random, and we can write

$$\mathbf{Y} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_5 \end{bmatrix}, \quad \mathbf{Z} = \begin{bmatrix} 1 & z_{11} \\ 1 & z_{21} \\ \vdots & \vdots \\ 1 & z_{51} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_5 \end{bmatrix}$$

Example: Fitting Straight Line Regression Model y as Dependent

The data for this model are contained in the observed response vector $\mathbf{y}$ and the design matrix $\mathbf{Z}$, where

$$\mathbf{y} = \begin{bmatrix} 1 \\ 4 \\ 3 \\ 8 \\ 9 \end{bmatrix}, \quad \mathbf{Z} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}$$

Note that we can handle a quadratic expression for the mean response by introducing the term $\beta_2 z_2$, with $z_2 = z_1^2$. The linear regression model for the jth trial in this latter case is

$$Y_j = \beta_0 + \beta_1 z_{j1} + \beta_2 z_{j2} + \varepsilon_j$$

or

$$Y_j = \beta_0 + \beta_1 z_{j1} + \beta_2 z_{j1}^2 + \varepsilon_j$$

■

Design Matrix for One-Way ANOVA as a Regression Model

Example 7.2 (The design matrix for one-way ANOVA as a regression model) Determine the design matrix if the linear regression model is applied to the one-way ANOVA situation in Example 6.6.

We create so-called *dummy* variables to handle the three population means: $\mu_1 = \mu + \tau_1$, $\mu_2 = \mu + \tau_2$, and $\mu_3 = \mu + \tau_3$. We set

$$z_1 = \begin{cases} 1 & \text{if the observation is from population 1} \\ 0 & \text{otherwise} \end{cases} \qquad z_2 = \begin{cases} 1 & \text{if the observation is from population 2} \\ 0 & \text{otherwise} \end{cases}$$

$$z_3 = \begin{cases} 1 & \text{if the observation is from population 3} \\ 0 & \text{otherwise} \end{cases}$$

and $\beta_0 = \mu, \beta_1 = \tau_1, \beta_2 = \tau_2, \beta_3 = \tau_3$. Then

$$Y_j = \beta_0 + \beta_1 z_{j1} + \beta_2 z_{j2} + \beta_3 z_{j3} + \varepsilon_j, \qquad j = 1, 2, \ldots, 8$$

where we arrange the observations from the three populations in sequence. Thus, we obtain the observed response vector and design matrix

$$\underset{(8\times 1)}{\mathbf{Y}} = \begin{bmatrix} 9 \\ 6 \\ 9 \\ 0 \\ 2 \\ 3 \\ 1 \\ 2 \end{bmatrix}; \qquad \underset{(8\times 4)}{\mathbf{Z}} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix}$$

■

The construction of dummy variables, as in Example 7.2, allows the whole of analysis of variance to be treated within the multiple linear regression framework.

Review: Simple Linear Regression

Testing Linear Relationship of an Assumption: Are Steeper Roller Coasters Faster?

© Rafael Macia/Science Source

The Highest Roller Coasters Are Fastest

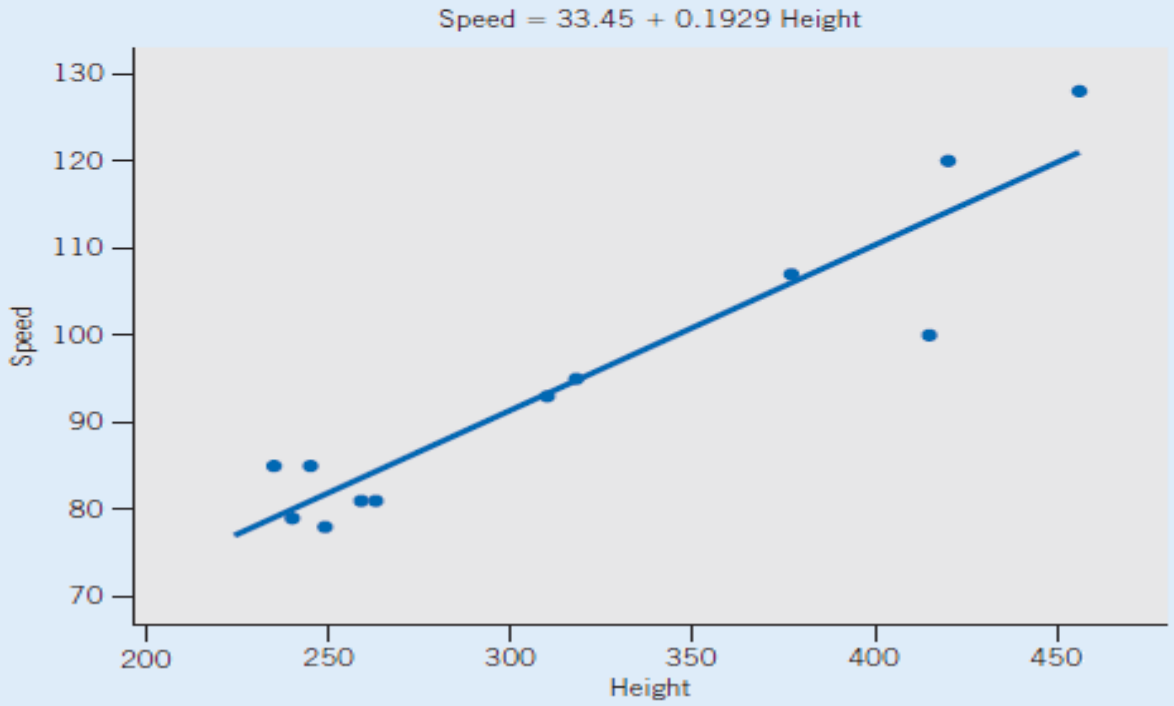
Some roller coasters are designed to twist riders and turn them upside down. Others are designed to provide fast rides over large drops. Among the 12 tallest roller coasters in the world, the maximum height (feet) is related to top speed (miles per hour). Each data point, consisting of the pair of values (height, speed), represents one roller coaster. The fitted line predicts an increase in top speed of .19 miles per hour for each foot of height, or 19 miles per hour for each 100 feet in height.

This linear relation can be used to predict the top speed of future roller coasters. For instance, that of the next 410 foot roller coaster.

Linear Relationship Between Height and Speed Rollercoasters: Regression Analysis

The Highest Roller Coasters Are Fastest

Some roller coasters are designed to twist riders and turn them upside down. Others are designed to provide fast rides over large drops. Among the 12 tallest roller coasters in the world, the maximum height (feet) is related to top speed (miles per hour). Each data point, consisting of the pair of values (height, speed), represents one roller coaster. The fitted line predicts an increase in top speed of .19 miles per hour for each foot of height, or 19 miles per hour for each 100 feet in height.
This linear relation can be used to predict the top speed of future roller coasters. For instance, that of the next 410 foot roller coaster.



Basic Questions of Regression

Except for the brief treatment in Sections 4 and 5 of Chapter 3, we have only discussed statistical inferences based on the sample measurements of a single variable. In many investigations, two or more variables are observed for each experimental unit in order to determine:

1. Whether the variables are related.
2. How strong the relationships appear to be.
3. Whether one variable of primary interest can be predicted from observations on the others.

To address these issues, we review and then expand the treatment in Chapter 3, which is restricted to a descriptive viewpoint.

In this chapter, we take the additional step of including the omnipresent random variation as an error term in the model. Then, on the basis of the model, we can test whether one variable actually influences the other. Further, we produce confidence interval answers when using the estimated straight line for prediction. The correlation coefficient is shown to measure the strength of the linear relationship.

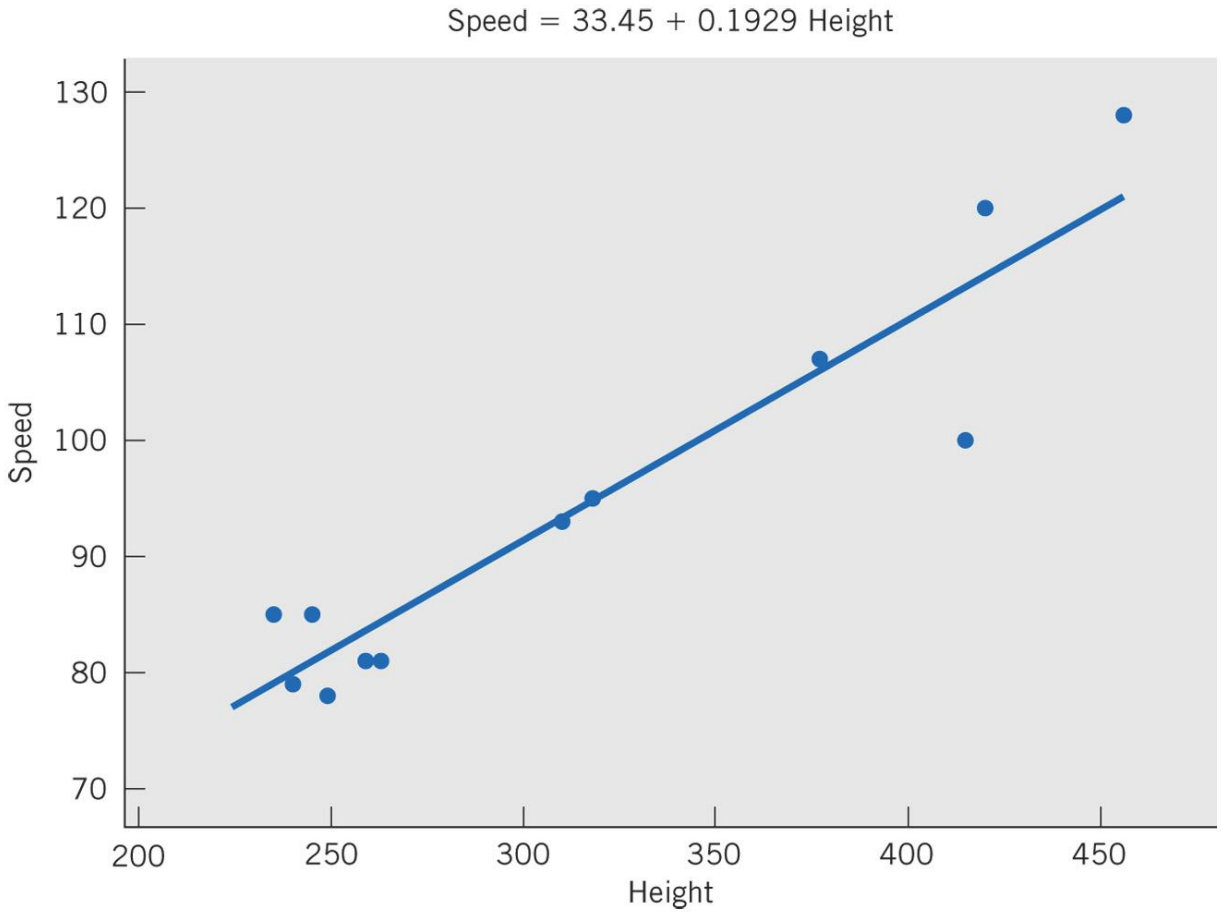


Figure on Page 449

Regression with Single Predictor Considered Bivariate (2 Variables)

1. Regression With a Single Predictor

Regression analysis concerns the study of relationships between quantitative variables with the object of identifying, estimating, and validating the relationship. The estimated relationship can then be used to predict one variable from the value of the other variables(s).
In this chapter, we confine our attention to study of the relation between two variables. A regression problem involving a single predictor is also called simple regression.

One may be curious about why the study of relationships of variables has been given the rather unusual name "regression." Historically, the word regression was first used in its present technical context by a British scientist, Sir Francis Galton, who analyzed the heights of sons and the average heights of their parents. From his observations, Galton concluded that sons of very tall (short) parents were generally taller (shorter) than the average but not as tall (short) as their parents. This result was published in 1885 under the title "Regression Toward Mediocrity in Hereditary Stature." In this context, "regression toward mediocrity" meant that the sons' heights tended to revert toward the average rather than progress to more extremes. However, in the course of time, the word regression became synonymous with the statistical study of relation among variables.

Underlying Assumption Regression: Randomness and Normal Distribution

1.1 HOW DOES RANDOMNESS IMPACT THE STUDY OF A RELATION?

Studies of relation among variables abound in virtually all disciplines of science and the humanities. We outline just a few illustrative situations in order to bring the object of regression analysis into sharp focus. The examples progress from a case where beforehand there is an underlying straight line model that is masked by random disturbances to a case where the data may or may not reveal some relationship along a line or curve.

Example 1

A Straight Line Model Masked by Random Disturbances

A factory manufactures items in batches and the production manager wishes to relate the production cost y of a batch to the batch size x. Certain costs are practically constant, regardless of the batch size x. Building costs and administrative and supervisory salaries are some examples. Let us denote the fixed costs collectively by F. Certain other costs may be directly proportional to the number of units produced. For example, both the raw materials and labor required to produce the product are included in this category. Let C denote the cost of producing one item. In the absence of any other factors, we can then expect to have the relation

Relationship Between Cost, Labor and Output or Production of Good

. Certain other costs may be directly proportional to the number of units produced. For example, both the raw materials and labor required to produce the product are included in this category. Let C denote the cost of producing one item. In the absence of any other factors, we can then expect to have the relation

$$y = F + \mathrm{C}x$$

In reality, other factors also affect the production cost, often in unpredictable ways. Machines occasionally break down and result in lost time and added expenses for repair. Variation of the quality of the raw materials may also cause occasional slowdown of the production process. Thus, an ideal relation can be masked by random disturbances. Consequently, the relationship between y and x must be investigated by a statistical analysis of the cost and batch-size data.

Fertilizer and Growth of Tomato Plants: May or May not be Linear

Example 2

Expect an Increasing Relation but Not Necessarily a Straight Line

Suppose that the yield y of tomato plants in an agricultural experiment is to be studied in relation to the dosage x of a certain fertilizer, while other contributing factors such as irrigation and soil dressing are to remain as constant as possible. The experiment consists of applying different dosages of the fertilizer, over the range of interest, in different plots and then recording the tomato yield from these plots. Different dosages of the fertilizer will typically produce different yields, but the relationship is not expected to follow a precise mathematical formula. Aside from unpredictable chance variations, the underlying form of the relation is not known.

Scatter Plot: Tighter Elliptical Shape Indicates Stronger Linear Relationship Between x and y

Example 3

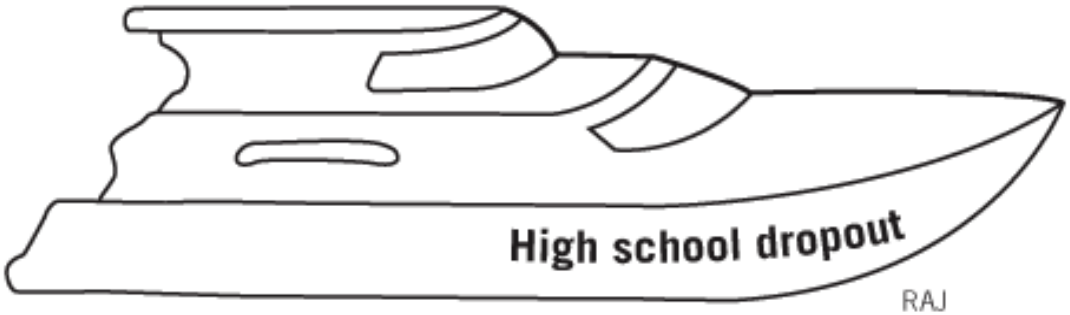
A Scatter Diagram May Reveal an Empirical Relation

The aptitude of a newly trained operator for performing a skilled job depends on both the duration of the training period and the nature of the training program. To evaluate the effectiveness of the training program, we must conduct an experimental study of the relation between growth in skill or learning y and duration x of the training. It is too much to expect a precise mathematical relation simply because no two human beings are exactly alike. However, an analysis of the data of the two variables could help us to assess the nature of the relation and utilize it in evaluating a training program.

For Any Fixed Value x or Variable, Randomness of x Affects the Response y

In each of these three cases, for any fixed value x, randomness causes the observed values of the response to spread over a range of values. Consequently, any relation with x must actually be formulated in terms of the expected value of the response variable and how it varies as x changes.

These examples illustrate the simplest settings for regression analysis where one wishes to determine how one variable is related to one other variable. In more complex situations several variables may be interrelated, or one variable of major interest may depend on several influencing variables. Regression analysis extends to these multivariate problems. (See Section 2, Chapter 12.)



Fine but you are an exception. Statistics,[1]Median weekly earnings in 2012, Bureau of Labor Statistics, Current Population Survey. based on extensive data, confirm that earnings typically increase with each additional step in education.

Even though randomness is omnipresent, regression analysis allows us to identify it and estimate relationships.

Regression Notation: x-variable is the Predictor or Independent; y-variables is the Dependent or Response

1.2 THE FIRST STEP IN REGRESSION ANALYSIS—PLOT THE DATA

Our aim is to study the relation between two variables x and y and use it to predict y from x. The variable x acts as an independent variable whose values are controlled by the experimenter. The variable y depends on x and is also subjected to unaccountable variations or errors.

Notation

x = **independent variable**, also called **predictor variable, explanatory variable, causal variable,** or **input variable**

y = **dependent** or **response variable**

For clarity, we introduce the main ideas of regression in the context of a specific experiment. This experiment, described in Example 4, and the data set of Table 1 will be referred to throughout this chapter. By so doing, we provide a flavor of the subject matter interpretation of the various inferences associated with a regression analysis.

Notation

x = **independent variable**, also called **predictor variable**, **explanatory variable**, **causal variable**, or **input variable**

y = **dependent** or **response variable**

Box on Page 452

Definitions of variables

Effect of x (dosage) on y (Duration of Relief): Unit of Analysis Human Subjects

For clarity, we introduce the main ideas of regression in the context of a specific experiment. This experiment, described in Example 4, and the data set of Table 1 will be referred to throughout this chapter. By so doing, we provide a flavor of the subject matter interpretation of the various inferences associated with a regression analysis.

TABLE 1 Dosage x (in milligrams) and the Number of Hours of Relief y from Allergy for Ten Patients

Dosage x	Duration of Relief y
3	9
3	5
4	12
5	9
6	14
6	16
7	22
8	18
8	24
9	22

TABLE 1 Dosage x (in Milligrams) and the Number of Days of Relief y from Allergy for Ten Patients

Dosage x	Duration of Relief y
3	9
3	5
4	12
5	9
6	14
6	16
7	22
8	18
8	24
9	22

Table 1 (p. 453)

Dosage *x* (in Milligrams) and the Number of Days of Relief *y* from Allergy for Ten Patients

Effect of x (dosage) on y (Duration of Relief): Sample Size = 10 Patients

Example 4

Relief from Symptoms of Allergy Related to Dosage

In one stage of the development of a new drug for an allergy, an experiment is conducted to study how different dosages of the drug affect the duration of relief from the allergic symptoms. Ten patients are included in the experiment. Each patient receives a specified dosage of the drug and is asked to report back as soon as the protection of the drug seems to wear off. The observations are recorded in Table 1, which shows the dosage x and duration of relief y for the 10 patients.

Seven different dosages are used in the experiment, and some of these are repeated for more than one patient. A glance at the table shows that y generally increases with x, but it is difficult to say much more about the form of the relation simply by looking at this tabular data.

Data Table: Effect of x (Independent Variable) on y (Response Variable)

For a generic experiment, we use n to denote the sample size or the number of runs of the experiment. Each run gives a pair of observations (x, y) in which x is the fixed setting of the independent variable and y denotes the corresponding response. See Table 2.

TABLE 2 Data Structure for a Simple Regression

Setting of the Independent Variable	Response
x_1	y_1
x_2	y_2
x_3	y_3
.	.
.	.
.	.
x_n	y_n

We always begin our analysis by plotting the data because the eye can easily detect patterns along a line or curve.

TABLE 2 Data Structure for a Simple Regression

Setting of the Independent Variable	Response
x_1	y_1
x_2	y_2
x_3	y_3
.	.
.	.
.	.
x_n	y_n

Table 2 (p. 453)

Data Structure for a Simple Regression

Scatter Diagram Reveals Linear Relationship: i.e. Based on y = mx + b

First Step in the Analysis

Plotting a **scatter diagram** is an important preliminary step prior to undertaking a formal statistical analysis of the relationship between two variables.

The existence of any increasing, or decreasing, relationship is readily apparent and preliminary judgments can be reached whether or not it is a straight line relation.

The scatter diagram of the observations in Table 1 appears in Figure 1. This scatter diagram reveals that the relationship is approximately linear in nature; that is, the points seem to cluster around a straight line. Because a linear relation is the simplest relationship to handle mathematically, we present the details of the statistical regression analysis for this case. Other situations can often be reduced to this case by applying a suitable transformation to one or both variables.

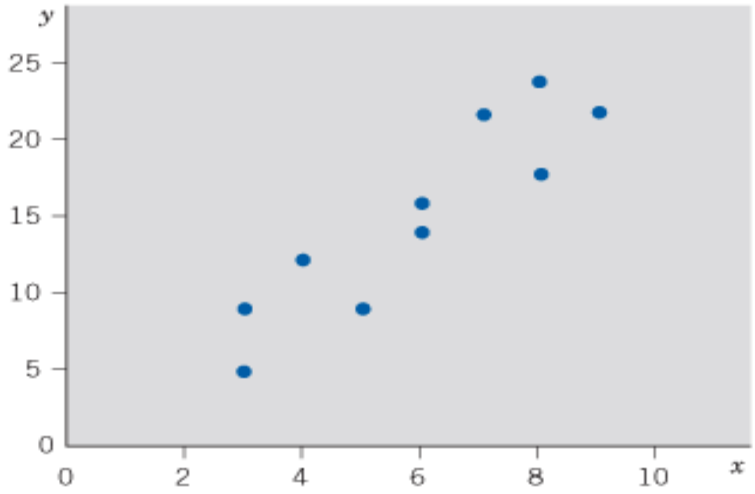


Figure 1 Scatter diagram of the data of Table 1.

First Step in the Analysis

Plotting a **scatter diagram** is an important preliminary step prior to undertaking a formal statistical analysis of the relationship between two variables.

Box on Page 454

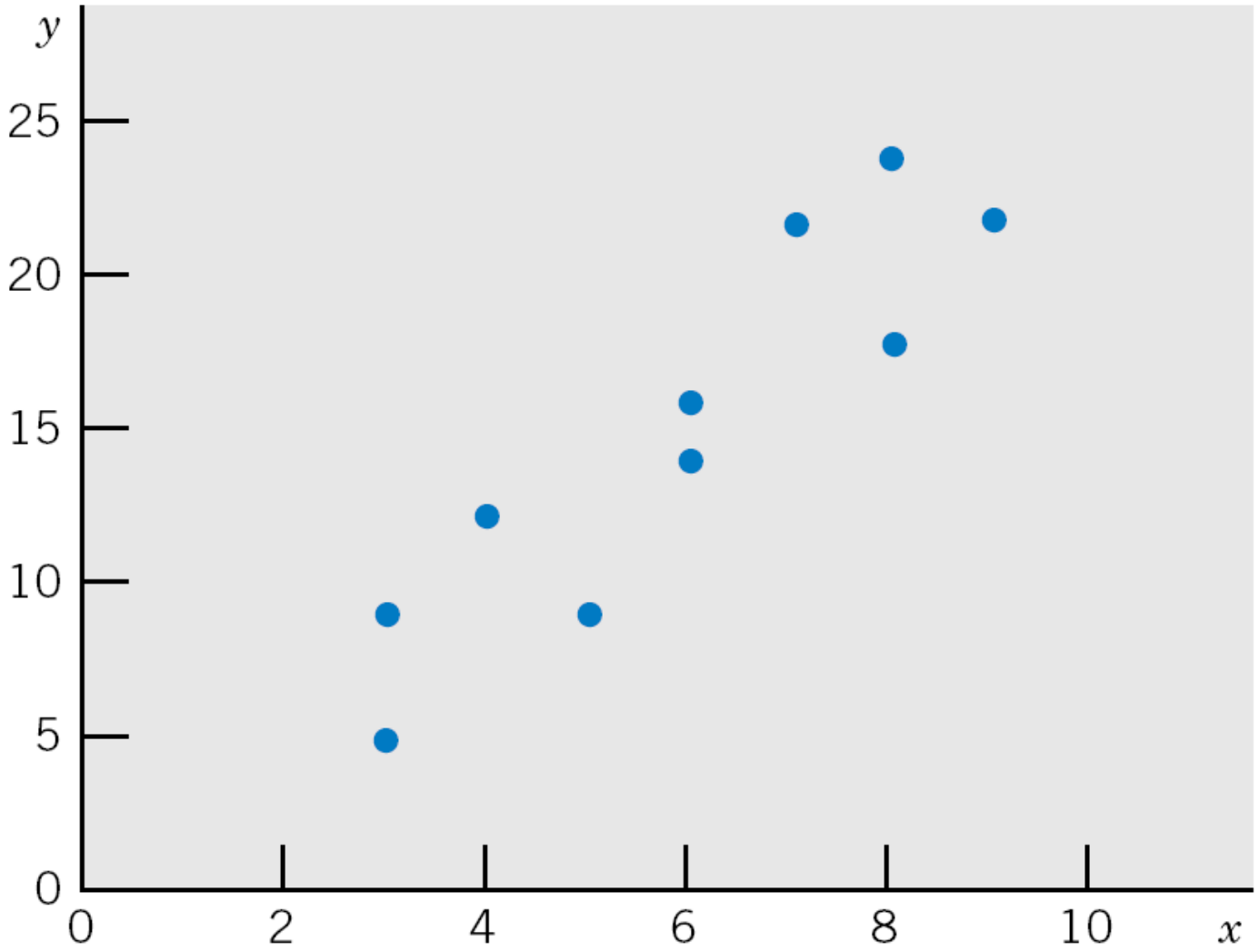


Figure 1 (p. 454)

Scatter diagram of the data of Table 1.

Straight Line Regression Equation $y = B_o + B_1X$ or $y = mx + b$

2. A Straight Line Regression Model

Recall that if the relation between y and x is exactly a straight line, then the variables are connected by the formula

$$y = \beta_0 + \beta_1 x$$

where β_0 indicates the intercept of the line with the y axis, and β_1 represents the slope of the line, or the change in y per unit change in x (see Figure 2).

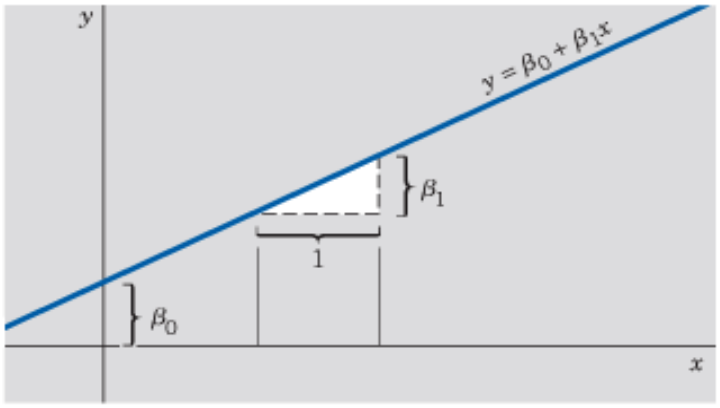


Figure 2 Graph of straight line $y = \beta_0 + \beta_1 x$.

Straight Line Regression Equation $y = B_o + B_1X$ or $y = mx + b$ Tentative Relationship Between x and y (Other Variables May Influence Not Measured In This Analysis)

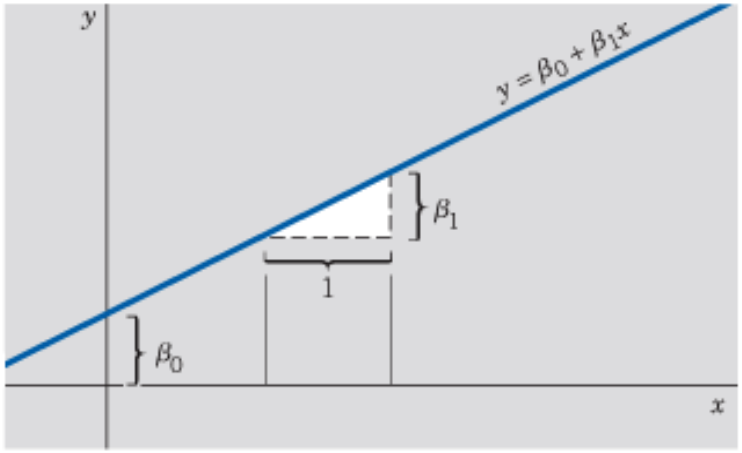


Figure 2 Graph of straight line $y = \beta_0 + \beta_1 x$.

Statistical ideas must be introduced into the study of relation when the points in a scatter diagram do not lie perfectly on a line, as in Figure 1. We think of these data as observations on an underlying linear relation that is being masked by random disturbances or experimental errors due in part to differences in severity of allergy, physical condition of subjects, their environment, and so on. All of the variables that influence the response, days of relief, are not even known, yet alone measured. The effects of all these variables are modeled as unobservable random variables. Given this viewpoint, we formulate the following linear regression model as a tentative representation of the mode of relationship between y and x.

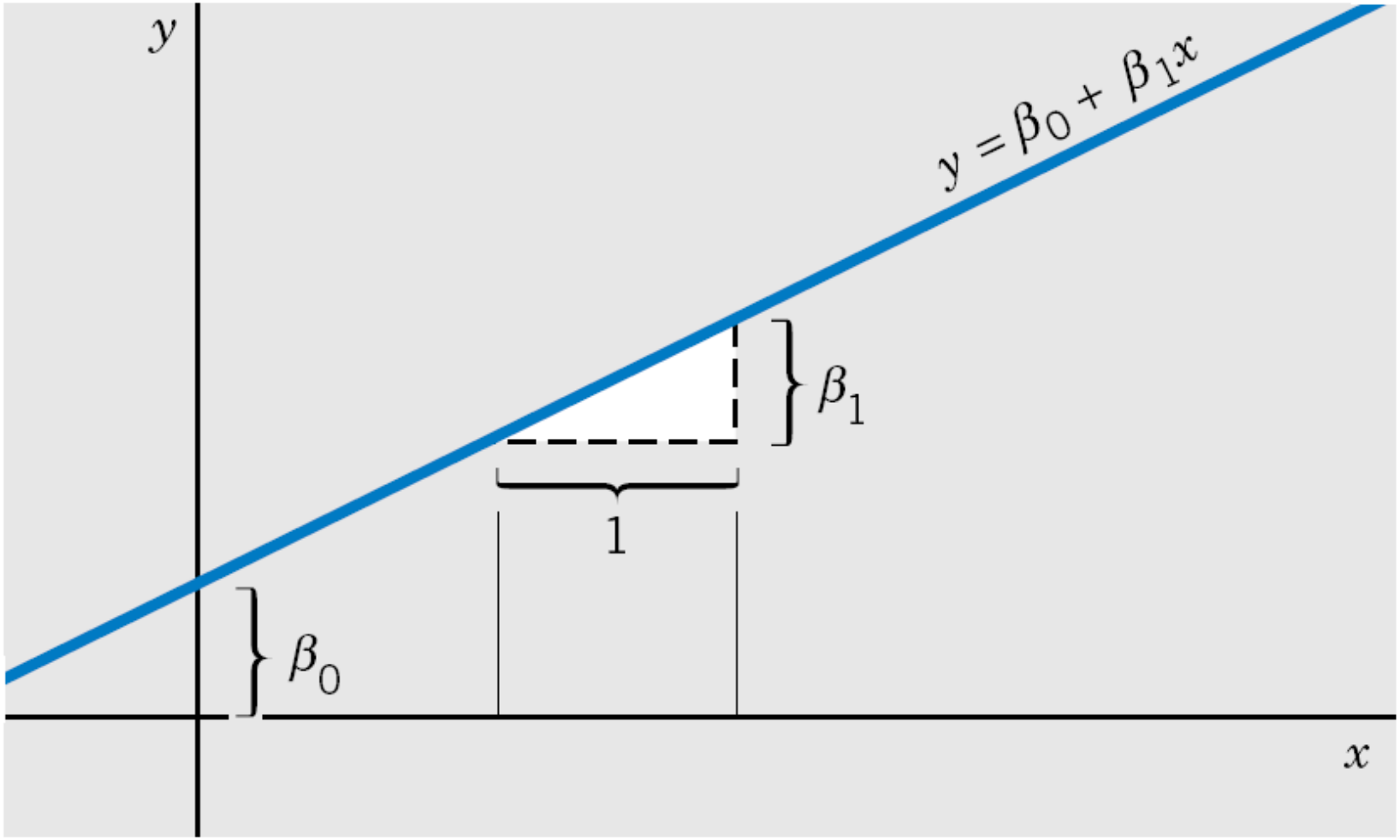


Figure 2 (p. 455)
Graph of straight line $y = \beta_0 + \beta_1 x$.

Regression is the Mean Line In the Scatter Plot Between x and y

Statistical Model for a Straight Line Regression

We assume that the response Y is a random variable that is related to the input variable x by

$$Y_i = \beta_0 + \beta_1 x_i + e_i \quad i = 1, \ldots, n$$

where:

1. Y_i denotes the response corresponding to the ith experimental run in which the input variable x is set at the value x_i.
2. $e_1, \ldots, e_n$ are the unknown error components that are superimposed on the true linear relation. These are **unobservable random variables**, which we assume are independently and normally distributed with mean zero and an unknown standard deviation σ.
3. The parameters β_0 and β_1, which together locate the straight line, are unknown.

The mean of the response Y_i, corresponding to the level x_i of the controlled variable, is $\beta_0 + \beta_1 x_i$.

Analogous to using the Greek letter μ to denote a single population mean, we use the Greek letters β_0 and β_1 to denote the slope and intercept of the unknown line on which the means lie as x varies.

Error or Residual is the Leftover Distance from the Mean Regression Line

Further, according to this model, the observation Y_i is one observation from the normal distribution with mean $\beta_0 + \beta_1 x_i$ and standard deviation σ. One interpretation of this is that as we attempt to observe the true value on the line, nature adds the random error e to this quantity. This statistical model is illustrated in Figure 3, which shows a few normal distributions for the response variable Y for different values of the input variable x. All these distributions have the same standard deviation and their means lie on the unknown true straight line $\beta_0 + \beta_1 x$. Aside from the fact that σ is unknown, the line on which the means of these normal distributions are located is also unknown. In fact, an important objective of the statistical analysis is to estimate this line.

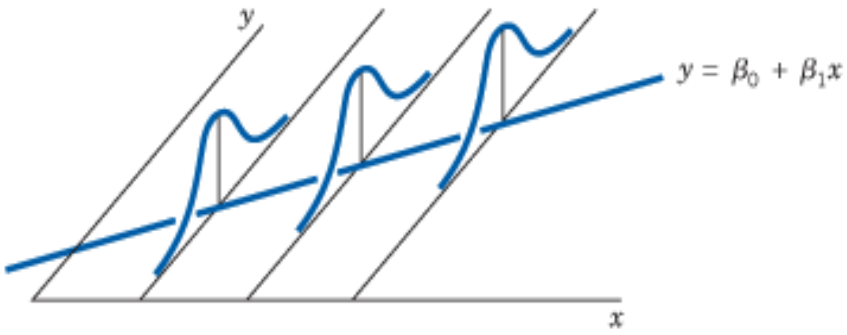


Figure 3 Normal distributions of Y with means on a straight line.

Statistical Model for a Straight Line Regression

We assume that the response Y is a random variable that is related to the input variable x by

$$Y_i = \beta_0 + \beta_1 x_i + e_i \qquad i = 1, \ldots, n$$

where:

1. Y_i denotes the response corresponding to the ith experimental run in which the input variable x is set at the value x_i.
2. $e_1, \ldots, e_n$ are the unknown error components that are superimposed on the true linear relation. These are **unobservable random variables,** which we assume are independently and normally distributed with mean zero and an unknown standard deviation σ.
3. The parameters β_0 and β_1, which together locate the straight line, are unknown.

Box on Page 455

Statistical Model for a Straight Line Regression

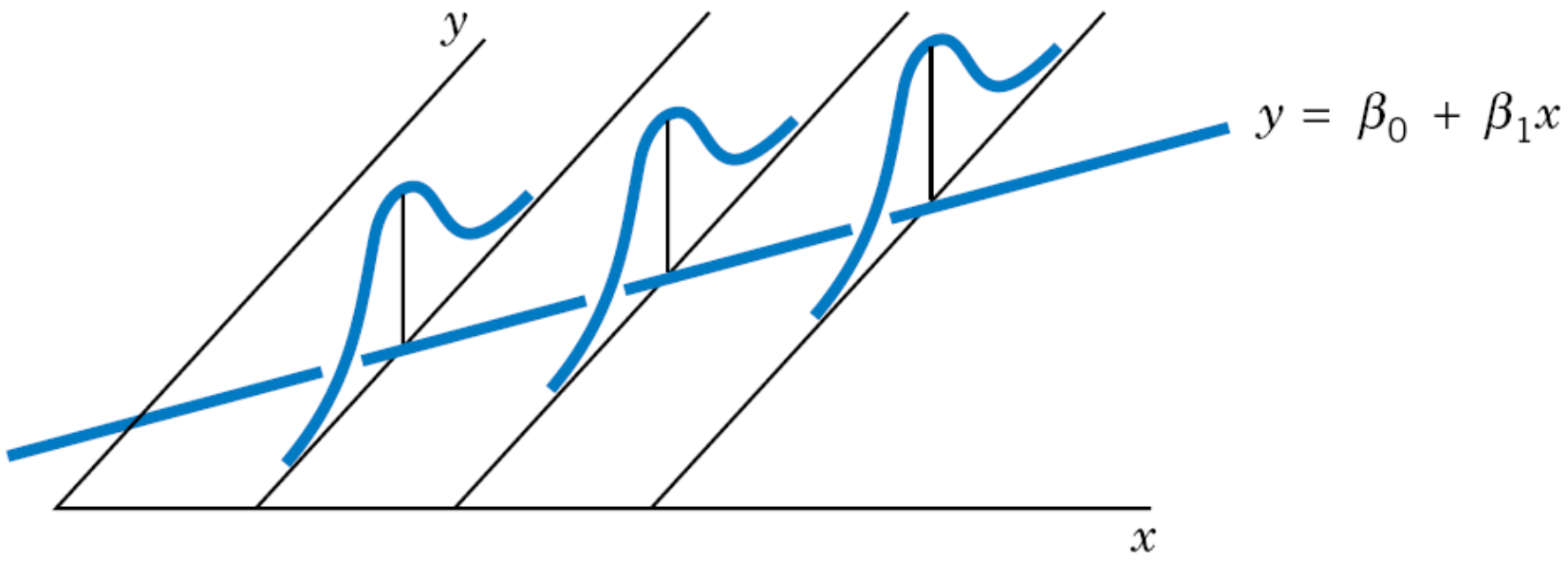


Figure 3 (p. 456)

Normal distributions of *Y* with means on a straight line.

Least Squares Estimation

7.3 Least Squares Estimation

One of the objectives of regression analysis is to develop an equation that will allow the investigator to predict the response for given values of the predictor variables. Thus, it is necessary to "fit" the model in (7-3) to the observed y_j corresponding to the known values $1, z_{j1}, \ldots, z_{jr}$. That is, we must determine the values for the *regression coefficients* $\boldsymbol{\beta}$ and the *error variance* σ^2 consistent with the available data.

Let $\mathbf{b}$ be trial values for $\boldsymbol{\beta}$. Consider the difference $y_j - b_0 - b_1 z_{j1} - \cdots - b_r z_{jr}$ between the observed response y_j and the value $b_0 + b_1 z_{j1} + \cdots + b_r z_{jr}$ that would be expected if $\mathbf{b}$ were the "true" parameter vector. Typically, the differences $y_j - b_0 - b_1 z_{j1} - \cdots - b_r z_{jr}$ will not be zero, because the response fluctuates (in a manner characterized by the error term assumptions) about its expected value. The *method of least squares* selects $\mathbf{b}$ so as to minimize the sum of the squares of the differences:

$$S(\mathbf{b}) = \sum_{j=1}^{n} (y_j - b_0 - b_1 z_{j1} - \cdots - b_r z_{jr})^2 = (\mathbf{y} - \mathbf{Z}\mathbf{b})'(\mathbf{y} - \mathbf{Z}\mathbf{b}) \tag{7-4}$$

The coefficients $\mathbf{b}$ chosen by the least squares criterion are called *least squares estimates* of the regression parameters $\boldsymbol{\beta}$. They will henceforth be denoted by $\hat{\boldsymbol{\beta}}$ to emphasize their role as estimates of $\boldsymbol{\beta}$.

Least Squares: Coefficients Beta-hat Produce Estimated (Fitted) Mean Responses, Sum of Whose Squares of the Differences from the Observed is as Small as Possible

The coefficients $\hat{\boldsymbol{\beta}}$ are consistent with the data in the sense that they produce estimated (fitted) mean responses, $\hat{\beta}_0 + \hat{\beta}_1 z_{j1} + \cdots + \hat{\beta}_r z_{jr}$, the sum of whose squares of the differences from the observed y_j is as small as possible. The deviations

$$\hat{\varepsilon}_j = y_j - \hat{\beta}_0 - \hat{\beta}_1 z_{j1} - \cdots - \hat{\beta}_r z_{jr}, \qquad j = 1, 2, \ldots, n \tag{7-5}$$

are called *residuals*. The vector of residuals $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{Z}\hat{\boldsymbol{\beta}}$ contains the information about the remaining unknown parameter σ^2. (See Result 7.2.)

