Applied Linear Statistical Models

Fifth Edition

Michael H. Kutner
Emory University

Christopher J. Nachtsheim
University of Minnesota

John Neter
University of Georgia

William Li
University of Minnesota

Boston Burr Ridge, IL Dubuque, IA Madison, WI New York San Francisco St Louis
Bangkok Bogotá Caracas Kuala Lumpur Lisbon London Madrid Mexico City
Milan Montreal New Delhi Santiago Seoul Singapore Sydney Taipei Toronto

APPLIED LINEAR STATISTICAL MODELS

This book is printed on acid-free paper.

1 2 3 4 5 6 7 8 9 0 DOC/DOC 0 9 8 7 6 5 4

ISBN 0-07-238688-6

Editorial director: *Brent Gordon*
Executive editor: *Richard T. Hercher, Jr.*
Editorial assistant: *Lee Stone*
Senior marketing manager: *Douglas Reiner*
Media producer: *Elizabeth Mavetz*
Project manager: *Jim Labeots*
Production supervisor: *Gina Hangos*
Lead designer: *Pam Verros*
Supplement producer: *Matthew Perry*
Senior digital content specialist: *Brian Nacik*
Cover design: *Kiera Pohl*
Typeface: *10/12 Times Roman*
Compositor: *Interactive Composition Corporation*
Printer: *R. R. Donnelley*

Library of Congress Cataloging-in-Publication Data

Kutner, Michael H.
Applied linear statistical models.—5th ed. / Michael H. Kutner . . . [et al.].
p. cm. — (McGraw-Hill/Irwin series Operations and decision sciences)
Rev. ed. of: Applied linear regression models. 4th ed. c2004.
Includes bibliographical references and index.
ISBN 0-07-238688-6 (acid-free paper)
1. Regression analysis. 2. Mathematical statistics. I. Kutner, Michael H. Applied linear regression models. II. Title. III. Series.
QA278.2.K87 2005
519.5'36—dc22 2004052447

www.mhhe.com

Contents

ility

Models and
1358

P

Simple Linear Regression

Chapter

Linear Regression with One Predictor Variable

Regression analysis is a statistical methodology that utilizes the relation between two or more quantitative variables so that a response or outcome variable can be predicted from the other, or others. This methodology is widely used in business, the social and behavioral sciences, the biological sciences, and many other disciplines. A few examples of applications are:

1. Sales of a product can be predicted by utilizing the relationship between sales and amount of advertising expenditures.
2. The performance of an employee on a job can be predicted by utilizing the relationship between performance and a battery of aptitude tests.
3. The size of the vocabulary of a child can be predicted by utilizing the relationship between size of vocabulary and age of the child and amount of education of the parents.
4. The length of hospital stay of a surgical patient can be predicted by utilizing the relationship between the time in the hospital and the severity of the operation.

In Part I we take up regression analysis when a single predictor variable is used for predicting the response or outcome variable of interest. In Parts II and III, we consider regression analysis when two or more variables are used for making predictions. In this chapter, we consider the basic ideas of regression analysis and discuss the estimation of the parameters of regression models containing a single predictor variable.

1.1 Relations between Variables

The concept of a relation between two variables, such as between family income and family expenditures for housing, is a familiar one. We distinguish between a *functional relation* and a *statistical relation,* and consider each of these in turn.

Functional Relation between Two Variables

A functional relation between two variables is expressed by a mathematical formula. If X denotes the *independent variable* and Y the *dependent variable,* a functional relation is

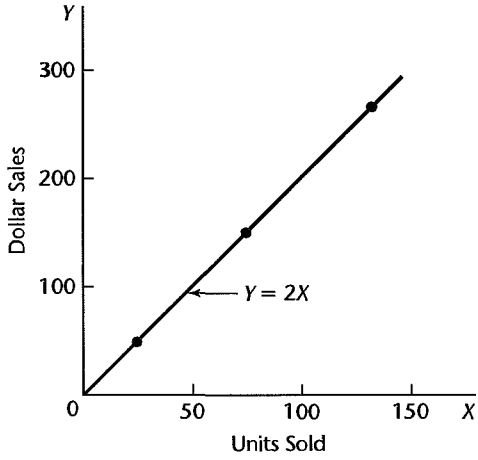


FIGURE 1.1
Example of Functional Relation.

of the form:

$$Y = f(X)$$

Given a particular value of X, the function f indicates the corresponding value of Y.

Example

Consider the relation between dollar sales (Y) of a product sold at a fixed price and number of units sold (X). If the selling price is \$2 per unit, the relation is expressed by the equation:

$$Y = 2X$$

This functional relation is shown in Figure 1.1. Number of units sold and dollar sales during three recent periods (while the unit price remained constant at \$2) were as follows:

Period	Number of Units Sold	Dollar Sales
1	75	\$150
2	25	50
3	130	260

These observations are plotted also in Figure 1.1. Note that all fall directly on the line of functional relationship. This is characteristic of all functional relations.

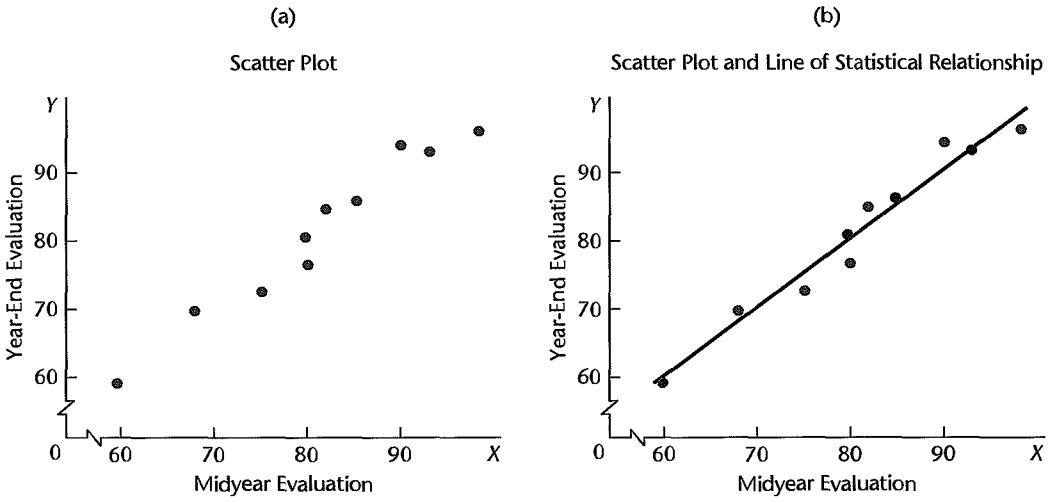
Statistical Relation between Two Variables

A statistical relation, unlike a functional relation, is not a perfect one. In general, the observations for a statistical relation do not fall directly on the curve of relationship.

Example 1

Performance evaluations for 10 employees were obtained at midyear and at year-end. These data are plotted in Figure 1.2a. Year-end evaluations are taken as the *dependent* or *response variable* Y, and midyear evaluations as the *independent, explanatory,* or *predictor*

FIGURE 1.2 Statistical Relation between Midyear Performance Evaluation and Year-End Evaluation.



variable X. The plotting is done as before. For instance, the midyear and year-end performance evaluations for the first employee are plotted at $X = 90$, $Y = 94$.