Result 7.1. Let $\mathbf{Z}$ have full rank $r + 1 \leq n$.[1] The least squares estimate of $\boldsymbol{\beta}$ in (7-3) is given by

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

Let $\hat{\mathbf{y}} = \mathbf{Z}\hat{\boldsymbol{\beta}} = \mathbf{H}\mathbf{y}$ denote the *fitted values* of $\mathbf{y}$, where $\mathbf{H} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$ is called "hat" matrix. Then the *residuals*

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \hat{\mathbf{y}} = [\mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}']\mathbf{y} = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

satisfy $\mathbf{Z}'\hat{\boldsymbol{\varepsilon}} = \mathbf{0}$ and $\hat{\mathbf{y}}'\hat{\boldsymbol{\varepsilon}} = 0$. Also, the

$$\begin{aligned} \textit{residual sum of squares} &= \sum_{j=1}^{n} (y_j - \hat{\beta}_0 - \hat{\beta}_1 z_{j1} - \cdots - \hat{\beta}_r z_{jr})^2 = \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} \\ &= \mathbf{y}'[\mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}']\mathbf{y} = \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{Z}\hat{\boldsymbol{\beta}} \end{aligned}$$

[1] If $\mathbf{Z}$ is not full rank, $(\mathbf{Z}'\mathbf{Z})^{-1}$ is replaced by $(\mathbf{Z}'\mathbf{Z})^{-}$, a *generalized inverse* of $\mathbf{Z}'\mathbf{Z}$. (See Exercise 7.6.)

Calculating Least Squares Estimates, Residuals and Residual Sum of Squares

Example 7.3 (Calculating the least squares estimates, the residuals, and the residual sum of squares) Calculate the least square estimates $\hat{\boldsymbol{\beta}}$, the residuals $\hat{\boldsymbol{\varepsilon}}$, and the residual sum of squares for a straight-line model

$$Y_j = \beta_0 + \beta_1 z_{j1} + \varepsilon_j$$

fit to the data

z_1	0	1	2	3	4
y	1	4	3	8	9

We have

$$\mathbf{Z}' = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 4 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} 1 \\ 4 \\ 3 \\ 8 \\ 9 \end{bmatrix} \quad \mathbf{Z}'\mathbf{Z} = \begin{bmatrix} 5 & 10 \\ 10 & 30 \end{bmatrix} \quad (\mathbf{Z}'\mathbf{Z})^{-1} = \begin{bmatrix} .6 & -.2 \\ -.2 & .1 \end{bmatrix} \quad \mathbf{Z}'\mathbf{y} = \begin{bmatrix} 25 \\ 70 \end{bmatrix}$$

Consequently,

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y} = \begin{bmatrix} .6 & -.2 \\ -.2 & .1 \end{bmatrix} \begin{bmatrix} 25 \\ 70 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

and the fitted equation is

$$\hat{y} = 1 + 2z$$

Vector of Fitted (Predicted Values) and Residual Sum of Squares

The vector of fitted (predicted) values is

$$\hat{\mathbf{y}} = \mathbf{Z}\hat{\boldsymbol{\beta}} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \\ 5 \\ 7 \\ 9 \end{bmatrix}$$

so

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \hat{\mathbf{y}} = \begin{bmatrix} 1 \\ 4 \\ 3 \\ 8 \\ 9 \end{bmatrix} - \begin{bmatrix} 1 \\ 3 \\ 5 \\ 7 \\ 9 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ -2 \\ 1 \\ 0 \end{bmatrix}$$

The residual sum of squares is

$$\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = \begin{bmatrix} 0 & 1 & -2 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ -2 \\ 1 \\ 0 \end{bmatrix} = 0^2 + 1^2 + (-2)^2 + 1^2 + 0^2 = 6 \quad \blacksquare$$

Sum-of-Squares Decomposition

Sum-of-Squares Decomposition

According to Result 7.1, $\hat{\mathbf{y}}'\hat{\boldsymbol{\varepsilon}} = 0$, so the total response sum of squares $\mathbf{y}'\mathbf{y} = \sum_{j=1}^{n} y_j^2$ satisfies

$$\mathbf{y}'\mathbf{y} = (\hat{\mathbf{y}} + \mathbf{y} - \hat{\mathbf{y}})'(\hat{\mathbf{y}} + \mathbf{y} - \hat{\mathbf{y}}) = (\hat{\mathbf{y}} + \hat{\boldsymbol{\varepsilon}})'(\hat{\mathbf{y}} + \hat{\boldsymbol{\varepsilon}}) = \hat{\mathbf{y}}'\hat{\mathbf{y}} + \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} \qquad (7\text{-}7)$$

Since the first column of $\mathbf{Z}$ is $\mathbf{1}$, the condition $\mathbf{Z}'\hat{\boldsymbol{\varepsilon}} = \mathbf{0}$ includes the requirement $0 = \mathbf{1}'\hat{\boldsymbol{\varepsilon}} = \sum_{j=1}^{n} \hat{\varepsilon}_j = \sum_{j=1}^{n} y_j - \sum_{j=1}^{n} \hat{y}_j$, or $\bar{y} = \bar{\hat{y}}$. Subtracting $n\bar{y}^2 = n(\bar{\hat{y}})^2$ from both sides of the decomposition in (7-7), we obtain the basic decomposition of the sum of squares about the mean:

$$\mathbf{y}'\mathbf{y} - n\bar{y}^2 = \hat{\mathbf{y}}'\hat{\mathbf{y}} - n(\bar{\hat{y}})^2 + \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$$

or

$$\sum_{j=1}^{n} (y_j - \bar{y})^2 = \sum_{j=1}^{n} (\hat{y}_j - \bar{y})^2 + \sum_{j=1}^{n} \hat{\varepsilon}_j^2 \qquad (7\text{-}8)$$

$$\begin{pmatrix} \text{total sum} \\ \text{of squares} \\ \text{about mean} \end{pmatrix} = \begin{pmatrix} \text{regression} \\ \text{sum of} \\ \text{squares} \end{pmatrix} + \begin{pmatrix} \text{residual (error)} \\ \text{sum of squares} \end{pmatrix}$$

Multiple Correlation Coefficient or R^2: Proportion of Total Variation Explained

$$\sum_{j=1}^{n} (y_j - \bar{y})^2 = \sum_{j=1}^{n} (\hat{y}_j - \bar{y})^2 + \sum_{j=1}^{n} \hat{\varepsilon}_j^2 \tag{7-8}$$

$$\begin{pmatrix} \text{total sum} \\ \text{of squares} \\ \text{about mean} \end{pmatrix} = \begin{pmatrix} \text{regression} \\ \text{sum of} \\ \text{squares} \end{pmatrix} + \begin{pmatrix} \text{residual (error)} \\ \text{sum of squares} \end{pmatrix}$$

The preceding sum of squares decomposition suggests that the quality of the models fit can be measured by the *coefficient of determination*

$$R^2 = 1 - \frac{\sum_{j=1}^{n} \hat{\varepsilon}_j^2}{\sum_{j=1}^{n} (y_j - \bar{y})^2} = \frac{\sum_{j=1}^{n} (\hat{y}_j - \bar{y})^2}{\sum_{j=1}^{n} (y_j - \bar{y})^2} \tag{7-9}$$

The quantity R^2 gives the proportion of the total variation in the y_j's "explained" by, or attributable to, the predictor variables $z_1, z_2, \ldots, z_r$. Here R^2 (or the *multiple correlation coefficient* $R = +\sqrt{R^2}$) equals 1 if the fitted equation passes through all the data points, so that $\hat{\varepsilon}_j = 0$ for all j. At the other extreme, R^2 is 0 if $\hat{\beta}_0 = \bar{y}$ and $\hat{\beta}_1 = \hat{\beta}_2 = \cdots = \hat{\beta}_r = 0$. In this case, the predictor variables $z_1, z_2, \ldots, z_r$ have no influence on the response.

Geometry Least Squares: E(Y) Linear Combination of Z. As Beta varies, ZB Spans Model Plane of All Linear Combinations

Geometry of Least Squares

A geometrical interpretation of the least squares technique highlights the nature of the concept. According to the classical linear regression model,

$$\text{Mean response vector} = E(\mathbf{Y}) = \mathbf{Z}\boldsymbol{\beta} = \beta_0 \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} + \beta_1 \begin{bmatrix} z_{11} \\ z_{21} \\ \vdots \\ z_{n1} \end{bmatrix} + \cdots + \beta_r \begin{bmatrix} z_{1r} \\ z_{2r} \\ \vdots \\ z_{nr} \end{bmatrix}$$

Thus, $E(\mathbf{Y})$ is a linear combination of the columns of $\mathbf{Z}$. As $\boldsymbol{\beta}$ varies, $\mathbf{Z}\boldsymbol{\beta}$ spans the model plane of all linear combinations. Usually, the observation vector $\mathbf{y}$ will not lie in the model plane, because of the random error $\boldsymbol{\epsilon}$; that is, $\mathbf{y}$ is not (exactly) a linear combination of the columns of $\mathbf{Z}$. Recall that

$$\underset{\left(\begin{array}{c}\text{response}\\ \text{vector}\end{array}\right)}{\mathbf{Y}} = \underset{\left(\begin{array}{c}\text{vector}\\ \text{in model}\\ \text{plane}\end{array}\right)}{\mathbf{Z}\boldsymbol{\beta}} + \underset{\left(\begin{array}{c}\text{error}\\ \text{vector}\end{array}\right)}{\boldsymbol{\epsilon}}$$

Least Squares Projection n = 3, r =1

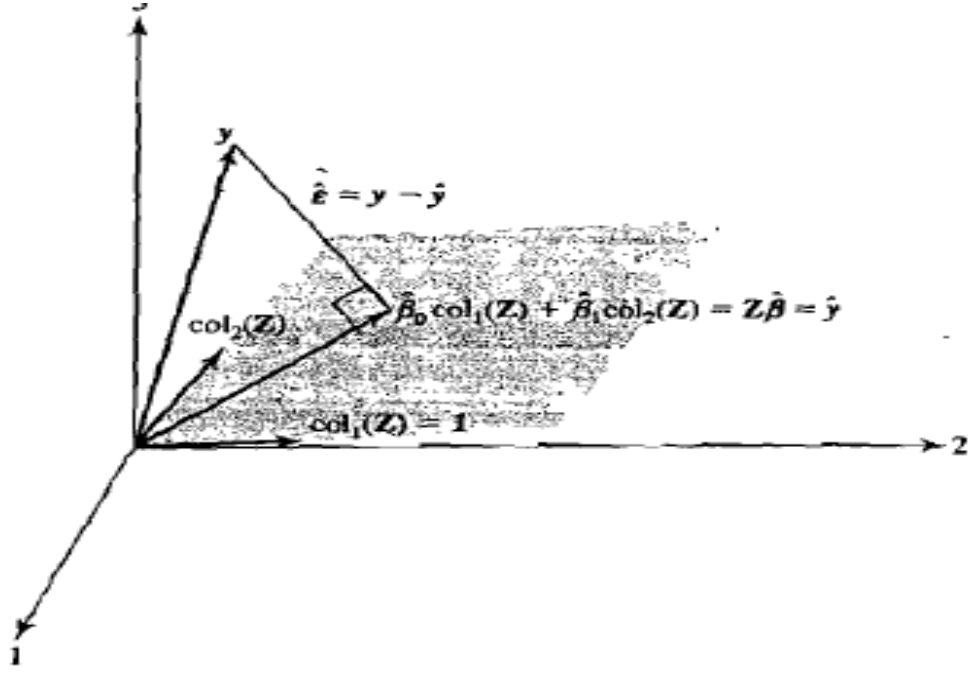


Figure 7.1 Least squares as a projection for $n = 3, r = 1$.

Once the observations become available, the least squares solution is derived from the deviation vector

$$\mathbf{y} - \mathbf{Zb} = (\text{observation vector}) - (\text{vector in model plane})$$

The squared length $(\mathbf{y} - \mathbf{Zb})'(\mathbf{y} - \mathbf{Zb})$ is the sum of squares $S(\mathbf{b})$. As illustrated in Figure 7.1, $S(\mathbf{b})$ is as small as possible when $\mathbf{b}$ is selected such that $\mathbf{Zb}$ is the point in the model plane closest to $\mathbf{y}$. This point occurs at the tip of the perpendicular projection of $\mathbf{y}$ on the plane. That is, for the choice $\mathbf{b} = \hat{\boldsymbol{\beta}}$, $\hat{\mathbf{y}} = \mathbf{Z}\hat{\boldsymbol{\beta}}$ is the projection of $\mathbf{y}$ on the plane consisting of all linear combinations of the columns of $\mathbf{Z}$. The residual vector $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \hat{\mathbf{y}}$ is perpendicular to that plane. This geometry holds even when $\mathbf{Z}$ is not of full rank.

Spectral Decomposition: Eigenvalues of Z'Z and e_1, e_2 etc. Are Corresponding Eigenvectors

When $\mathbf{Z}$ has full rank, the projection operation is expressed analytically as multiplication by the matrix $\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$. To see this, we use the spectral decomposition (2-16) to write

$$\mathbf{Z}'\mathbf{Z} = \lambda_1\mathbf{e}_1\mathbf{e}_1' + \lambda_2\mathbf{e}_2\mathbf{e}_2' + \cdots + \lambda_{r+1}\mathbf{e}_{r+1}\mathbf{e}_{r+1}'$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{r+1} > 0$ are the eigenvalues of $\mathbf{Z}'\mathbf{Z}$ and $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{r+1}$ are the corresponding eigenvectors. If $\mathbf{Z}$ is of full rank,

$$(\mathbf{Z}'\mathbf{Z})^{-1} = \frac{1}{\lambda_1}\mathbf{e}_1\mathbf{e}_1' + \frac{1}{\lambda_2}\mathbf{e}_2\mathbf{e}_2' + \cdots + \frac{1}{\lambda_{r+1}}\mathbf{e}_{r+1}\mathbf{e}_{r+1}'$$

Consider $\mathbf{q}_i = \lambda_i^{-1/2}\mathbf{Z}\mathbf{e}_i$, which is a linear combination of the columns of $\mathbf{Z}$. Then $\mathbf{q}_i'\mathbf{q}_k = \lambda_i^{-1/2}\lambda_k^{-1/2}\mathbf{e}_i'\mathbf{Z}'\mathbf{Z}\mathbf{e}_k = \lambda_i^{-1/2}\lambda_k^{-1/2}\mathbf{e}_i'\lambda_k\mathbf{e}_k = 0$ if $i \neq k$ or 1 if $i = k$. That is, the $r + 1$ vectors $\mathbf{q}_i$ are mutually perpendicular and have unit length. Their linear combinations span the space of all linear combinations of the columns of $\mathbf{Z}$. Moreover,

$$\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}' = \sum_{i=1}^{r+1}\lambda_i^{-1}\mathbf{Z}\mathbf{e}_i\mathbf{e}_i'\mathbf{Z}' = \sum_{i=1}^{r+1}\mathbf{q}_i\mathbf{q}_i'$$

Linear Combination of Columns of Z with r+1 Vectors q Mutually Perpendicular and Have Unit Length

Consider $\mathbf{q}_i = \lambda_i^{-1/2}\mathbf{Z}\mathbf{e}_i$, which is a linear combination of the columns of $\mathbf{Z}$. Then $\mathbf{q}_i'\mathbf{q}_k = \lambda_i^{-1/2}\lambda_k^{-1/2}\mathbf{e}_i'\mathbf{Z}'\mathbf{Z}\mathbf{e}_k = \lambda_i^{-1/2}\lambda_k^{-1/2}\mathbf{e}_i'\lambda_k\mathbf{e}_k = 0$ if $i \neq k$ or 1 if $i = k$. That is, the $r + 1$ vectors $\mathbf{q}_i$ are mutually perpendicular and have unit length. Their linear combinations span the space of all linear combinations of the columns of $\mathbf{Z}$. Moreover,

$$\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}' = \sum_{i=1}^{r+1} \lambda_i^{-1}\mathbf{Z}\mathbf{e}_i\mathbf{e}_i'\mathbf{Z}' = \sum_{i=1}^{r+1} \mathbf{q}_i\mathbf{q}_i'$$

According to Result 2A.2 and Definition 2A.12, the projection of $\mathbf{y}$ on a linear combination of $\{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_{r+1}\}$ is $\sum_{i=1}^{r+1} (\mathbf{q}_i'\mathbf{y})\mathbf{q}_i = \left(\sum_{i=1}^{r+1} \mathbf{q}_i\mathbf{q}_i'\right)\mathbf{y} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y} = \mathbf{Z}\hat{\boldsymbol{\beta}}$.

Thus, multiplication by $\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$ projects a vector onto the space spanned by the columns of $\mathbf{Z}$.[2]

Similarly, $[\mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}']$ is the matrix for the projection of $\mathbf{y}$ on the plane perpendicular to the plane spanned by the columns of $\mathbf{Z}$.

Sampling Properties Classical Least Squares Estimators

Sampling Properties of Classical Least Squares Estimators

The least squares estimator $\hat{\boldsymbol{\beta}}$ and the residuals $\hat{\boldsymbol{\varepsilon}}$ have the sampling properties detailed in the next result.

Result 7.2. Under the general linear regression model in (7-3), the least squares estimator $\hat{\boldsymbol{\beta}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}$ has

$$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta} \quad \text{and} \quad \text{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{Z}'\mathbf{Z})^{-1}$$

The residuals $\hat{\boldsymbol{\varepsilon}}$ have the properties

$$E(\hat{\boldsymbol{\varepsilon}}) = \mathbf{0} \quad \text{and} \quad \text{Cov}(\hat{\boldsymbol{\varepsilon}}) = \sigma^2[\mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'] = \sigma^2[\mathbf{I} - \mathbf{H}]$$

Also, $E(\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}) = (n - r - 1)\sigma^2$, so defining

$$s^2 = \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n - (r + 1)} = \frac{\mathbf{Y}'[\mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}']\mathbf{Y}}{n - r - 1} = \frac{\mathbf{Y}'[\mathbf{I} - \mathbf{H}]\mathbf{Y}}{n - r - 1}$$

we have

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

Moreover, $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\varepsilon}}$ are uncorrelated.

Proof. (See webpage: www.prenhall.com/statistics) ■

Review: Least Squares Regression

Least Squares Regression Extends Beyond Straight Line Regression Model

3. The Method of Least Squares

Let us tentatively assume that the preceding formulation of the straight line model is correct. We can then proceed to estimate the regression line and solve a few related inference problems. The problem of estimating the regression parameters β_0 and β_1 can be viewed as fitting the best straight line to the y to x relationship exhibited in the scatter diagram.

One can draw a line by eyeballing the scatter diagram, but such a judgment may be open to dispute. Moreover, statistical inferences cannot be based on a line that is estimated subjectively. On the other hand, the **method of least squares** is an objective and efficient method of determining the best fitting straight line. Moreover, this method is quite versatile because its application extends beyond the simple **straight line regression model**.

Deviations of the Observations from the Line of $y = b_o + b_1 x$ known as d_1

One can draw a line by eyeballing the scatter diagram, but such a judgment may be open to dispute. Moreover, statistical inferences cannot be based on a line that is estimated subjectively. On the other hand, the **method of least squares** is an objective and efficient method of determining the best fitting straight line. Moreover, this method is quite versatile because its application extends beyond the simple **straight line regression model**.

Suppose that an arbitrary line $y = b_0 + b_1 x$ is drawn on the scatter diagram as it is in Figure 4. In contrast to the Greek letters that locate the true line, we use the Latin letters b_0 and b_1 to denote the intercept and slope of an arbitrary, or trial, line. At the value x_i of the independent variable, the y value predicted by this line is $b_0 + b_1 x_i$ whereas the observed value is y_i. The discrepancy between the observed and predicted y values is then $y_i - b_0 - b_1 x_i = d_i$, which is the **vertical** distance of the point from the line.

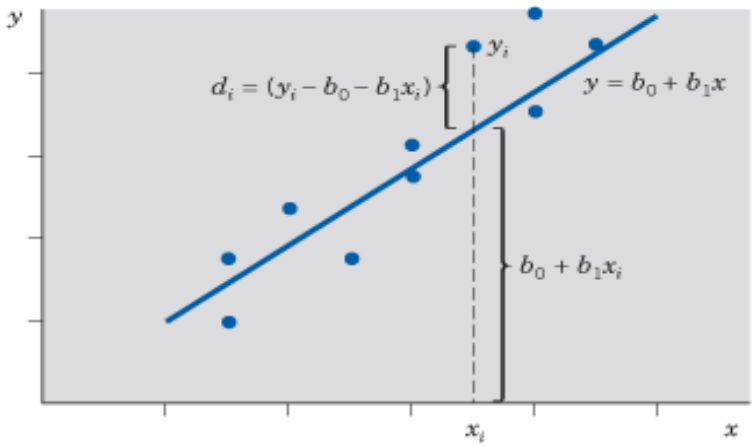


Figure 4 Deviations of the observations from a line $y = b_0 + b_1 x$.

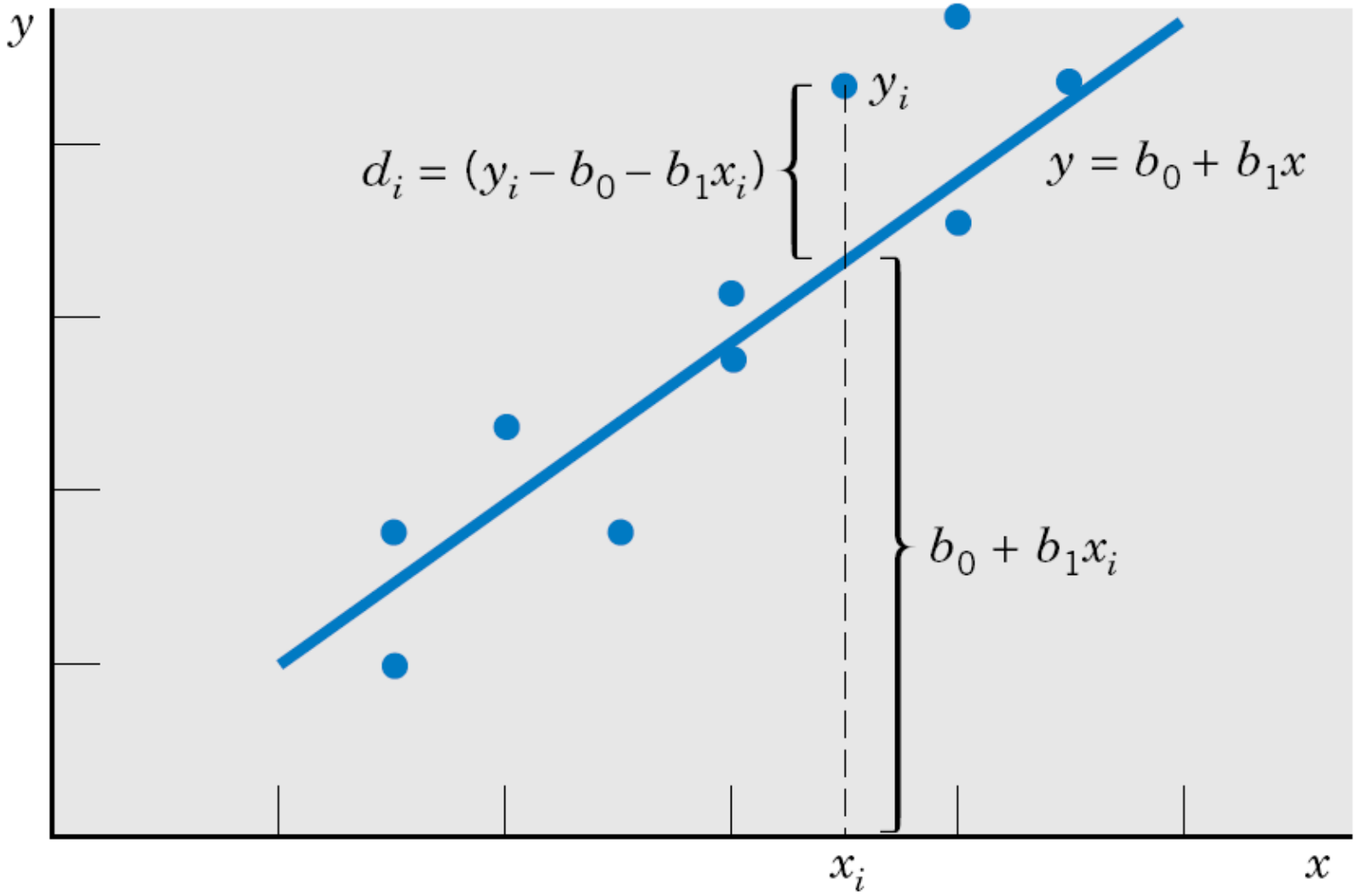


Figure 4 (p. 458)

Deviations of the observations from a line $y = b_0 + b_1 x$.

The Principle of Least Squares: D = (Observed Response – Predicted Response)^2

We square each discrepancy and then take their sum

$$D = \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

as an overall measure of the discrepancy of the observed points from the trial line $y = b_0 + b_1 x$. The magnitude of D obviously depends on the line that is drawn. In other words, it depends on the parameters b_0 and b_1, the two quantities that determine the trial line. A good fit will make D as small as possible. We now state the principle of least squares in general terms to indicate its usefulness to fitting many other models.

The Principle of Least Squares

Determine the values for the parameters so that the overall discrepancy

$$D = \sum (\text{Observed response} - \text{Predicted response})^2$$

is minimized.

The parameter values thus determined are called the **least squares estimates.**

The Principle of Least Squares

Determine the values for the parameters so that the overall discrepancy

$$D = \sum (\text{Observed response} - \text{Predicted response})^2$$

is minimized.

The parameter values thus determined are called the **least squares estimates.**

Box on Page 458

The Principle of Least Squares

Principle of Least Squares Corresponds to the Best Fitting Straight Line

The Principle of Least Squares

Determine the values for the parameters so that the overall discrepancy

$$D = \sum (\text{Observed response} - \text{Predicted response})^2$$

is minimized.

The parameter values thus determined are called the **least squares estimates.**

For the straight line model, the least squares principle involves the determination of b_0 and b_1 to minimize.

$$D = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

The particular values b_0 and b_1 that minimize the sum of squares are denoted by $\hat{\beta}_0$ and $\hat{\beta}_1$, respectively. The hat notation, or ^, over a parameter indicates that this random variable is an estimate of the corresponding parameter. They are called the **least squares estimates** of the regression parameters β_0 and β_1. The **best fitting straight line** or **best fitting regression line** is then given by the equation

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

where the hat over y indicates that it is an estimated quantity obtained by the least squares method.

Five (5) Summary Statistics for Calculating Regression Line

To describe the formulas for the least squares estimators, we first introduce some basic notation.

Basic Notation

$$\bar{x} = \frac{1}{n}\sum x \qquad \bar{y} = \frac{1}{n}\sum y$$

$$S_{xx} = \sum (x-\bar{x})^2 = \sum x^2 - \frac{\left(\sum x\right)^2}{n}$$

$$S_{yy} = \sum (y-\bar{y})^2 = \sum y^2 - \frac{\left(\sum y\right)^2}{n}$$

$$S_{xy} = \sum (x-\bar{x})\ (y-\bar{y}) = \sum xy - \frac{\left(\sum x\right)\left(\sum y\right)}{n}$$

The quantities $\bar{x}$ and $\bar{y}$ are the sample means of the x and y values, S_{xx} and S_{yy} are the sums of squared deviations from the means, and S_{xy} is the sum of the cross products of deviations. These five summary statistics are the key ingredients for calculating the least squares estimates and handling the inference problems associated with the linear regression model. (The reader may review Sections 4 and 5 of Chapter 3 where calculations of these statistics were illustrated.)

Least Squares Estimators:

The quantities $\bar{x}$ and $\bar{y}$ are the sample means of the x and y values, S_{xx} and S_{yy} are the sums of squared deviations from the means, and S_{xy} is the sum of the cross products of deviations. These five summary statistics are the key ingredients for calculating the least squares estimates and handling the inference problems associated with the linear regression model. (The reader may review Sections 4 and 5 of Chapter 3 where calculations of these statistics were illustrated.)

The formulas for the **least squares estimators** are

Least squares estimator of β_0

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

Least squares estimator of β_1

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

Basic Notation

$$\bar{x} = \frac{1}{n}\sum x \qquad \bar{y} = \frac{1}{n}\sum y$$

$$S_{xx} = \sum (x - \bar{x})^2 = \sum x^2 - \frac{\left(\sum x\right)^2}{n}$$

$$S_{yy} = \sum (y - \bar{y})^2 = \sum y^2 - \frac{\left(\sum y\right)^2}{n}$$

$$S_{xy} = \sum (x - \bar{x})(y - \bar{y}) = \sum xy - \frac{\left(\sum x\right)\left(\sum y\right)}{n}$$

Box on Page 459

Basic Notation

Reminder: Calculation of S_{xy} and S_{xx}

Example 5

Calculation of Sample Correlation

Calculate r for the $n = 4$ pairs of observations

$$(2, 5) \quad (1, 3) \quad (5, 6) \quad (0, 2)$$

SOLUTION

We first determine the mean $\overline{x}$ and deviations $x - \overline{x}$ and then $\overline{y}$ and the deviations $y - \overline{y}$. See Table 8.

TABLE 8 Calculation of r

	x	y	$x-\overline{x}$	$y-\overline{y}$	$(x-\overline{x})^2$	$(y-\overline{y})^2$	$(x-\overline{x})(y-\overline{y})$
	2	5	0	1	0	1	0
	1	3	−1	−1	1	1	1
	5	6	3	2	9	4	6
	0	2	−2	−2	4	4	4
Total	8	16	0	0	14	10	11
	$\overline{x} = 2$	$\overline{y} = 4$			S_{xx}	S_{yy}	S_{xy}

Equation of the Line Fitted by Least Squares

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

where

$$\text{Slope } \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$$

$$\text{Intercept } \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

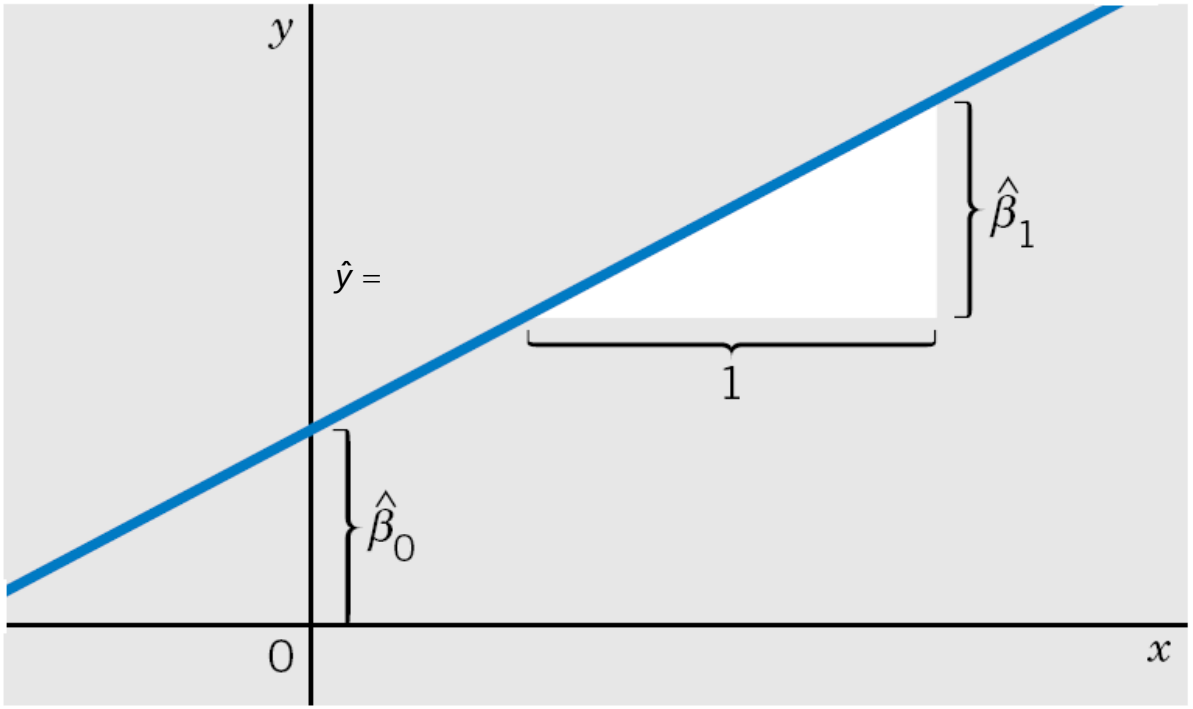


Figure 10, The line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \chi$, p. 108

Least squares estimator of β_0

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

Least squares estimator of β_1

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

Box 2 on Page 459

Best Fit Regression Line: Sum of Squares of the Deviations is the Smallest

The estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ can then be used to locate the best fitting line:

Fitted (or estimated) regression line

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

As we have already explained, this line provides the best fit to the data in the sense that the sum of squares of the deviations, or

$$\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2$$

is the smallest.

$E_i = y_i - B_o - B_1x_1$
Sum of the Residuals or Error Always Zero

The individual deviations of the observations y_i from the fitted values $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ are called the **residuals**, and we denote these by $\hat{e}_i$.

Residuals

$$\hat{e}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \quad i = 1, \ldots, n$$

Some residuals are positive and some negative, and a property of the least squares fit is that the **sum of the residuals is always zero**.

Residual Sum of Squares Due to Error (SSE = Sum of Each Error Squared)

In Chapter 12, we discuss how the residuals can be used to check the assumptions of a regression model. For now, the sum of squares of the residuals is a quantity of interest because it leads to an estimate of the variance σ^2 of the error distributions illustrated in Figure 3. The **residual sum of squares** is also called the **sum of squares due to error** and is abbreviated as SSE.

The **residual sum of squares** or the **sum of squares due to error** is

$$\text{SSE} = \sum_{i=1}^{n} \hat{e}_i^2 = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

The second expression for SSE, which follows after some algebraic manipulations (see Exercise 11.24), is handy for directly calculating SSE. However, we stress the importance of determining the individual residuals for their role in model checking (see Section 3, Chapter 12).

Fitted (or estimated) regression line

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Residuals

$$\hat{e}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \qquad i = 1, \ldots, n$$

The residual sum of squares or the sum of squares due to error is

$$\text{SSE} = \sum_{i=1}^{n} \hat{e}_i^2 = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

Boxes on Page 460

Error Variance or σ^2 = SSE/n-2

An **estimate of variance** σ^2 is obtained by dividing SSE by $n-2$. The reduction by 2 is because two degrees of freedom are lost from estimating the two parameters β_0 and β_1.

Estimate of Variance

The estimator of the error variance σ^2 is

$$s^2 = \frac{\text{SSE}}{n-2}$$

Estimate of Variance

The estimator of the error variance σ^2 is

$$S^2 = \frac{\text{SSE}}{n - 2}$$

Box on Page 461

Estimate of Variance

Regression Table Computation

In applying the least squares method to a given data set, we first compute the basic quantities $\overline{x}$, $\overline{y}$, S_{xx}, S_{yy}, and S_{xy}. Then the preceding formulas can be used to obtain the least squares regression line, the residuals, and the value of SSE. Computations for the data given in Table 1 are illustrated in Table 3.

TABLE 3 Computations for the Least Squares Line, SSE, and Residuals Using the Data of Table 1

x	y	x^2	y^2	xy	$\hat{\beta}_0 + \hat{\beta}_1 x$	Residual $\hat{e}$	
3	9	9	81	27	7.15	1.85	
3	5	9	25	15	7.15	−2.15	
4	12	16	144	48	9.89	2.11	
5	9	25	81	45	12.63	−3.63	
6	14	36	196	84	15.37	−1.37	
6	16	36	256	96	15.37	.63	
7	22	49	484	154	18.11	3.89	
8	18	64	324	144	20.85	−2.85	
8	24	64	576	192	20.85	3.15	
9	22	81	484	198	23.59	−1.59	
Total	59	151	389	2651	1003	.04	(rounding error)

$\overline{x} = 5.9, \quad \overline{y} = 15.1$

$\hat{\beta}_1 = \frac{112.1}{40.9} = 2.74$

Regression Table: Least Squares, Sum of Squares Errors and Residuals

	x	y	x^2	y^2	xy	$\beta_0 + \beta_1 x$	Residual $\hat{e}$	
	3	9	9	81	27	7.15	1.85	
	3	5	9	25	15	7.15	−2.15	
	4	12	16	144	48	9.89	2.11	
	5	9	25	81	45	12.63	−3.63	
	6	14	36	196	84	15.37	−1.37	
	6	16	36	256	96	15.37	.63	
	7	22	49	484	154	18.11	3.89	
	8	18	64	324	144	20.85	−2.85	
	8	24	64	576	192	20.85	3.15	
	9	22	81	484	198	23.59	−1.59	
Total	59	151	389	2651	1003		.04	(rounding error)

$\overline{x} = 5.9, \quad \overline{y} = 15.1$

$$S_{xx} = 389 - \frac{(59)^2}{10} = 40.9$$

$$S_{yy} = 2651 - \frac{(151)^2}{10} = 370.9$$

$$S_{xy} = 1003 - \frac{59 \times 151}{10} = 112.1$$

$$\hat{\beta}_1 = \frac{112.1}{40.9} = 2.74$$

$$\hat{\beta}_0 = 15.1 - 2.74 \times 5.9 = -1.07$$

$$\text{SSE} = 370.9 - \frac{(112.1)^2}{40.9} = 63.6528$$

TABLE 3 Computations for the Least Squares Line, SSE, and Residuals Using the Data of Table 1

x	y	x^2	y^2	xy	$\hat{\beta}_0 + \hat{\beta}_1 x$	Residual $\hat{e}$
3	9	9	81	27	7.15	1.85
3	5	9	25	15	7.15	−2.15
4	12	16	144	48	9.89	2.11
5	9	25	81	45	12.63	−3.63
6	14	36	196	84	15.37	−1.37
6	16	36	256	96	15.37	.63
7	22	49	484	154	18.11	3.89
8	18	64	324	144	20.85	−2.85
8	24	64	576	192	20.85	3.15
9	22	81	484	198	23.59	−1.59
Total	59	151	389	2651	1003	.04 (rounding error)

$$\bar{x} = 5.9, \quad \bar{y} = 15.1$$

$$S_{xx} = 389 - \frac{(59)^2}{10} = 40.9$$

$$S_{yy} = 2651 - \frac{(151)^2}{10} = 370.9$$

$$S_{xy} = 1003 - \frac{59 \times 151}{10} = 112.1$$

$$\hat{\beta}_1 = \frac{112.1}{40.9} = 2.74$$

$$\hat{\beta}_0 = 15.1 - 2.74 \times 5.9 = -1.07$$

$$\text{SSE} = 370.9 - \frac{(112.1)^2}{40.9} = 63.6528$$

Table 11.3 (p. 461)

Computation for the Least Squares Line, SSE, and Residuals Using the Data of Table 1

Equation of the Fitted Line Both in Notation and Graphical Format

The data in the first two columns yield the next three columns. Then, the sum of entries in a column are obtained so $\overline{x}$, $\overline{y}$, S_{xx}, S_{yy}, and S_{xy} can be calculated. From these, $\hat{\beta}_0$, $\hat{\beta}_1$, and SSE are obtained.

The equation of the line fitted by the least squares method is then

$$\hat{y} = -1.07 + 2.74x$$

Figure 5 shows a plot of the data along with the fitted regression line.

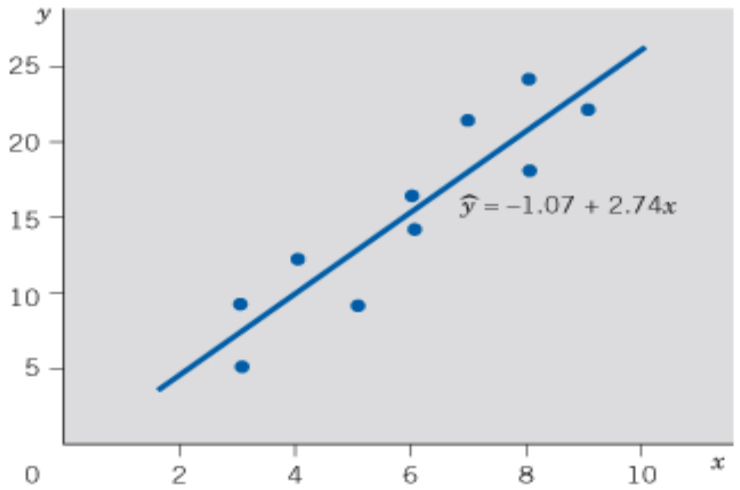


Figure 5 The least squares regression line for the data given in Table 1.

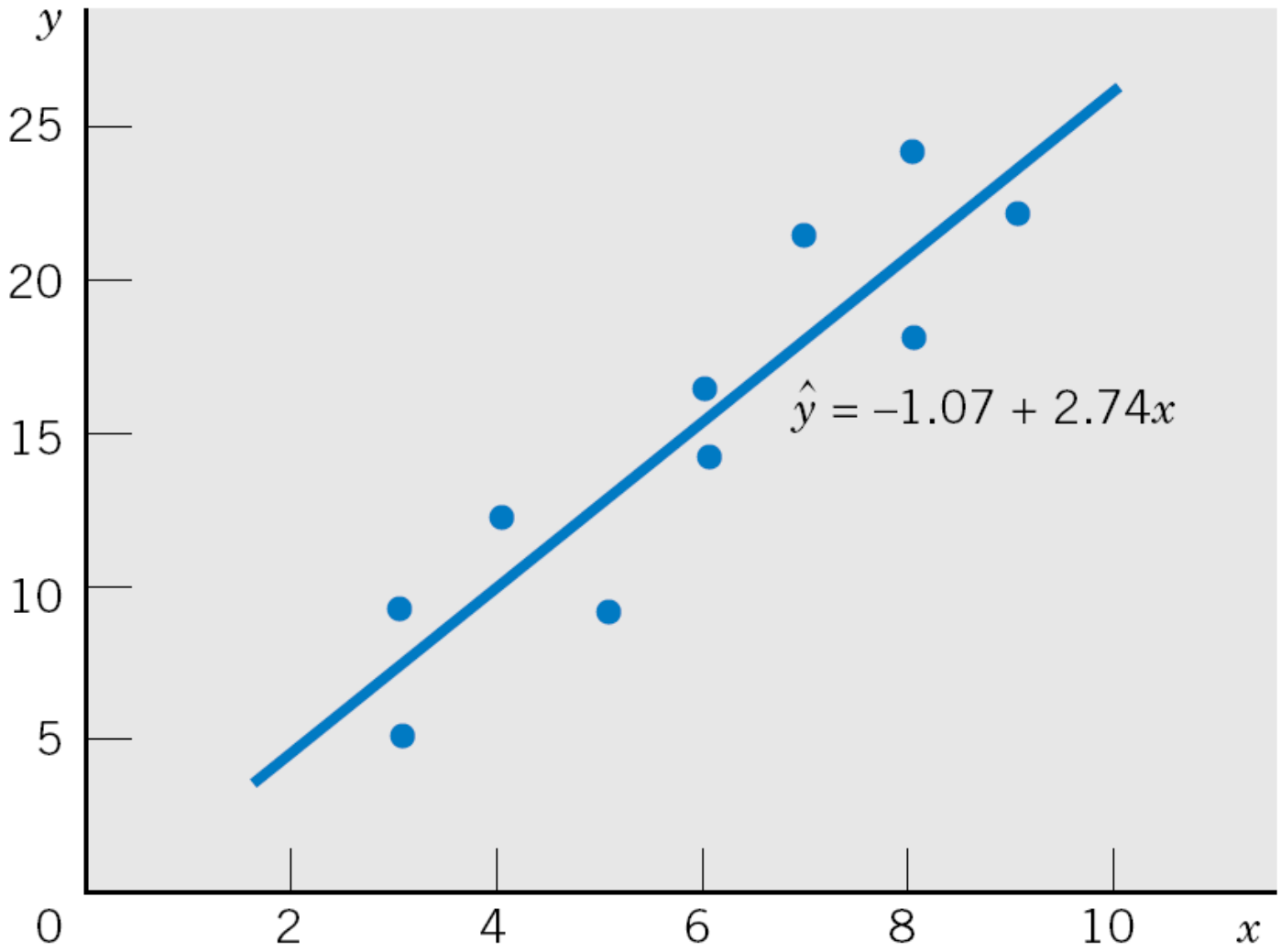


Figure 5 (p. 462)

The least squares regression line for the data given in Table 1.

Sum of Squares of the Residuals or Error Calculation

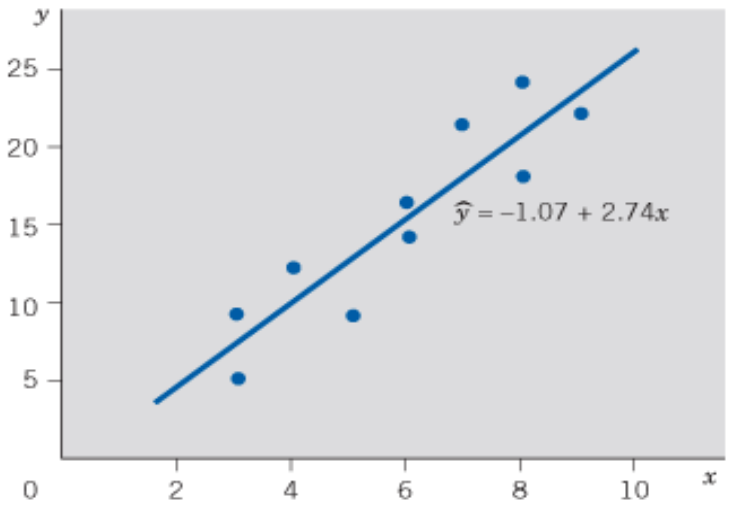


Figure 5 The least squares regression line for the data given in Table 1.

The residuals $\hat{e}_i = y_i - \hat{y}_i = y_i + 1.07 - 2.74\,x_i$ are computed in the last column of Table 3. The sum of squares of the residuals is

$$\sum_{i=1}^{n} \hat{e}_i^2 = (1.85)^2 + (-2.15)^2 + (2.11)^2 + \cdots + (-1.59)^2$$
$$= 63.653$$

which agrees with our previous calculations of SSE, except for the error due to rounding. Theoretically, the sum of the residuals should be zero, and the difference between the sum .04 and zero is also due to rounding.

Variance Calculation for the Regression Equation or σ^2 SSE/n-2

The estimate of the variance σ^2 is

$$s^2 = \frac{\text{SSE}}{n-2} = \frac{63.6528}{8} = 7.96$$

The calculations involved in a regression analysis become increasingly tedious with larger data sets. Access to a computer proves to be a considerable advantage. Table 4 illustrates a part of the computer-based analysis of linear regression using the data of Example 4 and the MINITAB package. For a more complete regression analysis, see Table 5.

TABLE 4 Regression Analysis of the Data in Table 1, Example 4, Using MINITAB

Data: C11T3
C1: 3 3 4 5 6 6 7 8 8 9
C2: 9 5 12 9 14 16 22 18 24 22
Dialog box:
Stat > **Regression** > **Regression**
Type *C2* in **Response**
Type *C1* in **Predictors**. Click **OK**.
Output:
Regression Analysis
The regression equation is
$y = -1.07 + 2.74x$

Using Weight and Girth (Width) at the Heart to Predict Body Weight of Polar Bears

Example 5

Predicting the Weight of Polar Bears from Body Measurements

Polar bears are among the world's largest land carnivores and males are typically larger than females. The U.S. Geological Survey monitors the health of the population. Monitoring procedures include making measurements on weight and girth at the heart, when the polar bear is sedated.
The measurements on $x = \text{girth}\ \ (\text{cm})$ and $y = \text{weight}\ \ (\text{lb})$ for $n = 30$ adult male bears, given in Table D.13 of the Data Bank, are summarized by

$$n = 30 \qquad \bar{x} = 149.737 \qquad \bar{y} = 837.533$$

$$\sum (x-\bar{x})^2 = 4367.99 \qquad \sum (x-\bar{x})\ (y-\bar{y}) = 41915.5 \qquad \sum (y-\bar{y})^2 = 492089$$

(a) Obtain the equation of the fitted regression line.

(b) Use the least squares line to obtain a point estimate of the weight of a bear which has girth 165 cm.

(c) Specify two variables that likely contribute the error term e in the model.

(d) Calculate the residual sum of squares and estimate σ^2.

Predicting Body Weight of Polar Bears Using Regression Analysis

SOLUTION

(a) We have

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{41915.5}{4367.99} = 9.596$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 837.533 - 9.596 \times 149.737 = -599.3$$

So the fitted regression line is $-599.3 + 9.596x$.

(b) According to the straight line regression model, the weight of the new polar bear with girth 165 cm is given by $\beta_0 + \beta_1 (165) + e^*$ where the error e^* has mean 0. Consequently, the expected value of the weight is $\beta_0 + \beta_1 (165)$ and its least squares estimate

$$\hat{\beta}_0 + \hat{\beta}_1 (165) = -599.3 + 9.596 (165) = 984.0 \text{ pounds}$$

is the predicted weight.

Residual Sum of Squares and Variance or σ^2 re: Body Weight of Polar Bears

$$\hat{\beta}_0 + \hat{\beta}_1 (165) = -599.3 + 9.596 (165) = 984.0 \text{ pounds}$$

is the predicted weight.

A good prediction equation permits the investigator to sometimes forgo weighing the polar bear in difficult circumstances when a large scale is not available.

(c) Any variable not specifically in the regression equation must be, by default, included in the error term. For instance, age would likely influence the weight as would the abundance of food in the previous six months.

(d)

$$\text{SSE} = S_{yy} - \frac{S_{xy}^2}{S_{xx}} = 492089 - \frac{41915.5^2}{4367.99} = 89865$$

and

$$s^2 = \frac{\text{SSE}}{n-2} = \frac{89865}{28} = 3209$$

Starting in Section 5, we use the estimate of σ^2 to set confidence intervals and conduct tests of hypotheses concerning β_0, β_1 and predicted values.

Least Squares Estimator is Only A Predication Tool of True Regression Line

4. The Sampling Variability of the Least Squares Estimators—Tools for Inference

It is important to remember that the line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ obtained by the principle of least squares is an **estimate** of the unknown true regression line $y = \beta_0 + \beta_1 x$. In our drug evaluation problem (Example 4), the estimated line is

$$\hat{y} = -1.07 + 2.74\,x$$

Its slope $\hat{\beta}_1 = 2.74$ suggests that the mean duration of relief increases by 2.74 hours for each unit dosage of the drug. Also, if we were to estimate the expected duration of relief for a specified dosage $x^* = 4.5$ milligrams, we would naturally use the fitted regression line to calculate the estimate $-1.07 + 2.74 \times 4.5 = 11.26$ hours. A few questions concerning these estimates naturally arise at this point.

1. In light of the value 2.74 for $\hat{\beta}_1$, could the slope β_1 of the true regression line be as much as 4? Could it be zero so that the true regression line is $y = \beta_0$, which does not depend on x? What are the plausible values for β_1?
2. How much uncertainty should be attached to the estimated duration of 11.26 hours corresponding to the given dosage $x^* = 4.5$?

Standard Deviation or the Standard Errors of the Least Squares Estimators

To answer these and related questions, we must know something about the sampling distributions of the least squares estimators. These sampling distributions will enable us to test hypotheses and set confidence intervals for the parameters β_0 and β_1 that determine the straight line and for the straight line itself. Again, the t distribution is relevant.

1. The standard deviations (also called standard errors) of the least squares estimators are

$$\text{S.E.}\left(\hat{\beta}_1\right) = \frac{\sigma}{\sqrt{S_{xx}}} \quad \text{S.E.}\left(\hat{\beta}_0\right) = \sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$$

To estimate the standard error, use

$$S = \sqrt{\frac{\text{SSE}}{n-2}} \text{ in place of } \sigma$$

1. The standard deviations (also called standard errors) of the least squares estimators are

$$\text{S.E.}(\hat{\beta}_1) = \frac{\sigma}{\sqrt{S_{xx}}} \qquad \text{S.E.}(\hat{\beta}_0) = \sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$$

To estimate the standard error, use

$$S = \sqrt{\frac{\text{SSE}}{n-2}} \qquad \text{in place of } \sigma$$

Box on Page 467

Use t distribution for Inferences About the Slope or B_o

1. The standard deviations (also called standard errors) of the least squares estimators are

$$\text{S.E.}\left(\hat{\beta}_1\right) = \frac{\sigma}{\sqrt{S_{xx}}} \quad \text{S.E.}\left(\hat{\beta}_0\right) = \sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$$

To estimate the standard error, use

$$S = \sqrt{\frac{\text{SSE}}{n-2}} \text{ in place of } \sigma$$

2. **Inferences about the slope** β_1 are based on the t distribution

$$T = \frac{\hat{\beta}_1 - \beta_1}{S/\sqrt{S_{xx}}} \quad \text{d.f.} = n-2$$

Inferences about the intercept β_0 are based on the t distribution

$$T = \frac{\hat{\beta}_0 - \beta_0}{S\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}} \quad \text{d.f.} = n-2$$

2. **Inferences about the slope β_1** are based on the t distribution

$$T = \frac{\hat{\beta}_1 - \beta_1}{S/\sqrt{S_{xx}}} \qquad \text{d.f.} = n - 2$$

Inferences about the intercept β_0 are based on the t distribution

$$T = \frac{\hat{\beta}_0 - \beta_0}{S\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}} \qquad \text{d.f.} = n - 2$$

Box on Pages 467

Use t distribution for Inference About the Regression Equation

$$T = \frac{\hat{\beta}_0 - \beta_0}{S\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}} \quad \text{d.f.} = n - 2$$

3. At a specified value $x = x^*$, the **expected response** is $\beta_0 + \beta_1 x^*$. This is estimated by $\hat{\beta}_0 + \hat{\beta}_1 x^*$ with

Estimated standard error

$$S\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$

Inferences about $\beta_0 + \beta_1 x^*$ are based on the t distribution

$$T = \frac{\left(\hat{\beta}_0 + \hat{\beta}_1 x^*\right) - \left(\beta_0 + \beta_1 x^*\right)}{S\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}} \quad \text{d.f.} = n - 2$$

3. At a specified value $x = x^*$, the **expected response** is $\beta_0 + \beta_1 x^*$. This is estimated by $\hat{\beta}_0 + \hat{\beta}_1 x^*$ with

Estimated standard error

$$S\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$

Inferences about $\beta_0 + \beta_1 x^*$ are based on the t distribution

$$T = \frac{(\hat{\beta}_0 + \hat{\beta}_1 x^*) - (\beta_0 + \beta_1 x^*)}{S\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}} \qquad \text{d.f.} = n - 2$$

Box on Page 467

Hypothesis Testing of the Slope B_o: Whether Greater Than or Less Than Zero

5. Important Inference Problems

We are now prepared to test hypotheses, construct confidence intervals, and make predictions in the context of straight line regression.

5.1 INFERENCE CONCERNING THE SLOPE β_1

In a regression analysis problem, it is of special interest to determine whether the expected response does or does not vary with the magnitude of the input variable x. According to the linear regression model,

$$\text{Expected response} = \beta_0 + \beta_1 x$$

This does not change with a change in x if and only if $\beta_1 = 0$. We can therefore test the null hypothesis $H_0 : \beta_1 = 0$ against a one- or a two-sided alternative, depending on the nature of the relation that is anticipated. If we refer to the boxed statement (2) of Section 4, the null hypothesis $H_0 : \beta_1 = 0$ is to be tested using the test statistic

$$T = \frac{\hat{\beta}_1}{S/\sqrt{S_{xx}}} \quad \text{d.f.} = n - 2$$

Hypothesis Test: Slope $B_o > 0$

Example 6

A Test to Establish that Duration of Relief Increases with Dosage

Do the data given in Table 1 constitute strong evidence that the mean duration of relief increases with higher dosages of the drug?

SOLUTION

For an increasing relation, we must have $\beta_1 > 0$. Therefore, we are to test the null hypothesis $H_0: \beta_1 = 0$ versus the one-sided alternative $H_1: \beta_1 > 0$.

We select $\alpha = .05$. Since $t_{.05} = 1.860$, with d.f. $= 8$ we set the rejection region $R: T \geq 1.860$ as shown in Figure 6.

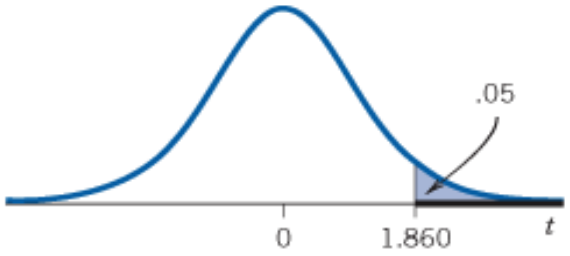


Figure 6 The rejection region $R: T \geq 1.860$.

Hypothesis Test: Slope $B_o > 0$

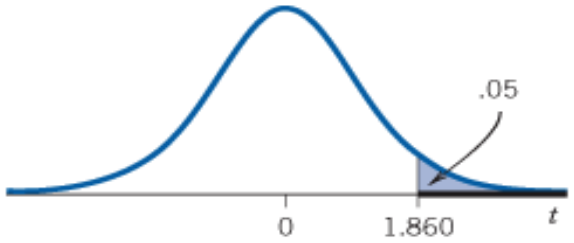


Figure 6 The rejection region
$R: T \geq 1.860$.

Using the calculations that follow Table 3, we have

$$\hat{\beta}_1 = 2.74$$

$$s^2 = \frac{\text{SSE}}{n-2} = \frac{63.6528}{8} = 7.9566,\ s = 2.8207$$

$$\text{Estimated S.E.}\left(\hat{\beta}_1\right) = \frac{s}{\sqrt{S_{xx}}} = \frac{2.8207}{\sqrt{40.90}} = .441$$

$$\text{Test statistic } t = \frac{2.74}{.441} = 6.213$$

The observed t value is in the rejection region, so H_0 is rejected. Moreover, 6.213 is much larger than $t_{.005} = 3.355$, so the P-value is much smaller than .005.

A computer calculation gives $P[T > 6.213] = .0001$. There is strong evidence that larger dosages of the drug tend to increase the duration of relief over the range covered in the study.

How to Interpret When Slope or B_1 = Zero—No Linear Relation? or Just Not Enough Variation in the x Variable

A warning is in order here concerning the interpretation of the test of $H_0 : \beta_1 = 0$. If H_0 is not rejected, we may be tempted to conclude that y does not depend on x. Such an unqualified statement may be erroneous. First, the absence of a linear relation has only been established over the range of the x values in the experiment. It may be that x was just not varied enough to influence y. Second, the interpretation of lack of dependence on x is valid only if our model formulation is correct. If the scatter diagram depicts a relation on a curve but we inadvertently formulate a straight line model and test $H_0 : \beta_1 = 0$, the conclusion that H_0 is not rejected should be interpreted to mean "no linear relation," rather than "no relation." We elaborate on this point further in Section 6. Our present viewpoint is to assume that the model is correctly formulated and discuss the various inference problems associated with it.

Test of the Null Hypothesis: Does Slope B_1 = Specified Value of Beta; Relies on t Distribution

More generally, we may test whether or not β_1 is equal to some specified value β_{10}, not necessarily zero.

The **test of the null hypothesis**

$$H_0 : \beta_1 = \beta_{10}$$

is based on

$$T = \frac{\hat{\beta}_1 - \beta_{10}}{S/\sqrt{S_{xx}}} \quad \text{d.f.} = n - 2$$

Confidence Interval for Parameter or Slope or Beta Coefficient B_1 Relies on t Distribution

In addition to testing hypotheses, we can provide a confidence interval for the parameter β_1 using the t distribution.

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

$$\left(\hat{\beta}_1 - t_{\alpha/2}\frac{S}{\sqrt{S_{xx}}},\ \hat{\beta}_1 + t_{\alpha/2}\frac{S}{\sqrt{S_{xx}}}\right)$$

where $t_{\alpha/2}$ is the upper $\alpha/2$ point of the t distribution with d.f. $= n-2$.

Example Confidence Interval for Parameter or Slope or Beta Coefficient B_1

Example 7

A Confidence Interval for β_1

Construct a 95% confidence interval for the slope of the regression line in reference to the drug trial data of Table 1.

SOLUTION

In Example 6, we found that $\hat{\beta}_1 = 2.74$ and $s/\sqrt{S_{xx}} = .441$. The required confidence interval is given by

$$2.74 \pm 2.306 \times .441 = 2.74 \pm 1.02 \quad \text{or} \quad (1.72, 3.76)$$

We are 95% confident that by adding one extra milligram to the dosage, the mean duration of relief would increase somewhere between 1.72 and 3.76 hours.

Example Confidence Interval for Y-intercept or Beta Coefficient B_0

5.2 INFERENCE ABOUT THE INTERCEPT β_0

Although somewhat less important in practice, inferences similar to those outlined in Section 5.1 can be provided for the parameter β_0. The procedures are again based on the t distribution with d.f. $= n-2$, stated for $\hat{\beta}_0$ in Section 4. In particular,

A $100(1-\alpha)\%$ **confidence interval** for β_0 is

$$\left(\hat{\beta}_0 - t_{\alpha/2} S\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}},\quad \hat{\beta}_0 + t_{\alpha/2} S\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}}\right)$$

To illustrate this formula, let us consider the data of Table 1. In Table 3, we find $\hat{\beta}_0 = -1.07$, $\bar{x} = 5.9$, and $S_{xx} = 40.9$. Also, $s = 2.8207$.

Therefore, a 95% confidence interval for β_0 is calculated as

$$-1.07 \pm 2.306 \times 2.8207\sqrt{\frac{1}{10}+\frac{(5.9)^2}{40.9}}$$

$$= -1.07 \pm 6.34 \quad \text{or} \quad (-7.41, 5.27)$$

Example Confidence Interval for Y-intercept or Beta Coefficient B_0

A $100(1-\alpha)\%$ **confidence interval** for β_0 is

$$\left(\hat{\beta}_0 - t_{\alpha/2} S \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}},\quad \hat{\beta}_0 + t_{\alpha/2} S \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}\right)$$

To illustrate this formula, let us consider the data of Table 1. In Table 3, we find $\hat{\beta}_0 = -1.07$, $\bar{x} = 5.9$, and $S_{xx} = 40.9$. Also, $s = 2.8207$.

Therefore, a 95% confidence interval for β_0 is calculated as

$$-1.07 \pm 2.306 \times 2.8207 \sqrt{\frac{1}{10} + \frac{(5.9)^2}{40.9}}$$
$$= -1.07 \pm 6.34 \quad \text{or} \quad (-7.41, 5.27)$$

Note that β_0 represents the mean response corresponding to the value 0 for the input variable x. In the drug evaluation problem the parameter β_0 is of little practical interest because the range of x values covered in the experiment was 3 to 9 and it would be unrealistic to extend the line to $x = 0$. In fact, the estimate $\hat{\beta}_0 = -1.07$ does not have an interpretation as a (time) duration of relief.

The **test of the null hypothesis**

$$H_0: \beta_1 = \beta_{10}$$

is based on

$$T = \frac{\hat{\beta}_1 - \beta_{10}}{S/\sqrt{S_{xx}}} \qquad \text{d.f.} = n - 2$$

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

$$\left(\hat{\beta}_1 - t_{\alpha/2}\frac{S}{\sqrt{S_{xx}}}, \quad \hat{\beta}_1 + t_{\alpha/2}\frac{S}{\sqrt{S_{xx}}}\right)$$

where $t_{\alpha/2}$ is the upper $\alpha/2$ point of the t distribution with d.f. $= n - 2$.

A $100(1 - \alpha)\%$ **confidence interval** for β_0 is

$$\left(\hat{\beta}_0 - t_{\alpha/2}\, S\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}, \quad \hat{\beta}_0 + t_{\alpha/2}\, S\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}\right)$$

Boxes on Page 469-470

Confidence Interval for the Expected Response $B_0 + B_1X$

5.3 ESTIMATION OF THE MEAN RESPONSE FOR A SPECIFIED *X* VALUE

Often, the objective in a regression study is to employ the fitted regression in estimating the expected response corresponding to a specified level of the input variable. For example, we may want to estimate the expected duration of relief for a specified dosage x^* of the drug. According to the linear model described in Section 2, the expected response at a value x^* of the input variable x is given by $\beta_0 + \beta_1 x^*$. The expected response is estimated by $\hat{\beta}_0 + \hat{\beta}_1 x^*$ which is the ordinate of the fitted regression line at $x = x^*$. Referring to statement (3) of Section 4, we determine that the t distribution can be used to construct confidence intervals or test hypotheses.

A $100(1-\alpha)\%$ **confidence interval for the expected response** $\beta_0 + \beta_1 x^*$ is

$$\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2} S \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$

Example: Confidence Interval for the Expected Response $B_0 + B_1X$

A $100(1-\alpha)\%$ **confidence interval for the expected response** $\beta_0 + \beta_1 x^*$ is

$$\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2} S \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$

To **test the hypothesis** that $\beta_0 + \beta_1 x^* = \mu_0$, some specified value, we use

$$T = \frac{\hat{\beta}_0 + \hat{\beta}_1 x^* - \mu_0}{S \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}} \quad \text{d.f.} = n - 2$$

Example 8

A Confidence Interval for the Expected Duration of Relief

Again consider the data given in Table 1 and the calculations for the regression analysis given in Table 3. Obtain a 95% confidence interval for the expected duration of relief when the dosage is (a) $x^* = 6$ and (b) $x^* = 9.5$.

Example: Confidence Interval for the Expected Response $B_0 + B_1X$ Where X=6

Example 8

A Confidence Interval for the Expected Duration of Relief

Again consider the data given in Table 1 and the calculations for the regression analysis given in Table 3. Obtain a 95% confidence interval for the expected duration of relief when the dosage is (a) $x^* = 6$ and (b) $x^* = 9.5$.

SOLUTION

(a) The fitted regression line is

$$\hat{y} = -1.07 + 2.74x$$

The expected duration of relief corresponding to the dosage $x^* = 6$ milligrams of the drug is estimated as

$$\hat{\beta}_0 + \hat{\beta}_1 x^* = -1.07 + 2.74 \times 6 = 15.37 \text{ hours}$$

$$\text{Estimated standard error} = s\sqrt{\frac{1}{10} + \frac{(6-5.9)^2}{40.9}}$$

$$= 2.8207 \times .3166 = .893$$

A 95% confidence interval for the mean duration of relief with the dosage $x^* = 6$ is therefore

$$15.37 \pm t_{.025} \times .893 = 15.37 \pm 2.306 \times .893$$

$$= 15.37 \pm 2.06 \quad \text{or} \quad (13.31, 17.43)$$

Example: Confidence Interval for the Expected Response $B_0 + B_1X$ Where x = 9.5

A 95% confidence interval for the mean duration of relief with the dosage $x^* = 6$ is therefore

$$\begin{aligned} 15.37 \pm t_{.025} \times .893 &= 15.37 \pm 2.306 \times .893 \\ &= 15.37 \pm 2.06 \quad \text{or} \quad (13.31, 17.43) \end{aligned}$$

We are 95% confident that 6 milligrams of the drug produces an average duration of relief that is between about 13.31 and 17.43 hours.

(b) Suppose that we also wish to estimate the mean duration of relief under the dosage $x^* = 9.5$. We follow the same steps to calculate the point estimate.

$$\hat{\beta}_0 + \hat{\beta}_1 x^* = -1.07 + 2.74 \times 9.5 = 24.96 \text{ hours}$$

$$\begin{aligned} \text{Estimated standard error} &= 2.8207\sqrt{\frac{1}{10} + \frac{(9.5 - 5.9)^2}{40.9}} \\ &= 1.821 \end{aligned}$$

A 95% confidence interval is

$$24.96 \pm 2.306 \times 1.821 = 24.96 \pm 4.20 \quad \text{or} \quad (20.76, 29.16)$$

We are 95% confident that 9.5 milligrams of the drug produces an average of 20.76 to 29.16 hours of relief. Note that the interval is much larger than the one for 6 milligrams.

Confidence Interval for the Expected Response $B_0 + B_1X$ Better Estimate When x-Value Closer to x-bar or Mean

(b) Suppose that we also wish to estimate the mean duration of relief under the dosage $x^* = 9.5$. We follow the same steps to calculate the point estimate.

$$\hat{\beta}_0 + \hat{\beta}_1 x^* = -1.07 + 2.74 \times 9.5 = 24.96 \text{ hours}$$

$$\text{Estimated standard error} = 2.8207\sqrt{\frac{1}{10} + \frac{(9.5 - 5.9)^2}{40.9}}$$

$$= 1.821$$

A 95% confidence interval is

$$24.96 \pm 2.306 \times 1.821 = 24.96 \pm 4.20 \quad \text{or} \quad (20.76, 29.16)$$

We are 95% confident that 9.5 milligrams of the drug produces an average of 20.76 to 29.16 hours of relief. Note that the interval is much larger than the one for 6 milligrams.

The formula for the standard error shows that when x^* is close to $\overline{x}$, the standard error is smaller than it is when x^* is far removed from $\overline{x}$. This is confirmed by Example 8, where the standard error at $x^* = 9.5$ can be seen to be more than twice as large as the value at $x^* = 6$. Consequently, the confidence interval for the former is also wider. In general, estimation is more precise near the mean $\overline{x}$ than it is for values of the x variable that lie far from the mean.

Fitted Regression Line Based on the Data/Observations Collected – Not Good for Estimating Values Outside Range of x Values in the Experiment

Caution concerning extrapolation: Extreme caution should be exercised in extending a fitted regression line to make long-range predictions far away from the range of x values covered in the experiment. Not only does the confidence interval become so wide that predictions based on it can be extremely unreliable, but an even greater danger exists. If the pattern of the relationship between the variables changes drastically at a distant value of x, the data provide no information with which to detect such a change. Figure 7 illustrates this situation. We would observe a good linear relationship if we experimented with x values in the 5 to 10 range, but if the fitted line were extended to estimate the response at $x^* = 20$, then our estimate would drastically miss the mark.

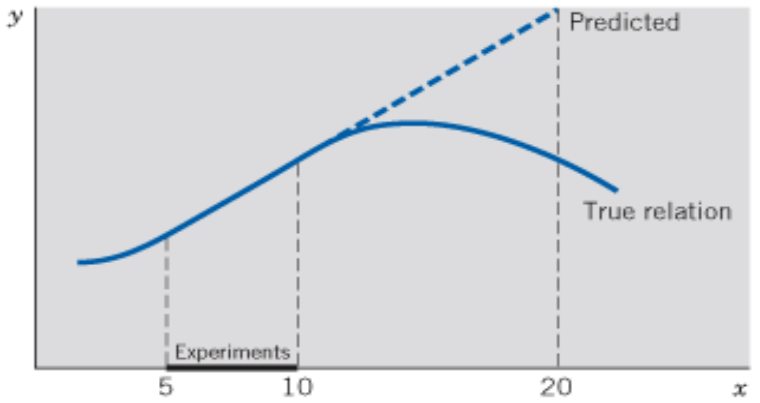


Figure 7 Danger in long-range prediction.

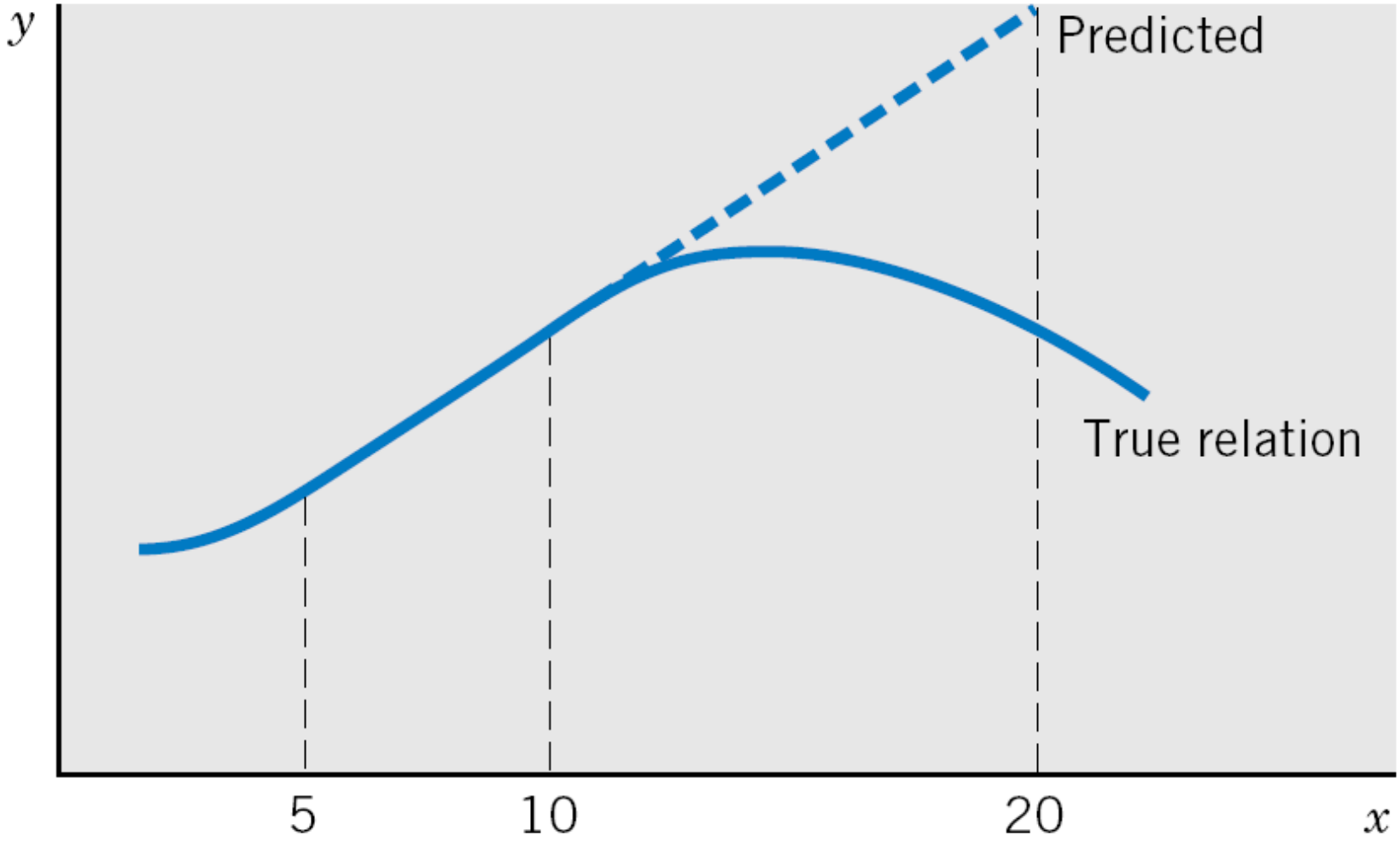


Figure 7 (p. 473)

Danger in long-range prediction.

Estimated Standard Error or Standard Deviation for Confidence Interval for the Expected Response $B_0 + B_1X$

5.4 PREDICTION OF A SINGLE RESPONSE FOR A SPECIFIED *X* VALUE

Suppose that we give a specified dosage x^* of the drug to a **single** patient and we want to predict the duration of relief from the symptoms of allergy. This problem is different from the one considered in Section 5.3, where we were interested in estimating the mean duration of relief for the population of **all** patients given the dosage x^*. The prediction is still determined from the fitted line; that is, the predicted value of the response is $\hat{\beta}_0 + \hat{\beta}_1 x^*$ as it was in the preceding case. However, the standard error of the prediction here is larger, because a single observation is more uncertain than the mean of the population distribution. We now give the formula of the estimated standard error for this case.

The **estimated standard error when predicting a single observation** y at a given x^* is

$$s\sqrt{1+\frac{1}{n}+\frac{(x^*-\bar{x})^2}{S_{xx}}}$$

The formula for the confidence interval must be modified accordingly. We call the resulting interval a **prediction interval** because it pertains to a future observation.

The **estimated standard error when predicting a single observation** y at a given x^* is

$$S\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$

Box on Page 473

Prediction of Y-Value After Calculating Regression Equation

Example 10

Prediction after Fitting a Straight Line Relation of a Human Development Index to Internet Usage

One measure of the development of a country is the Human Development Index (HDI) which combines life expectancy, literacy, educational attainment, and gross domestic product per capita into an index whose values lie between 0 and 1, inclusive.
We randomly selected fifteen countries, of the 152 countries, below the top twenty-five most developed countries on the list. HDI is the response variable y, and Internet usage per 100 persons, x, is the predictor variable. The data, given in Exercise 11.31, have the summary statistics

$$n = 15 \qquad \overline{x} = 30.133 \qquad \overline{y} = .6100$$
$$S_{xx} = 8062.39 \quad S_{yy} = .35350 \quad S_{xy} = 49.729$$

(a) Determine the equation of the best fitting straight line.

(b) Do the data substantiate the claim that Internet usage per 100 persons is a good predictor of HDI and that large values of both variables tend to occur together?

(c) Estimate the mean value of HDI for 25 Internet users per 100 persons and construct a 95% confidence interval.

(d) Find the predicted y for $x = 95$ Internet users per 100 persons.

Example: Prediction of Y-Value After Calculating Regression Equation

SOLUTION

(a)

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{49.729}{8062.39} = .00617$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = .6100 - .00617 \times 30.193 = .4241$$

So, the equation of the fitted line is

$$\hat{y} = .4241 + .00617x$$

(b) To answer this question, we decide to test $H_0: \beta_1 = 0$ versus $H_1: \beta_1 > 0$. The test statistic is

$$T = \frac{\hat{\beta}_1}{S/\sqrt{S_{xx}}}$$

We select $\alpha = .01$. Since $t_{.01} = 2.650$ with d.f. $= 13$, we set the right-sided rejection region $R: T \geq 2.650$ as shown in Figure 8.

Example: Testing H_0: $B_1 = 0$ vs. H_1: $B_1>0$ for Regression Equation

(b) To answer this question, we decide to test $H_0: \beta_1 = 0$ versus $H_1: \beta_1 > 0$. The test statistic is

$$T = \frac{\hat{\beta}_1}{S/\sqrt{S_{xx}}}$$

We select $\alpha = .01$. Since $t_{.01} = 2.650$ with d.f. $= 13$, we set the right-sided rejection region $R: T \geq 2.650$ as shown in Figure 8.

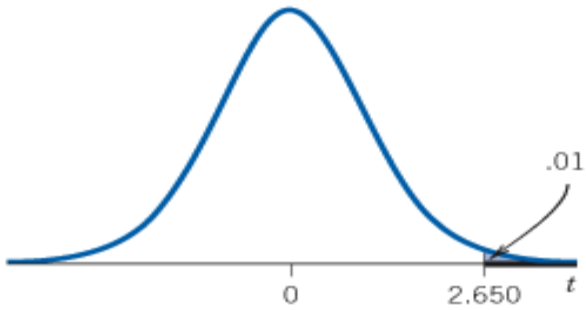


Figure 8 The rejection region $R: T \geq 2.650$.

We calculate

$$\text{SSE} = S_{yy} - \frac{S_{xy}^2}{S_{xx}} = .35350 - \frac{(49.729)^2}{8062.39} = .04678$$

$$s = \sqrt{\frac{\text{SSE}}{n-2}} = \sqrt{\frac{.04678}{13}} = .0600$$

Example: Testing H_0: $B_1 = 0$ vs. H_1: $B_1 > 0$ for Regression Equation

We calculate

$$\text{SSE} = S_{yy} - \frac{S_{xy}^2}{S_{xx}} = .35350 - \frac{(49.729)^2}{8062.39} = .04678$$

$$s = \sqrt{\frac{\text{SSE}}{n-2}} = \sqrt{\frac{.04678}{13}} = .0600$$

$$\text{Estimated S.E.}\left(\hat{\beta}_1\right) = \frac{s}{\sqrt{S_{xx}}} = \frac{.0600}{\sqrt{8062.39}} = .000668$$

The t statistic has the value

$$t = \frac{.00617}{.000668} = 9.24$$

Since the observed $t = 9.24$ is greater than 2.650, H_0 is rejected with $\alpha = .01$. The P-value is much less than .0001.

We conclude that larger values of Internet users per 100 persons significantly increases the expected HDI, within the range of values of x included in the data.

Confidence Interval for An Expected Value Substituted into the Regression Equation

(c) The expected value of the HDI corresponding to $x^* = 25$ Internet users per 100 is estimated as

$$\hat{\beta}_0 + \hat{\beta}_1 x^* = .4241 + .00617\,(25) = .5784$$

and its

$$\text{Estimated S.E.} = s\sqrt{\frac{1}{15} + \frac{(25 - 30.133)^2}{8062.39}} = .0159$$

Since $t_{.025} = 2.160$ for d.f. $= 13$, the required confidence interval is

$$.5734 \pm .0159 \times .0238 = .5734 \pm .0343 \quad \text{or} \quad (.514, .613)$$

We are 95% confident that the expected value of HDI, for $x^* = 18$, is between .514 and .613.

Fitted Regression Not Good for Estimating Values Outside Range of x Values in the Experiment: Thus, Extrapolation Not Always Accurate

(d) Since $x = 95$ is far above the largest value of 77.0 users per 100, it is not sensible to predict y at $x = 95$ using the fitted regression line. Here a formal calculation gives

$$\text{Predicted HDI} = .4241 + .00617\,(95) = 1.010$$

which is a nonsensical result for an index that should not exceed 1. As mentioned earlier, extrapolation typically gives unreliable results.

Computer Output for Regression Equation for Pill Dose-Response Relief

Regression analyses are most conveniently done on a computer. A more complete selection of the output from the computer software package MINITAB, for the data in Example 4, is given in Table 5.

TABLE 5 MINITAB Computer Output for the Data in Example 4

The regression equation is
Relief = -1.07 + 2.74 Dosage

Predictor	Coef	SE Coef	T	P
Constant	-1.071	2.751	-0.39	0.707
Dosage	2.7408	0.4411	6.21	0.000

S = 2.82074 R-Sq = 82.8%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	307.25	307.25	38.62	0.000
Residual Error	8	63.65	7.96		
Total	9	370.90			

TABLE 5 MINITAB Computer Output for the Data in Example 4

```
The regression equation is
Relief = -1.07 + 2.74 Dosage

Predictor        Coef      SE Coef          T          P
Constant       -1.071        2.751      -0.39      0.707
Dosage         2.7408       0.4411       6.21      0.000

S = 2.82074      R-Sq = 82.8%

Analysis of Variance

Source            DF          SS          MS          F          P
Regression         1      307.25      307.25      38.62      0.000
Residual Error     8       63.65        7.96
Total              9      370.90
```

Table 5 (p. 476)

MINITAB Computer Output for the Data in Example 4

Computer Software SAS Output

The output of the computer software package SAS for the data in Example 4 is given in Table 6. Notice the similarity of information in Tables 5 and 6. Both include the least squares estimates of the coefficients, their estimated standard deviations, and the t test for testing that the coefficient is zero. The estimate of σ^2 is presented as the mean square error in the analysis of variance table.