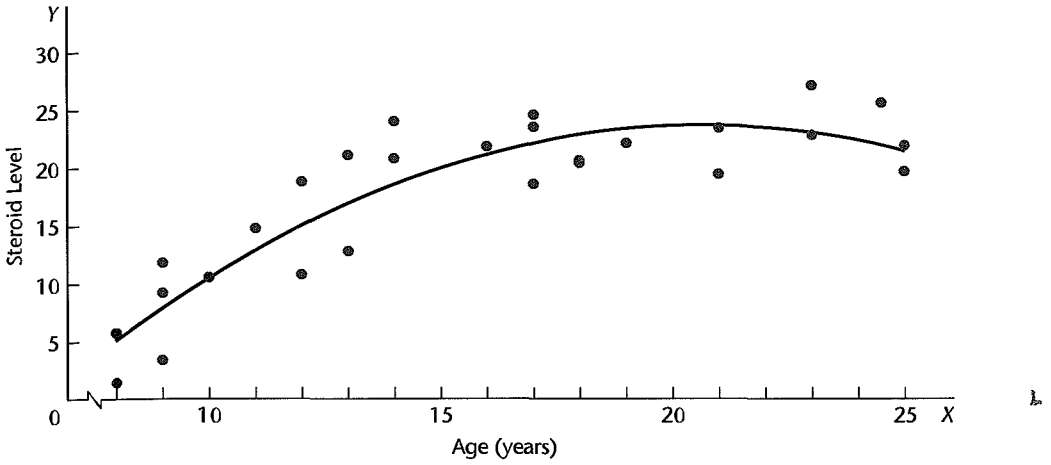
Figure 1.2a clearly suggests that there is a relation between midyear and year-end evaluations, in the sense that the higher the midyear evaluation, the higher tends to be the year-end evaluation. However, the relation is not a perfect one. There is a scattering of points, suggesting that some of the variation in year-end evaluations is not accounted for by midyear performance assessments. For instance, two employees had midyear evaluations of $X = 80$, yet they received somewhat different year-end evaluations. Because of the scattering of points in a statistical relation, Figure 1.2a is called a *scatter diagram* or *scatter plot*. In statistical terminology, each point in the scatter diagram represents a *trial* or a *case*.

In Figure 1.2b, we have plotted a line of relationship that describes the statistical relation between midyear and year-end evaluations. It indicates the general tendency by which year-end evaluations vary with the level of midyear performance evaluation. Note that most of the points do not fall directly on the line of statistical relationship. This scattering of points around the line represents variation in year-end evaluations that is not associated with midyear performance evaluation and that is usually considered to be of a random nature. Statistical relations can be highly useful, even though they do not have the exactitude of a functional relation.

Example 2

Figure 1.3 presents data on age and level of a steroid in plasma for 27 healthy females between 8 and 25 years old. The data strongly suggest that the statistical relationship is *curvilinear* (not linear). The curve of relationship has also been drawn in Figure 1.3. It implies that, as age increases, steroid level increases up to a point and then begins to level off. Note again the scattering of points around the curve of statistical relationship, typical of all statistical relations.

FIGURE 1.3 Curvilinear Statistical Relation between Age and Steroid Level in Healthy Females Aged 8 to 25.



1.2 Regression Models and Their Uses

Historical Origins

Regression analysis was first developed by Sir Francis Galton in the latter part of the 19th century. Galton had studied the relation between heights of parents and children and noted that the heights of children of both tall and short parents appeared to "revert" or "regress" to the mean of the group. He considered this tendency to be a regression to "mediocrity." Galton developed a mathematical description of this regression tendency, the precursor of today's regression models.

The term *regression* persists to this day to describe statistical relations between variables.

Basic Concepts

A regression model is a formal means of expressing the two essential ingredients of a statistical relation:

1. A tendency of the response variable Y to vary with the predictor variable X in a systematic fashion.
2. A scattering of points around the curve of statistical relationship.

These two characteristics are embodied in a regression model by postulating that:

1. There is a probability distribution of Y for each level of X.
2. The means of these probability distributions vary in some systematic fashion with X.

Example

Consider again the performance evaluation example in Figure 1.2. The year-end evaluation Y is treated in a regression model as a random variable. For each level of midyear performance evaluation, there is postulated a probability distribution of Y. Figure 1.4 shows such a probability distribution for $X = 90$, which is the midyear evaluation for the first employee.

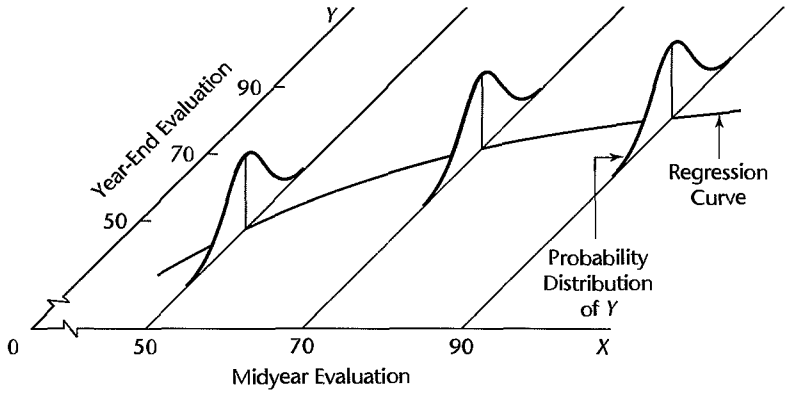


FIGURE 1.4 Pictorial Representation of Regression Model.

The actual year-end evaluation of this employee, $Y = 94$, is then viewed as a random selection from this probability distribution.

Figure 1.4 also shows probability distributions of Y for midyear evaluation levels $X = 50$ and $X = 70$. Note that the means of the probability distributions have a systematic relation to the level of X. This systematic relationship is called the *regression function of Y on X*. The graph of the regression function is called the *regression curve*. Note that in Figure 1.4 the regression function is slightly curvilinear. This would imply for our example that the increase in the expected (mean) year-end evaluation with an increase in midyear performance evaluation is retarded at higher levels of midyear performance.

Regression models may differ in the form of the regression function (linear, curvilinear), in the shape of the probability distributions of Y (symmetrical, skewed), and in other ways. Whatever the variation, the concept of a probability distribution of Y for any given X is the formal counterpart to the empirical scatter in a statistical relation. Similarly, the regression curve, which describes the relation between the means of the probability distributions of Y and the level of X, is the counterpart to the general tendency of Y to vary with X systematically in a statistical relation.

Regression Models with More than One Predictor Variable. Regression models may contain more than one predictor variable. Three examples follow.

1. In an efficiency study of 67 branch offices of a consumer finance chain, the response variable was direct operating cost for the year just ended. There were four predictor variables: average size of loan outstanding during the year, average number of loans outstanding, total number of new loan applications processed, and an index of office salaries.

2. In a tractor purchase study, the response variable was volume (in horsepower) of tractor purchases in a sales territory of a farm equipment firm. There were nine predictor variables, including average age of tractors on farms in the territory, number of farms in the territory, and a quantity index of crop production in the territory.

3. In a medical study of short children, the response variable was the peak plasma growth hormone level. There were 14 predictor variables, including age, gender, height, weight, and 10 skinfold measurements.

The model features represented in Figure 1.4 must be extended into further dimensions when there is more than one predictor variable. With two predictor variables X_1 and X_2,

for instance, a probability distribution of Y for each (X_1, X_2) combination is assumed by the regression model. The systematic relation between the means of these probability distributions and the predictor variables X_1 and X_2 is then given by a regression surface.

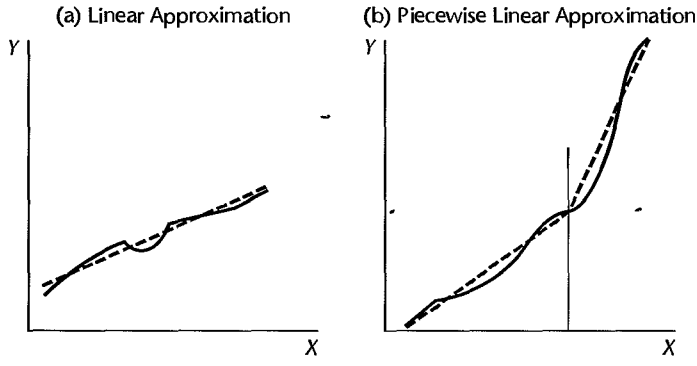
Construction of Regression Models

Selection of Predictor Variables. Since reality must be reduced to manageable proportions whenever we construct models, only a limited number of explanatory or predictor variables can—or should—be included in a regression model for any situation of interest. A central problem in many exploratory studies is therefore that of choosing, for a regression model, a set of predictor variables that is "good" in some sense for the purposes of the analysis. A major consideration in making this choice is the extent to which a chosen variable contributes to reducing the remaining variation in Y after allowance is made for the contributions of other predictor variables that have tentatively been included in the regression model. Other considerations include the importance of the variable as a causal agent in the process under analysis; the degree to which observations on the variable can be obtained more accurately, or quickly, or economically than on competing variables; and the degree to which the variable can be controlled. In Chapter 9, we will discuss procedures and problems in choosing the predictor variables to be included in the regression model.

Functional Form of Regression Relation. The choice of the functional form of the regression relation is tied to the choice of the predictor variables. Sometimes, relevant theory may indicate the appropriate functional form. Learning theory, for instance, may indicate that the regression function relating unit production cost to the number of previous times the item has been produced should have a specified shape with particular asymptotic properties.

More frequently, however, the functional form of the regression relation is not known in advance and must be decided upon empirically once the data have been collected. Linear or quadratic regression functions are often used as satisfactory first approximations to regression functions of unknown nature. Indeed, these simple types of regression functions may be used even when theory provides the relevant functional form, notably when the known form is highly complex but can be reasonably approximated by a linear or quadratic regression function. Figure 1.5a illustrates a case where the complex regression function

FIGURE 1.5 Uses of Linear Regression Functions to Approximate Complex Regression Functions—Bold Line Is the True Regression Function and Dotted Line Is the Regression Approximation.



may be reasonably approximated by a linear regression function. Figure 1.5b provides an example where two linear regression functions may be used "piecewise" to approximate a complex regression function.

Scope of Model. In formulating a regression model, we usually need to restrict the coverage of the model to some interval or region of values of the predictor variable(s). The scope is determined either by the design of the investigation or by the range of data at hand. For instance, a company studying the effect of price on sales volume investigated six price levels, ranging from \$4.95 to \$6.95. Here, the scope of the model is limited to price levels ranging from near \$5 to near \$7. The shape of the regression function substantially outside this range would be in serious doubt because the investigation provided no evidence as to the nature of the statistical relation below \$4.95 or above \$6.95.

Uses of Regression Analysis

Regression analysis serves three major purposes: (1) description, (2) control, and (3) prediction. These purposes are illustrated by the three examples cited earlier. The tractor purchase study served a descriptive purpose. In the study of branch office operating costs, the main purpose was administrative control; by developing a usable statistical relation between cost and the predictor variables, management was able to set cost standards for each branch office in the company chain. In the medical study of short children, the purpose was prediction. Clinicians were able to use the statistical relation to predict growth hormone deficiencies in short children by using simple measurements of the children.

The several purposes of regression analysis frequently overlap in practice. The branch office example is a case in point. Knowledge of the relation between operating cost and characteristics of the branch office not only enabled management to set cost standards for each office but management could also predict costs, and at the end of the fiscal year it could compare the actual branch cost against the expected cost.

Regression and Causality

The existence of a statistical relation between the response variable Y and the explanatory or predictor variable X does not imply in any way that Y depends causally on X. No matter how strong is the statistical relation between X and Y, no cause-and-effect pattern is necessarily implied by the regression model. For example, data on size of vocabulary (X) and writing speed (Y) for a sample of young children aged 5–10 will show a positive regression relation. This relation does not imply, however, that an increase in vocabulary causes a faster writing speed. Here, other explanatory variables, such as age of the child and amount of education, affect both the vocabulary (X) and the writing speed (Y). Older children have a larger vocabulary and a faster writing speed.

Even when a strong statistical relationship reflects causal conditions, the causal conditions may act in the opposite direction, from Y to X. Consider, for instance, the calibration of a thermometer. Here, readings of the thermometer are taken at different known temperatures, and the regression relation is studied so that the accuracy of predictions made by using the thermometer readings can be assessed. For this purpose, the thermometer reading is the predictor variable X, and the actual temperature is the response variable Y to be predicted. However, the causal pattern here does not go from X to Y, but in the opposite direction: the actual temperature (Y) affects the thermometer reading (X).

These examples demonstrate the need for care in drawing conclusions about causal relations from regression analysis. Regression analysis by itself provides no information about causal patterns and must be supplemented by additional analyses to obtain insights about causal relations.

Use of Computers

Because regression analysis often entails lengthy and tedious calculations, computers are usually utilized to perform the necessary calculations. Almost every statistics package for computers contains a regression component. While packages differ in many details, their basic regression output tends to be quite similar.

After an initial explanation of required regression calculations, we shall rely on computer calculations for all subsequent examples. We illustrate computer output by presenting output and graphics from BMDP (Ref. 1.1), MINITAB (Ref. 1.2), SAS (Ref. 1.3), SPSS (Ref. 1.4), SYSTAT (Ref. 1.5), JMP (Ref. 1.6), S-Plus (Ref. 1.7), and MATLAB (Ref. 1.8).