TABLE 6 SAS Computer Output for the Data in Example 4

Model: MODEL 1
Dependent Variable: y

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	307.24719	307.24719	38.62	0.0003
Error	8	63.65281	7.95660		
Corrected Total	9	370.90000			

Root MSE	2.82074	R-Square	0.8284

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr>-t-
Intercep	1	-1.07090	2.75091	-0.39	0.7072
x	1	2.74083	0.44106	6.21	0.0003

TABLE 6 SAS Computer Output for the Data in Example 4

Model: MODEL 1
Dependent Variable: y

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	307.24719	307.24719	38.62	0.0003
Error	8	63.65281	7.95660		
Corrected Total	9	370.90000			

Root MSE 2.82074 R-Square 0.8284

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr>|t|
Intercep	1	-1.07090	2.75091	-0.39	0.7072
x	1	2.74083	0.44106	6.21	0.0003

Table 6 (p. 477)

SAS Computer Output for the Data in Example 4

Example Regression Analysis: Fitting Least Squares Line and Finding Confidence Interval Based on Given X-Value

Example 11

Predicting the Number of Situps after a Semester of Conditioning

University students taking a physical fitness class were asked to count the number of situps they could do at the start of the class and again at the end of the semester.
Refer to the physical fitness data, on numbers of situps, in Table D.5 of the Data Bank.

(a) Find the least squares fitted line to predict the posttest number of situps from the pretest number at the start of the conditioning class.

(b) Find a 95% confidence interval for the mean number of posttest situps for persons who can perform 35 situps in the pretest. Also find a 95% prediction interval for the number of posttest situps that will be performed by a new person this semester who does 35 situps in the pretest.

(c) Repeat part b, but replace the number of pretest situps with 20.

Regression Analysis: Pretest Sit Ups (x) Predicting Posttest Sit Ups (y)

SOLUTION

The scatter plot in Figure 9, suggests that a straight line may model the expected value of posttest situps given the number of pretest situps. Here x is the number of pretest situps and y is the number of posttest situps. We use MINITAB statistical software to obtain the output.

Regression Analysis: Post Situps versus Pre Situps

The regression equation is

Post Situps = 10.3 + 0.899 Pre Situps

Predictor	Coef	SE Coef	T	P
Constant	10.331	2.533	4.08	0.000
Pre Situps	0.89904	0.06388	14.07	0.000

S = 5.17893 R-Sq = 71.5%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	5312.9	5312.9	198.09	0.000
Residual Error	79	2118.9	26.8		
Total	80	7431.8			

ANOVA Table Illustrates Two Sources of Deviation: Explained by the Regression and Unexplained by the Regression Equation

Source	SS	df
Regression (Explained)	Sum the squares of the explained deviations $\sum(\hat{y}-\bar{y})^2$	# of parameters - 1 always 1 for simple regression
Residual / Error (Unexplained)	Sum the squares of the unexplained deviations $\sum(y-\hat{y})^2$	sample size - # of parameters n - 2 for simple regression
Total	Sum the squares of the deviations from the mean $\sum(y-\bar{y})^2$	sample size - 1 n - 1

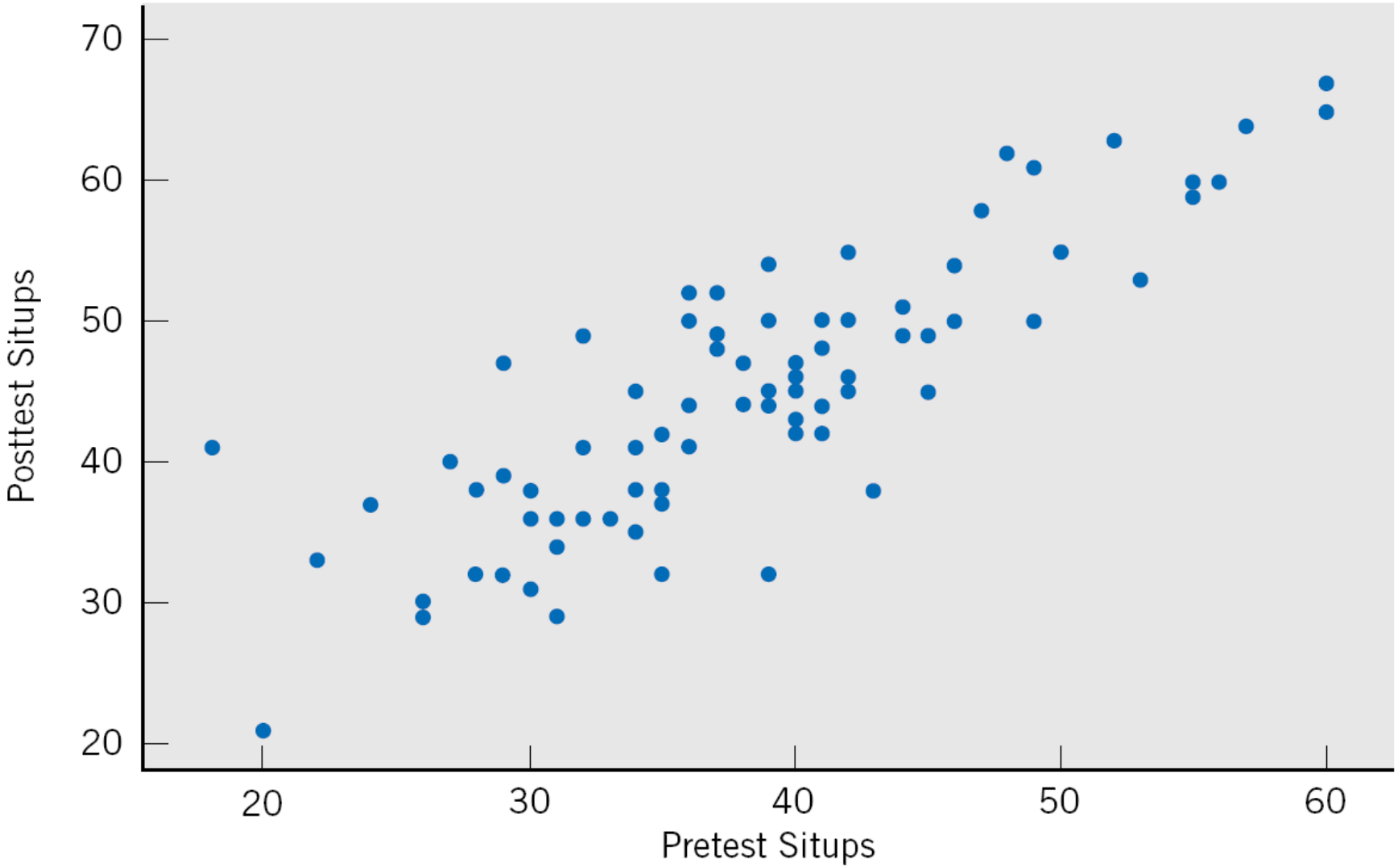


Figure 9 (p. 478)

Scatter plot of number of sit-ups.

Confidence Intervals for x= 35 Sit Ups and x = 20 Sit Ups; and for Corresponding y Values Predicted from Respective X-Values

New

Obs	Pre Sit	Fit	SE Fit	95% CI	95% PI
1	35.0	41.797	0.620	(40.563, 43.032)	(31.415, 52.179)
2	20.0	28.312	1.321	(25.682, 30.941)	(17.673, 38.950)

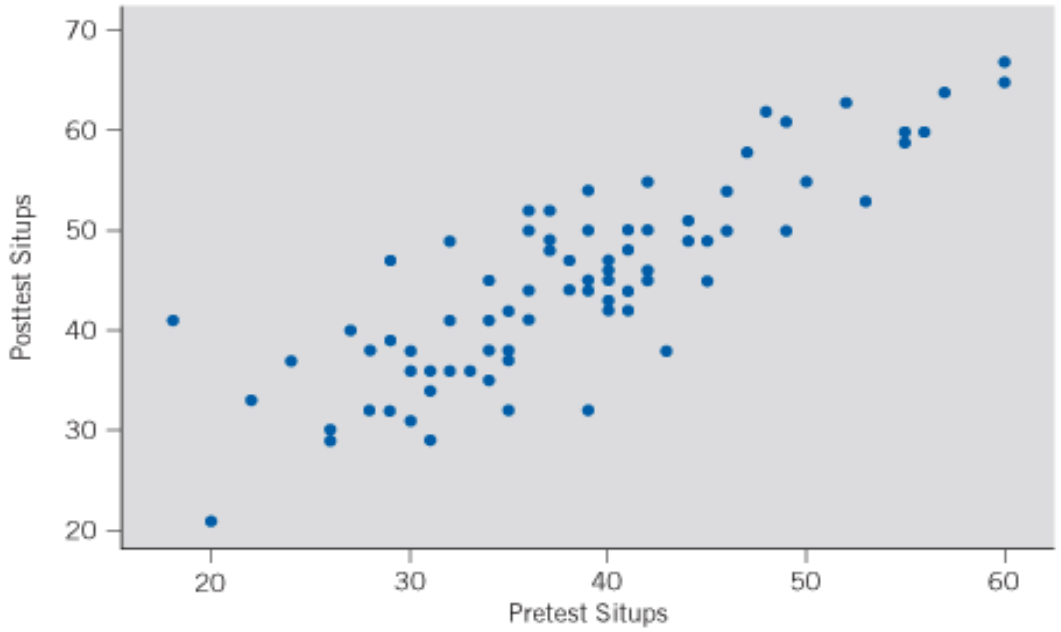


Figure 9 Scatter plot of number of situps.

Confidence Intervals for Y-Values Predicted from X-Values Wider Given σ^2

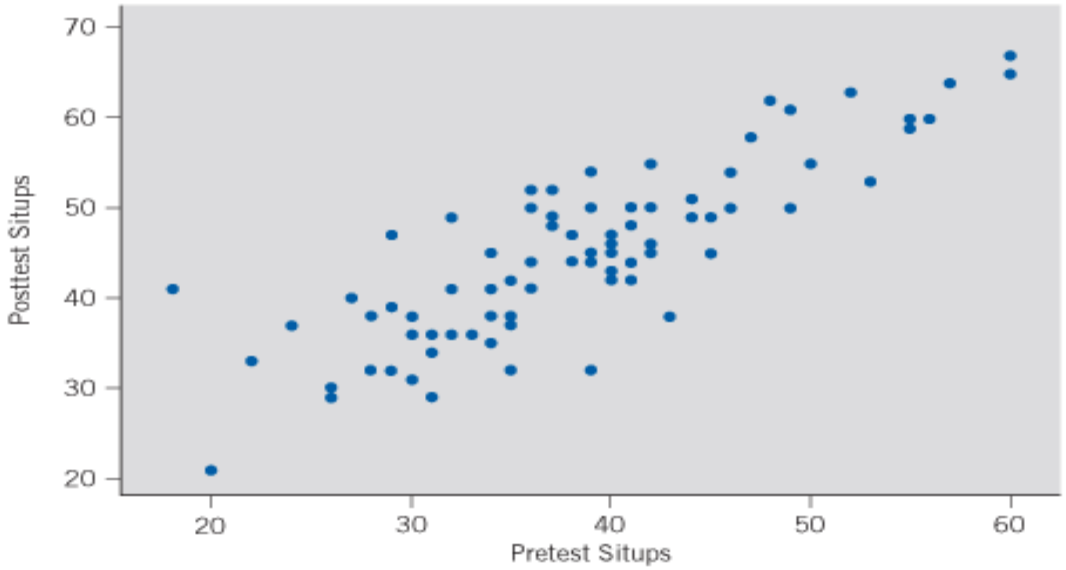


Figure 9 Scatter plot of number of situps.

From the output $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 10.3 + 0.899\ x$ and $s^2 = (5.1789)^2 = 26.8$ is the estimate of σ^2.

We have selected the option in MINITAB to obtain the two confidence intervals and prediction intervals given in the output. The prediction intervals pertain to the posttest number of situps performed by a specific new person. The first is for a person who performed 35 situps in the pretest. The prediction intervals are wider than the corresponding confidence intervals for the expected number of posttest situps for the population of all students who would do 35 situps in the pretest. The same relation holds, as it must, for 20 pretest situps.

Strength of Linear Relationship: Y Values Explained by the Linear Dependence on X

6. The Strength of a Linear Relation

To arrive at a measure of adequacy of the straight line model, we examine how much of the variation in the response variable is explained by the fitted regression line. To this end, we view an observed y_i as consisting of two components.

$$y_i = \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right) + \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)$$

y_i		$\left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)$	$\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)$
Observed y value		Explained by linear relation	Residual or deviation from linear relation

In an ideal situation where all the points lie exactly on the line, the residuals are all zero, and the y values are completely accounted for or **explained** by the linear dependence on x.

$$y_i = (\hat{\beta}_0 + \hat{\beta}_1 x_i) + (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$$

Observed y value	Explained by linear relation	Residual or deviation from linear relation

Box on Page 482

Relationship Between Sum of Squares Residuals and Total Sum of Squares

We can consider the sum of squares of the residuals

$$\text{SSE} = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2 = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

to be an overall measure of the discrepancy or departure from linearity. The total variability of the y values is reflected in the **total sum of squares**

$$S_{yy} = \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2$$

Regression Measures Variability Given X Predicting Y and Residuals Leftover

$$S_{yy} = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

of which SSE forms a part. The difference

$$\begin{aligned} S_{yy} - \text{SSE} &= S_{yy} - \left(S_{yy} - \frac{S_{xy}^2}{S_{xx}} \right) \\ &= \frac{S_{xy}^2}{S_{xx}} \end{aligned}$$

forms the other part. Paralleling the decomposition of the observation y_i, as a residual plus a part due to regression, we consider a decomposition of the variability of the y values.

$$S_{yy} = (S_{yy} - \text{SSE}) + \text{SSE}$$

Decomposition of Variability

S_{yy}	=	$\frac{S_{xy}^2}{S_{xx}}$	+	SSE
Total variability of Y		Variability explained by the linear relation		Residual or unexplained variability

Decomposition of Variability: Sum of Squares Due to Regression of x on y Plus Sum of Squares Due to Error

Decomposition of Variability

$$S_{yy} = \frac{S_{xy}^2}{S_{xx}} + \text{SSE}$$

Total variability of Y	Variability explained by the linear relation	Residual or unexplained variability

The first term on the right-hand side of this equality is called the **sum of squares (SS) due to regression**. Likewise, the total variability S_{yy} is also called the **total SS** of y. In order for the straight line model to be considered as providing a good fit to the data, the SS due to the linear regression should comprise a major portion of S_{yy}. In an ideal situation in which all points lie on the line, SSE is zero, so S_{yy} is completely explained by the fact that the x values vary in the experiment. That is, the linear relationship between y and x is solely responsible for the variability in the y values.

Strength of Linear Relationship Between x and y Explained by R-Square (Range of Values -1 to 1)

As an index of how well the straight line model fits, it is then reasonable to consider the **proportion of the y variability explained by the linear relation**

$$R^2 = \frac{\text{SS due to linear regression}}{\text{Total SS of } y} = \frac{S_{xy}^2 / S_{xx}}{S_{yy}} = \frac{S_{xy}^2}{S_{xx} S_{yy}}$$

From Section 4 of Chapter 3, recall that the quantity

$$r = \frac{S_{xy}}{\sqrt{S_{xx}\, S_{yy}}}$$

is named the **sample correlation coefficient**. Thus, the square of the sample correlation coefficient represents the proportion of the y variability explained by the linear relation.

The **strength of a linear relation** is measured by

$$R^2 = r^2 = \frac{S_{xy}^2}{S_{xx} S_{yy}}$$

which is the square of the sample correlation coefficient r.

Decomposition of Variability

$$S_{yy} = \frac{S_{xy}^2}{S_{xx}} + \text{SSE}$$

Total variability of y	Variability explained by the linear relation	Residual or unexplained variability

The **strength of a linear relation** is measured by

$$r^2 = \frac{S_{xy}^2}{S_{xx} S_{yy}}$$

which is the square of the sample correlation coefficient r.

Boxes on Page 483

Decomposition of Variability

Strength of Linear Relationship Between x and y Explained by R-Square (Range of Values -1 to 1)

The value of r is always between -1 and 1, inclusive whereas r^2 is always between 0 and 1. Also, the estimated slope and the correlation coefficient are related (see Exercise 11.47).

$$r = \hat{\beta}_1 \frac{\sqrt{S_{xx}}}{\sqrt{S_{yy}}}$$

so that the estimated slope $\hat{\beta}_1$ and r have the same sign.

Example 12

The Proportion of Variability in Duration Explained by Dosage

Let us consider the drug trial data in Table 1. From the calculations provided in Table 3,

$$S_{xx} = 40.9 \quad S_{yy} = 370.9 \quad S_{xy} = 112.1$$

Fitted regression line

$$\hat{y} = -1.07 + 2.74x$$

How much of the variability in y is explained by the linear regression model?

Example: Strength of Linear Relationship Between x and y Explained by R-Square

Example 12

The Proportion of Variability in Duration Explained by Dosage

Let us consider the drug trial data in Table 1. From the calculations provided in Table 3,

$$S_{xx} = 40.9 \quad S_{yy} = 370.9 \quad S_{xy} = 112.1$$

Fitted regression line

$$\hat{y} = -1.07 + 2.74x$$

How much of the variability in y is explained by the linear regression model?

SOLUTION

To answer this question, we calculate

$$r^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}} = \frac{(112.1)^2}{40.9 \times 370.9} = .83$$

This means that 83% of the variability in the duration of relief y is explained by the linear regression on dosage. The linear model seems quite satisfactory in this respect.

Example: Proportion of Variation in Post-Sit Ups Obtained Explained by R-Square

Example 13

Proportion of Variation Explained in Number of Situps

Refer to physical fitness data in Table D.5 of the Data Bank. Using the data on numbers of situps, find the proportion of variation in the posttest number of situps explained by the pretest number that was obtained at the beginning of the conditioning class.

SOLUTION

Repeating the relevant part of the computer output from Example 11,

```
The regression equation is
Post Situps = 10.3 + 0.899 Pre Situps

Predictor       Coef    SE Coef       T       P
Constant      10.331      2.533    4.08   0.000
Pre Situps   0.89904    0.06388   14.07   0.000

S = 5.17893    R-Sq = 71.5%

Analysis of Variance

Source           DF       SS      MS       F       P
Regression        1   5312.9  5312.9  198.09   0.000
Residual Error   79   2118.9    26.8
Total            80   7431.8
```

we find R-Sq = 71.5 % , or proportion .715. From the analysis-of-variance table we could also have

Example: Proportion of Variation in Post-Sit Ups Obtained Explained by R-Square

Repeating the relevant part of the computer output from Example 11,

The regression equation is
Post Situps = 10.3 + 0.899 Pre Situps

Predictor	Coef	SE Coef	T	P
Constant	10.331	2.533	4.08	0.000
Pre Situps	0.89904	0.06388	14.07	0.000

S = 5.17893 R-Sq = 71.5%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	5312.9	5312.9	198.09	0.000
Residual Error	79	2118.9	26.8		
Total	80	7431.8			

we find R-Sq = 71.5 % , or proportion .715. From the analysis-of-variance table we could also have calculated

$$\frac{\text{Sum of squares regression}}{\text{Total sum of squares}} = \frac{5312.9}{7431.8} = .715$$

Using a person's pretest number of situps to predict their posttest number of situps explains 71.5% of the variation in the posttest number.

Value of R-Square Small, Conclude Not a Strong Linear Relation Between x and y

When the value of r^2 is small, we can only conclude that a straight line relation does not give a good fit to the data. Such a case may arise due to the following reasons:

1. There is little relation between the variables in the sense that the scatter diagram fails to exhibit any pattern, as illustrated in Figure 10*a*. In this case, the use of a different regression model is not likely to reduce the SSE or explain a substantial part of S_{yy}.

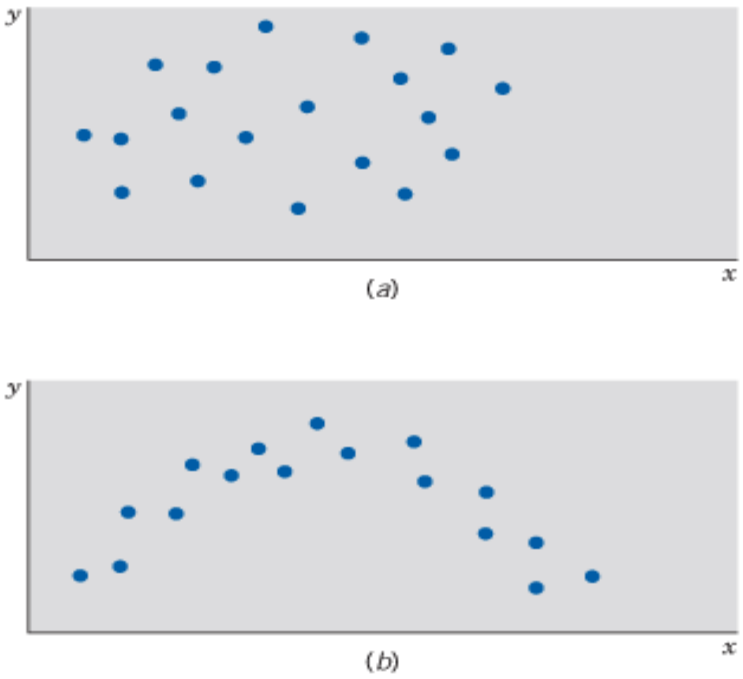
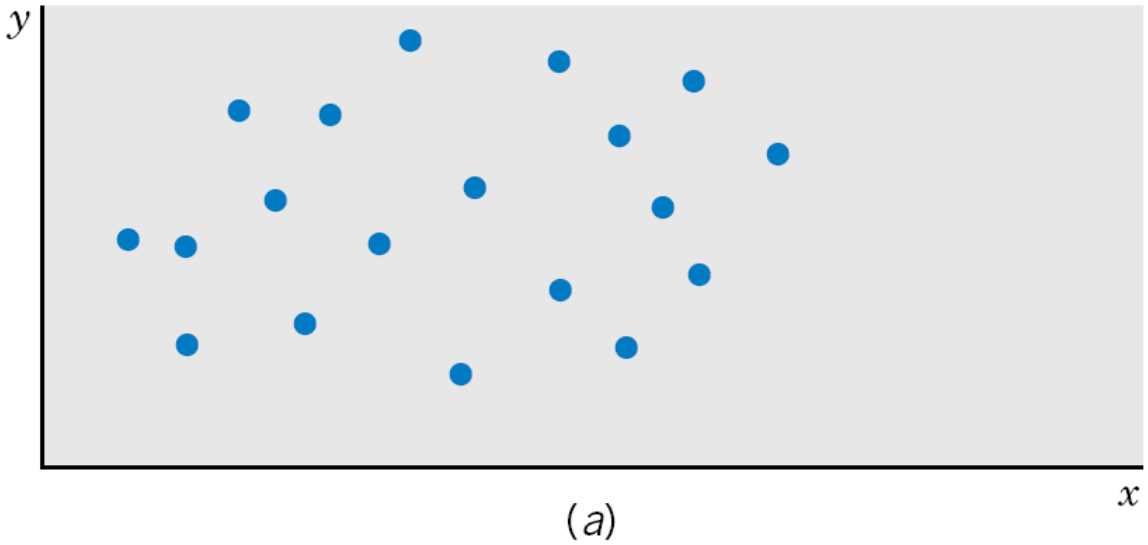
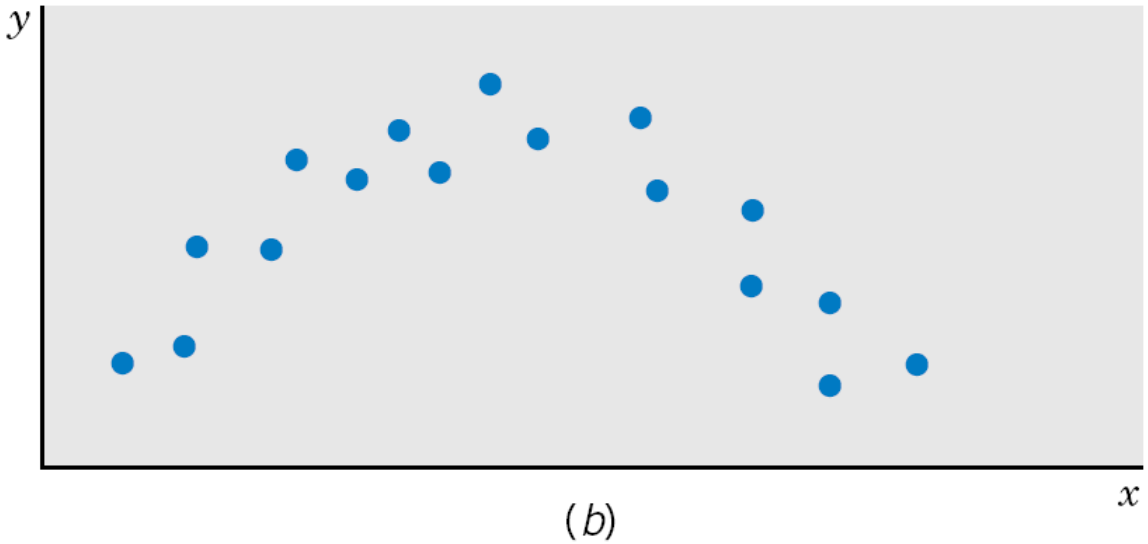


Figure 10 Scatter diagram patterns: (*a*) No relation. (*b*) A nonlinear relation.

Figure 10 (p. 485)

Scatter diagram patterns:
(a) No relation;
(b) A non-linear relation.





Scatter Plot Helps Determine Type of Relationship Between x and y – Maybe Curve Thus Another Relationship for a Better Fit

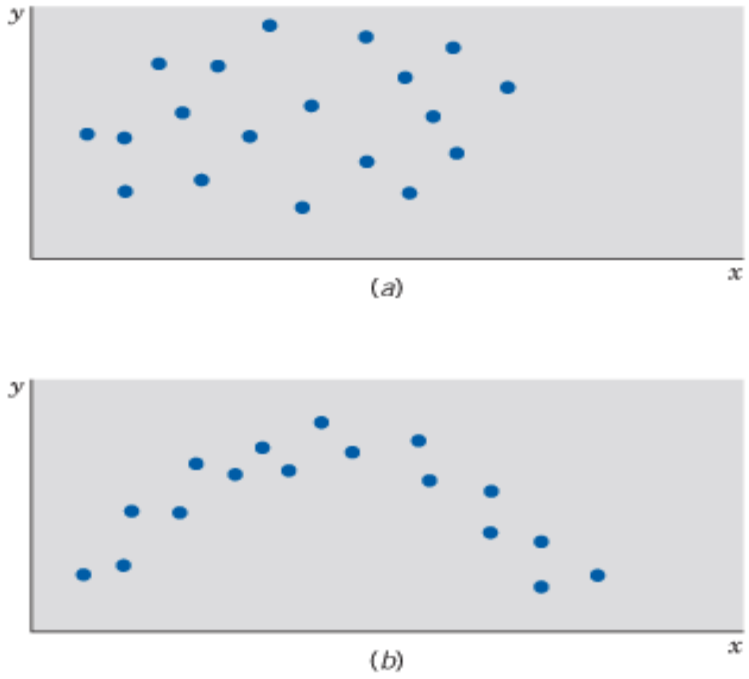


Figure 10 Scatter diagram patterns: (*a*) No relation. (*b*) A nonlinear relation.

2. There is a prominent relation but it is nonlinear in nature; that is, the scatter is banded around a curve rather than a line. The part of S_{yy} that is explained by straight line regression is small because the model is inappropriate. Some other relationship may improve the fit substantially. Figure 10*b* illustrates such a case, where the SSE can be reduced by fitting a suitable curve to the data.

Assumptions for Formulation Straight Line: Both x and y Normally Distributed

7. Remarks About the Straight Line Model Assumptions

A regression study is not completed by performing a few routine hypothesis tests and constructing confidence intervals for parameters on the basis of the formulas given in Section 4. Such conclusions can be seriously misleading if the assumptions made in the model formulations are grossly incompatible with the data. It is therefore essential to check the data carefully for indications of any violation of the assumptions. To review, the assumptions involved in the formulation of our straight line model are briefly stated again.

1. The underlying relation is linear.
2. Independence of errors.
3. Constant variance.
4. Normal distribution.

Of course, when the general nature of the relationship between y and x forms a curve rather than a straight line, the prediction obtained from fitting a straight line model to the data may produce nonsensical results. Often, a suitable transformation of the data reduces a nonlinear relation to one that is approximately linear in form. A few simple transformations are discussed in Chapter 12. Violating the assumption of independence is perhaps the most serious matter, because this can drastically distort the conclusions drawn from the t tests and the confidence

Assumptions for Formulation Straight Line: Independence of Errors and Constant Variance

1. The underlying relation is linear.
2. Independence of errors.
3. Constant variance.
4. Normal distribution.

Of course, when the general nature of the relationship between y and x forms a curve rather than a straight line, the prediction obtained from fitting a straight line model to the data may produce nonsensical results. Often, a suitable transformation of the data reduces a nonlinear relation to one that is approximately linear in form. A few simple transformations are discussed in Chapter 12. Violating the assumption of independence is perhaps the most serious matter, because this can drastically distort the conclusions drawn from the t tests and the confidence statements associated with interval estimation. The implications of assumptions 3 and 4 were illustrated earlier in Figure 3. If the scatter diagram shows different amounts of variability in the y values for different levels of x, then the assumption of constant variance may have been violated. Here, again, an appropriate transformation of the data often helps to stabilize the variance. Finally, using the t distribution in hypothesis testing and confidence interval estimation is valid as long as the errors are approximately normally distributed. A moderate departure from normality does not impair the conclusions, especially when the data set is large. In other words, a violation of assumption 4 alone is not as serious as a violation of any of the other assumptions. Methods of checking the residuals to detect any serious violation of the model assumptions are discussed in Chapter 12.

Gauss' Least Squares Theorem

Result 7.3 (Gauss'[3] least squares theorem). Let $\mathbf{Y} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $E(\boldsymbol{\varepsilon}) = \mathbf{0}$, $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$, and $\mathbf{Z}$ has full rank $r + 1$. For any $\mathbf{c}$, the estimator

$$\mathbf{c}'\hat{\boldsymbol{\beta}} = c_0\hat{\beta}_0 + c_1\hat{\beta}_1 + \cdots + c_r\hat{\beta}_r$$

[2]If $\mathbf{Z}$ is not of full rank, we can use the *generalized inverse* $(\mathbf{Z}'\mathbf{Z})^- = \sum_{i=1}^{r_1+1} \lambda_i^{-1}\mathbf{e}_i\mathbf{e}_i'$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{r_1+1} > 0 = \lambda_{r_1+2} = \cdots = \lambda_{r+1}$, as described in Exercise 7.6. Then $\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^-\mathbf{Z}' = \sum_{i=1}^{r_1+1} \mathbf{q}_i\mathbf{q}_i'$ has rank $r_1 + 1$ and generates the unique projection of $\mathbf{y}$ on the space spanned by the linearly independent columns of $\mathbf{Z}$. This is true for any choice of the generalized inverse. (See [23].)

[3]Much later, Markov proved a less general result, which misled many writers into attaching his name to this theorem.

of $\mathbf{c}'\boldsymbol{\beta}$ has the smallest possible variance among all linear estimators of the form

$$\mathbf{a}'\mathbf{Y} = a_1Y_1 + a_2Y_2 + \cdots + a_nY_n$$

that are unbiased for $\mathbf{c}'\boldsymbol{\beta}$.

Inferences About the Regression Model: Z Has Full Rank r=1 and Normally Distributed Error

7.4 Inferences About the Regression Model

We describe inferential procedures based on the classical linear regression model in (7-3) with the additional (tentative) assumption that the errors $\boldsymbol{\varepsilon}$ have a normal distribution. Methods for checking the general adequacy of the model are considered in Section 7.6.

Inferences Concerning the Regression Parameters

Before we can assess the importance of particular variables in the *regression function*

$$E(Y) = \beta_0 + \beta_1 z_1 + \cdots + \beta_r z_r \tag{7-10}$$

we must determine the sampling distributions of $\hat{\boldsymbol{\beta}}$ and the residual sum of squares, $\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$. To do so, we shall assume that the errors $\boldsymbol{\varepsilon}$ have a normal distribution.

Result 7.4. Let $\mathbf{Y} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\mathbf{Z}$ has full rank $r + 1$ and $\boldsymbol{\varepsilon}$ is distributed as $N_n(\mathbf{0}, \sigma^2\mathbf{I})$. Then the maximum likelihood estimator of $\boldsymbol{\beta}$ is the same as the least squares estimator $\hat{\boldsymbol{\beta}}$. Moreover,

$$\hat{\boldsymbol{\beta}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y} \quad \text{is distributed as} \quad N_{r+1}(\boldsymbol{\beta}, \sigma^2(\mathbf{Z}'\mathbf{Z})^{-1})$$

Confidence Ellipsoid for Beta Expressed in Terms of Estimated Covariance Matrix

and is distributed independently of the residuals $\hat{\boldsymbol{\epsilon}} = \mathbf{Y} - \mathbf{Z}\hat{\boldsymbol{\beta}}$. Further,

$$n\hat{\sigma}^2 = \hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}} \quad \text{is distributed as} \quad \sigma^2\chi^2_{n-r-1}$$

where $\hat{\sigma}^2$ is the maximum likelihood estimator of σ^2.

Proof. (See webpage: www.prenhall.com/statistics) ■

A confidence ellipsoid for $\boldsymbol{\beta}$ is easily constructed. It is expressed in terms of the estimated covariance matrix $s^2(\mathbf{Z}'\mathbf{Z})^{-1}$, where $s^2 = \hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}/(n - r - 1)$.

Confidence Ellipsoid for Beta Coefficient or B_i

A confidence ellipsoid for $\boldsymbol{\beta}$ is easily constructed. It is expressed in terms of the estimated covariance matrix $s^2(\mathbf{Z}'\mathbf{Z})^{-1}$, where $s^2 = \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}/(n - r - 1)$.

Result 7.5. Let $\mathbf{Y} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\mathbf{Z}$ has full rank $r + 1$ and $\boldsymbol{\varepsilon}$ is $N_n(\mathbf{0}, \sigma^2\mathbf{I})$. Then a $100(1 - \alpha)$ percent confidence region for $\boldsymbol{\beta}$ is given by

$$(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'\mathbf{Z}'\mathbf{Z}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \leq (r + 1)s^2 F_{r+1,n-r-1}(\alpha)$$

where $F_{r+1,n-r-1}(\alpha)$ is the upper (100α)th percentile of an F-distribution with $r + 1$ and $n - r - 1$ d.f.

Also, *simultaneous* $100(1 - \alpha)$ percent confidence intervals for the β_i are given by

$$\hat{\beta}_i \pm \sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_i)}\sqrt{(r + 1)F_{r+1,n-r-1}(\alpha)}, \quad i = 0, 1, \ldots, r$$

where $\widehat{\mathrm{Var}}(\hat{\beta}_i)$ is the diagonal element of $s^2(\mathbf{Z}'\mathbf{Z})^{-1}$ corresponding to $\hat{\beta}_i$.

Confidence Ellipsoid Centered at Maximum Likelihood Estimate Estimated Beta or Beta-hat

The confidence ellipsoid is centered at the maximum likelihood estimate $\hat{\boldsymbol{\beta}}$, and its orientation and size are determined by the eigenvalues and eigenvectors of $\mathbf{Z}'\mathbf{Z}$. If an eigenvalue is nearly zero, the confidence ellipsoid will be very long in the direction of the corresponding eigenvector.

Practitioners often ignore the "simultaneous" confidence property of the interval estimates in Result 7.5. Instead, they replace $(r + 1)F_{r+1,n-r-1}(\alpha)$ with the one-at-a-time t value $t_{n-r-1}(\alpha/2)$ and use the intervals

$$\hat{\beta} \pm t_{n-r-1}\left(\frac{\alpha}{2}\right)\sqrt{\widehat{\text{Var}}(\hat{\beta}_i)} \qquad (7\text{-}11)$$

when searching for important predictor variables.

Fitting Regression Model to Real-Estate Data n = 20 homes

Example 7.4 (Fitting a regression model to real-estate data) The assessment data in Table 7.1 were gathered from 20 homes in a Milwaukee, Wisconsin, neighborhood. Fit the regression model

$$Y_j = \beta_0 + \beta_1 z_{j1} + \beta_2 z_{j2} + \varepsilon_j$$

where z_1 = total dwelling size (in hundreds of square feet), z_2 = assessed value (in thousands of dollars), and Y = selling price (in thousands of dollars), to these data using the method of least squares. A computer calculation yields

$$(\mathbf{Z}'\mathbf{Z})^{-1} = \begin{bmatrix} 5.1523 & & \\ .2544 & .0512 & \\ -.1463 & -.0172 & .0067 \end{bmatrix}$$

Real Estate Data: Predictor 1 – Dwelling Size; Predictor 2 – Assessed Value; Response – Selling Price

Table 7.1 Real-Estate Data

z_1 Total dwelling size ($100\ ft^2$)	z_2 Assessed value ($1000)	Y Selling price ($1000)
15.31	57.3	74.8
15.20	63.8	74.0
16.25	65.4	72.9
14.33	57.0	70.0
14.57	63.8	74.9
17.33	63.2	76.0
14.48	60.2	72.0
14.91	57.7	73.5
15.25	56.4	74.5
13.89	55.6	73.5
15.18	62.6	71.5
14.44	63.4	71.0
14.87	60.2	78.9
18.63	67.2	86.5
15.20	57.1	68.0
25.76	89.6	102.0
19.05	68.6	84.0
15.37	60.1	69.0
18.06	66.3	88.0
16.35	65.8	76.0

Beta Coefficients for Real Estate Data

and

$$\hat{\boldsymbol{\beta}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y} = \begin{bmatrix} 30.967 \\ 2.634 \\ .045 \end{bmatrix}$$

Thus, the fitted equation is

$$\hat{y} = \underset{(7.88)}{30.967} + \underset{(.785)}{2.634}z_1 + \underset{(.285)}{.045}z_2$$

with $s = 3.473$. The numbers in parentheses are the estimated standard deviations of the least squares coefficients. Also, $R^2 = .834$, indicating that the data exhibit a strong regression relationship. (See Panel 7.1, which contains the regression analysis of these data using the SAS statistical software package.) If the residuals $\hat{\boldsymbol{\epsilon}}$ pass the diagnostic checks described in Section 7.6, the fitted equation could be used to predict the selling price of another house in the neighborhood from its size

SAS Output: Real Estate Data: Predictor 1 – Dwelling Size; Predictor 2 – Assessed Value; Response – Selling Price

PANEL 7.1 SAS ANALYSIS FOR EXAMPLE 7.4 USING PROC REG.

```
title 'Regression Analysis';
data estate;
infile 'T7-1.dat';
input z1 z2 y;
proc reg data = estate;
model y = z1 z2;
```

PROGRAM COMMANDS

OUTPUT

Model: MODEL 1
Dependent Variable:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F value	Prob > F
Model	2	1032.87506	516.43753	42.828	0.0001
Error	17	204.99494	12.05853		
C Total	19	1237.87000			

Root MSE	3.47254	R-square	0.8344
Deep Mean	76.55000	Adj R-sq	0.8149
C.V.	4.53630		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter = 0	Prob > |T|
INTERCEP	1	30.966566	7.88220844	3.929	0.0011
z1	1	2.634400	0.78559872	3.353	0.0038
z2	1	0.045184	0.28518271	0.158	0.8760

Confidence Interval for B_2 Or Predictor 2, i.e. or Assessed Value

and assessed value. We note that a 95% confidence interval for β_2 [see (7-14)] is given by

$$\hat{\beta}_2 \pm t_{17}(.025)\sqrt{\widehat{\text{Var}}(\hat{\beta}_2)} = .045 \pm 2.110(.285)$$

or

$$(-.556, .647)$$

Since the confidence interval includes $\beta_2 = 0$, the variable z_2 might be dropped from the regression model and the analysis repeated with the single predictor variable z_1. Given dwelling size, assessed value seems to add little to the prediction of selling price. ■

Likelihood Ratio Tests for Effect of Regression Parameters, i.e. Certain Predictors on the Response

Likelihood Ratio Tests for the Regression Parameters

Part of regression analysis is concerned with assessing the effects of particular predictor variables on the response variable. One null hypothesis of interest states that certain of the z_i's do not influence the response Y. These predictors will be labeled $z_{q+1}, z_{q+2}, \ldots, z_r$. The statement that $z_{q+1}, z_{q+2}, \ldots, z_r$ do not influence Y translates into the statistical hypothesis

$$H_0: \beta_{q+1} = \beta_{q+2} = \cdots = \beta_r = 0 \quad \text{or} \quad H_0: \boldsymbol{\beta}_{(2)} = \mathbf{0} \tag{7-12}$$

where $\boldsymbol{\beta}_{(2)}' = [\beta_{q+1}, \beta_{q+2}, \ldots, \beta_r]$.