1.3 Simple Linear Regression Model with Distribution of Error Terms Unspecified

Formal Statement of Model

In Part I we consider a basic regression model where there is only one predictor variable and the regression function is linear. The model can be stated as follows:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{1.1}$$

where:

Y_i is the value of the response variable in the ith trial

β_0 and β_1 are parameters

X_i is a known constant, namely, the value of the predictor variable in the ith trial

ε_i is a random error term with mean $E\{\varepsilon_i\} = 0$ and variance $\sigma^2\{\varepsilon_i\} = \sigma^2$; ε_i and ε_j are uncorrelated so that their covariance is zero (i.e., $\sigma\{\varepsilon_i, \varepsilon_j\} = 0$ for all $i, j; i \neq j$)

$i = 1, \ldots, n$

Regression model (1.1) is said to be *simple, linear in the parameters,* and *linear in the predictor variable.* It is "simple" in that there is only one predictor variable, "linear in the parameters," because no parameter appears as an exponent or is multiplied or divided by another parameter, and "linear in the predictor variable," because this variable appears only in the first power. A model that is linear in the parameters and in the predictor variable is also called a *first-order model.*

Important Features of Model

1. The response Y_i in the ith trial is the sum of two components: (1) the constant term $\beta_0 + \beta_1 X_i$ and (2) the random term ε_i. Hence, Y_i is a random variable.
2. Since $E\{\varepsilon_i\} = 0$, it follows from (A.13c) in Appendix A that:

$$E\{Y_i\} = E\{\beta_0 + \beta_1 X_i + \varepsilon_i\} = \beta_0 + \beta_1 X_i + E\{\varepsilon_i\} = \beta_0 + \beta_1 X_i$$

Note that $\beta_0 + \beta_1 X_i$ plays the role of the constant a in (A.13c).

Thus, the response Y_i, when the level of X in the ith trial is X_i, comes from a probability distribution whose mean is:

$$E\{Y_i\} = \beta_0 + \beta_1 X_i \tag{1.2}$$

We therefore know that the regression function for model (1.1) is:

$$E\{Y\} = \beta_0 + \beta_1 X \tag{1.3}$$

since the regression function relates the means of the probability distributions of Y for given X to the level of X.

3. The response Y_i in the ith trial exceeds or falls short of the value of the regression function by the error term amount ε_i.

4. The error terms ε_i are assumed to have constant variance σ^2. It therefore follows that the responses Y_i have the same constant variance:

$$\sigma^2\{Y_i\} = \sigma^2 \tag{1.4}$$

since, using (A.16a), we have:

$$\sigma^2\{\beta_0 + \beta_1 X_i + \varepsilon_i\} = \sigma^2\{\varepsilon_i\} = \sigma^2$$

Thus, regression model (1.1) assumes that the probability distributions of Y have the same variance σ^2, regardless of the level of the predictor variable X.

5. The error terms are assumed to be uncorrelated. Since the error terms ε_i and ε_j are uncorrelated, so are the responses Y_i and Y_j.

6. In summary, regression model (1.1) implies that the responses Y_i come from probability distributions whose means are $E\{Y_i\} = \beta_0 + \beta_1 X_i$ and whose variances are σ^2, the same for all levels of X. Further, any two responses Y_i and Y_j are uncorrelated.

Example

A consultant for an electrical distributor is studying the relationship between the number of bids requested by construction contractors for basic lighting equipment during a week and the time required to prepare the bids. Suppose that regression model (1.1) is applicable and is as follows:

$$Y_i = 9.5 + 2.1X_i + \varepsilon_i$$

where X is the number of bids prepared in a week and Y is the number of hours required to prepare the bids. Figure 1.6 contains a presentation of the regression function:

$$E\{Y\} = 9.5 + 2.1X$$

Suppose that in the ith week, $X_i = 45$ bids are prepared and the actual number of hours required is $Y_i = 108$. In that case, the error term value is $\varepsilon_i = 4$, for we have

$$E\{Y_i\} = 9.5 + 2.1(45) = 104$$

and

$$Y_i = 108 = 104 + 4$$

Figure 1.6 displays the probability distribution of Y when $X = 45$ and indicates from where in this distribution the observation $Y_i = 108$ came. Note again that the error term ε_i is simply the deviation of Y_i from its mean value $E\{Y_i\}$.

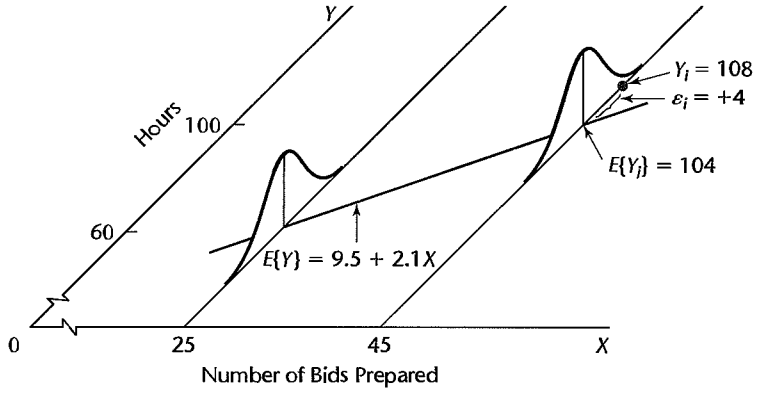


FIGURE 1.6
Illustration of Simple Linear Regression Model (1.1).

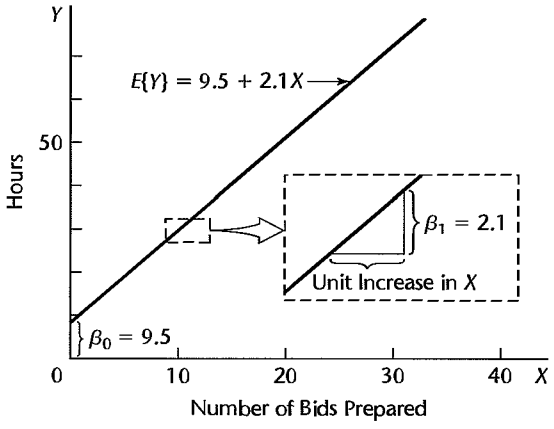


FIGURE 1.7
Meaning of Parameters of Simple Linear Regression Model (1.1).

Figure 1.6 also shows the probability distribution of Y when $X = 25$. Note that this distribution exhibits the same variability as the probability distribution when $X = 45$, in conformance with the requirements of regression model (1.1).

Meaning of Regression Parameters

The parameters β_0 and β_1 in regression model (1.1) are called *regression coefficients*. β_1 is the slope of the regression line. It indicates the change in the mean of the probability distribution of Y per unit increase in X. The parameter β_0 is the Y intercept of the regression line. When the scope of the model includes $X = 0$, β_0 gives the mean of the probability distribution of Y at $X = 0$. When the scope of the model does not cover $X = 0$, β_0 does not have any particular meaning as a separate term in the regression model.

Example

Figure 1.7 shows the regression function:

$$E\{Y\} = 9.5 + 2.1X$$

for the electrical distributor example. The slope $\beta_1 = 2.1$ indicates that the preparation of one additional bid in a week leads to an increase in the mean of the probability distribution of Y of 2.1 hours.

The intercept $\beta_0 = 9.5$ indicates the value of the regression function at $X = 0$. However, since the linear regression model was formulated to apply to weeks where the number of

bids prepared ranges from 20 to 80, β_0 does not have any intrinsic meaning of its own here. If the scope of the model were to be extended to X levels near zero, a model with a curvilinear regression function and some value of β_0 different from that for the linear regression function might well be required.

Alternative Versions of Regression Model

Sometimes it is convenient to write the simple linear regression model (1.1) in somewhat different, though equivalent, forms. Let X_0 be a constant identically equal to 1. Then, we can write (1.1) as follows:

$$Y_i = \beta_0 X_0 + \beta_1 X_i + \varepsilon_i \qquad \text{where } X_0 \equiv 1 \tag{1.5}$$

This version of the model associates an X variable with each regression coefficient.

An alternative modification is to use for the predictor variable the deviation $X_i - \bar{X}$ rather than X_i. To leave model (1.1) unchanged, we need to write:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1(X_i - \bar{X}) + \beta_1 \bar{X} + \varepsilon_i \\ &= (\beta_0 + \beta_1 \bar{X}) + \beta_1(X_i - \bar{X}) + \varepsilon_i \\ &= \beta_0^* + \beta_1(X_i - \bar{X}) + \varepsilon_i \end{aligned}$$

Thus, this alternative model version is:

$$Y_i = \beta_0^* + \beta_1(X_i - \bar{X}) + \varepsilon_i \tag{1.6}$$

where:

$$\beta_0^* = \beta_0 + \beta_1 \bar{X} \tag{1.6a}$$

We use models (1.1), (1.5), and (1.6) interchangeably as convenience dictates.

1.4 Data for Regression Analysis

Ordinarily, we do not know the values of the regression parameters β_0 and β_1 in regression model (1.1), and we need to estimate them from relevant data. Indeed, as we noted earlier, we frequently do not have adequate *a priori* knowledge of the appropriate predictor variables and of the functional form of the regression relation (e.g., linear or curvilinear), and we need to rely on an analysis of the data for developing a suitable regression model.

Data for regression analysis may be obtained from nonexperimental or experimental studies. We consider each of these in turn.

Observational Data

Observational data are data obtained from nonexperimental studies. Such studies do not control the explanatory or predictor variable(s) of interest. For example, company officials wished to study the relation between age of employee (X) and number of days of illness last year (Y). The needed data for use in the regression analysis were obtained from personnel records. Such data are observational data since the explanatory variable, age, is not controlled.

Regression analyses are frequently based on observational data, since often it is not feasible to conduct controlled experimentation. In the company personnel example just mentioned, for instance, it would not be possible to control age by assigning ages to persons.

A major limitation of observational data is that they often do not provide adequate information about cause-and-effect relationships. For example, a positive relation between age of employee and number of days of illness in the company personnel example may not imply that number of days of illness is the direct result of age. It might be that younger employees of the company primarily work indoors while older employees usually work outdoors, and that work location is more directly responsible for the number of days of illness than age.

Whenever a regression analysis is undertaken for purposes of description based on observational data, one should investigate whether explanatory variables other than those considered in the regression model might more directly explain cause-and-effect relationships.

Experimental Data

Frequently, it is possible to conduct a controlled experiment to provide data from which the regression parameters can be estimated. Consider, for instance, an insurance company that wishes to study the relation between productivity of its analysts in processing claims and length of training. Nine analysts are to be used in the study. Three of them will be selected at random and trained for two weeks, three for three weeks, and three for five weeks. The productivity of the analysts during the next 10 weeks will then be observed. The data so obtained will be experimental data because control is exercised over the explanatory variable, length of training.

When control over the explanatory variable(s) is exercised through random assignments, as in the productivity study example, the resulting experimental data provide much stronger information about cause-and-effect relationships than do observational data. The reason is that randomization tends to balance out the effects of any other variables that might affect the response variable, such as the effect of aptitude of the employee on productivity.

In the terminology of experimental design, the length of training assigned to an analyst in the productivity study example is called a *treatment*. The analysts to be included in the study are called the *experimental units*. Control over the explanatory variable(s) then consists of assigning a treatment to each of the experimental units by means of randomization.

Completely Randomized Design

The most basic type of statistical design for making randomized assignments of treatments to experimental units (or vice versa) is the *completely randomized design*. With this design, the assignments are made completely at random. This complete randomization provides that all combinations of experimental units assigned to the different treatments are equally likely, which implies that every experimental unit has an equal chance to receive any one of the treatments.

A completely randomized design is particularly useful when the experimental units are quite homogeneous. This design is very flexible; it accommodates any number of treatments and permits different sample sizes for different treatments. Its chief disadvantage is that, when the experimental units are heterogeneous, this design is not as efficient as some other statistical designs.

1.5 Overview of Steps in Regression Analysis

The regression models considered in this and subsequent chapters can be utilized either for observational data or for experimental data from a completely randomized design. (Regression analysis can also utilize data from other types of experimental designs, but

the regression models presented here will need to be modified.) Whether the data are observational or experimental, it is essential that the conditions of the regression model be appropriate for the data at hand for the model to be applicable.

We begin our discussion of regression analysis by considering inferences about the regression parameters for the simple linear regression model (1.1). For the rare occasion where prior knowledge or theory alone enables us to determine the appropriate regression model, inferences based on the regression model are the first step in the regression analysis. In the usual situation, however, where we do not have adequate knowledge to specify the appropriate regression model in advance, the first step is an exploratory study of the data, as shown in the flowchart in Figure 1.8. On the basis of this initial exploratory analysis, one or more preliminary regression models are developed. These regression models are then examined for their appropriateness for the data at hand and revised, or new models

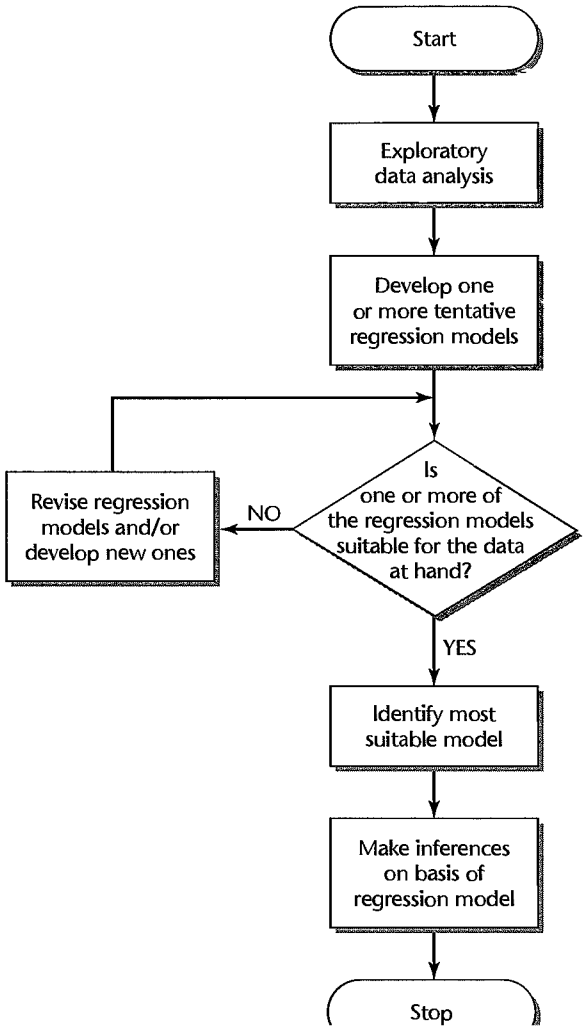


FIGURE 1.8 Typical Strategy for Regression Analysis.

are developed, until the investigator is satisfied with the suitability of a particular regression model. Only then are inferences made on the basis of this regression model, such as inferences about the regression parameters of the model or predictions of new observations.

We begin, for pedagogic reasons, with inferences based on the regression model that is finally considered to be appropriate. One must have an understanding of regression models and how they can be utilized before the issues involved in the development of an appropriate regression model can be fully explained.