Setting

$$\mathbf{Z} = \underset{n\times(r+1)}{\left[\begin{array}{c|c} \underset{n\times(q+1)}{\mathbf{Z}_1} & \underset{n\times(r-q)}{\mathbf{Z}_2} \end{array}\right]}, \qquad \boldsymbol{\beta} = \left[\begin{array}{c} \underset{((q+1)\times 1)}{\boldsymbol{\beta}_{(1)}} \\ \hline \underset{((r-q)\times 1)}{\boldsymbol{\beta}_{(2)}} \end{array}\right]$$

we can express the general linear model as

$$\mathbf{Y} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = [\mathbf{Z}_1 \mid \mathbf{Z}_2]\left[\begin{array}{c} \boldsymbol{\beta}_{(1)} \\ \hline \boldsymbol{\beta}_{(2)} \end{array}\right] + \boldsymbol{\varepsilon} = \mathbf{Z}_1\boldsymbol{\beta}_{(1)} + \mathbf{Z}_2\boldsymbol{\beta}_{(2)} + \boldsymbol{\varepsilon}$$

Under the null hypothesis $H_0: \boldsymbol{\beta}_{(2)} = \mathbf{0}$, $\mathbf{Y} = \mathbf{Z}_1\boldsymbol{\beta}_{(1)} + \boldsymbol{\varepsilon}$. The likelihood ratio test of H_0 is based on the

$$\begin{aligned} \text{Extra sum of squares} &= SS_{\text{res}}(\mathbf{Z}_1) - SS_{\text{res}}(\mathbf{Z}) \\ &= (\mathbf{y} - \mathbf{Z}_1\hat{\boldsymbol{\beta}}_{(1)})'(\mathbf{y} - \mathbf{Z}_1\hat{\boldsymbol{\beta}}_{(1)}) - (\mathbf{y} - \mathbf{Z}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{Z}\hat{\boldsymbol{\beta}}) \end{aligned} \tag{7-13}$$

where $\hat{\boldsymbol{\beta}}_{(1)} = (\mathbf{Z}_1'\mathbf{Z}_1)^{-1}\mathbf{Z}_1'\mathbf{y}$.

Likelihood Ratio Test of Ho Based on Extra Sum of Squares (Use F-Distribution as Critical Value)

Under the null hypothesis $H_0: \boldsymbol{\beta}_{(2)} = \mathbf{0}$, $\mathbf{Y} = \mathbf{Z}_1\boldsymbol{\beta}_{(1)} + \boldsymbol{\varepsilon}$. The likelihood ratio test of H_0 is based on the

$$\text{Extra sum of squares} = SS_{\text{res}}(\mathbf{Z}_1) - SS_{\text{res}}(\mathbf{Z}) \tag{7-13}$$
$$= (\mathbf{y} - \mathbf{Z}_1\hat{\boldsymbol{\beta}}_{(1)})'(\mathbf{y} - \mathbf{Z}_1\hat{\boldsymbol{\beta}}_{(1)}) - (\mathbf{y} - \mathbf{Z}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{Z}\hat{\boldsymbol{\beta}})$$

where $\hat{\boldsymbol{\beta}}_{(1)} = (\mathbf{Z}_1'\mathbf{Z}_1)^{-1}\mathbf{Z}_1'\mathbf{y}$.

Result 7.6. Let $\mathbf{Z}$ have full rank $r + 1$ and $\boldsymbol{\varepsilon}$ be distributed as $N_n(\mathbf{0}, \sigma^2\mathbf{I})$. The likelihood ratio test of $H_0: \boldsymbol{\beta}_{(2)} = \mathbf{0}$ is equivalent to a test of H_0 based on the extra sum of squares in (7-13) and $s^2 = (\mathbf{y} - \mathbf{Z}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{Z}\hat{\boldsymbol{\beta}})/(n - r - 1)$. In particular, the likelihood ratio test rejects H_0 if

$$\frac{(SS_{\text{res}}(\mathbf{Z}_1) - SS_{\text{res}}(\mathbf{Z}))/(r - q)}{s^2} > F_{r-q,n-r-1}(\alpha)$$

where $F_{r-q,n-r-1}(\alpha)$ is the upper (100α)th percentile of an F-distribution with $r - q$ and $n - r - 1$ d.f.

Likelihood Ratio Test: Fit the Model With and Without the Predictor In Question: If the Residual Sum of Squares Improves, Then Retain if Not Do Not Retain; Assess With F-Ratio

Comment. The likelihood ratio test is implemented as follows. To test whether all coefficients in a subset are zero, fit the model with and without the terms corresponding to these coefficients. The improvement in the residual sum of squares (the extra sum of squares) is compared to the residual sum of squares for the full model via the F-ratio. The same procedure applies even in analysis of variance situations where $\mathbf{Z}$ is not of full rank.[4]

More generally, it is possible to formulate null hypotheses concerning $r - q$ linear combinations of $\boldsymbol{\beta}$ of the form $H_0: \mathbf{C}\boldsymbol{\beta} = \mathbf{A}_0$. Let the $(r - q) \times (r + 1)$ matrix $\mathbf{C}$ have full rank, let $\mathbf{A}_0 = \mathbf{0}$, and consider

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

(This null hypothesis reduces to the previous choice when $\mathbf{C} = \left[\mathbf{0} \vdots \underset{(r-q)\times(r-q)}{\mathbf{I}}\right]$.)

Maximum Likelihood Test: Does the Angle of the Variable Lie in the Confidence Ellipsoid

Under the full model, $\mathbf{C}\hat{\boldsymbol{\beta}}$ is distributed as $N_{r-q}(\mathbf{C}\boldsymbol{\beta}, \sigma^2\mathbf{C}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{C}')$. We reject $H_0: \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$ at level α if $\mathbf{0}$ does *not* lie in the $100(1-\alpha)\%$ confidence ellipsoid for $\mathbf{C}\boldsymbol{\beta}$. Equivalently, we reject $H_0: \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$ if

$$\frac{(\mathbf{C}\hat{\boldsymbol{\beta}})'(\mathbf{C}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{C}')^{-1}(\mathbf{C}\hat{\boldsymbol{\beta}})}{s^2} > (r-q)F_{r-q,n-r-1}(\alpha) \qquad (7\text{-}14)$$

where $s^2 = (\mathbf{y} - \mathbf{Z}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{Z}\hat{\boldsymbol{\beta}})/(n-r-1)$ and $F_{r-q,n-r-1}(\alpha)$ is the upper (100α)th percentile of an F-distribution with $r-q$ and $n-r-1$ d.f. The test in (7-14) is the likelihood ratio test, and the numerator in the F-ratio is the extra residual sum of squares incurred by fitting the model, subject to the restriction that $\mathbf{C}\boldsymbol{\beta} = \mathbf{0}$. (See [23]).

The next example illustrates how unbalanced experimental designs are easily handled by the general theory just described.

Additional Predictors: Do They Add Value to the Model?

Example 7.5 (Testing the importance of additional predictors using the extra sum-of-squares approach) Male and female patrons rated the service in three establishments (locations) of a large restaurant chain. The service ratings were converted into an index. Table 7.2 contains the data for $n = 18$ customers. Each data point in the table is categorized according to location (1, 2, or 3) and gender (male = 0 and female = 1). This categorization has the format of a two-way table with unequal numbers of observations per cell. For instance, the combination of location 1 and male has 5 responses, while the combination of location 2 and female has 2 responses. Introducing three dummy variables to account for location and two dummy variables to account for gender, we can develop a regression model linking the service index Y to location, gender, and their "interaction" using the design matrix

Table 7.2 Restaurant-Service Data

Location	Gender	Service (Y)
1	0	15.2
1	0	21.2
1	0	27.3
1	0	21.2
1	0	21.2
1	1	36.4
1	1	92.4
2	0	27.3
2	0	15.2
2	0	9.1
2	0	18.2
2	0	50.0
2	1	44.0
2	1	63.6
3	0	15.2
3	0	30.3
3	1	36.4
3	1	40.9

Restaurant Service Data

Table 7.2 Restaurant-Service Data

Location	Gender	Service (Y)
1	0	15.2
1	0	21.2
1	0	27.3
1	0	21.2
1	0	21.2
1	1	36.4
1	1	92.4
2	0	27.3
2	0	15.2
2	0	9.1
2	0	18.2
2	0	50.0
2	1	44.0
2	1	63.6
3	0	15.2
3	0	30.3
3	1	36.4
3	1	40.9

Interaction Term: Location-Gender Variable

	constant	location			gender		interaction						
	1	1	0	0	1	0	1	0	0	0	0	0	5 responses
	1	1	0	0	1	0	1	0	0	0	0	0	
	1	1	0	0	1	0	1	0	0	0	0	0	
	1	1	0	0	1	0	1	0	0	0	0	0	
	1	1	0	0	1	0	1	0	0	0	0	0	
	1	1	0	0	0	1	0	1	0	0	0	0	2 responses
	1	1	0	0	0	1	0	1	0	0	0	0	
	1	0	1	0	1	0	0	0	1	0	0	0	5 responses
	1	0	1	0	1	0	0	0	1	0	0	0	
$\mathbf{Z} =$	1	0	1	0	1	0	0	0	1	0	0	0	
	1	0	1	0	1	0	0	0	1	0	0	0	
	1	0	1	0	1	0	0	0	1	0	0	0	
	1	0	1	0	0	1	0	0	0	1	0	0	2 responses
	1	0	1	0	0	1	0	0	0	1	0	0	
	1	0	0	1	1	0	0	0	0	0	1	0	2 responses
	1	0	0	1	1	0	0	0	0	0	1	0	
	1	0	0	1	0	1	0	0	0	0	0	1	2 responses
	1	0	0	1	0	1	0	0	0	0	0	1	

The coefficient vector can be set out as

$$\boldsymbol{\beta}' = [\beta_0, \beta_1, \beta_2, \beta_3, \tau_1, \tau_2, \gamma_{11}, \gamma_{12}, \gamma_{21}, \gamma_{22}, \gamma_{31}, \gamma_{32}]$$

where the β_i's ($i > 0$) represent the effects of the locations on the determination service, the τ_i's represent the effects of gender on the service index, and the γ_{ij} represent the location-gender interaction effects.

Calculating the Likelihood Ratio and Comparing to F-value

The coefficient vector can be set out as

$$\boldsymbol{\beta}' = [\beta_0, \beta_1, \beta_2, \beta_3, \tau_1, \tau_2, \gamma_{11}, \gamma_{12}, \gamma_{21}, \gamma_{22}, \gamma_{31}, \gamma_{32}]$$

where the β_i's ($i > 0$) represent the effects of the locations on the determination of service, the τ_i's represent the effects of gender on the service index, and the γ_{ik}'s represent the location-gender interaction effects.

The design matrix $\mathbf{Z}$ is not of full rank. (For instance, column 1 equals the sum of columns 2–4 or columns 5–6.) In fact, $\text{rank}(\mathbf{Z}) = 6$.

For the complete model, results from a computer program give

$$\text{SS}_{\text{res}}(\mathbf{Z}) = 2977.4$$

and $n - \text{rank}(\mathbf{Z}) = 18 - 6 = 12$.

The model without the interaction terms has the design matrix $\mathbf{Z}_1$ consisting of the first six columns of $\mathbf{Z}$. We find that

$$\text{SS}_{\text{res}}(\mathbf{Z}_1) = 3419.1$$

with $n - \text{rank}(\mathbf{Z}_1) = 18 - 4 = 14$. To test H_0: $\gamma_{11} = \gamma_{12} = \gamma_{21} = \gamma_{22} = \gamma_{31} = \gamma_{32} = 0$ (no location–gender interaction), we compute

$$F = \frac{(\text{SS}_{\text{res}}(\mathbf{Z}_1) - \text{SS}_{\text{res}}(\mathbf{Z}))/(6 - 4)}{s^2} = \frac{(\text{SS}_{\text{res}}(\mathbf{Z}_1) - \text{SS}_{\text{res}}(\mathbf{Z}))/2}{\text{SS}_{\text{res}}(\mathbf{Z})/12}$$

$$= \frac{(3419.1 - 2977.4)/2}{2977.4/12} = .89$$

Interaction Term Not Significant; However Gender Does Influence Restaurant Service Rating

$$F = \frac{(SS_{res}(\mathbf{Z}_1) - SS_{res}(\mathbf{Z}))/(6-4)}{s^2} = \frac{(SS_{res}(\mathbf{Z}_1) - SS_{res}(\mathbf{Z}))/2}{SS_{res}(\mathbf{Z})/12}$$

$$= \frac{(3419.1 - 2977.4)/2}{2977.4/12} = .89$$

The F-ratio may be compared with an appropriate percentage point of an F-distribution with 2 and 12 d.f. This F-ratio is not significant for any reasonable significance level α. Consequently, we conclude that the service index does not depend upon any location–gender interaction, and these terms can be dropped from the model.

Using the extra sum-of-squares approach, we may verify that there is no difference between locations (no location effect), but that gender is significant; that is, males and females do not give the same ratings to service.

In analysis-of-variance situations where the cell counts are unequal, the variation in the response attributable to different predictor variables and their interactions cannot usually be separated into independent amounts. To evaluate the relative influences of the predictors on the response in this case, it is necessary to fit the model with and without the terms in question and compute the appropriate F-test statistics. ■

Inferences from the Estimated Regression Function

7.5 Inferences from the Estimated Regression Function

Once an investigator is satisfied with the fitted regression model, it can be used to solve two prediction problems. Let $\mathbf{z}_0' = [1, z_{01}, \ldots, z_{0r}]$ be selected values for the predictor variables. Then $\mathbf{z}_0$ and $\hat{\boldsymbol{\beta}}$ can be used (1) to estimate the regression function $\beta_0 + \beta_1 z_{01} + \cdots + \beta_r z_{0r}$ at $\mathbf{z}_0$ and (2) to estimate the value of the response Y at $\mathbf{z}_0$.

Estimating the Regression Function at z_0

Let Y_0 denote the value of the response when the predictor variables have values $\mathbf{z}_0' = [1, z_{01}, \ldots, z_{0r}]$. According to the model in (7-3), the expected value of Y_0 is

$$E(Y_0 | \mathbf{z}_0) = \beta_0 + \beta_1 z_{01} + \cdots + \beta_r z_{0r} = \mathbf{z}_0' \boldsymbol{\beta} \tag{7-15}$$

Its least squares estimate is $\mathbf{z}_0' \hat{\boldsymbol{\beta}}$.

Confidence Interval for Bo or Intercept

Result 7.7. For the linear regression model in (7-3), $\mathbf{z}_0'\hat{\boldsymbol{\beta}}$ is the unbiased linear estimator of $E(Y_0|\mathbf{z}_0)$ with minimum variance, $\text{Var}(\mathbf{z}_0'\hat{\boldsymbol{\beta}}) = \mathbf{z}_0'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{z}_0\sigma^2$. If the errors $\boldsymbol{\varepsilon}$ are normally distributed, then a $100(1-\alpha)\%$ confidence interval for $E(Y_0|\mathbf{z}_0) = \mathbf{z}_0'\boldsymbol{\beta}$ is provided by

$$\mathbf{z}_0'\hat{\boldsymbol{\beta}} \pm t_{n-r-1}\left(\frac{\alpha}{2}\right)\sqrt{(\mathbf{z}_0'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{z}_0)s^2}$$

where $t_{n-r-1}(\alpha/2)$ is the upper $100(\alpha/2)$th percentile of a t-distribution with $n-r-1$ d.f.

Predicting a New Observation More Uncertain Then Estimating Expected Value of Y_0

Forecasting a New Observation at $\mathbf{z}_0$

Prediction of a new observation, such as Y_0, at $\mathbf{z}_0' = [1, z_{01}, \ldots, z_{0r}]$ is more uncertain than estimating the *expected value* of Y_0. According to the regression model of (7-3),

$$Y_0 = \mathbf{z}_0'\boldsymbol{\beta} + \varepsilon_0$$

or

$$(\text{new response } Y_0) = (\text{expected value of } Y_0 \text{ at } \mathbf{z}_0) + (\text{new error})$$

where ε_0 is distributed as $N(0, \sigma^2)$ and is independent of $\boldsymbol{\varepsilon}$ and, hence, of $\hat{\boldsymbol{\beta}}$ and s^2. The errors $\boldsymbol{\varepsilon}$ influence the estimators $\hat{\boldsymbol{\beta}}$ and s^2 through the responses $\mathbf{Y}$, but ε_0 does not.

Result 7.8. Given the linear regression model of (7-3), a new observation Y_0 has the *unbiased predictor*

$$\mathbf{z}_0'\hat{\boldsymbol{\beta}} = \hat{\beta}_0 + \hat{\beta}_1 z_{01} + \cdots + \hat{\beta}_r z_{0r}$$

The variance of the *forecast error* $Y_0 - \mathbf{z}_0'\hat{\boldsymbol{\beta}}$ is

$$\text{Var}(Y_0 - \mathbf{z}_0'\hat{\boldsymbol{\beta}}) = \sigma^2(1 + \mathbf{z}_0'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{z}_0)$$

When the errors $\boldsymbol{\varepsilon}$ have a normal distribution, a $100(1 - \alpha)\%$ *prediction interval* for Y_0 is given by

$$\mathbf{z}_0'\hat{\boldsymbol{\beta}} \pm t_{n-r-1}\left(\frac{\alpha}{2}\right)\sqrt{s^2(1 + \mathbf{z}_0'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{z}_0)}$$

where $t_{n-r-1}(\alpha/2)$ is the upper $100(\alpha/2)$th percentile of a t-distribution with $n - r - 1$ degrees of freedom.

Confidence Interval for Y or CPU Time

Example 7.6 (Interval estimates for a mean response and a future response) Companies considering the purchase of a computer must first assess their future needs in order to determine the proper equipment. A computer scientist collected data from seven similar company sites so that a forecast equation of computer-hardware requirements for inventory management could be developed. The data are given in Table 7.3 for

z_1 = customer orders (in thousands)
z_2 = add-delete item count (in thousands)
Y = CPU (central processing unit) time (in hours)

Construct a 95% confidence interval for the mean CPU time, $E(Y_0|\mathbf{z}_0) = \beta_0 + \beta_1 z_{01} + \beta_2 z_{02}$ at $\mathbf{z}_0' = [1, 130, 7.5]$. Also, find a 95% prediction interval for a new facility's CPU requirement corresponding to the same $\mathbf{z}_0$.

A computer program provides the estimated regression function

$$\hat{y} = 8.42 + 1.08z_1 + .42z_2$$

$$(\mathbf{Z}'\mathbf{Z})^{-1} = \begin{bmatrix} 8.17969 & & \\ -.06411 & .00052 & \\ .08831 & -.00107 & .01440 \end{bmatrix}$$

and $s = 1.204$. Consequently,

$$\mathbf{z}_0'\hat{\boldsymbol{\beta}} = 8.42 + 1.08(130) + .42(7.5) = 151.97$$

and $s\sqrt{\mathbf{z}_0'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{z}_0} = 1.204(.58928) = .71$. We have $t_4(.025) = 2.776$, so the 95% confidence interval for the mean CPU time at $\mathbf{z}_0$ is

$$\mathbf{z}_0'\hat{\boldsymbol{\beta}} \pm t_4(.025)s\sqrt{\mathbf{z}_0'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{z}_0} = 151.97 \pm 2.776(.71)$$

or (150.00, 153.94).

Predictor and Response Values: If Add or Delete Does Regression Equation Change, i.e. does Confidence Interval Change for Y_0

and $s = 1.204$. Consequently,

$$\mathbf{z}_0'\hat{\boldsymbol{\beta}} = 8.42 + 1.08(130) + .42(7.5) = 151.97$$

and $s\sqrt{\mathbf{z}_0'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{z}_0} = 1.204(.58928) = .71$. We have $t_4(.025) = 2.776$, so the 95% confidence interval for the mean CPU time at $\mathbf{z}_0$ is

$$\mathbf{z}_0'\hat{\boldsymbol{\beta}} \pm t_4(.025)s\sqrt{\mathbf{z}_0'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{z}_0} = 151.97 \pm 2.776(.71)$$

or (150.00, 153.94).

Table 7.3 Computer Data

z_1 (Orders)	z_2 (Add–delete items)	Y (CPU time)
123.5	2.108	141.5
146.1	9.213	168.9
133.9	1.905	154.8
128.5	.815	146.5
151.5	1.061	172.8
136.2	8.603	160.1
92.0	1.125	108.5

Source: Data taken from H. P. Artis, *Forecasting Computer Requirements: A Forecaster's Dilemma* (Piscataway, NJ: Bell Laboratories, 1979).

CPU Time or Y-response Variable At New Facility Changes Confidence Interval

Table 7.3 Computer Data

z_1 (Orders)	z_2 (Add–delete items)	Y (CPU time)
123.5	2.108	141.5
146.1	9.213	168.9
133.9	1.905	154.8
128.5	.815	146.5
151.5	1.061	172.8
136.2	8.603	160.1
92.0	1.125	108.5

Source: Data taken from H. P. Artis, *Forecasting Computer Requirements: A Forecaster's Dilemma* (Piscataway, NJ: Bell Laboratories, 1979).

Since $s\sqrt{1 + \mathbf{z}_0'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{z}_0} = (1.204)(1.16071) = 1.40$, a 95% prediction interval for the CPU time at a new facility with conditions $\mathbf{z}_0$ is

$$\mathbf{z}_0'\hat{\boldsymbol{\beta}} \pm t_4(.025)s\sqrt{1 + \mathbf{z}_0'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{z}_0} = 151.97 \pm 2.776(1.40)$$

or (148.08, 155.86). ■

Model Checking and Other Aspects of Regression: Error and Model Fit

7.6 Model Checking and Other Aspects of Regression

Does the Model Fit?

Assuming that the model is "correct," we have used the estimated regression function to make inferences. Of course, it is imperative to examine the adequacy of the model *before* the estimated function becomes a permanent part of the decision-making apparatus.

All the sample information on lack of fit is contained in the residuals

$$\begin{aligned}
\hat{\varepsilon}_1 &= y_1 - \hat{\beta}_0 - \hat{\beta}_1 z_{11} - \cdots - \hat{\beta}_r z_{1r} \\
\hat{\varepsilon}_2 &= y_2 - \hat{\beta}_0 - \hat{\beta}_1 z_{21} - \cdots - \hat{\beta}_r z_{2r} \\
&\vdots \qquad\qquad\qquad \vdots \\
\hat{\varepsilon}_n &= y_n - \hat{\beta}_0 - \hat{\beta}_1 z_{n1} - \cdots - \hat{\beta}_r z_{nr}
\end{aligned}$$

or

$$\hat{\boldsymbol{\varepsilon}} = [\mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}']\mathbf{y} = [\mathbf{I} - \mathbf{H}]\mathbf{y} \tag{7-16}$$

If the model is valid, each residual $\hat{\varepsilon}_j$ is an estimate of the error ε_j, which is assumed to be a normal random variable with mean zero and variance σ^2. Although the residuals $\hat{\boldsymbol{\varepsilon}}$ have expected value $\mathbf{0}$, their covariance matrix $\sigma^2[\mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'] = \sigma^2[\mathbf{I} - \mathbf{H}]$ is not diagonal. Residuals have unequal variances and nonzero correlations. Fortunately, the correlations are often small and the variances are nearly equal.

"Studentized" Residuals Are Expected to Follow Normal Distribution

Because the residuals $\hat{\boldsymbol{\varepsilon}}$ have covariance matrix $\sigma^2[\mathbf{I} - \mathbf{H}]$, the variances of the ε_j can vary greatly if the diagonal elements of $\mathbf{H}$, the *leverages* h_{jj}, are substantially different. Consequently, many statisticians prefer graphical diagnostics based on studentized residuals. Using the residual mean square s^2 as an estimate of σ^2, we have

$$\widehat{\mathrm{Var}}(\hat{\varepsilon}_j) = s^2(1 - h_{jj}), \qquad j = 1, 2, \ldots, n \tag{7-17}$$

and the *studentized residuals* are

$$\hat{\varepsilon}_j^* = \frac{\hat{\varepsilon}_j}{\sqrt{s^2(1 - h_{jj})}}, \qquad j = 1, 2, \ldots, n \tag{7-18}$$

We expect the studentized residuals to look, approximately, like independent drawings from an $N(0, 1)$ distribution. Some software packages go one step further and studentize $\hat{\varepsilon}_j$ using the delete-one estimated variance $s^2(j)$, which is the residual mean square when the jth observation is dropped from the analysis.

Steps to Follow When Plotting Residuals as Diagnostic Tool

Residuals should be plotted in various ways to detect possible anomalies. For general diagnostic purposes, the following are useful graphs:

1. *Plot the residuals $\hat{\varepsilon}_j$ against the predicted values $\hat{y}_j = \hat{\beta}_0 + \hat{\beta}_1 z_{j1} + \cdots + \hat{\beta}_r z_{jr}$.* Departures from the assumptions of the model are typically indicated by two types of phenomena:
   - **(a)** *A dependence of the residuals on the predicted value.* This is illustrated in Figure 7.2(a). The numerical calculations are incorrect, or a β_0 term has been omitted from the model.
   - **(b)** *The variance is not constant.* The pattern of residuals may be funnel shaped, as in Figure 7.2(b), so that there is large variability for large $\hat{y}$ and small variability for small $\hat{y}$. If this is the case, the variance of the error is not constant, and transformations or a weighted least squares approach (or both) are required. (See Exercise 7.3.) In Figure 7.2(d), the residuals form a horizontal band. This is ideal and indicates equal variances and no dependence on $\hat{y}$.

Steps to Follow When Plotting Residuals as Diagnostic Tool

2. *Plot the residuals $\hat{\epsilon}_j$ against a predictor variable, such as z_1, or products of predictor variables, such as z_1^2 or $z_1 z_2$.* A systematic pattern in these plots suggests the need for more terms in the model. This situation is illustrated in Figure 7.2(c).
3. *Q–Q plots and histograms.* Do the errors appear to be normally distributed? To answer this question, the residuals $\hat{\epsilon}_j$ or $\hat{\epsilon}_j^*$ can be examined using the techniques discussed in Section 4.6. The *Q–Q* plots, histograms, and dot diagrams help to detect the presence of unusual observations or severe departures from normality that may require special attention in the analysis. If n is large, minor departures from normality will not greatly affect inferences about $\boldsymbol{\beta}$.

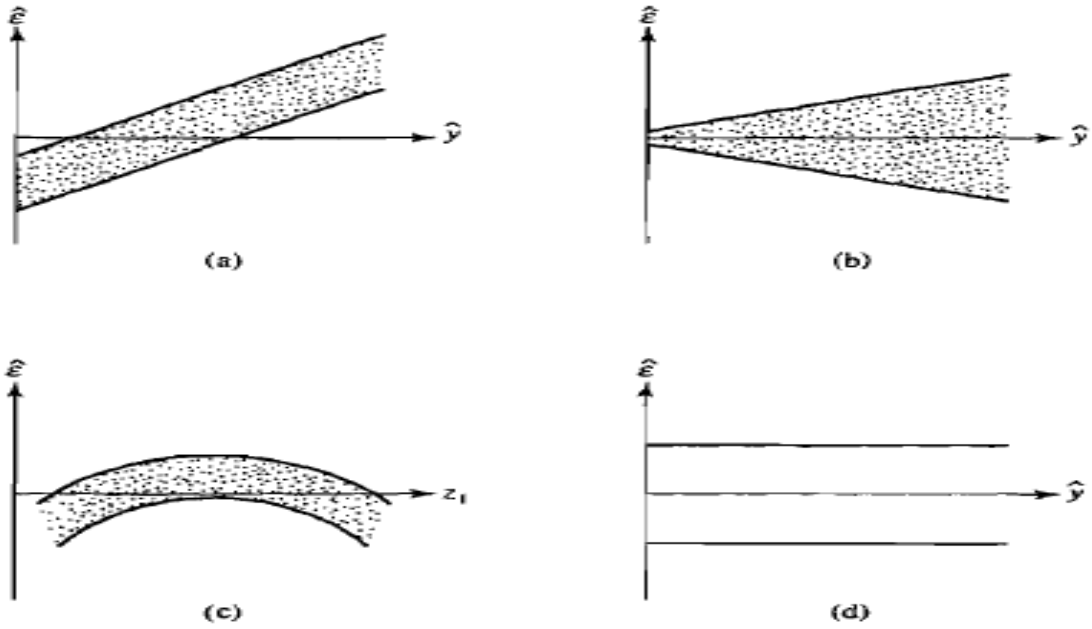


Figure 7.2 Residual plots.

Sample Residual Plots

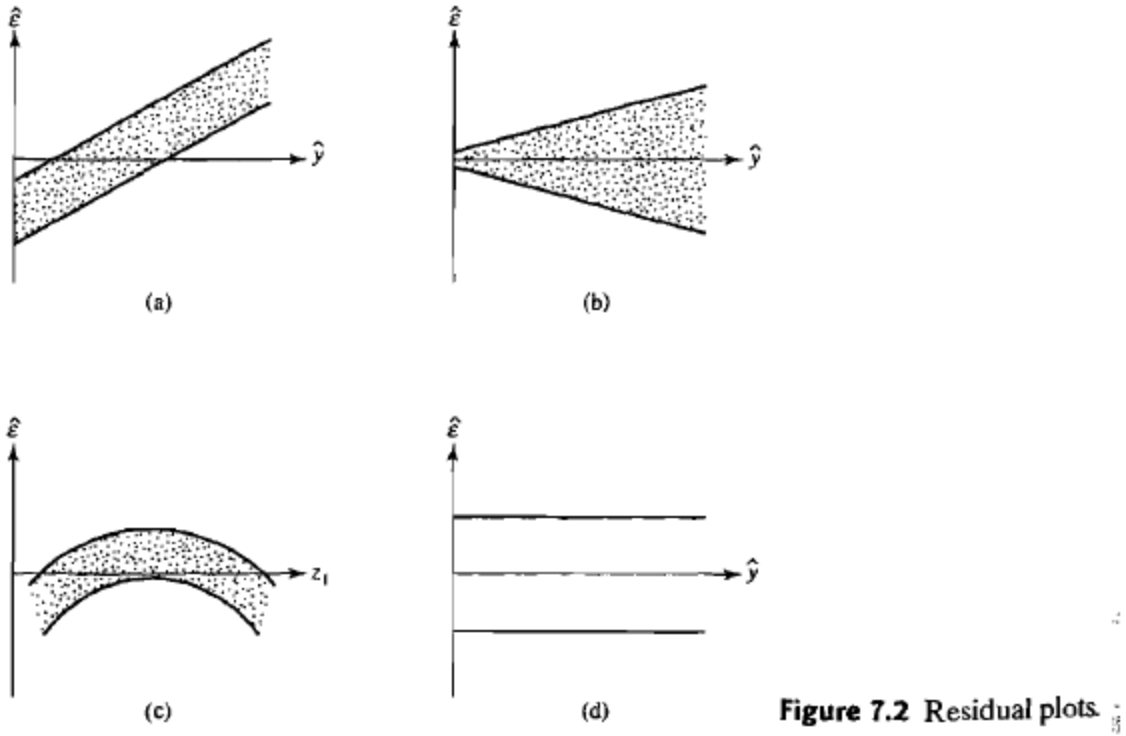


Figure 7.2 Residual plots.

Plot Residuals Versus Time (Chronological Sequence) to Demonstrate Independence

4. *Plot the residuals versus time.* The assumption of independence is crucial, but hard to check. If the data are naturally chronological, a plot of the residuals versus time may reveal a systematic pattern. (A plot of the positions of the residuals in space may also reveal associations among the errors.) For instance, residuals that increase over time indicate a strong positive dependence. A statistical test of independence can be constructed from the first autocorrelation,

$$r_1 = \frac{\sum_{j=2}^{n} \hat{\varepsilon}_j \hat{\varepsilon}_{j-1}}{\sum_{j=1}^{n} \hat{\varepsilon}_j^2} \tag{7-19}$$

of residuals from adjacent periods. A popular test based on the statistic $\sum_{j=2}^{n} (\hat{\varepsilon}_j - \hat{\varepsilon}_{j-1})^2 \Big/ \sum_{j=1}^{n} \hat{\varepsilon}_j^2 \doteq 2(1 - r_1)$ is called the *Durbin–Watson test*. (See [14] for a description of this test and tables of critical values.)

Residual Plots for the CPU Data

Example 7.7 (Residual plots) Three residual plots for the computer data discussed in Example 7.6 are shown in Figure 7.3. The sample size $n = 7$ is really too small to allow definitive judgments; however, it appears as if the regression assumptions are tenable. ■

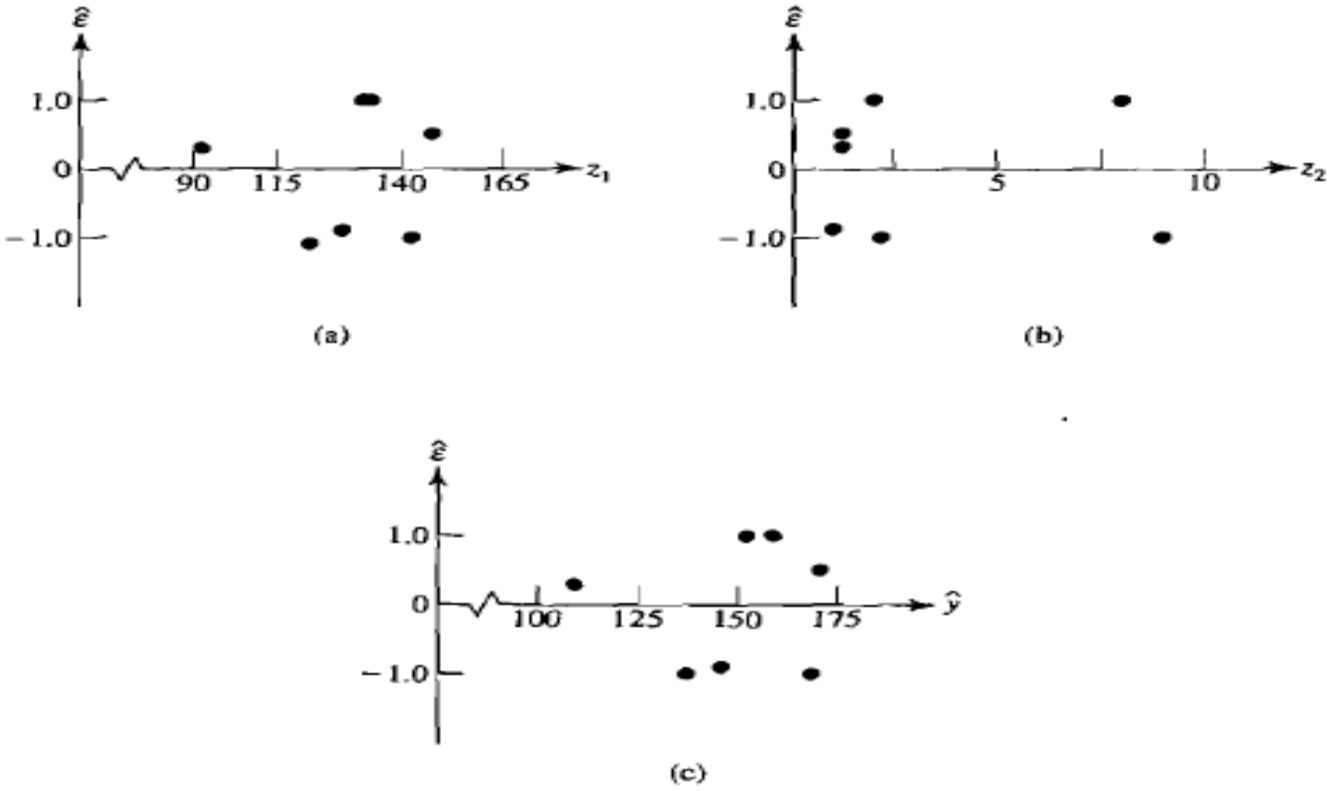


Figure 7.3 Residual plots for the computer data of Example 7.6.

Leverage and Influence

Leverage and Influence

Although a residual analysis is useful in assessing the fit of a model, departures from the regression model are often hidden by the fitting process. For example, there may be "outliers" in either the response or explanatory variables that can have a considerable effect on the analysis yet are not easily detected from an examination of residual plots. In fact, these outliers may *determine* the fit.

The leverage h_{jj} the (j, j) diagonal element of $\mathbf{H} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}$, can be interpreted in two related ways. First, the leverage is associated with the jth data point measures, in the space of the explanatory variables, how far the jth observation is from the other $n - 1$ observations. For simple linear regression with one explanatory variable z,

$$h_{jj} = \frac{1}{n} + \frac{(z_j - \bar{z})^2}{\sum_{j=1}^{n} (z_j - \bar{z})^2}$$

The average leverage is $(r + 1)/n$. (See Exercise 7.8.)

Second, the leverage h_{jj}, is a measure of pull that a single case exerts on the fit. The vector of predicted values is

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

Leverage Points Major Impact on Predicted Value of Y

The vector of predicted values is

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

where the jth row expresses the fitted value $\hat{y}_j$ in terms of the observations as

$$\hat{y}_j = h_{jj}y_j + \sum_{k \neq j} h_{jk}\, y_k$$

Provided that all other y values are held fixed

$$(\text{change in } \hat{y}_j) = h_{jj}\,(\text{change in } y_j)$$

If the leverage is large relative to the other h_{jk}, then y_j will be a major contributor to the predicted value $\hat{y}_j$.

Observations that significantly affect inferences drawn from the data are said to be *influential*. Methods for assessing influence are typically based on the change in the vector of parameter estimates, $\hat{\boldsymbol{\beta}}$, when observations are deleted. Plots based upon leverage and influence statistics and their use in diagnostic checking of regression models are described in [3], [5], and [10]. These references are recommended for anyone involved in an analysis of regression models.

If, after the diagnostic checks, no serious violations of the assumptions are detected, we can make inferences about $\boldsymbol{\beta}$ and the future Y values with some assurance that we will not be misled.

Determining How Many Predictors to Use: Calculate R_2 or Mallow's C_p

Selecting predictor variables from a large set. In practice, it is often difficult to formulate an appropriate regression function immediately. Which predictor variables should be included? What form should the regression function take?

When the list of possible predictor variables is very large, not all of the variables can be included in the regression function. Techniques and computer programs designed to select the "best" subset of predictors are now readily available. The good ones try all subsets: z_1 alone, z_2 alone, . . . , z_1 and z_2, ... The best choice is decided by examining some criterion quantity like R^2. [See (7-9).] However, R^2 always increases with the inclusion of additional predictor variables. Although this problem can be circumvented by using the adjusted R^2, $\bar{R}^2 = 1 - (1 - R^2)(n - 1)/(n - r - 1)$, a better statistic for selecting variables seems to be Mallow's C_p statistic (see [12]),

$$C_p = \left(\frac{\text{residual sum of squares for subset model with } p \text{ parameters, including an intercept}}{(\text{residual variance for full model})} \right) - (n - 2p)$$

A plot of the pairs (p, C_p), one for each subset of predictors, will indicate models that forecast the observed responses well. Good models typically have (p, C_p) coordinates near the 45° line. In Figure 7.4, we have circled the point corresponding to the "best" subset of predictor variables.

Stepwise Regression Technique: Add One Variable At a Time and Evaluate R^2

If the list of predictor variables is very long, cost considerations limit the number of models that can be examined. Another approach, called *stepwise regression* (see [13]), attempts to select important predictors without considering all the possibilities.

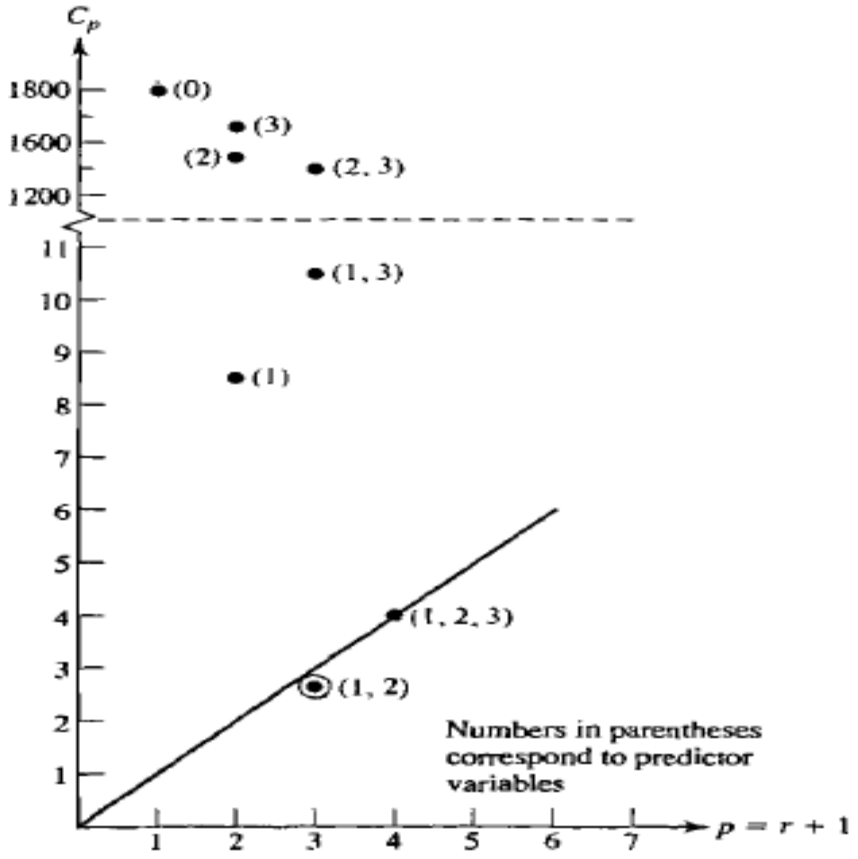


Figure 7.4 C_p plot for computer data from Example 7.6 with three predictor variables (z_1 = orders, z_2 = add–delete count, z_3 = number of items; see the example and original source).