1.6 Estimation of Regression Function

The observational or experimental data to be used for estimating the parameters of the regression function consist of observations on the explanatory or predictor variable X and the corresponding observations on the response variable Y. For each trial, there is an X observation and a Y observation. We denote the (X, Y) observations for the first trial as (X_1, Y_1), for the second trial as (X_2, Y_2), and in general for the ith trial as (X_i, Y_i), where $i = 1, \ldots, n$.

Example

In a small-scale study of persistence, an experimenter gave three subjects a very difficult task. Data on the age of the subject (X) and on the number of attempts to accomplish the task before giving up (Y) follow:

Subject i:	1	2	3
Age X_i:	20	55	30
Number of attempts Y_i:	5	12	10

In terms of the notation to be employed, there were $n = 3$ subjects in this study, the observations for the first subject were $(X_1, Y_1) = (20, 5)$, and similarly for the other subjects.

Method of Least Squares

To find "good" estimators of the regression parameters β_0 and β_1, we employ the method of least squares. For the observations (X_i, Y_i) for each case, the method of least squares considers the deviation of Y_i from its expected value:

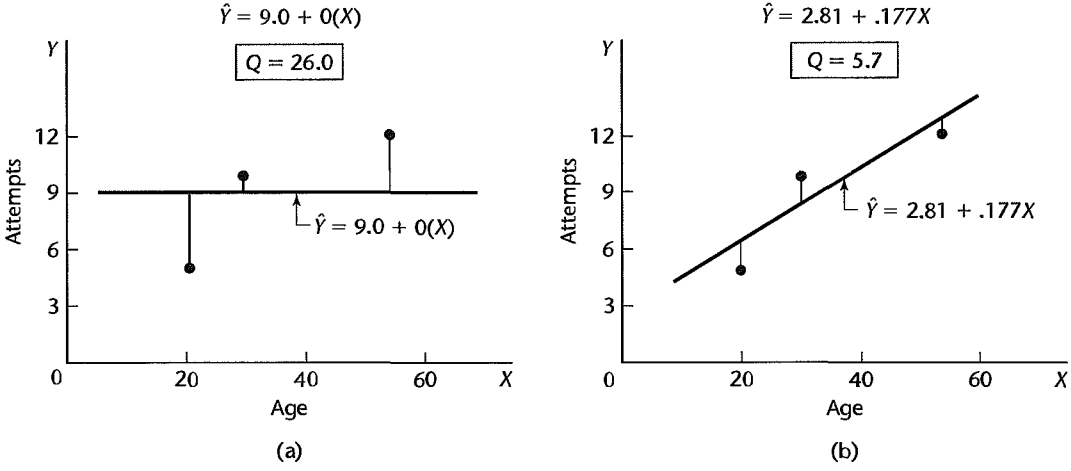
$$Y_i - (\beta_0 + \beta_1 X_i) \tag{1.7}$$

In particular, the method of least squares requires that we consider the sum of the n squared deviations. This criterion is denoted by Q:

$$Q = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 \tag{1.8}$$

According to the method of least squares, the estimators of β_0 and β_1 are those values b_0 and b_1, respectively, that minimize the criterion Q for the given sample observations $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$.

FIGURE 1.9 Illustration of Least Squares Criterion Q for Fit of a Regression Line—Persistence Study Example.



Example

Figure 1.9a presents the scatter plot of the data for the persistence study example and the regression line that results when we use the mean of the responses (9.0) as the predictor and ignore X:

$$\hat{Y} = 9.0 + 0(X)$$

Note that this regression line uses estimates $b_0 = 9.0$ and $b_1 = 0$, and that $\hat{Y}$ denotes the ordinate of the estimated regression line. Clearly, this regression line is not a good fit, as evidenced by the large vertical deviations of two of the Y observations from the corresponding ordinates $\hat{Y}$ of the regression line. The deviation for the first subject, for which $(X_1, Y_1) = (20, 5)$, is:

$$Y_1 - (b_0 + b_1 X_1) = 5 - [9.0 + 0(20)] = 5 - 9.0 = -4$$

The sum of the squared deviations for the three cases is:

$$Q = (5 - 9.0)^2 + (12 - 9.0)^2 + (10 - 9.0)^2 = 26.0$$

Figure 1.9b shows the same data with the regression line:

$$\hat{Y} = 2.81 + .177X$$

The fit of this regression line is clearly much better. The vertical deviation for the first case now is:

$$Y_1 - (b_0 + b_1 X_1) = 5 - [2.81 + .177(20)] = 5 - 6.35 = -1.35$$

and the criterion Q is much reduced:

$$Q = (5 - 6.35)^2 + (12 - 12.55)^2 + (10 - 8.12)^2 = 5.7$$

Thus, a better fit of the regression line to the data corresponds to a smaller sum Q.

The objective of the method of least squares is to find estimates b_0 and b_1 for β_0 and β_1, respectively, for which Q is a minimum. In a certain sense, to be discussed shortly, these

estimates will provide a "good" fit of the linear regression function. The regression line in Figure 1.9b is, in fact, the least squares regression line.

Least Squares Estimators. The estimators b_0 and b_1 that satisfy the least squares criterion can be found in two basic ways:

1. Numerical search procedures can be used that evaluate in a systematic fashion the least squares criterion Q for different estimates b_0 and b_1 until the ones that minimize Q are found. This approach was illustrated in Figure 1.9 for the persistence study example.
2. Analytical procedures can often be used to find the values of b_0 and b_1 that minimize Q. The analytical approach is feasible when the regression model is not mathematically complex.

Using the analytical approach, it can be shown for regression model (1.1) that the values b_0 and b_1 that minimize Q for any particular set of sample data are given by the following simultaneous equations:

$$\sum Y_i = nb_0 + b_1 \sum X_i \tag{1.9a}$$

$$\sum X_i Y_i = b_0 \sum X_i + b_1 \sum X_i^2 \tag{1.9b}$$

Equations (1.9a) and (1.9b) are called *normal equations;* b_0 and b_1 are called *point estimators* of β_0 and β_1, respectively.

The normal equations (1.9) can be solved simultaneously for b_0 and b_1:

$$b_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} \tag{1.10a}$$

$$b_0 = \frac{1}{n}\left(\sum Y_i - b_1 \sum X_i\right) = \bar{Y} - b_1 \bar{X} \tag{1.10b}$$

where $\bar{X}$ and $\bar{Y}$ are the means of the X_i and the Y_i observations, respectively. Computer calculations generally are based on many digits to obtain accurate values for b_0 and b_1.

Comment

The normal equations (1.9) can be derived by calculus. For given sample observations (X_i, Y_i), the quantity Q in (1.8) is a function of β_0 and β_1. The values of β_0 and β_1 that minimize Q can be derived by differentiating (1.8) with respect to β_0 and β_1. We obtain:

$$\frac{\partial Q}{\partial \beta_0} = -2\sum(Y_i - \beta_0 - \beta_1 X_i)$$

$$\frac{\partial Q}{\partial \beta_1} = -2\sum X_i(Y_i - \beta_0 - \beta_1 X_i)$$

We then set these partial derivatives equal to zero, using b_0 and b_1 to denote the particular values of β_0 and β_1 that minimize Q:

$$-2\sum(Y_i - b_0 - b_1 X_i) = 0$$

$$-2\sum X_i(Y_i - b_0 - b_1 X_i) = 0$$

Simplifying, we obtain:

$$\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i) = 0$$

$$\sum_{i=1}^{n} X_i(Y_i - b_0 - b_1 X_i) = 0$$

Expanding, we have:

$$\sum Y_i - nb_0 - b_1 \sum X_i = 0$$

$$\sum X_i Y_i - b_0 \sum X_i - b_1 \sum X_i^2 = 0$$

from which the normal equations (1.9) are obtained by rearranging terms.

A test of the second partial derivatives will show that a minimum is obtained with the least squares estimators b_0 and b_1. ■

Properties of Least Squares Estimators. An important theorem, called the *Gauss-Markov theorem,* states:

> Under the conditions of regression model (1.1), the least squares estimators b_0 and b_1 in (1.10) are unbiased and have minimum variance among all unbiased linear estimators. **(1.11)**

This theorem, proven in the next chapter, states first that b_0 and b_1 are unbiased estimators. Hence:

$$E\{b_0\} = \beta_0 \qquad E\{b_1\} = \beta_1$$

so that neither estimator tends to overestimate or underestimate systematically.

Second, the theorem states that the estimators b_0 and b_1 are more precise (i.e., their sampling distributions are less variable) than any other estimators belonging to the class of unbiased estimators that are linear functions of the observations $Y_1, \ldots, Y_n$. The estimators b_0 and b_1 are such linear functions of the Y_i. Consider, for instance, b_1. We have from (1.10a):

$$b_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

It will be shown in Chapter 2 that this expression is equal to:

$$b_1 = \frac{\sum(X_i - \bar{X})Y_i}{\sum(X_i - \bar{X})^2} = \sum k_i Y_i$$

where:

$$k_i = \frac{X_i - \bar{X}}{\sum(X_i - \bar{X})^2}$$

Since the k_i are known constants (because the X_i are known constants), b_1 is a linear combination of the Y_i and hence is a linear estimator.

In the same fashion, it can be shown that b_0 is a linear estimator. Among all linear estimators that are unbiased then, b_0 and b_1 have the smallest variability in repeated samples in which the X levels remain unchanged.

Example

The Toluca Company manufactures refrigeration equipment as well as many replacement parts. In the past, one of the replacement parts has been produced periodically in lots of varying sizes. When a cost improvement program was undertaken, company officials wished to determine the optimum lot size for producing this part. The production of this part involves setting up the production process (which must be done no matter what is the lot size) and machining and assembly operations. One key input for the model to ascertain the optimum lot size was the relationship between lot size and labor hours required to produce the lot. To determine this relationship, data on lot size and work hours for 25 recent production runs were utilized. The production conditions were stable during the six-month period in which the 25 runs were made and were expected to continue to be the same during the next three years, the planning period for which the cost improvement program was being conducted.

Table 1.1 contains a portion of the data on lot size and work hours in columns 1 and 2. Note that all lot sizes are multiples of 10, a result of company policy to facilitate the administration of the parts production. Figure 1.10a shows a SYSTAT scatter plot of the data. We see that the lot sizes ranged from 20 to 120 units and that none of the production runs was outlying in the sense of being either unusually small or large. The scatter plot also indicates that the relationship between lot size and work hours is reasonably linear. We also see that no observations on work hours are unusually small or large, with reference to the relationship between lot size and work hours.

To calculate the least squares estimates b_0 and b_1 in (1.10), we require the deviations $X_i - \bar{X}$ and $Y_i - \bar{Y}$. These are given in columns 3 and 4 of Table 1.1. We also require the cross-product terms $(X_i - \bar{X})(Y_i - \bar{Y})$ and the squared deviations $(X_i - \bar{X})^2$; these are shown in columns 5 and 6. The squared deviations $(Y_i - \bar{Y})^2$ in column 7 are for later use.

TABLE 1.1 **Data on Lot Size and Work Hours and Needed Calculations for Least Squares Estimates—Toluca Company Example.**

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Run i	Lot Size X_i	Work Hours Y_i	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X})(Y_i - \bar{Y})$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})^2$
1	80	399	10	86.72	867.2	100	7,520.4
2	30	121	−40	−191.28	7,651.2	1,600	36,588.0
3	50	221	−20	−91.28	1,825.6	400	8,332.0
...	...	...	...	...	...	...	...
23	40	244	−30	−68.28	2,048.4	900	4,662.2
24	80	342	10	29.72	297.2	100	883.3
25	70	323	0	10.72	0.0	0	114.9
Total	1,750	7,807	0	0	70,690	19,800	307,203
Mean	70.0	312.28					

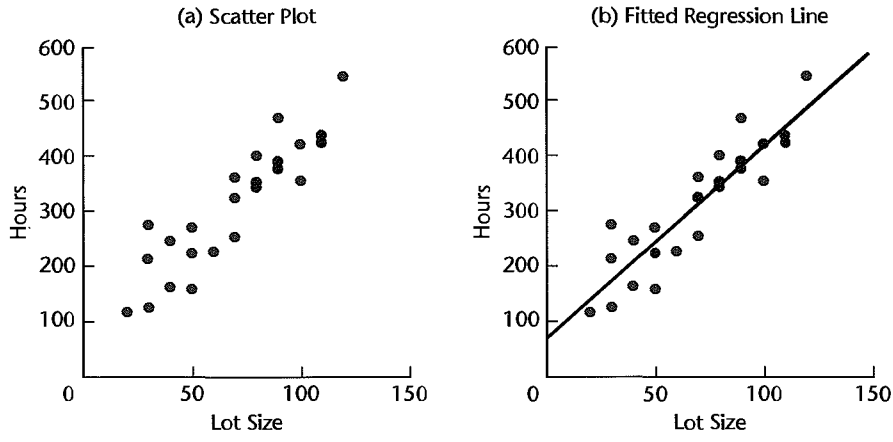


FIGURE 1.10 SYSTAT Scatter Plot and Fitted Regression Line—Toluca Company Example.

FIGURE 1.11 Portion of MINITAB Regression Output—Toluca Company Example.

```
The regression equation is
Y = 62.4 + 3.57 X

Predictor       Coef      Stdev    t-ratio       p
Constant       62.37      26.18       2.38   0.026
X             3.5702     0.3470      10.29   0.000

s = 48.82       R-sq = 82.2%     R-sq(adj) = 81.4%
```

We see from Table 1.1 that the basic quantities needed to calculate the least squares estimates are as follows:

$$\sum(X_i - \bar{X})(Y_i - \bar{Y}) = 70{,}690$$
$$\sum(X_i - \bar{X})^2 = 19{,}800$$
$$\bar{X} = 70.0$$
$$\bar{Y} = 312.28$$

Using (1.10) we obtain:

$$b_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} = \frac{70{,}690}{19{,}800} = 3.5702$$

$$b_0 = \bar{Y} - b_1\bar{X} = 312.28 - 3.5702(70.0) = 62.37$$

Thus, we estimate that the mean number of work hours increases by 3.57 hours for each additional unit produced in the lot. This estimate applies to the range of lot sizes in the data from which the estimates were derived, namely to lot sizes ranging from about 20 to about 120.