Stepwise Regression Rubric

The procedure can be described by listing the basic steps (algorithm) involved in the computations:

Step 1. All possible *simple* linear regressions are considered. The predictor variable that explains the largest significant proportion of the variation in Y (the variable that has the largest correlation with the response) is the first variable to enter the regression function.

Step 2. The next variable to enter is the one (out of those not yet included) that makes the largest significant contribution to the regression sum of squares. The significance of the contribution is determined by an F-test. (See Result 7.6.) The value of the F-statistic that must be exceeded before the contribution of a variable is deemed significant is often called the *F to enter.*

Step 3. Once an additional variable has been included in the equation, the individual contributions to the regression sum of squares of the other variables already in the equation are checked for significance using F-tests. If the F-statistic is less than the one (called the *F to remove*) corresponding to a prescribed significance level, the variable is deleted from the regression function.

Step 4. Steps 2 and 3 are repeated until all possible additions are nonsignificant and all possible deletions are significant. At this point the selection stops.

Akaike's Information Criterion: Balances Size of Residual Sum of Squares with Number of Model Parameters

Because of the step-by-step procedure, there is no guarantee that this approach will select, for example, the best three variables for prediction. A second drawback is that the (automatic) selection methods are not capable of indicating when transformations of variables are useful.

Another popular criterion for selecting an appropriate model, called an information criterion, also balances the size of the residual sum of squares with the number of parameters in the model.

Akaike's information criterion (AIC) is

$$\text{AIC} = n \ln\left(\frac{\text{residual sum of squares for subset model with } p \text{ parameters, including an intercept}}{n}\right) + 2p$$

It is desirable that residual sum of squares be small, but the second term penalizes for too many parameters. Overall, we want to select models from those having the smaller values of AIC.

Collinearity: If Z not of Full Rank, then Columns Collinear, thus no inverse. To Offset Collinearity, Delete One or Two Predictors Strongly Correlated with Response

Colinearity. If $\mathbf{Z}$ is not of full rank, some linear combination, such as $\mathbf{Za}$, must equal $\mathbf{0}$. In this situation, the columns are said to be *colinear*. This implies that $\mathbf{Z'Z}$ does not have an inverse. For most regression analyses, it is unlikely that $\mathbf{Za} = \mathbf{0}$ exactly. Yet, if linear combinations of the columns of $\mathbf{Z}$ exist that are nearly $\mathbf{0}$, the calculation of $(\mathbf{Z'Z})^{-1}$ is numerically unstable. Typically, the diagonal entries of $(\mathbf{Z'Z})^{-1}$ will be large. This yields large estimated variances for the $\hat{\beta}_i$'s and it is then difficult to detect the "significant" regression coefficients $\hat{\beta}_i$. The problems caused by colinearity can be overcome somewhat by (1) deleting one of a pair of predictor variables that are *strongly* correlated or (2) relating the response Y to the *principal components* of the predictor variables—that is, the rows $\mathbf{z}_j'$ of $\mathbf{Z}$ are treated as a sample, and the first few principal components are calculated as is subsequently described in Section 8.3. The response Y is then regressed on these new predictor variables.

Bias Caused by Misspecified Model: Omitting Important Predictors

Bias caused by a misspecified model. Suppose some important predictor variables are omitted from the proposed regression model. That is, suppose the true model has $\mathbf{Z} = [\mathbf{Z}_1 \vdots \mathbf{Z}_2]$ with rank $r + 1$ and

$$\underset{(n\times 1)}{\mathbf{Y}} = \left[\underset{(n\times(q+1))}{\mathbf{Z}_1} \vdots \underset{(n\times(r-q))}{\mathbf{Z}_2}\right] \begin{bmatrix} \underset{((q+1)\times 1)}{\boldsymbol{\beta}_{(1)}} \\ \cdots\cdots \\ \underset{((r-q)\times 1)}{\boldsymbol{\beta}_{(2)}} \end{bmatrix} + \underset{(n\times 1)}{\boldsymbol{\varepsilon}} \tag{7-20}$$

$$= \mathbf{Z}_1\boldsymbol{\beta}_{(1)} + \mathbf{Z}_2\boldsymbol{\beta}_{(2)} + \boldsymbol{\varepsilon}$$

where $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $\text{Var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$. However, the investigator unknowingly fits a model using only the first q predictors by minimizing the error sum of squares $(\mathbf{Y} - \mathbf{Z}_1\boldsymbol{\beta}_{(1)})'(\mathbf{Y} - \mathbf{Z}_1\boldsymbol{\beta}_{(1)})$. The least squares estimator of $\boldsymbol{\beta}_{(1)}$ is $\hat{\boldsymbol{\beta}}_{(1)} = (\mathbf{Z}_1'\mathbf{Z}_1)^{-1}\mathbf{Z}_1'\mathbf{Y}$. Then, unlike the situation when the model is correct,

$$E(\hat{\boldsymbol{\beta}}_{(1)}) = (\mathbf{Z}_1'\mathbf{Z}_1)^{-1}\mathbf{Z}_1'E(\mathbf{Y}) = (\mathbf{Z}_1'\mathbf{Z}_1)^{-1}\mathbf{Z}_1'(\mathbf{Z}_1\boldsymbol{\beta}_{(1)} + \mathbf{Z}_2\boldsymbol{\beta}_{(2)} + E(\boldsymbol{\varepsilon}))$$

$$= \boldsymbol{\beta}_{(1)} + (\mathbf{Z}_1'\mathbf{Z}_1)^{-1}\mathbf{Z}_1'\mathbf{Z}_2\boldsymbol{\beta}_{(2)} \tag{7-21}$$

That is, $\hat{\boldsymbol{\beta}}_{(1)}$ is a biased estimator of $\boldsymbol{\beta}_{(1)}$ unless the columns of $\mathbf{Z}_1$ are perpendicular to those of $\mathbf{Z}_2$ (that is, $\mathbf{Z}_1'\mathbf{Z}_2 = \mathbf{0}$). If important variables are missing from the model, the least squares estimates $\hat{\boldsymbol{\beta}}_{(1)}$ may be misleading.

Multivariate Multiple Regression

7.7 Multivariate Multiple Regression

In this section, we consider the problem of modeling the relationship between m responses $Y_1, Y_2, \ldots, Y_m$ and a single set of predictor variables $z_1, z_2, \ldots, z_r$. Each response is assumed to follow its own regression model, so that

$$\begin{aligned} Y_1 &= \beta_{01} + \beta_{11}z_1 + \cdots + \beta_{r1}z_r + \varepsilon_1 \\ Y_2 &= \beta_{02} + \beta_{12}z_1 + \cdots + \beta_{r2}z_r + \varepsilon_2 \\ &\vdots \qquad\qquad \vdots \\ Y_m &= \beta_{0m} + \beta_{1m}z_1 + \cdots + \beta_{rm}z_r + \varepsilon_m \end{aligned} \tag{7-22}$$

The error term $\boldsymbol{\varepsilon}' = [\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_m]$ has $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $\text{Var}(\boldsymbol{\varepsilon}) = \boldsymbol{\Sigma}$. Thus, the error terms associated with different responses may be correlated.

To establish notation conforming to the classical linear regression model, let $[z_{j0}, z_{j1}, \ldots, z_{jr}]$ denote the values of the predictor variables for the jth trial, let $\mathbf{Y}_j' = [Y_{j1}, Y_{j2}, \ldots, Y_{jm}]$ be the responses, and let $\boldsymbol{\varepsilon}_j' = [\varepsilon_{j1}, \varepsilon_{j2}, \ldots, \varepsilon_{jm}]$ be the errors. In matrix notation, the design matrix

$$\underset{(n\times(r+1))}{\mathbf{Z}} = \begin{bmatrix} z_{10} & z_{11} & \cdots & z_{1r} \\ z_{20} & z_{21} & \cdots & z_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n0} & z_{n1} & \cdots & z_{nr} \end{bmatrix}$$

Matrix Notation Multivariate Regression

$$\underset{(n\times(r+1))}{\mathbf{Z}} = \begin{bmatrix} z_{10} & z_{11} & \cdots & z_{1r} \\ z_{20} & z_{21} & \cdots & z_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n0} & z_{n1} & \cdots & z_{nr} \end{bmatrix}$$

is the same as that for the single-response regression model. [See (7-3).] The other matrix quantities have multivariate counterparts. Set

$$\underset{(n\times m)}{\mathbf{Y}} = \begin{bmatrix} Y_{11} & Y_{12} & \cdots & Y_{1m} \\ Y_{21} & Y_{22} & \cdots & Y_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{n1} & Y_{n2} & \cdots & Y_{nm} \end{bmatrix} = [\mathbf{Y}_{(1)} \vdots \mathbf{Y}_{(2)} \vdots \cdots \vdots \mathbf{Y}_{(m)}]$$

$$\underset{((r+1)\times m)}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_{01} & \beta_{02} & \cdots & \beta_{0m} \\ \beta_{11} & \beta_{12} & \cdots & \beta_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{r1} & \beta_{r2} & \cdots & \beta_{rm} \end{bmatrix} = [\boldsymbol{\beta}_{(1)} \vdots \boldsymbol{\beta}_{(2)} \vdots \cdots \vdots \boldsymbol{\beta}_{(m)}]$$

$$\underset{(n\times m)}{\boldsymbol{\varepsilon}} = \begin{bmatrix} \varepsilon_{11} & \varepsilon_{12} & \cdots & \varepsilon_{1m} \\ \varepsilon_{21} & \varepsilon_{22} & \cdots & \varepsilon_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \varepsilon_{n1} & \varepsilon_{n2} & \cdots & \varepsilon_{nm} \end{bmatrix} = [\boldsymbol{\varepsilon}_{(1)} \vdots \boldsymbol{\varepsilon}_{(2)} \vdots \cdots \vdots \boldsymbol{\varepsilon}_{(m)}]$$

$$= \begin{bmatrix} \boldsymbol{\varepsilon}_1' \\ \boldsymbol{\varepsilon}_2' \\ \vdots \\ \boldsymbol{\varepsilon}_n' \end{bmatrix}$$

Multivariate Linear Regression Model

The *multivariate linear regression model* is

$$\underset{(n\times m)}{\mathbf{Y}} = \underset{(n\times(r+1))}{\mathbf{Z}}\ \underset{((r+1)\times m)}{\boldsymbol{\beta}} + \underset{(n\times m)}{\boldsymbol{\varepsilon}}$$

with (7-23)

$$E(\boldsymbol{\varepsilon}_{(i)}) = \mathbf{0} \quad \text{and} \quad \text{Cov}(\boldsymbol{\varepsilon}_{(i)}, \boldsymbol{\varepsilon}_{(k)}) = \sigma_{ik}\mathbf{I} \qquad i,k = 1,2,\ldots,m$$

The m observations on the jth trial have covariance matrix $\boldsymbol{\Sigma} = \{\sigma_{ik}\}$, but observations from different trials are uncorrelated. Here $\boldsymbol{\beta}$ and σ_{ik} are unknown parameters; the design matrix $\mathbf{Z}$ has jth row $[z_{j0}, z_{j1}, \ldots, z_{jr}]$.

Simply stated, the ith response $\mathbf{Y}_{(i)}$ follows the linear regression model

$$\mathbf{Y}_{(i)} = \mathbf{Z}\boldsymbol{\beta}_{(i)} + \boldsymbol{\varepsilon}_{(i)}, \qquad i = 1,2,\ldots,m \tag{7-24}$$

with $\text{Cov}(\boldsymbol{\varepsilon}_{(i)}) = \sigma_{ii}\mathbf{I}$. However, the errors for *different* responses on the *same* trial can be correlated.

Given the outcomes $\mathbf{Y}$ and the values of the predictor variables $\mathbf{Z}$ with full column rank, we determine the least squares estimates $\hat{\boldsymbol{\beta}}_{(i)}$ exclusively from the observations $\mathbf{Y}_{(i)}$ on the ith response. In conformity with the single-response solution, we take

$$\hat{\boldsymbol{\beta}}_{(i)} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}_{(i)} \tag{7-25}$$

Generalized Variance Minimized by Least Squares Estimates of the Beta Coefficients or Predictors

Collecting these univariate least squares estimates, we obtain

$$\hat{\boldsymbol{\beta}} = [\hat{\boldsymbol{\beta}}_{(1)} \mid \hat{\boldsymbol{\beta}}_{(2)} \mid \cdots \mid \hat{\boldsymbol{\beta}}_{(m)}] = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'[\mathbf{Y}_{(1)} \mid \mathbf{Y}_{(2)} \mid \cdots \mid \mathbf{Y}_{(m)}]$$

or

$$\hat{\boldsymbol{\beta}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y} \tag{7-26}$$

For any choice of parameters $\mathbf{B} = [\mathbf{b}_{(1)} \mid \mathbf{b}_{(2)} \mid \cdots \mid \mathbf{b}_{(m)}]$, the matrix of errors is $\mathbf{Y} - \mathbf{ZB}$. The error sum of squares and cross products matrix is

$$(\mathbf{Y} - \mathbf{ZB})'(\mathbf{Y} - \mathbf{ZB}) = \begin{bmatrix} (\mathbf{Y}_{(1)} - \mathbf{Z}\mathbf{b}_{(1)})'(\mathbf{Y}_{(1)} - \mathbf{Z}\mathbf{b}_{(1)}) & \cdots & (\mathbf{Y}_{(1)} - \mathbf{Z}\mathbf{b}_{(1)})'(\mathbf{Y}_{(m)} - \mathbf{Z}\mathbf{b}_{(m)}) \\ \vdots & & \vdots \\ (\mathbf{Y}_{(m)} - \mathbf{Z}\mathbf{b}_{(m)})'(\mathbf{Y}_{(1)} - \mathbf{Z}\mathbf{b}_{(1)}) & \cdots & (\mathbf{Y}_{(m)} - \mathbf{Z}\mathbf{b}_{(m)})'(\mathbf{Y}_{(m)} - \mathbf{Z}\mathbf{b}_{(m)}) \end{bmatrix} \tag{7-27}$$

The selection $\mathbf{b}_{(i)} = \hat{\boldsymbol{\beta}}_{(i)}$ minimizes the ith diagonal sum of squares $(\mathbf{Y}_{(i)} - \mathbf{Z}\mathbf{b}_{(i)})'(\mathbf{Y}_{(i)} - \mathbf{Z}\mathbf{b}_{(i)})$. Consequently, $\operatorname{tr}[(\mathbf{Y} - \mathbf{ZB})'(\mathbf{Y} - \mathbf{ZB})]$ is minimized by the choice $\mathbf{B} = \hat{\boldsymbol{\beta}}$. Also, the generalized variance $|(\mathbf{Y} - \mathbf{ZB})'(\mathbf{Y} - \mathbf{ZB})|$ is minimized by the least squares estimates $\hat{\boldsymbol{\beta}}$. (See Exercise 7.11 for an additional generalized sum of squares property.)

Orthogonality Conditions Must Hold Among Residuals, Predicted Values and Columns of Z

Using the least squares estimates $\boldsymbol{\beta}$, we can form the matrices of

$$\text{Predicted values:} \quad \hat{\mathbf{Y}} = \mathbf{Z}\hat{\boldsymbol{\beta}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}$$

$$\text{Residuals:} \quad \hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \hat{\mathbf{Y}} = [\mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}']\mathbf{Y} \tag{7-28}$$

The orthogonality conditions among the residuals, predicted values, and columns of $\mathbf{Z}$, which hold in classical linear regression, hold in multivariate multiple regression. They follow from $\mathbf{Z}'[\mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'] = \mathbf{Z}' - \mathbf{Z}' = \mathbf{0}$. Specifically,

$$\mathbf{Z}'\hat{\boldsymbol{\varepsilon}} = \mathbf{Z}'[\mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}']\mathbf{Y} = \mathbf{0} \tag{7-29}$$

so the residuals $\hat{\boldsymbol{\varepsilon}}_{(i)}$ are perpendicular to the columns of $\mathbf{Z}$. Also,

$$\hat{\mathbf{Y}}'\hat{\boldsymbol{\varepsilon}} = \hat{\boldsymbol{\beta}}'\mathbf{Z}'[\mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}']\mathbf{Y} = \mathbf{0} \tag{7-30}$$

confirming that the predicted values $\hat{\mathbf{Y}}_{(i)}$ are perpendicular to all residual vectors $\hat{\boldsymbol{\varepsilon}}_{(k)}$. Because $\mathbf{Y} = \hat{\mathbf{Y}} + \hat{\boldsymbol{\varepsilon}}$,

$$\mathbf{Y}'\mathbf{Y} = (\hat{\mathbf{Y}} + \hat{\boldsymbol{\varepsilon}})'(\hat{\mathbf{Y}} + \hat{\boldsymbol{\varepsilon}}) = \hat{\mathbf{Y}}'\hat{\mathbf{Y}} + \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} + \mathbf{0} + \mathbf{0}'$$

or

$$\begin{matrix} \mathbf{Y}'\mathbf{Y} & = & \hat{\mathbf{Y}}'\hat{\mathbf{Y}} & + & \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} \\ \begin{pmatrix}\text{total sum of squares}\\ \text{and cross products}\end{pmatrix} & = & \begin{pmatrix}\text{predicted sum of squares}\\ \text{and cross products}\end{pmatrix} & + & \begin{pmatrix}\text{residual (error) sum}\\ \text{of squares and}\\ \text{cross products}\end{pmatrix} \end{matrix} \tag{7-31}$$

Fitting Multivariate Straight-Line Regression Model

The residual sum of squares and cross products can also be written as

$$\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = \mathbf{Y}'\mathbf{Y} - \hat{\mathbf{Y}}'\hat{\mathbf{Y}} = \mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\beta}}'\mathbf{Z}'\mathbf{Z}\hat{\boldsymbol{\beta}} \tag{7-32}$$

Example 7.8 (Fitting a multivariate straight-line regression model) To illustrate the calculations of $\hat{\boldsymbol{\beta}}$, $\hat{\mathbf{Y}}$, and $\hat{\boldsymbol{\varepsilon}}$, we fit a straight-line regression model (see Panel 7.2),

$$Y_{j1} = \beta_{01} + \beta_{11}z_{j1} + \varepsilon_{j1}$$
$$Y_{j2} = \beta_{02} + \beta_{12}z_{j1} + \varepsilon_{j2}, \qquad j = 1, 2, \ldots, 5$$

to two responses Y_1 and Y_2 using the data in Example 7.3. These data, augmented by observations on an additional response, are as follows:

z_1	0	1	2	3	4
y_1	1	4	3	8	9
y_2	−1	−1	2	3	2

The design matrix $\mathbf{Z}$ remains unchanged from the single-response problem. We find that

$$\mathbf{Z}' = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 4 \end{bmatrix} \qquad (\mathbf{Z}'\mathbf{Z})^{-1} = \begin{bmatrix} .6 & -.2 \\ -.2 & .1 \end{bmatrix}$$

SAS Output for General Linear Model: Regression Equation

PANEL 7.2 SAS ANALYSIS FOR EXAMPLE 7.8 USING PROC. GLM.

```
title 'Multivariate Regression Analysis';
data mra;
infile 'Example 7-8 data;
input y1 y2 z1;
proc glm data = mra;
model y1 y2 = z1/ss3;
manova h = z1/printe;
```

PROGRAM COMMANDS

General Linear Models Procedure

Dependent Variable: Y1

OUTPUT

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	40.00000000	40.00000000	20.00	0.0208
Error	3	6.00000000	2.00000000		
Corrected Total	4	46.00000000			

R-Square	C.V.	Root MSE	Y1 Mean
0.869565	28.28427	1.414214	5.00000000

(continues on next page)

SAS Output for General Linear Model: Regression Equation (cont).

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Z1	1	40.00000000	40.00000000	20.00	0.0208

Parameter	Estimate	T for H0: Parameter = 0	Pr > |T|	Std Error of Estimate
INTERCEPT	1.000000000	0.91	0.4286	1.09544512
Z1	2.000000000	4.47	0.0208	0.44721360

Dependent Variable: Y2

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	10.00000000	10.00000000	7.50	0.0714
Error	3	4.00000000	1.33333333		
Corrected Total	4	14.00000000			

R-Square	C.V.	Root MSE	Y2 Mean
0.714286	115.4701	1.154701	1.00000000

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Z1	1	10.00000000	10.00000000	7.50	0.0714

Parameter	Estimate	T for H0: Parameter = 0	Pr > |T|	Std Error of Estimate
INTERCEPT	−1.000000000	−1.12	0.3450	0.89442719
Z1	1.000000000	2.74	0.0714	0.36514837

SAS Output for Regression Equation: Fit Statistics

E = Error SS & CP Matrix

	Y1	Y2
Y1	6	−2
Y2	−2	4

Manova Test Criteria and Exact F Statistics for
the Hypothesis of no Overall Z1 Effect
H = Type III SS&CP Matrix for Z1 E = Error SS&CP Matrix
S = 1 M = 0 N = 0

Statistic	Value	F	Num DF	Den DF	Pr > F
Wilks' Lambda	0.06250000	15.0000	2	2	0.0625
Pillai's Trace	0.93750000	15.0000	2	2	0.0625
Hotelling-Lawley Trace	15.00000000	15.0000	2	2	0.0625
Roy's Greatest Root	15.00000000	15.0000	2	2	0.0625

Linear Algebra for Multivariate Regression Equation

and

$$\mathbf{Z}'\mathbf{y}_{(2)} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 4 \end{bmatrix} \begin{bmatrix} -1 \\ -1 \\ 2 \\ 3 \\ 2 \end{bmatrix} = \begin{bmatrix} 5 \\ 20 \end{bmatrix}$$

so

$$\hat{\boldsymbol{\beta}}_{(2)} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}_{(2)} = \begin{bmatrix} .6 & -.2 \\ -.2 & .1 \end{bmatrix} \begin{bmatrix} 5 \\ 20 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

From Example 7.3,

$$\hat{\boldsymbol{\beta}}_{(1)} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}_{(1)} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

Hence,

$$\hat{\boldsymbol{\beta}} = [\hat{\boldsymbol{\beta}}_{(1)} \mid \hat{\boldsymbol{\beta}}_{(2)}] = \begin{bmatrix} 1 & -1 \\ 2 & 1 \end{bmatrix} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'[\mathbf{y}_{(1)} \mid \mathbf{y}_{(2)}]$$

The fitted values are generated from $\hat{y}_1 = 1 + 2z_1$ and $\hat{y}_2 = -1 + z_2$. Collectively,

$$\hat{\mathbf{Y}} = \mathbf{Z}\hat{\boldsymbol{\beta}} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 2 & 1 \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ 3 & 0 \\ 5 & 1 \\ 7 & 2 \\ 9 & 3 \end{bmatrix}$$

and

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \hat{\mathbf{Y}} = \begin{bmatrix} 0 & 1 & -2 & 1 & 0 \\ 0 & -1 & 1 & 1 & -1 \end{bmatrix}'$$

Linear Algebra for Multivariate Regression Equation

and

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \hat{\mathbf{Y}} = \begin{bmatrix} 0 & 1 & -2 & 1 & 0 \\ 0 & -1 & 1 & 1 & -1 \end{bmatrix}'$$

Note that

$$\hat{\boldsymbol{\varepsilon}}'\hat{\mathbf{Y}} = \begin{bmatrix} 0 & 1 & -2 & 1 & 0 \\ 0 & -1 & 1 & 1 & -1 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 3 & 0 \\ 5 & 1 \\ 7 & 2 \\ 9 & 3 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

Since

$$\mathbf{Y}'\mathbf{Y} = \begin{bmatrix} 1 & 4 & 3 & 8 & 9 \\ -1 & -1 & 2 & 3 & 2 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 4 & -1 \\ 3 & 2 \\ 8 & 3 \\ 9 & 2 \end{bmatrix} = \begin{bmatrix} 171 & 43 \\ 43 & 19 \end{bmatrix}$$

$$\hat{\mathbf{Y}}'\hat{\mathbf{Y}} = \begin{bmatrix} 165 & 45 \\ 45 & 15 \end{bmatrix} \quad \text{and} \quad \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = \begin{bmatrix} 6 & -2 \\ -2 & 4 \end{bmatrix}$$

the sum of squares and cross-products decomposition

$$\mathbf{Y}'\mathbf{Y} = \hat{\mathbf{Y}}'\hat{\mathbf{Y}} + \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$$

is easily verified. ■

For Least Squares Estimator: Predicted Error and Beta Coefficients Both Uncorrelated

Result 7.9. For the least squares estimator $\hat{\boldsymbol{\beta}} = [\hat{\boldsymbol{\beta}}_{(1)} \mid \hat{\boldsymbol{\beta}}_{(2)} \mid \cdots \mid \hat{\boldsymbol{\beta}}_{(m)}]$ determined under the multivariate multiple regression model (7-23) with full rank $(\mathbf{Z}) = r + 1 < n$,

$$E(\hat{\boldsymbol{\beta}}_{(i)}) = \boldsymbol{\beta}_{(i)} \quad \text{or} \quad E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$$

and

$$\text{Cov}(\hat{\boldsymbol{\beta}}_{(i)}, \hat{\boldsymbol{\beta}}_{(k)}) = \sigma_{ik}(\mathbf{Z}'\mathbf{Z})^{-1}, \qquad i, k = 1, 2, \ldots, m$$

The residuals $\hat{\boldsymbol{\varepsilon}} = [\hat{\boldsymbol{\varepsilon}}_{(1)} \mid \hat{\boldsymbol{\varepsilon}}_{(2)} \mid \cdots \mid \hat{\boldsymbol{\varepsilon}}_{(m)}] = \mathbf{Y} - \mathbf{Z}\hat{\boldsymbol{\beta}}$ satisfy $E(\hat{\boldsymbol{\varepsilon}}_{(i)}) = \mathbf{0}$ and $E(\hat{\boldsymbol{\varepsilon}}'_{(i)}\hat{\boldsymbol{\varepsilon}}_{(k)}) = (n - r - 1)\sigma_{ik}$, so

$$E(\hat{\boldsymbol{\varepsilon}}) = \mathbf{0} \quad \text{and} \quad E\left(\frac{1}{n - r - 1}\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}\right) = \boldsymbol{\Sigma}$$

Also, $\hat{\boldsymbol{\varepsilon}}$ and $\hat{\boldsymbol{\beta}}$ are uncorrelated.

Mean Vectors and Covariance Matrices Are Unbiased Estimators of Fitted Regression Relationship

The mean vectors and covariance matrices determined in Result 7.9 enable us to obtain the sampling properties of the least squares predictors.

We first consider the problem of estimating the mean vector when the predictor variables have the values $\mathbf{z}_0' = [1, z_{01}, \ldots, z_{0r}]$. The mean of the ith response variable is $\mathbf{z}_0'\boldsymbol{\beta}_{(i)}$, and this is estimated by $\mathbf{z}_0'\hat{\boldsymbol{\beta}}_{(i)}$, the ith component of the fitted regression relationship. Collectively,

$$\mathbf{z}_0'\hat{\boldsymbol{\beta}} = [\mathbf{z}_0'\hat{\boldsymbol{\beta}}_{(1)} \vdots \mathbf{z}_0'\hat{\boldsymbol{\beta}}_{(2)} \vdots \cdots \vdots \mathbf{z}_0'\hat{\boldsymbol{\beta}}_{(m)}] \tag{7-34}$$

is an unbiased estimator $\mathbf{z}_0'\boldsymbol{\beta}$ since $E(\mathbf{z}_0'\hat{\boldsymbol{\beta}}_{(i)}) = \mathbf{z}_0'E(\hat{\boldsymbol{\beta}}_{(i)}) = \mathbf{z}_0'\boldsymbol{\beta}_{(i)}$ for each component. From the covariance matrix for $\hat{\boldsymbol{\beta}}_{(i)}$ and $\hat{\boldsymbol{\beta}}_{(k)}$, the estimation errors $\mathbf{z}_0'\boldsymbol{\beta}_{(i)} - \mathbf{z}_0'\hat{\boldsymbol{\beta}}_{(i)}$ have covariances

$$\begin{aligned} E[\mathbf{z}_0'(\boldsymbol{\beta}_{(i)} - \hat{\boldsymbol{\beta}}_{(i)})(\boldsymbol{\beta}_{(k)} - \hat{\boldsymbol{\beta}}_{(k)})'\mathbf{z}_0] &= \mathbf{z}_0'(E(\boldsymbol{\beta}_{(i)} - \hat{\boldsymbol{\beta}}_{(i)})(\boldsymbol{\beta}_{(k)} - \hat{\boldsymbol{\beta}}_{(k)})')\mathbf{z}_0 \\ &= \sigma_{ik}\mathbf{z}_0'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{z}_0 \end{aligned} \tag{7-35}$$

Maximum Likelihood Estimators and Their Distributions Obtained When Errors Have Normal Distribution

The related problem is that of forecasting a *new* observation vector $\mathbf{Y}_0' = [Y_{01}, Y_{02}, \ldots, Y_{0m}]$ at $\mathbf{z}_0$. According to the regression model, $Y_{0i} = \mathbf{z}_0'\boldsymbol{\beta}_{(i)} + \varepsilon_{0i}$ where the "new" error $\boldsymbol{\varepsilon}_0' = [\varepsilon_{01}, \varepsilon_{02}, \ldots, \varepsilon_{0m}]$ is independent of the errors $\boldsymbol{\varepsilon}$ and satisfies $E(\varepsilon_{0i}) = 0$ and $E(\varepsilon_{0i}\varepsilon_{0k}) = \sigma_{ik}$. The *forecast error* for the ith component of $\mathbf{Y}_0$ is

$$\begin{aligned} Y_{0i} - \mathbf{z}_0'\hat{\boldsymbol{\beta}}_{(i)} &= Y_{0i} - \mathbf{z}_0'\boldsymbol{\beta}_{(i)} + \mathbf{z}_0'\boldsymbol{\beta}_{(i)} - \mathbf{z}_0'\hat{\boldsymbol{\beta}}_{(i)} \\ &= \varepsilon_{0i} - \mathbf{z}_0'(\hat{\boldsymbol{\beta}}_{(i)} - \boldsymbol{\beta}_{(i)}) \end{aligned}$$

so $E(Y_{0i} - \mathbf{z}_0'\hat{\boldsymbol{\beta}}_{(i)}) = E(\varepsilon_{0i}) - \mathbf{z}_0'E(\hat{\boldsymbol{\beta}}_{(i)} - \boldsymbol{\beta}_{(i)}) = 0$, indicating that $\mathbf{z}_0'\hat{\boldsymbol{\beta}}_{(i)}$ is an *unbiased predictor* of Y_{0i}. The forecast errors have covariances

$$\begin{aligned} &E(Y_{0i} - \mathbf{z}_0'\hat{\boldsymbol{\beta}}_{(i)})(Y_{0k} - \mathbf{z}_0'\hat{\boldsymbol{\beta}}_{(k)}) \\ &\quad = E(\varepsilon_{0i} - \mathbf{z}_0'(\hat{\boldsymbol{\beta}}_{(i)} - \boldsymbol{\beta}_{(i)}))(\varepsilon_{0k} - \mathbf{z}_0'(\hat{\boldsymbol{\beta}}_{(k)} - \boldsymbol{\beta}_{(k)})) \\ &\quad = E(\varepsilon_{0i}\varepsilon_{0k}) + \mathbf{z}_0'E(\hat{\boldsymbol{\beta}}_{(i)} - \boldsymbol{\beta}_{(i)})(\hat{\boldsymbol{\beta}}_{(k)} - \boldsymbol{\beta}_{(k)})'\mathbf{z}_0 \\ &\qquad - \mathbf{z}_0'E((\hat{\boldsymbol{\beta}}_{(i)} - \boldsymbol{\beta}_{(i)})\varepsilon_{0k}) - E(\varepsilon_{0i}(\hat{\boldsymbol{\beta}}_{(k)} - \boldsymbol{\beta}_{(k)})')\mathbf{z}_0 \\ &\quad = \sigma_{ik}(1 + \mathbf{z}_0'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{z}_0) \end{aligned} \tag{7-36}$$

Note that $E((\hat{\boldsymbol{\beta}}_{(i)} - \boldsymbol{\beta}_{(i)})\varepsilon_{0k}) = \mathbf{0}$ since $\hat{\boldsymbol{\beta}}_{(i)} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{\varepsilon}_{(i)} + \boldsymbol{\beta}_{(i)}$ is independent of $\boldsymbol{\varepsilon}_0$. A similar result holds for $E(\varepsilon_{0i}(\hat{\boldsymbol{\beta}}_{(k)} - \boldsymbol{\beta}_{(k)})')$.

Maximum likelihood estimators and their distributions can be obtained when the errors $\boldsymbol{\varepsilon}$ have a normal distribution.

Multivariate Multiple Regression Holds With Full Rank (Z)

Result 7.10. Let the multivariate multiple regression model in (7-23) hold with full rank $(\mathbf{Z}) = r + 1$, $n \geq (r + 1) + m$, and let the errors $\boldsymbol{\varepsilon}$ have a normal distribution. Then

$$\hat{\boldsymbol{\beta}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}$$

is the maximum likelihood estimator of $\boldsymbol{\beta}$ and $\hat{\boldsymbol{\beta}}$ has a normal distribution with $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ and $\text{Cov}(\hat{\boldsymbol{\beta}}_{(i)}, \hat{\boldsymbol{\beta}}_{(k)}) = \sigma_{ik}(\mathbf{Z}'\mathbf{Z})^{-1}$. Also, $\hat{\boldsymbol{\beta}}$ is independent of the maximum likelihood estimator of the positive definite $\boldsymbol{\Sigma}$ given by

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = \frac{1}{n}(\mathbf{Y} - \mathbf{Z}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{Z}\hat{\boldsymbol{\beta}})$$

and

$$n\hat{\boldsymbol{\Sigma}} \quad \text{is distributed as} \quad W_{p,n-r-1}(\boldsymbol{\Sigma})$$

The maximized likelihood $L(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) = (2\pi)^{-mn/2}|\hat{\boldsymbol{\Sigma}}|^{-n/2}e^{-mn/2}$.

Univariate and Bivariate Diagnostics Implemented in Advance of Regression Model Fitting – Check for Normality

Result 7.10 provides additional support for using least squares estimates. When the errors are normally distributed, $\hat{\boldsymbol{\beta}}$ and $n^{-1}\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}$ are the maximum likelihood estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$, respectively. Therefore, for large samples, they have nearly the smallest possible variances.

Comment. The multivariate multiple regression model poses no new computational problems. Least squares (maximum likelihood) estimates, $\hat{\boldsymbol{\beta}}_{(i)} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}_{(i)}$, are computed individually for each response variable. Note, however, that the model requires that the *same* predictor variables be used for all responses.

Once a multivariate multiple regression model has been fit to the data, it should be subjected to the diagnostic checks described in Section 7.6 for the single-response model. The residual vectors $[\hat{\epsilon}_{j1}, \hat{\epsilon}_{j2}, \ldots, \hat{\epsilon}_{jm}]$ can be examined for normality or outliers using the techniques in Section 4.6.

The remainder of this section is devoted to brief discussions of inference for the normal theory multivariate multiple regression model. Extended accounts of these procedures appear in [2] and [18].

Likelihood Ratio Tests: Responses do not Depend on Predictors (Hypothesis to Reject)

Likelihood Ratio Tests for Regression Parameters

The multiresponse analog of (7-12), the hypothesis that the responses do not depend on $z_{q+1}, z_{q+2}, \ldots, z_r$, becomes

$$H_0: \underset{((r-q)\times m)}{\boldsymbol{\beta}_{(2)}} = \mathbf{0} \quad \text{where} \quad \boldsymbol{\beta} = \left[\begin{array}{c} \underset{((q+1)\times m)}{\boldsymbol{\beta}_{(1)}} \\ \hline \underset{((r-q)\times m)}{\boldsymbol{\beta}_{(2)}} \end{array}\right] \tag{7-37}$$

Setting $\mathbf{Z} = \left[\underset{(n\times(q+1))}{\mathbf{Z}_1} \vdots \underset{(n\times(r-q))}{\mathbf{Z}_2}\right]$, we can write the general model as

$$E(\mathbf{Y}) = \mathbf{Z}\boldsymbol{\beta} = [\mathbf{Z}_1 \vdots \mathbf{Z}_2]\left[\begin{array}{c} \boldsymbol{\beta}_{(1)} \\ \hline \boldsymbol{\beta}_{(2)} \end{array}\right] = \mathbf{Z}_1\boldsymbol{\beta}_{(1)} + \mathbf{Z}_2\boldsymbol{\beta}_{(2)}$$

Likelihood Ratio Test: Use Wilks' Lambda Statistic

Under $H_0: \boldsymbol{\beta}_{(2)} = \mathbf{0}$, $\mathbf{Y} = \mathbf{Z}_1\boldsymbol{\beta}_{(1)} + \boldsymbol{\varepsilon}$ and the likelihood ratio test of H_0 is based on the quantities involved in the

extra sum of squares and cross products

$$= (\mathbf{Y} - \mathbf{Z}_1\hat{\boldsymbol{\beta}}_{(1)})'(\mathbf{Y} - \mathbf{Z}_1\hat{\boldsymbol{\beta}}_{(1)}) - (\mathbf{Y} - \mathbf{Z}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{Z}\hat{\boldsymbol{\beta}})$$
$$= n(\hat{\boldsymbol{\Sigma}}_1 - \hat{\boldsymbol{\Sigma}})$$

where $\hat{\boldsymbol{\beta}}_{(1)} = (\mathbf{Z}_1'\mathbf{Z}_1)^{-1}\mathbf{Z}_1'\mathbf{Y}$ and $\hat{\boldsymbol{\Sigma}}_1 = n^{-1}(\mathbf{Y} - \mathbf{Z}_1\hat{\boldsymbol{\beta}}_{(1)})'(\mathbf{Y} - \mathbf{Z}_1\hat{\boldsymbol{\beta}}_{(1)})$.

From Result 7.10, the likelihood ratio, Λ, can be expressed in terms of generalized variances:

$$\Lambda = \frac{\max_{\boldsymbol{\beta}_{(1)},\boldsymbol{\Sigma}} L(\boldsymbol{\beta}_{(1)}, \boldsymbol{\Sigma})}{\max_{\boldsymbol{\beta},\boldsymbol{\Sigma}} L(\boldsymbol{\beta}, \boldsymbol{\Sigma})} = \frac{L(\hat{\boldsymbol{\beta}}_{(1)}, \hat{\boldsymbol{\Sigma}}_1)}{L(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Sigma}})} = \left(\frac{|\hat{\boldsymbol{\Sigma}}|}{|\hat{\boldsymbol{\Sigma}}_1|}\right)^{n/2} \tag{7-38}$$

Equivalently, *Wilks' lambda statistic*

$$\Lambda^{2/n} = \frac{|\hat{\boldsymbol{\Sigma}}|}{|\hat{\boldsymbol{\Sigma}}_1|}$$

can be used.

Wilks Lambda Uses Chi-Square Distribution

Result 7.11. Let the multivariate multiple regression model of (7-23) hold with $\mathbf{Z}$ of full rank $r + 1$ and $(r + 1) + m \leq n$. Let the errors $\boldsymbol{\varepsilon}$ be normally distributed. Under $H_0: \boldsymbol{\beta}_{(2)} = \mathbf{0}$, $n\hat{\boldsymbol{\Sigma}}$ is distributed as $W_{p,n-r-1}(\boldsymbol{\Sigma})$ independently of $n(\hat{\boldsymbol{\Sigma}}_1 - \hat{\boldsymbol{\Sigma}})$, which, in turn, is distributed as $W_{p,r-q}(\boldsymbol{\Sigma})$. The likelihood ratio test of H_0 is equivalent to rejecting H_0 for large values of

$$-2 \ln \Lambda = -n \ln\left(\frac{|\hat{\boldsymbol{\Sigma}}|}{|\hat{\boldsymbol{\Sigma}}_1|}\right) = -n \ln \frac{|n\hat{\boldsymbol{\Sigma}}|}{|n\hat{\boldsymbol{\Sigma}} + n(\hat{\boldsymbol{\Sigma}}_1 - \hat{\boldsymbol{\Sigma}})|}$$

For n large,[5] the modified statistic

$$-\left[n - r - 1 - \frac{1}{2}(m - r + q + 1)\right] \ln\left(\frac{|\hat{\boldsymbol{\Sigma}}|}{|\hat{\boldsymbol{\Sigma}}_1|}\right)$$

has, to a close approximation, a chi-square distribution with $m(r - q)$ d.f.

Other Multivariate Test Statistics

Other Multivariate Test Statistics

Tests other than the likelihood ratio test have been proposed for testing $H_0: \boldsymbol{\beta}_{(2)} = \mathbf{0}$ in the multivariate multiple regression model.

Popular computer-package programs routinely calculate four multivariate test statistics. To connect with their output, we introduce some alternative notation. Let $\mathbf{E}$ be the $p \times p$ *error*, or residual, sum of squares and cross products matrix

$$\mathbf{E} = n\hat{\boldsymbol{\Sigma}}$$

that results from fitting the full model. The $p \times p$ *hypothesis*, or extra, sum of squares and cross-products matrix

$$\mathbf{H} = n(\hat{\boldsymbol{\Sigma}}_1 - \hat{\boldsymbol{\Sigma}})$$

The statistics can be defined in terms of $\mathbf{E}$ and $\mathbf{H}$ directly, or in terms of the nonzero eigenvalues $\eta_1 \geq \eta_2 \geq \cdots \geq \eta_s$ of $\mathbf{HE}^{-1}$, where $s = \min(p, r - q)$. Equivalently, they are the roots of $|(\hat{\boldsymbol{\Sigma}}_1 - \hat{\boldsymbol{\Sigma}}) - \eta\hat{\boldsymbol{\Sigma}}| = 0$. The definitions are

$$\text{Wilks' lambda} = \prod_{i=1}^{s} \frac{1}{1 + \eta_i} = \frac{|\mathbf{E}|}{|\mathbf{E} + \mathbf{H}|}$$

$$\text{Pillai's trace} = \sum_{i=1}^{s} \frac{\eta_i}{1 + \eta_i} = tr[\mathbf{H}(\mathbf{H} + \mathbf{E})^{-1}]$$

$$\text{Hotelling–Lawley trace} = \sum_{i=1}^{s} \eta_i = \text{tr}[\mathbf{HE}^{-1}]$$

$$\text{Roy's greatest root} = \frac{\eta_1}{1 + \eta_1}$$

Regression Fit Statistics Similar to Wilks Lambda or Likelihood Ratio Test

The statistics can be defined in terms of $\mathbf{E}$ and $\mathbf{H}$ directly, or in terms of the nonzero eigenvalues $\eta_1 \geq \eta_2 \geq \ldots \geq \eta_s$ of $\mathbf{HE}^{-1}$, where $s = \min(p, r-q)$. Equivalently, they are the roots of $|(\hat{\boldsymbol{\Sigma}}_1 - \hat{\boldsymbol{\Sigma}}) - \eta\hat{\boldsymbol{\Sigma}}| = 0$. The definitions are

$$\text{Wilks' lambda} = \prod_{i=1}^{s} \frac{1}{1+\eta_i} = \frac{|\mathbf{E}|}{|\mathbf{E}+\mathbf{H}|}$$

$$\text{Pillai's trace} = \sum_{i=1}^{s} \frac{\eta_i}{1+\eta_i} = tr[\mathbf{H}(\mathbf{H}+\mathbf{E})^{-1}]$$

$$\text{Hotelling–Lawley trace} = \sum_{i=1}^{s} \eta_i = \text{tr}[\mathbf{HE}^{-1}]$$

$$\text{Roy's greatest root} = \frac{\eta_1}{1+\eta_1}$$

Roy's test selects the coefficient vector $\mathbf{a}$ so that the univariate F-statistic based on $\mathbf{a}'\mathbf{Y}_j$ has its maximum possible value. When several of the eigenvalues η_i are moderately large, Roy's test will perform poorly relative to the other three. Simulation studies suggest that its power will be best when there is only one large eigenvalue.

Charts and tables of critical values are available for Roy's test. (See [21] and [17].) Wilks' lambda, Roy's greatest root, and the Hotelling–Lawley trace test are nearly equivalent for large sample sizes.

If there is a large discrepancy in the reported P-values for the four tests, the eigenvalues and vectors may lead to an interpretation. In this text, we report Wilks' lambda, which is the likelihood ratio test.

Concept of Linear Regression With Covariance Matrix at Full Rank

7.8 The Concept of Linear Regression

The classical linear regression model is concerned with the association between a single dependent variable Y and a collection of predictor variables $z_1, z_2, \ldots, z_r$. The regression model that we have considered treats Y as a random variable whose mean depends upon *fixed* values of the z_i's. This mean is assumed to be a linear function of the regression *coefficients* $\beta_0, \beta_1, \ldots, \beta_r$.

The linear regression model also arises in a different setting. Suppose all the variables $Y, Z_1, Z_2, \ldots, Z_r$ are random and have a joint distribution, not necessarily normal, with mean vector $\underset{(r+1)\times 1}{\boldsymbol{\mu}}$ and covariance matrix $\underset{(r+1)\times(r+1)}{\boldsymbol{\Sigma}}$. Partitioning $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in an obvious fashion, we write

$$\boldsymbol{\mu} = \begin{bmatrix} \underset{(1\times 1)}{\mu_Y} \\ \cdots\cdots \\ \underset{(r\times 1)}{\boldsymbol{\mu}_Z} \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \left[\begin{array}{c|c} \underset{(1\times 1)}{\sigma_{YY}} & \underset{(1\times r)}{\boldsymbol{\sigma}'_{ZY}} \\ \hline \underset{(r\times 1)}{\boldsymbol{\sigma}_{ZY}} & \underset{(r\times r)}{\boldsymbol{\Sigma}_{ZZ}} \end{array}\right]$$

with

$$\boldsymbol{\sigma}'_{ZY} = [\sigma_{YZ_1}, \sigma_{YZ_2}, \ldots, \sigma_{YZ_r}] \tag{7-44}$$

$\boldsymbol{\Sigma}_{ZZ}$ can be taken to have full rank.[6] Consider the problem of predicting Y using the

$$\textit{linear predictor} = b_0 + b_1 Z_1 + \cdots + b_r Z_r = b_0 + \mathbf{b}'\mathbf{Z} \tag{7-45}$$

[6] If $\boldsymbol{\Sigma}_{ZZ}$ is not of full rank, one variable—for example, Z_k—can be written as a linear combination of the other Z_i's and thus is redundant in forming the linear regression function $\mathbf{Z}'\boldsymbol{\beta}$. That is, $\mathbf{Z}$ may be replaced by any subset of components whose nonsingular covariance matrix has the same rank as $\boldsymbol{\Sigma}_{ZZ}$.

B_0 and B_1 Minimize Mean Square Error: Yield Optimal Linear Prediction

$\mathbf{\Sigma}_{ZZ}$ can be taken to have full rank.[6] Consider the problem of predicting Y using the

$$\textit{linear predictor} = b_0 + b_1 Z_1 + \cdots + b_r Z_r = b_0 + \mathbf{b}'\mathbf{Z} \qquad (7\text{-}45)$$

[6] If $\mathbf{\Sigma}_{ZZ}$ is not of full rank, one variable—for example, Z_k—can be written as a linear combination of the other Z_i's and thus is redundant in forming the linear regression function $\mathbf{Z}'\boldsymbol{\beta}$. That is, $\mathbf{Z}$ may be replaced by any subset of components whose nonsingular covariance matrix has the same rank as $\mathbf{\Sigma}_{ZZ}$.

For a given predictor of the form of (7-45), the error in the prediction of Y is

$$\textit{prediction error} = Y - b_0 - b_1 Z_1 - \cdots - b_r Z_r = Y - b_0 - \mathbf{b}'\mathbf{Z} \qquad (7\text{-}46)$$

Because this error is random, it is customary to select b_0 and $\mathbf{b}$ to minimize the

$$\textit{mean square error} = E(Y - b_0 - \mathbf{b}'\mathbf{Z})^2 \qquad (7\text{-}47)$$

Now the mean square error depends on the joint distribution of Y and $\mathbf{Z}$ only through the parameters $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$. It is possible to express the "optimal" linear predictor in terms of these latter quantities.

Linear Regression Line is Correlation Among Predictors and Response

Result 7.12. The linear predictor $\beta_0 + \boldsymbol{\beta}'\mathbf{Z}$ with coefficients

$$\boldsymbol{\beta} = \boldsymbol{\Sigma}_{\mathbf{ZZ}}^{-1}\boldsymbol{\sigma}_{\mathbf{ZY}}, \qquad \beta_0 = \mu_Y - \boldsymbol{\beta}'\boldsymbol{\mu}_{\mathbf{Z}}$$

has minimum mean square among all *linear* predictors of the response Y. Its mean square error is

$$E(Y - \beta_0 - \boldsymbol{\beta}'\mathbf{Z})^2 = E(Y - \mu_Y - \boldsymbol{\sigma}_{\mathbf{ZY}}'\boldsymbol{\Sigma}_{\mathbf{ZZ}}^{-1}(\mathbf{Z} - \boldsymbol{\mu}_{\mathbf{Z}}))^2 = \sigma_{YY} - \boldsymbol{\sigma}_{\mathbf{ZY}}'\boldsymbol{\Sigma}_{\mathbf{ZZ}}^{-1}\boldsymbol{\sigma}_{\mathbf{ZY}}$$

Also, $\beta_0 + \boldsymbol{\beta}'\mathbf{Z} = \mu_Y + \boldsymbol{\sigma}_{\mathbf{ZY}}'\boldsymbol{\Sigma}_{\mathbf{ZZ}}^{-1}(\mathbf{Z} - \boldsymbol{\mu}_{\mathbf{Z}})$ is the linear predictor having maximum correlation with Y; that is,

$$\operatorname{Corr}(Y, \beta_0 + \boldsymbol{\beta}'\mathbf{Z}) = \max_{b_0, \mathbf{b}} \operatorname{Corr}(Y, b_0 + \mathbf{b}'\mathbf{Z})$$

$$= \sqrt{\frac{\boldsymbol{\beta}'\boldsymbol{\Sigma}_{\mathbf{ZZ}}\boldsymbol{\beta}}{\sigma_{YY}}} = \sqrt{\frac{\boldsymbol{\sigma}_{\mathbf{ZY}}'\boldsymbol{\Sigma}_{\mathbf{ZZ}}^{-1}\boldsymbol{\sigma}_{\mathbf{ZY}}}{\sigma_{YY}}}$$

Population Coefficient of Determination is the R^2

The correlation between Y and its best linear predictor is called the *population multiple correlation coefficient*

$$\rho_{Y(\mathbf{Z})} = +\sqrt{\frac{\boldsymbol{\sigma}_{\mathbf{Z}Y}'\boldsymbol{\Sigma}_{\mathbf{Z}\mathbf{Z}}^{-1}\boldsymbol{\sigma}_{\mathbf{Z}Y}}{\sigma_{YY}}} \tag{7-48}$$

The square of the population multiple correlation coefficient, $\rho_{Y(\mathbf{Z})}^2$, is called the *population coefficient of determination*. Note that, unlike other correlation coefficients, the multiple correlation coefficient is a *positive* square root, so $0 \leq \rho_{Y(\mathbf{Z})} \leq 1$.

The population coefficient of determination has an important interpretation. From Result 7.12, the mean square error in using $\beta_0 + \boldsymbol{\beta}'\mathbf{Z}$ to forecast Y is

$$\sigma_{YY} - \boldsymbol{\sigma}_{\mathbf{Z}Y}'\boldsymbol{\Sigma}_{\mathbf{Z}\mathbf{Z}}^{-1}\boldsymbol{\sigma}_{\mathbf{Z}Y} = \sigma_{YY} - \sigma_{YY}\left(\frac{\boldsymbol{\sigma}_{\mathbf{Z}Y}'\boldsymbol{\Sigma}_{\mathbf{Z}\mathbf{Z}}^{-1}\boldsymbol{\sigma}_{\mathbf{Z}Y}}{\sigma_{YY}}\right) = \sigma_{YY}(1 - \rho_{Y(\mathbf{Z})}^2) \tag{7-49}$$

If $\rho_{Y(\mathbf{Z})}^2 = 0$, there is no predictive power in $\mathbf{Z}$. At the other extreme, $\rho_{Y(\mathbf{Z})}^2 = 1$ implies that Y can be predicted with no error.

Determining Best Linear Predictor, Mean Square Error and Multiple Correlation Coefficient

Example 7.11 (Determining the best linear predictor, its mean square error, and the multiple correlation coefficient) Given the mean vector and covariance matrix of Y, Z_1, Z_2,

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_Y \\ \hline \boldsymbol{\mu}_{\mathbf{Z}} \end{bmatrix} = \begin{bmatrix} 5 \\ \hline 2 \\ 0 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{YY} & \boldsymbol{\sigma}'_{\mathbf{Z}Y} \\ \hline \boldsymbol{\sigma}_{\mathbf{Z}Y} & \boldsymbol{\Sigma}_{\mathbf{ZZ}} \end{bmatrix} = \left[\begin{array}{c|cc} 10 & 1 & -1 \\ \hline 1 & 7 & 3 \\ -1 & 3 & 2 \end{array}\right]$$

determine (a) the best linear predictor $\beta_0 + \beta_1 Z_1 + \beta_2 Z_2$, (b) its mean square error, and (c) the multiple correlation coefficient. Also, verify that the mean square error equals $\sigma_{YY}(1 - \rho^2_{Y(\mathbf{Z})})$.

First,

$$\boldsymbol{\beta} = \boldsymbol{\Sigma}_{\mathbf{ZZ}}^{-1}\boldsymbol{\sigma}_{\mathbf{Z}Y} = \begin{bmatrix} 7 & 3 \\ 3 & 2 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} .4 & -.6 \\ -.6 & 1.4 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} 1 \\ -2 \end{bmatrix}$$

$$\beta_0 = \mu_Y - \boldsymbol{\beta}'\boldsymbol{\mu}_{\mathbf{Z}} = 5 - [1, -2] \begin{bmatrix} 2 \\ 0 \end{bmatrix} = 3$$

so the best linear predictor is $\beta_0 + \boldsymbol{\beta}'\mathbf{Z} = 3 + Z_1 - 2Z_2$. The mean square error is

$$\sigma_{YY} - \boldsymbol{\sigma}'_{\mathbf{Z}Y}\boldsymbol{\Sigma}_{\mathbf{ZZ}}^{-1}\boldsymbol{\sigma}_{\mathbf{Z}Y} = 10 - [1, -1] \begin{bmatrix} .4 & -.6 \\ -.6 & 1.4 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} = 10 - 3 = 7$$

Multiple Correlation Coefficient

so the best linear predictor is $\beta_0 + \boldsymbol{\beta}'\mathbf{Z} = 3 + Z_1 - 2Z_2$. The mean square error is

$$\sigma_{YY} - \boldsymbol{\sigma}'_{\mathbf{Z}Y}\boldsymbol{\Sigma}^{-1}_{\mathbf{ZZ}}\boldsymbol{\sigma}_{\mathbf{Z}Y} = 10 - [1, -1]\begin{bmatrix} .4 & -.6 \\ -.6 & 1.4 \end{bmatrix}\begin{bmatrix} 1 \\ -1 \end{bmatrix} = 10 - 3 = 7$$

and the multiple correlation coefficient is

$$\rho_{Y(\mathbf{Z})} = \sqrt{\frac{\boldsymbol{\sigma}'_{\mathbf{Z}Y}\boldsymbol{\Sigma}^{-1}_{\mathbf{ZZ}}\boldsymbol{\sigma}_{\mathbf{Z}Y}}{\sigma_{YY}}} = \sqrt{\frac{3}{10}} = .548$$

Note that $\sigma_{YY}(1 - \rho^2_{Y(\mathbf{Z})}) = 10\left(1 - \frac{3}{10}\right) = 7$ is the mean square error. ■

Mean of the Conditional Distribution is Best Linear Predictor Equation

It is possible to show (see Exercise 7.5) that

$$1 - \rho^2_{Y(\mathbf{Z})} = \frac{1}{\rho^{YY}} \tag{7-50}$$

where ρ^{YY} is the upper-left-hand corner of the inverse of the correlation matrix determined from $\boldsymbol{\Sigma}$.

The restriction to linear predictors is closely connected to the assumption of normality. Specifically, if we take

$$\begin{bmatrix} Y \\ Z_1 \\ Z_2 \\ \vdots \\ Z_r \end{bmatrix} \quad \text{to be distributed as} \quad N_{r+1}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

then the conditional distribution of Y with $z_1, z_2, \ldots, z_r$ fixed (see Result 4.6) is

$$N(\mu_Y + \boldsymbol{\sigma}'_{\mathbf{Z}Y}\boldsymbol{\Sigma}^{-1}_{\mathbf{ZZ}}(\mathbf{Z} - \boldsymbol{\mu}_{\mathbf{Z}}), \sigma_{YY} - \boldsymbol{\sigma}'_{\mathbf{Z}Y}\boldsymbol{\Sigma}^{-1}_{\mathbf{ZZ}}\boldsymbol{\sigma}_{\mathbf{Z}Y})$$

The mean of this conditional distribution is the linear predictor in Result 7.12.

Regression Function is Best Linear Predictor of Y when Population is Normal

The mean of this conditional distribution is the linear predictor in Result 7.12. That is,

$$E(Y|z_1, z_2, \ldots, z_r) = \mu_Y + \boldsymbol{\sigma}'_{\mathbf{ZY}}\boldsymbol{\Sigma}_{\mathbf{ZZ}}^{-1}(\mathbf{z} - \boldsymbol{\mu}_{\mathbf{Z}}) \qquad (7\text{-}51)$$
$$= \beta_0 + \boldsymbol{\beta}'\mathbf{z}$$

and we conclude that $E(Y|z_1, z_2, \ldots, z_r)$ is the best linear predictor of Y when the population is $N_{r+1}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The conditional expectation of Y in (7-51) is called the *regression function*. For normal populations, it is linear.

When the population is *not* normal, the regression function $E(Y|z_1, z_2, \ldots, z_r)$ need not be of the form $\beta_0 + \boldsymbol{\beta}'\mathbf{z}$. Nevertheless, it can be shown (see [22]) that $E(Y|z_1, z_2, \ldots, z_r)$, whatever its form, predicts Y with the smallest mean square error. Fortunately, this wider optimality among all estimators is possessed by the *linear* predictor when the population is normal.

Maximum Likelihood Estimator of Mean Square Error

Result 7.13. Suppose the joint distribution of Y and $\mathbf{Z}$ is $N_{r+1}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Let

$$\hat{\boldsymbol{\mu}} = \begin{bmatrix} \overline{Y} \\ \hline \overline{\mathbf{Z}} \end{bmatrix} \quad \text{and} \quad \mathbf{S} = \left[\begin{array}{c|c} s_{YY} & \mathbf{s}_{\mathbf{Z}Y}' \\ \hline \mathbf{s}_{\mathbf{Z}Y} & \mathbf{S}_{\mathbf{ZZ}} \end{array}\right]$$

be the sample mean vector and sample covariance matrix, respectively, for a random sample of size n from this population. Then the maximum likelihood estimators of the coefficients in the linear predictor are

$$\hat{\boldsymbol{\beta}} = \mathbf{S}_{\mathbf{ZZ}}^{-1}\mathbf{s}_{\mathbf{Z}Y}, \qquad \hat{\beta}_0 = \overline{Y} - \mathbf{s}_{\mathbf{Z}Y}'\mathbf{S}_{\mathbf{ZZ}}^{-1}\overline{\mathbf{Z}} = \overline{Y} - \hat{\boldsymbol{\beta}}'\overline{\mathbf{Z}}$$

Consequently, the maximum likelihood estimator of the linear regression function is

$$\hat{\beta}_0 + \hat{\boldsymbol{\beta}}'\mathbf{z} = \overline{Y} + \mathbf{s}_{\mathbf{Z}Y}'\mathbf{S}_{\mathbf{ZZ}}^{-1}(\mathbf{z} - \overline{\mathbf{Z}})$$

and the maximum likelihood estimator of the mean square error $E[Y - \beta_0 - \boldsymbol{\beta}'\mathbf{Z}]^2$ is

$$\hat{\sigma}_{YY\cdot\mathbf{Z}} = \frac{n-1}{n}(s_{YY} - \mathbf{s}_{\mathbf{Z}Y}'\mathbf{S}_{\mathbf{ZZ}}^{-1}\mathbf{s}_{\mathbf{Z}Y})$$

It is customary to change the divisor from n to $n - (r + 1)$ in the estimator of the mean square error, $\sigma_{YY\cdot\mathbf{Z}} = E(Y - \beta_0 - \boldsymbol{\beta}'\mathbf{Z})^2$, in order to obtain the *unbiased* estimator

$$\left(\frac{n-1}{n-r-1}\right)(s_{YY} - \mathbf{s}_{\mathbf{Z}Y}'\mathbf{S}_{\mathbf{ZZ}}^{-1}\mathbf{s}_{\mathbf{Z}Y}) = \frac{\sum_{j=1}^{n}(Y_j - \hat{\beta}_0 - \hat{\boldsymbol{\beta}}'\mathbf{Z}_j)^2}{n-r-1} \qquad (7\text{-}52)$$

Maximum Likelihood Estimate of Regression Function

Example 7.12 (Maximum likelihood estimate of the regression function—single response) For the computer data of Example 7.6, the $n = 7$ observations on Y (CPU time), Z_1 (orders), and Z_2 (add–delete items) give the sample mean vector and sample covariance matrix:

$$\hat{\boldsymbol{\mu}} = \begin{bmatrix} \bar{y} \\ \hline \bar{\mathbf{z}} \end{bmatrix} = \begin{bmatrix} 150.44 \\ \hline 130.24 \\ 3.547 \end{bmatrix}$$

$$\mathbf{S} = \begin{bmatrix} s_{YY} & \mathbf{s}_{\mathbf{z}Y}' \\ \hline \mathbf{s}_{\mathbf{z}Y} & \mathbf{S}_{\mathbf{zz}} \end{bmatrix} = \left[\begin{array}{c|cc} 467.913 & 418.763 & 35.983 \\ \hline 418.763 & 377.200 & 28.034 \\ 35.983 & 28.034 & 13.657 \end{array}\right]$$

Assuming that Y, Z_1, and Z_2 are jointly normal, obtain the estimated regression function and the estimated mean square error.

Result 7.13 gives the maximum likelihood estimates

$$\hat{\boldsymbol{\beta}} = \mathbf{S}_{\mathbf{zz}}^{-1}\mathbf{s}_{\mathbf{z}Y} = \begin{bmatrix} .003128 & -.006422 \\ -.006422 & .086404 \end{bmatrix} \begin{bmatrix} 418.763 \\ 35.983 \end{bmatrix} = \begin{bmatrix} 1.079 \\ .420 \end{bmatrix}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\boldsymbol{\beta}}'\bar{\mathbf{z}} = 150.44 - [1.079, .420] \begin{bmatrix} 130.24 \\ 3.547 \end{bmatrix} = 150.44 - 142.019$$

$$= 8.421$$

and the estimated regression function

$$\hat{\beta}_0 + \hat{\boldsymbol{\beta}}'\mathbf{z} = 8.42 - 1.08z_1 + .42z_2$$

Maximum Likelihood Estimate of Regression Function

and the estimated regression function

$$\hat{\beta}_0 + \hat{\boldsymbol{\beta}}'\mathbf{z} = 8.42 - 1.08z_1 + .42z_2$$

The maximum likelihood estimate of the mean square error arising from the prediction of Y with this regression function is

$$\left(\frac{n-1}{n}\right)(s_{YY} - \mathbf{s}_{ZY}'\mathbf{S}_{ZZ}^{-1}\mathbf{s}_{ZY})$$

$$= \left(\frac{6}{7}\right)\left(467.913 - [418.763, 35.983]\begin{bmatrix} .003128 & -.006422 \\ -.006422 & .086404 \end{bmatrix}\begin{bmatrix} 418.763 \\ 35.983 \end{bmatrix}\right)$$

$$= .894$$

■

Review: Multiple Linear Regression

Least Squares Fit With Several Independent Variables, Predicting the Dependent Variable

Micronutrients and Kelp Cultures: Evidence for Cobalt and Manganese Deficiency in Southern California Deep Seawater

Scientists have quantified the manner in which naturally occurring copper and zinc concentrations are toxic to marine organisms. They studied the effects of micronutrients on the growth of commercially significant giant kelp.

A regression analysis based on the least squares method can lead to models that relate growth to specific toxic substances.

The results are that toxic copper (Cu) and Zinc (Zn) ion concentrations may partially control growth. Also, cobalt (Co) and manganese (Mn) deficiencies exhibit some control in the growth.[1] J. S. Kuwabara, "*Micronutrients and kelp cultures: evidence for cobalt and manganese deficiency in southern California deep sea waters*," *Science* 216 (June 11, 1982), pp. 1219-1221. Copyright © 1982 by AAAS.

A least squares fit of gametophytic growth data in the defined medium generated the expression

$$\begin{aligned} Y = {} & 136 + 8x_{\mathrm{Mn}} - 5x_{\mathrm{Cu}} + 7x_{\mathrm{Co}} \\ & - 7x_{\mathrm{Zn}}x_{\mathrm{Cu}} - 15x_{\mathrm{Zn}}^2 - 27x_{\mathrm{Mn}}^2 - 12x_{\mathrm{Cu}}^2 \\ & - 18x_{\mathrm{Co}}^2 - 6x_{\mathrm{Cu}}x_{\mathrm{Zn}}^2 - 6x_{\mathrm{Cu}}x_{\mathrm{Mn}}^2 \end{aligned}$$

Independent Variables Consume More Variance in the Model i.e. Larger R^2

A least squares fit of gametophytic growth data in the defined medium generated the expression

$$\begin{aligned} Y = {} & 136 + 8x_{\mathrm{Mn}} - 5x_{\mathrm{Cu}} + 7x_{\mathrm{Co}} \\ & - 7x_{\mathrm{Zn}}x_{\mathrm{Cu}} - 15x_{\mathrm{Zn}}^2 - 27x_{\mathrm{Mn}}^2 - 12x_{\mathrm{Cu}}^2 \\ & - 18x_{\mathrm{Co}}^2 - 6x_{\mathrm{Cu}}x_{\mathrm{Zn}}^2 - 6x_{\mathrm{Cu}}x_{\mathrm{Mn}}^2 \end{aligned}$$

where Y is mean gametophytic length in micrometers. The authors consider the fit of the experimental data to this equation as excellent.

Here, several variables are important for predicting growth.

The basic ideas of regression analysis have a much broader scope of application than the straight line model of Chapter 11. In this chapter, our goal is to extend the ideas of regression analysis in two important directions.

1. To handle nonlinear relations by means of appropriate transformations applied to one or both variables.
2. To accommodate several predictor variables into a regression model.

These extensions enable the reader to appreciate the breadth of regression techniques that are applicable to real-life problems. We then discuss some graphical procedures that are helpful in detecting any serious violation of the assumptions that underlie a regression analysis.

Regression Question: How Do the Variables Affect the Response?

2. Multiple Linear Regression

A response variable y may depend on a predictor variable x but, after a straight line fit, it may turn out that the unexplained variation is large, so r^2 is small and a poor fit is indicated. At the same time, an attempt to transform one or both of the variables may fail to dramatically improve the value of r^2. This difficulty may well be due to the fact that the response depends on not just x but other factors as well. When used alone, x fails to be a good predictor of y because of the effects of those other influencing variables. For instance, the yield of a crop depends on not only the amount of fertilizer but also on the rainfall and average temperature during the growing season. Cool weather and no rain could completely cancel the choice of a correct fertilizer.

To obtain a useful prediction model, one should record the observations of all variables that may significantly affect the response. These other variables may then be incorporated explicitly into the regression analysis. The name **multiple regression** refers to a model of relationship where the response depends on two or more predictor variables. Here, we discuss the main ideas of a multiple regression analysis in the setting of two predictor variables.

Data File Format For Analysis

Suppose that the response variable y in an experiment is expected to be influenced by two input variables x_1 and x_2, and the data relevant to these input variables are recorded along with the measurements of y. With n runs of an experiment, we have a data set of the form shown in Table 5.

TABLE 5 Data Structure for Multiple Regression with Two Input Variables

Experimental	Input Variables		Response
Run	x_1	x_2	y
1	x_{11}	x_{12}	y_1
2	x_{21}	x_{22}	y_2
.	.	.	.
.	.	.	.
.	.	.	.
i	x_{i1}	x_{i2}	y_i
.	.	.	.
.	.	.	.
.	.	.	.
n	x_{n1}	x_{n2}	y_n

By analogy with the simple linear regression model, we can then tentatively formulate:

TABLE 5 Data Structure for Multiple Regression with Two Input Variables

Experimental Run	Input Variables x_1	Input Variables x_2	Response y
1	x_{11}	x_{12}	y_1
2	x_{21}	x_{22}	y_2
.	.	.	.
.	.	.	.
.	.	.	.
i	x_{i1}	x_{i2}	y_i
.	.	.	.
.	.	.	.
.	.	.	.
n	x_{n1}	x_{n2}	y_n

Table 5 (p. 504)

Data Structure for Multiple Regression with Two Input Variables

2.1 Data from an Environmental Survey

The data set shown in Figure 2.3 represents 30 responses from a questionnaire concerning the president's environmental policies. (See the file **Questionnaire Data.xlsx**.) Identify the variables and observations.

	A	B	C	D	E	F	G
1	Person	Age	Gender	State	Children	Salary	Opinion
2	1	35	Male	Minnesota	1	$65,400	5
3	2	61	Female	Texas	2	$62,000	1
4	3	35	Male	Ohio	0	$63,200	3
5	4	37	Male	Florida	2	$52,000	5
6	5	32	Female	California	3	$81,400	1
7	6	33	Female	New York	3	$46,300	5
8	7	65	Female	Minnesota	2	$49,600	1
9	8	45	Male	New York	1	$45,900	5
10	9	40	Male	Texas	3	$47,700	4
11	10	32	Female	Texas	1	$59,900	4
12	11	57	Male	New York	1	$48,100	4
13	12	38	Female	Virginia	0	$58,100	3
14	13	37	Female	Illinois	2	$56,000	1
15	14	42	Female	Virginia	2	$53,400	1
16	15	38	Female	New York	2	$39,000	2
17	16	48	Male	Michigan	1	$61,500	2
18	17	40	Male	Ohio	0	$37,700	1
19	18	57	Female	Michigan	2	$36,700	4
20	19	44	Male	Florida	2	$45,200	3
21	20	40	Male	Michigan	0	$59,000	4
22	21	21	Female	Minnesota	2	$54,300	2
23	22	49	Male	New York	1	$62,100	4
24	23	34	Male	New York	0	$78,000	3
25	24	49	Male	Arizona	0	$43,200	5
26	25	40	Male	Arizona	1	$44,500	3
27	26	38	Male	Ohio	1	$43,300	1
28	27	27	Male	Illinois	3	$45,400	2
29	28	63	Male	Michigan	2	$53,900	1
30	29	52	Male	California	1	$44,100	3
31	30	48	Female	New York	2	$31,000	4

Principle of Least Squares: Vary the b_o, b_1, b_2 and Simultaneously Minimize Sum of Squared Deviations

By analogy with the simple linear regression model, we can then tentatively formulate:

A Multiple Regression Model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i \quad i = 1, \ldots, n$$

where x_{i1} and x_{i2} are the values of the two input variables for the ith experimental run and Y_i is the corresponding response.

The error components e_i are assumed to be independent normal variables with mean 0 and variance σ^2.

The regression parameters β_0, β_1, and β_2 are unknown and so is σ^2.

This model suggests that aside from the random error, the response varies linearly with each of the independent variables when the other remains fixed.

The principle of least squares is again useful in estimating the regression parameters. For this model, we are required to vary b_0, b_1, and b_2 simultaneously to minimize the sum of squared deviations

$$\sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i1} - b_2 x_{i2})^2$$

Regression Analysis: Linear Algebra to Fit the Straight Line, Where S_{11} an S_{12} Sums of Squares and Cross Products Deviations

The least squares estimates $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ are the solutions to the following equations, which are extensions of the corresponding equations for fitting the straight line model (see Section 3 of Chapter 11.)

$$\hat{\beta}_1 S_{11} + \hat{\beta}_2 S_{12} = S_{1y}$$
$$\hat{\beta}_1 S_{12} + \hat{\beta}_2 S_{22} = S_{2y}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2$$

where S_{11}, S_{12}, and so on, are the sums of squares and cross products of deviations of the variables indicated in the suffix. They are computed just as in a straight line regression model. Methods are available for interval estimation, hypothesis testing, and examining the adequacy of fit. In principle, these methods are similar to those used in the simple regression model, but the algebraic formulas are more complex and hand computations become more tedious. However, a multiple regression analysis is easily performed on a computer with the aid of the standard packages such as MINITAB, SAS, or SPSS. We illustrate the various aspects of a multiple regression analysis with the data of Example 3 and computer-based calculations.

A Multiple Regression Model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i \qquad i = 1, \ldots, n$$

where x_{i1} and x_{i2} are the values of the input variables for the ith experimental run and Y_i is the corresponding response.

The error components e_i are assumed to be independent normal variables with mean 0 and variance σ^2.

The regression parameters β_0, β_1, and β_2 are unknown and so is σ^2.

Box on Page 504

A Multiple Regression Model

Multiple Regression Analysis Example

Example 3

Interpreting the Regression of Blood Pressure on Weight and Age

We are interested in studying the systolic blood pressure y in relation to weight x_1 and age x_2 in a class of males of approximately the same height. From 13 subjects preselected according to weight and age, the data set listed in Table 6 was obtained.

TABLE 6 The Data of x_1 = Weight in Pounds, x_2 = Age, and y = Blood Pressure of 13 Males

x_1	x_2	y
152	50	120
183	20	141
171	20	124
165	30	126
158	30	117
161	50	129
149	60	123
158	50	125
170	40	132
153	55	123
164	40	132
190	40	155
185	20	147

Use a computer package to perform a regression analysis using the model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i$$

Table 6 (p. 505)

The Data of x_1 = Weight in Pounds, x_2 = Age, and y = Blood Pressure of 13 Males

TABLE 6 The Data of x_1 = Weight in Pounds, x_2 = Age, and y = Blood Pressure of 13 Males

x_1	x_2	y
152	50	120
183	20	141
171	20	124
165	30	126
158	30	117
161	50	129
149	60	123
158	50	125
170	40	132
153	55	123
164	40	132
190	40	155
185	20	147

Regression Analysis: Computer Syntax in Minitab (Similar to Most Software)

Use a computer package to perform a regression analysis using the model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i$$

SOLUTION

To use MINITAB, we first enter the data of x_1, x_2, and y in three different columns and then use the regression command,

Data: C12T5

C1:	152	183	171	···	185
C2:	50	20	20	···	20
C3:	120	141	124	···	147

Dialog box:

Stat > **Regression** > **Regression**

Type C3 in **Response**.

Type C1 and C2 in **Predictors**.

Click **OK**.

Data: C12T5.txt
C1: 152 183 171 ⋯185
C2: 50 20 20 ⋯ 20
C3: 120 141 124 ⋯147
Dialog box:
Stat > Regression > Regression
Type *C3* in **Response.**
Type *C1* and *C2* in **Predictors.**
Click **OK.**.

Box p. 506

Regression Analysis Output – Interpreting Results

With the last command, the computer executes a multiple regression analysis. We focus our attention on the principal aspects of the output, as shown in Table 7.

TABLE 7 Regression Analysis of the Data in Table 6: Selected MINITAB Output

① The regression equation is
Y = −65.1 + 1.08 Weight + 0.425 Age

Predictor	Coef	SE Coef	T	P
Constant	−65.10	14.94	−4.36	0.001
Weight	② 1.07710	0.07707	13.98	0.000
Age	0.42541	0.07315	④ 5.82	0.000

③ S = 2.50861 ⑤ R−SQ = 95.8%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	1423.84	711.92	113.13	0.000
Residual Error	10	⑥ 62.93	6.29		
Total	12	1486.77			

We now proceed to interpret the results in Table 7 and use them to make further statistical inferences.

Regression Analysis Output – Interpreting Results

We now proceed to interpret the results in Table 7 and use them to make further statistical inferences.

i. The equation of the fitted linear regression is

$$\text{①} \quad \hat{y} = -65.1 + 1.08x_1 + .425x_2$$

This means that the mean blood pressure increases by 1.08 if weight x_1 increases by one pound and age x_2 remains fixed. Similarly, a 1-year increase in age with the weight held fixed will only increase the mean blood pressure by .425.

ii. The estimated regression coefficient and the corresponding estimated standard errors are

$$\text{②} \quad \begin{aligned} \hat{\beta}_0 &= -65.10 \quad & \text{Estimated S.E.}(\hat{\beta}_0) &= 14.94 \\ \hat{\beta}_1 &= 1.07710 \quad & \text{Estimated S.E.}(\hat{\beta}_1) &= .07707 \\ \hat{\beta}_2 &= .42541 \quad & \text{Estimated S.E.}(\hat{\beta}_2) &= .07315 \end{aligned}$$

③ Further, the error standard deviation σ is estimated by $s = 2.50861$ with

$$\begin{aligned} \text{Degrees of freedom} &= n - (\text{No. of input variables}) - 1 \\ &= 13 - 2 - 1 \\ &= 10 \end{aligned}$$

TABLE 7 Regression Analysis of the Data in Table 6: Selected MINITAB Output

① The regression equation is
Y = −65.1 + 1.08 Weight + 0.425 Age

Predictor		Coef	SE Coef		T	P
Constant		−65.10	14.94		−4.36	0.001
Weight	②	1.07710	0.07707		13.98	0.000
Age		0.42541	0.07315	④	5.82	0.000

③ S = 2.50861 ⑤ R−SQ = 95.8%

Analysis of Variance

Source	DF		SS	MS	F	P
Regression	2		1423.84	711.92	113.13	0.000
Residual Error	10	⑥	62.93	6.29		
Total	12		1486.77			

Table 7 (p. 506)

Regression Analysis of the Data in Table 6: Selected MINITAB Output

TABLE 8 A Regression Analysis of the Data in Example 2 Using SAS

Model: MODEL1
Dependent Variable: y

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1423.83797	711.91898	113.13	<.0001
⑥ Error	10	62.93126	6.29313		
Corrected Total	12	1486.76923			

③ Root MSE 2.50861 ⑤ R-Square 0.9577

Parameter Estimates

Variable	DF		Parameter Estimate	Standard Error		t Value	Pr > |t|
Intercept	1		−65.09968	14.94458		−4.36	0.0014
x1	1	②	1.07710	0.07707	④	13.98	<.0001
x2	1		0.42541	0.07315		5.82	0.0002

Table 8 (p. 508)

A Regression Analysis of the Data in Example 2 Using SAS.

Calculating Confidence Interval from the Computer Output

These results are useful in interval estimation and hypothesis tests about the regression coefficients. In particular, a $100\,(1-\alpha)\,\%$ confidence interval for a coefficient β is given by

$$\text{Estimated coefficient} \pm t_{\alpha/2}\ (\text{Estimated S.E.})$$

where $t_{\alpha/2}$ is the upper $\alpha/2$ point of the distribution with d.f. $= 10$.

For instance, a 95% confidence interval for β_1 is

$$1.07710 \pm 2.228 \times .07707$$
$$= 1.07710 \pm .17171 \quad \text{or} \quad (.905,\ 1.249)$$

To test the null hypothesis that a particular coefficient β is zero, we employ the test statistic

$$t = \frac{\text{Estimated coefficient} - 0}{\text{Estimated S.E.}} \quad \text{d.f.} = 10$$

ANOVA Reveals Decomposition of the Total Variability Equals 1486.77 = $\Sigma (y_i - \bar{y})^2$

$$t = \frac{\text{Estimated coefficient} - 0}{\text{Estimated S.E.}} \quad \text{d.f.} = 10$$

These t-ratios appear in Table 7. Suppose that we wish to examine whether the mean blood pressure significantly increases with age. In the language of hypothesis testing, this problem translates to one of testing $H_0: \beta_2 = 0$ versus $H_1: \beta_2 > 0$. The observed value of the ④ test statistic is $t = 5.82$ with d.f. $= 10$. Since this is larger than the tabulated value $t_{.01} = 2.764$, the null hypothesis is rejected in favor of H_1, with $\alpha = .01$. In fact, it is rejected even with $\alpha = .005$.

iii. In Table 7, the result "R-SQ = 95.8 % " or

⑤ $R^2 = .958$

tells us that 95.8% of the variability of y is explained by the fitted multiple regression of y on x_1 and x_2. The "analysis of variance" shows the decomposition of the total variability $\Sigma (y_i - \bar{y})^2 = 1486.77$ into the two components.

⑥ 1486.77	=	1423.84	+	62.93
Total variability of y		Variability explained by the regression of y on x_1 and x_2		Residual or unexplained variability

Calculation of R^2 = Variability Explained by the Regression of y on x_1 and x_2 Divided by Total Variability of y

iii. In Table 7, the result "R-SQ = 95.8 % " or

$$⑤ \quad R^2 = .958$$

tells us that 95.8% of the variability of y is explained by the fitted multiple regression of y on x_1 and x_2. The "analysis of variance" shows the decomposition of the total variability $\Sigma (y_i - \bar{y})^2 = 1486.77$ into the two components.

⑥ 1486.77	=	1423.84	+	62.93
Total variability of y		Variability explained by the regression of y on x_1 and x_2		Residual or unexplained variability

Thus,

$$R^2 = \frac{1423.84}{1486.77} = .958$$

and σ^2 is estimated by $s^2 = 62.93 / 10 = 6.293$, so $s = 2.509$ [which checks with s from ii].

The **square of the multiple correlation coefficient** R^2 gives the proportion of variability in y explained by the fitted multiple regression.

The output from the SAS package is given in Table 8. The quantities needed in our analysis have been labeled with the same circled numbers as in the MINITAB output.

TABLE 8 A Regression Analysis of the Data in Example 3 Using SAS

Model: MODEL1
Dependent Variable: y

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1423.83797	711.91898	113.13	<.0001
Error	10	⑥ 62.93126	6.29313		
Corrected Total	12	1486.76923			

③ Root MSE 2.50861 ⑤ R-Square 0.9577

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	−65.09968	14.94458	−4.36	0.0014
x1	1	② 1.07710	0.07707	④ 13.98	<.0001
x2	1	0.42541	0.07315	5.82	0.0002

Example Multiple Linear Regression

Example 4

Computer-Aided Regression Analysis—Two Predictors

The times for 81 students to complete a rowing test both before and after completing a one-semester conditioning course are given in Table D.5 in the Data Bank. It may be that not only the pretest rowing time but also gender would be useful for predicting the posttest rowing time. Perform a regression analysis.

SOLUTION

We use MINITAB to obtain the output

```
Regression Analysis: Post row versus Pre row, Gender


The regression equation is
Post row = 97.3 + 0.726 Pre row + 32.1 Gender


Predictor     Coef   SE Coef        T       P
Constant     97.33     31.68     3.07   0.003
Pre row    0.72573   0.05487    13.23   0.000
Gender      32.083     9.756     3.29   0.002


S = 31.8137   R-Sq = 85.7%
```

Regression Analysis Output- Which Variables to Retain and Which to Discard

Regression Analysis: Post row versus Pre row, Gender

The regression equation is
Post row = 97.3 + 0.726 Pre row + 32.1 Gender

Predictor	Coef	SE Coef	T	P
Constant	97.33	31.68	3.07	0.003
Pre row	0.72573	0.05487	13.23	0.000
Gender	32.083	9.756	3.29	0.002

S = 31.8137 R-Sq = 85.7%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	471547	235774	232.95	0.000
Residual Error	78	78945	1012		
Total	80	550492			

Which variables should be used to predict posttest rowing time? Reading from the column of P -values for the individual coefficients, the largest is only .003. The constant term and the coefficients of pretest rowing time and gender are significantly different from 0. All three terms are needed in the model.
The plot of residuals versus fit in Figure 4 reveals a constant width band so there is no evidence against the assumption of constant variance. The one large negative residual is case 17 and the two large positive residuals are cases 29 and 70.

Plot of Residuals (y-y-hat) vs. y-hat or predicted values of y

Which variables should be used to predict posttest rowing time? Reading from the column of P -values for the individual coefficients, the largest is only .003. The constant term and the coefficients of pretest rowing time and gender are significantly different from 0. All three terms are needed in the model.
The plot of residuals versus fit in Figure 4 reveals a constant width band so there is no evidence against the assumption of constant variance. The one large negative residual is case 17 and the two large positive residuals are cases 29 and 70.

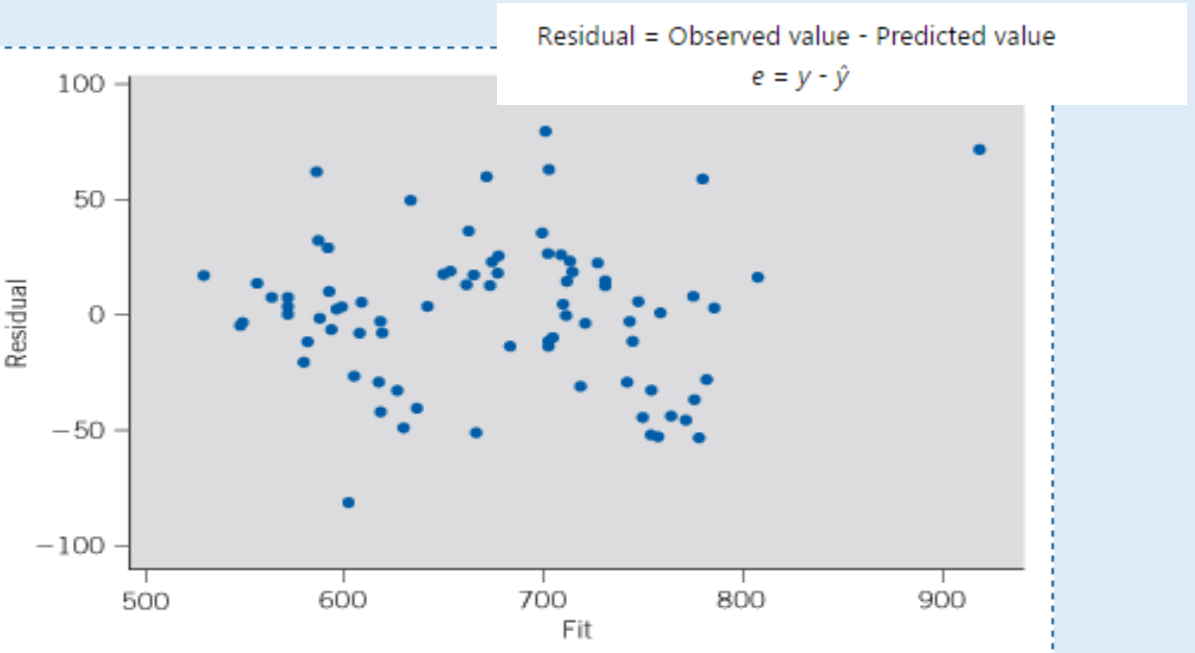


Figure 4 Residuals of posttest row times versus fits.

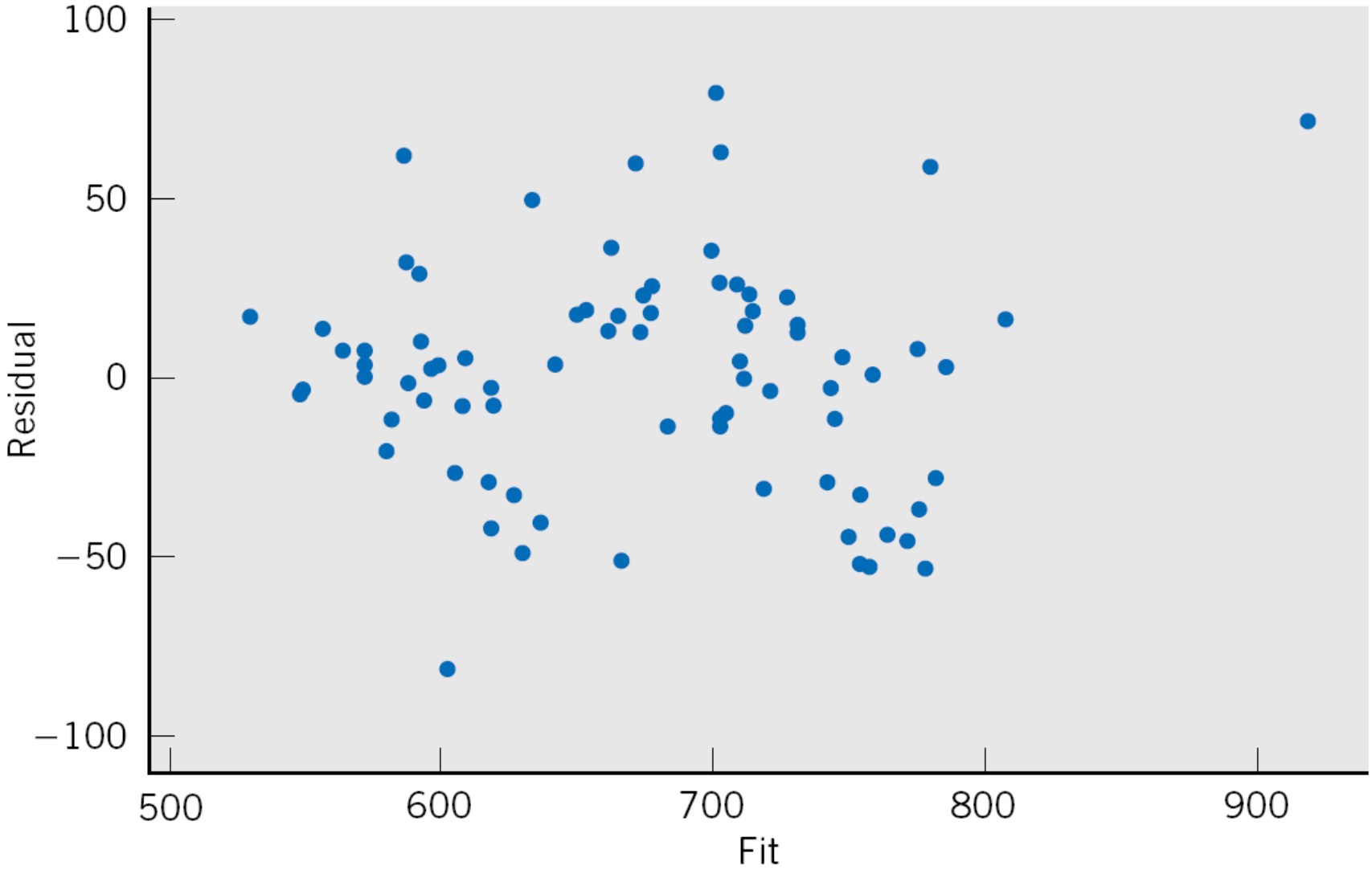


Figure 4 (p. 510)

Residuals of posttest row times versus fits.

Sample Problem 12.11

12.11 Consider the response variable miles per gallon on highways and the two predictor variables $x_1 =$ engine volume (liter) and $x_2 =$ size of battery (volt). Using a recent government Fuel Economy Guide, and the data on hybrid-electric cars and SUVs, we obtain the regression analysis given in Table 9.

TABLE 9 Computer Output of a Regression Analysis to Be Used for Exercise 12.11

```
Regression Analysis: y versus x1, x2

The Regression equation is
y = 45.3 - 3.22 x1 - 0.0207 X2

Predictor      Coef      SE Coef         T         P
Constant     45.299        2.345     19.32     0.000
x1          -3.2243       0.4562     -7.07     0.000
x2        -0.020661     0.008574     -2.41     0.026

S = 3.40356     R-SQ = 79.5%

Analysis of Variance

Source         DF         SS         MS          F         P
Regression      2     897.53     448.77      38.74     0.000
Error          20     231.68      11.58
Total          22    1129.22
```

Table 9 (p. 514)

Computer Output of a Regression Analysis to Be Used for Exercise 12.11

Sample Problem 12.11 - Solution

(a) Identify the least squares estimates $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$.

ANSWER

$\hat{\beta}_1 = -3.22$, $\hat{\beta}_2 = -.0207$, and $\hat{\beta}_0 = 45.3$

SOLUTION

$\hat{\beta}_0 = 45.3$, $\hat{\beta}_1 = -3.22$, $\hat{\beta}_2 = -0.02066$

(b) What model is suggested from this analysis?

SOLUTION

$\hat{y} = 45.3 - 3.22x_1 - 0.02066x_2$.

(c) What is the proportion of y variability explained by the regression on x_1 and x_2?

ANSWER

.795

SOLUTION

The proportion of y variability explained is $R^2 = 0.795$.

(d) Estimate σ^2.

ANSWER

11.58

SOLUTION

$s^2 = \frac{\text{SSE}}{n-2} = \text{MSE} = 11.58$

Residual Plots to Test Regression Model Least Squares Fit Adequacy

3. Residual Plots to Check the Adequacy of a Statistical Model

General Attitude Toward a Statistical Model

A regression analysis is not completed by fitting a model by least squares, providing confidence intervals, and testing various hypotheses. These steps tell only half the story: the statistical inferences that can be made when the postulated model is adequate. In most studies, we cannot be sure that a particular model is nearly correct. Therefore, we should adopt the following strategy.

1. **Tentatively entertain a model**.
2. **Obtain least squares estimates and compute the residuals**.
3. **Review the model by examining the residuals**.

Step 3 often suggests methods of appropriately modifying the model. We then return to step 1, where the modified model is entertained, and this **iteration** is continued until a model is obtained for which the data do not seem to contradict the assumptions made about the model.

Once a model is fitted by least squares, all the information on variation that cannot be explained by the model is contained in the residuals

$$\hat{e}_i = y_i - \hat{y}_i \quad i = 1, 2, \ldots, n$$

where y_i is the observed value and $\hat{y}_i$ denotes the corresponding value predicted by the fitted model. For example, in the case of a simple linear regression model, $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

Underlying Assumption of Each Variable—Whether Predictor or Response is Normal Distribution

Once a model is fitted by least squares, all the information on variation that cannot be explained by the model is contained in the residuals

$$\hat{e}_i = y_i - \hat{y}_i \quad i = 1, 2, \ldots, n$$

where y_i is the observed value and $\hat{y}_i$ denotes the corresponding value predicted by the fitted model. For example, in the case of a simple linear regression model, $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

Recall from our discussion of the straight line model in Chapter 11 that we have made the assumptions of independence, constant variance, and a normal distribution for the error components e_i. The inference procedures are based on these assumptions. When the model is correct, the residuals can be considered as estimates of the errors e_i that are distributed as $N(0, \sigma)$.

To determine the merits of the tentatively entertained model, we can examine the residuals by plotting them in various ways. Then if we recognize any systematic pattern formed by the plotted residuals, we would suspect that some assumptions regarding the model are invalid. There are many ways to plot the residuals, depending on what aspect is to be examined. We mention a few of these here to illustrate the techniques.

General Attitude Toward a Statistical Model

A regression analysis is not completed by fitting a model by least squares, providing confidence intervals, and testing various hypotheses. These steps tell only half the story: the statistical inferences that can be made when the postulated model is adequate. In most studies, we cannot be sure that a particular model is correct. Therefore, we should adopt the following strategy.

1. **Tentatively entertain a model.**
2. **Obtain least squares estimates and compute the residuals.**
3. **Review the model by examining the residuals.**

Box on Page 516

General Attitude Toward a Statistical Model

Residual Plot Using Dot Plot or Diagram

3.1 HISTOGRAM OR DOT DIAGRAM OF RESIDUALS

To picture the overall behavior of the residuals, we can plot a **histogram** for a large number of observations or a **dot diagram** for fewer observations. For example, in a dot diagram like the one in Figure 6*a*, the residuals seem to behave like a sample from a normal population and there do not appear to be any "wild" observations. In contrast, Figure 6*b* illustrates a situation in which the distribution appears to be quite normal except for a single residual that lies far to the right of the others. The circumstances that produced the associated observation demand a close scrutiny.

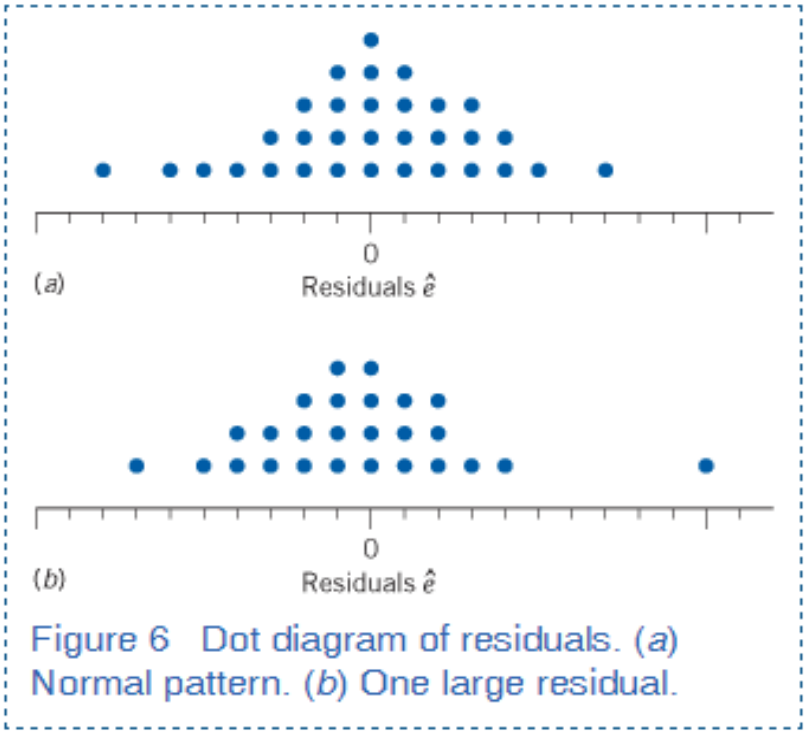


Figure 6 Dot diagram of residuals. (*a*) Normal pattern. (*b*) One large residual.

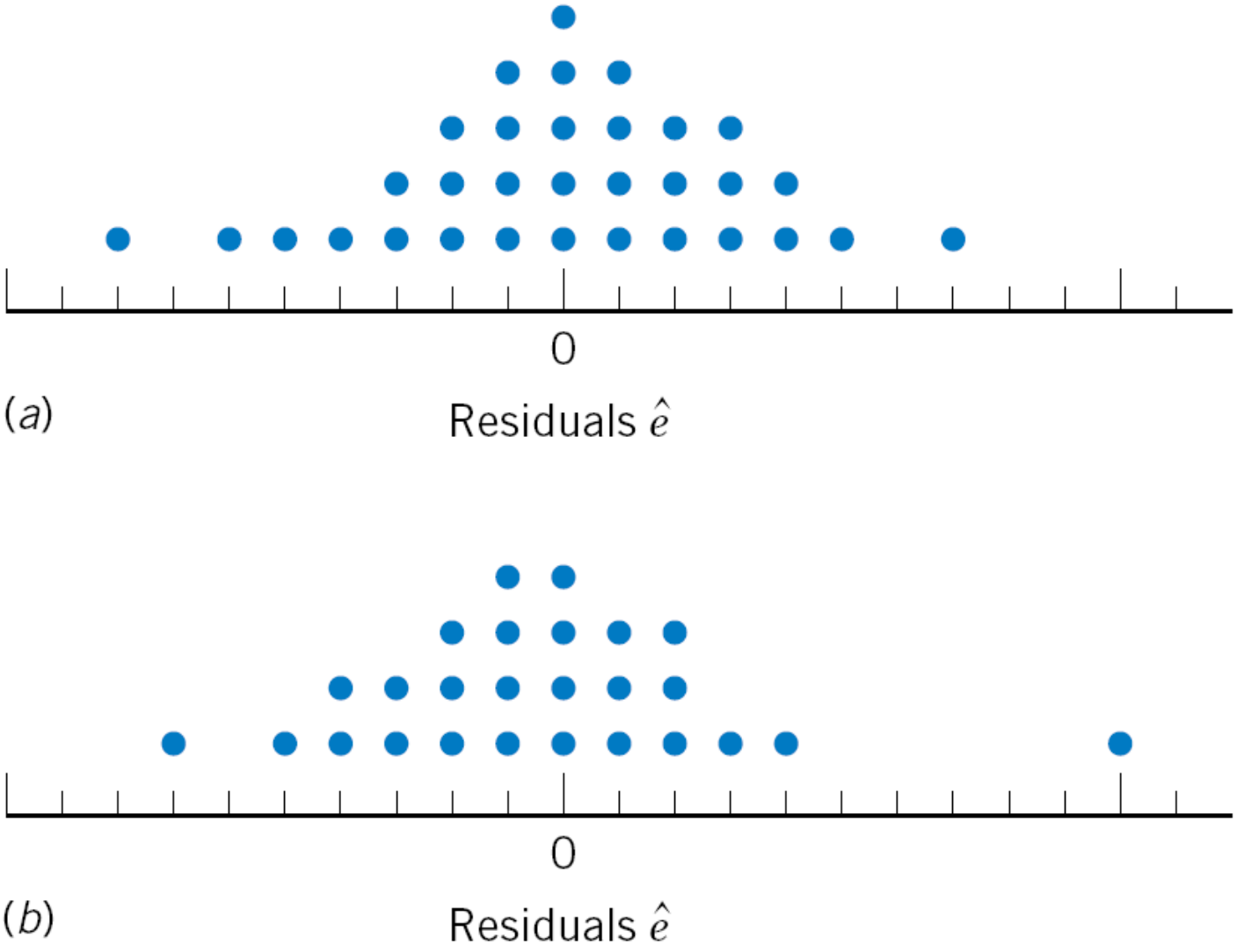


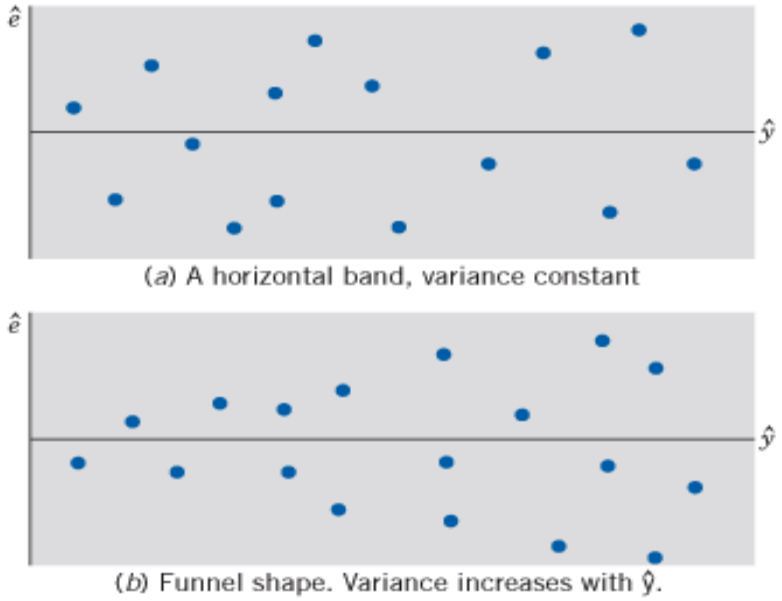
Figure 6 (p. 517)

Dot diagram of residuals. (a) Normal pattern. (b) One large residual.

Plot of Residuals vs. Predicted Value or y-hat: (a) Constant Spread; (b) Increasing Spread

3.2 PLOT OF RESIDUAL VERSUS PREDICTED VALUE

A plot of the residuals $\hat{e}_i$ versus the **predicted value** $\hat{y}_i$ often helps to detect the inadequacies of an assumed relation or a violation of the assumption of constant error variance. Figure 7a illustrates some typical phenomena.

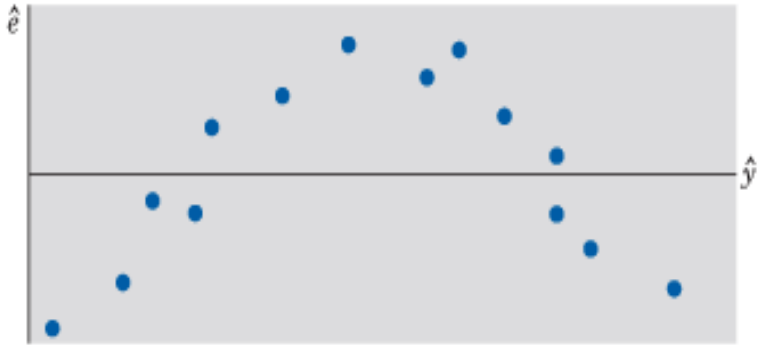


(*a*) A horizontal band, variance constant

(*b*) Funnel shape. Variance increases with $\hat{y}$.

Figure 7a Plot of residual versus predicted value. (*a*) Constant spread. (*b*) Increasing spread.

Plot of Residuals vs. Predicted Value or y-hat or y-hat: (c) Systematic Curved Pattern



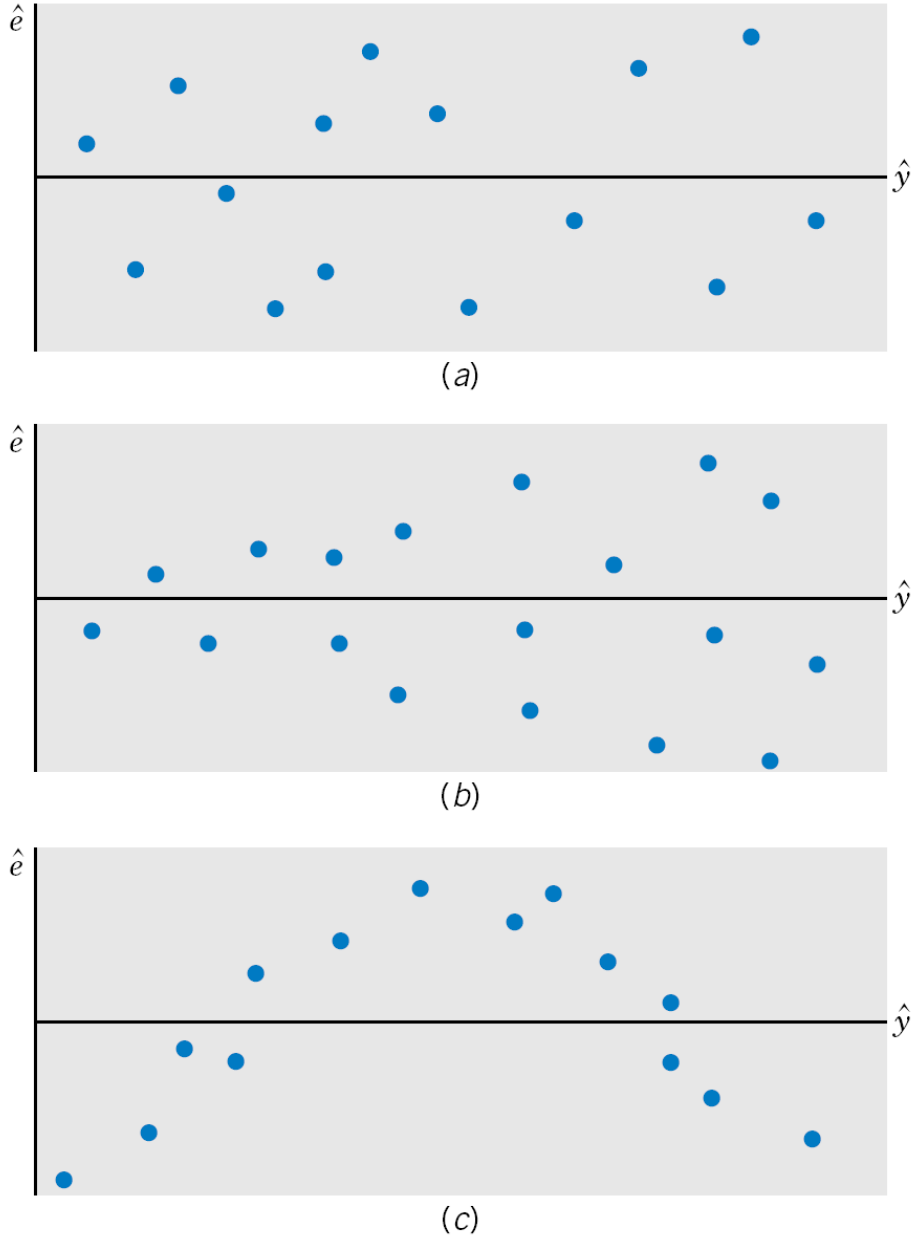
(*c*) Systematic curved pattern.

Figure 7b Plot of residual versus predicted value. (*c*) Systematic curved pattern.

If the points form a horizontal band around zero, as in Figure 7a, then no abnormality is indicated. In Figure 7b, the width of the band increases noticeably with increasing values of $\hat{y}$. This indicates that the error variance σ^2 tends to increase with an increasing level of response. We would then suspect the validity of the assumption of constant variance in the model. Figure 7c shows residuals that form a systematic pattern. Instead of being randomly distributed around the $\hat{y}$ axis, they tend first to increase steadily and then decrease. This would lead us to suspect that the model is inadequate and a squared term or some other nonlinear x term should be considered.

Figure 7 (p. 517 - 518)

Plot of residual versus predicted value.
(a) Constant spread.
(b) Increasing spread.
(c) Curved pattern.



Plot of Residual vs. Chronological Time Order to Detect Violation of Independence

3.3 PLOT OF RESIDUAL VERSUS TIME ORDER

The most crucial assumption in a regression analysis is that the errors e_i are independent. Lack of independence frequently occurs in business and economic applications, where the observations are collected in a time sequence with the intention of using regression techniques to predict future trends. In many other experiments, trials are conducted successively in time. In any event, a plot of the residuals versus **time order** often detects a violation of the assumption of independence. For example, the plot in Figure 8 exhibits a systematic pattern in that a string of high values is followed by a string of low values. This indicates that consecutive residuals are (positively) correlated, and we would suspect a violation of the independence assumption.

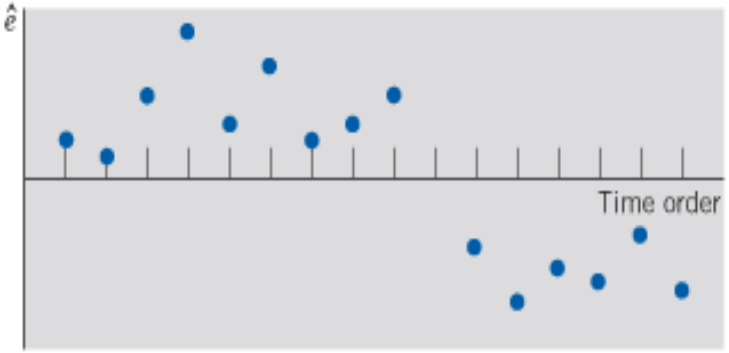


Figure 8 Plot of residual versus time order.

No Pattern to the Cluster, Then Independence; If Points Clustered Along the Line then Lacks Independence

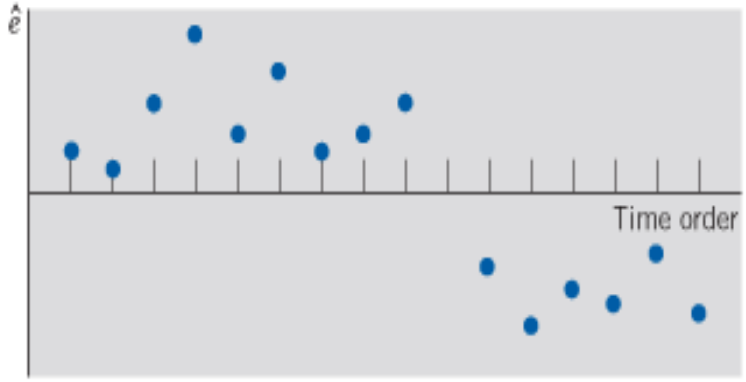


Figure 8 Plot of residual versus time order.

Independence can also be checked by plotting the successive pairs $(\hat{e}_i, \hat{e}_{i-1})$, where $\hat{e}_1$ indicates the residual from the first y value observed, $\hat{e}_2$ indicates the second, and so on. Independence is suggested if the scatter diagram is a patternless cluster, whereas points clustered along a line suggest a lack of independence between adjacent observations.

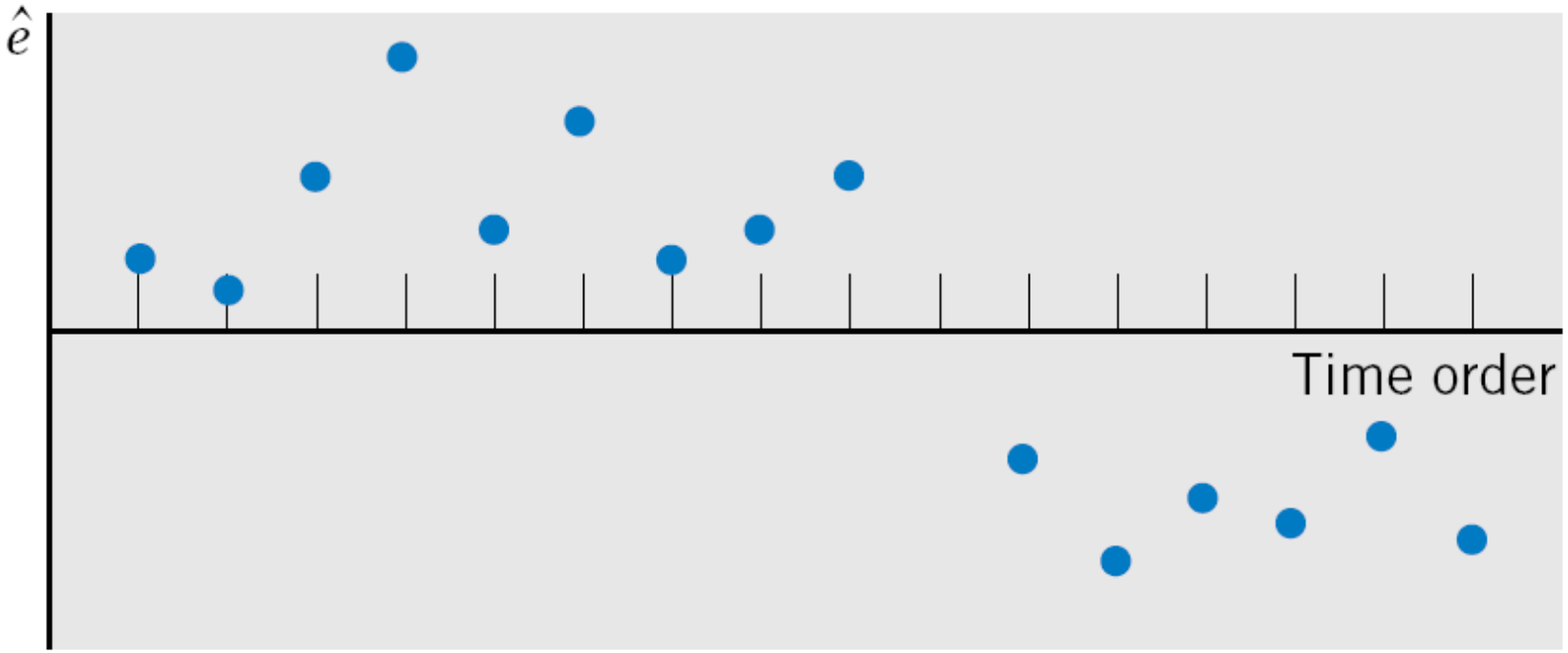


Figure 8 (p. 518)
Plot of residual versus time order.

Checking the Regression Model for Adequacy of Least Squares Fit

Example 6

Checking the Model for Rowing Time

In Example 4, we obtained the regression equation

$$\text{Post row} = 97.3 + .726\ \text{Pre row} + 32.1\ \text{Gender}$$

and $s = 31.8137$ Before making any inferences, the model assumptions need to be checked. Construct the three graphs

a. Histogram of residuals.

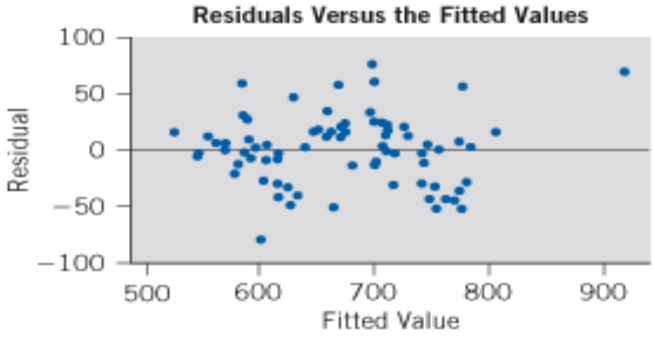
b. Residuals versus predicted values.

c. Residuals versus observation order.

Comment on any patterns that reveal violations of the assumptions concerning the model.

SOLUTION

The MINITAB regression *four-in-one* graphics option creates four graphs including the three requested.



Histogram of the Residuals- Check for Normal Distribution; Residuals Vs. Order of Data, If No Pattern Then Independent

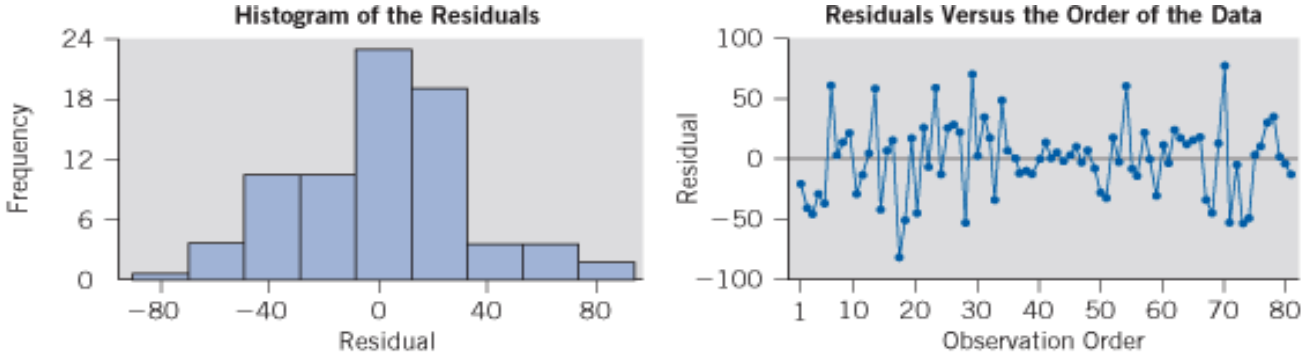


Figure 9 Three residual plots for posttest rowing time using MINITAB.

There are no obvious patterns that suggest a systematic departure from the model or time dependence in the order recorded. A normal-scores plot would better assess the normal assumption but some lack of normality is not a problem when the sample size is this large.

It is important to remember that our confidence in statistical inference procedures is related to the validity of the assumptions about them. A mechanically made inference may be misleading if some model assumption is grossly violated. An **examination of the residuals** is an important part of regression analysis, because it helps to detect any inconsistency between the data and the postulated model.

If no serious violation of the assumption is exposed in the process of examining residuals, we consider the model adequate and proceed with the relevant inferences. Otherwise, we must search for a more appropriate model.

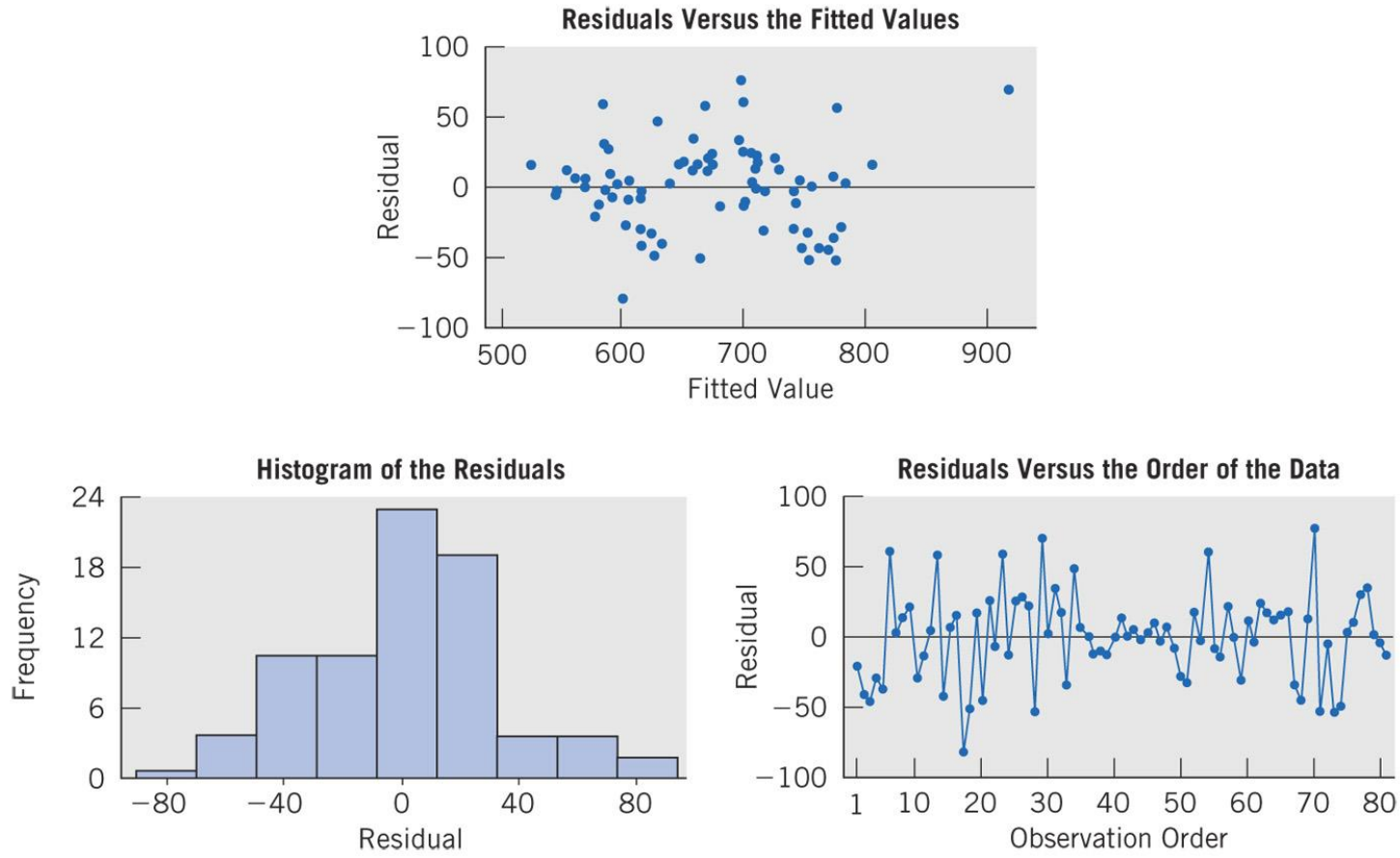


Figure 9 (p. 519)

Three residual plots for posttest rowing time using MINTAB.

It is important to remember that our confidence in statistical inference procedures is related to the validity of the assumptions about them. A mechanically made inference may be misleading if some model assumption is grossly violated. An **examination of the residuals** is an important part of regression analysis, because it helps to detect any inconsistency between the data and the postulated model.

Box on Page 520