Figure 1.11 contains a portion of the MINITAB regression output for the Toluca Company example. The estimates b_0 and b_1 are shown in the column labeled Coef, corresponding to

the lines Constant and X, respectively. The additional information shown in Figure 1.11 will be explained later.

Point Estimation of Mean Response

Estimated Regression Function. Given sample estimators b_0 and b_1 of the parameters in the regression function (1.3):

$$E\{Y\} = \beta_0 + \beta_1 X$$

we estimate the regression function as follows:

$$\hat{Y} = b_0 + b_1 X \tag{1.12}$$

where $\hat{Y}$ (read Y hat) is the value of the estimated regression function at the level X of the predictor variable.

We call a *value* of the response variable a *response* and $E\{Y\}$ the *mean response*. Thus, the mean response stands for the mean of the probability distribution of Y corresponding to the level X of the predictor variable. $\hat{Y}$ then is a point estimator of the mean response when the level of the predictor variable is X. It can be shown as an extension of the Gauss-Markov theorem (1.11) that $\hat{Y}$ is an unbiased estimator of $E\{Y\}$, with minimum variance in the class of unbiased linear estimators.

For the cases in the study, we will call $\hat{Y}_i$:

$$\hat{Y}_i = b_0 + b_1 X_i \qquad i = 1, \ldots, n \tag{1.13}$$

the *fitted value* for the ith case. Thus, the fitted value $\hat{Y}_i$ is to be viewed in distinction to the *observed value* Y_i.

Example

For the Toluca Company example, we found that the least squares estimates of the regression coefficients are:

$$b_0 = 62.37 \qquad b_1 = 3.5702$$

Hence, the estimated regression function is:

$$\hat{Y} = 62.37 + 3.5702X$$

This estimated regression function is plotted in Figure 1.10b. It appears to be a good description of the statistical relationship between lot size and work hours.

To estimate the mean response for any level X of the predictor variable, we simply substitute that value of X in the estimated regression function. Suppose that we are interested in the mean number of work hours required when the lot size is $X = 65$; our point estimate is:

$$\hat{Y} = 62.37 + 3.5702(65) = 294.4$$

Thus, we estimate that the mean number of work hours required for production runs of $X = 65$ units is 294.4 hours. We interpret this to mean that if many lots of 65 units are produced under the conditions of the 25 runs on which the estimated regression function is based, the mean labor time for these lots is about 294 hours. Of course, the labor time for any one lot of size 65 is likely to fall above or below the mean response because of inherent variability in the production system, as represented by the error term in the model.

TABLE 1.2 **Fitted Values, Residuals, and Squared Residuals—Toluca Company Example.**

Run i	(1) Lot Size X_i	(2) Work Hours Y_i	(3) Estimated Mean Response $\hat{Y}_i$	(4) Residual $Y_i - \hat{Y}_i = e_i$	(5) Squared Residual $(Y_i - \hat{Y}_i)^2 = e_i^2$
1	80	399	347.98	51.02	2,603.0
2	30	121	169.47	−48.47	2,349.3
3	50	221	240.88	−19.88	395.2
...	...	...	...	...	...
23	40	244	205.17	38.83	1,507.8
24	80	342	347.98	−5.98	35.8
25	70	323	312.28	10.72	114.9
Total	1,750	7,807	7,807	0	54,825

Fitted values for the sample cases are obtained by substituting the appropriate X values into the estimated regression function. For the first sample case, we have $X_1 = 80$. Hence, the fitted value for the first case is:

$$\hat{Y}_1 = 62.37 + 3.5702(80) = 347.98$$

This compares with the observed work hours of $Y_1 = 399$. Table 1.2 contains the observed and fitted values for a portion of the Toluca Company data in columns 2 and 3, respectively.

Alternative Model (1.6). When the alternative regression model (1.6):

$$Y_i = \beta_0^* + \beta_1(X_i - \bar{X}) + \varepsilon_i$$

is to be utilized, the least squares estimator b_1 of β_1 remains the same as before. The least squares estimator of $\beta_0^* = \beta_0 + \beta_1\bar{X}$ becomes, from (1.10b):

$$b_0^* = b_0 + b_1\bar{X} = (\bar{Y} - b_1\bar{X}) + b_1\bar{X} = \bar{Y} \tag{1.14}$$

Hence, the estimated regression function for alternative model (1.6) is:

$$\hat{Y} = \bar{Y} + b_1(X - \bar{X}) \tag{1.15}$$

In the Toluca Company example, $\bar{Y} = 312.28$ and $\bar{X} = 70.0$ (Table 1.1). Hence, the estimated regression function in alternative form is:

$$\hat{Y} = 312.28 + 3.5702(X - 70.0)$$

For the first lot in our example, $X_1 = 80$; hence, we estimate the mean response to be:

$$\hat{Y}_1 = 312.28 + 3.5702(80 - 70.0) = 347.98$$

which, of course, is identical to our earlier result.

Residuals

The ith *residual* is the difference between the observed value Y_i and the corresponding fitted value $\hat{Y}_i$. This residual is denoted by e_i and is defined in general as follows:

$$e_i = Y_i - \hat{Y}_i \tag{1.16}$$

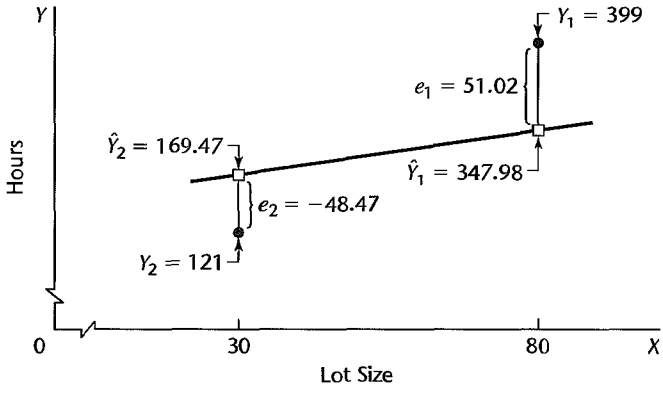


FIGURE 1.12 Illustration of Residuals—Toluca Company Example (not drawn to scale).

For regression model (1.1), the residual e_i becomes:

$$e_i = Y_i - (b_0 + b_1 X_i) = Y_i - b_0 - b_1 X_i \qquad \textbf{(1.16a)}$$

The calculation of the residuals for the Toluca Company example is shown for a portion of the data in Table 1.2. We see that the residual for the first case is:

$$e_1 = Y_1 - \hat{Y}_1 = 399 - 347.98 = 51.02$$

The residuals for the first two cases are illustrated graphically in Figure 1.12. Note in this figure that the magnitude of a residual is represented by the vertical deviation of the Y_i observation from the corresponding point on the estimated regression function (i.e., from the corresponding fitted value $\hat{Y}_i$).

We need to distinguish between the model error term value $\varepsilon_i = Y_i - E\{Y_i\}$ and the residual $e_i = Y_i - \hat{Y}_i$. The former involves the vertical deviation of Y_i from the unknown true regression line and hence is unknown. On the other hand, the residual is the vertical deviation of Y_i from the fitted value $\hat{Y}_i$ on the estimated regression line, and it is known.

Residuals are highly useful for studying whether a given regression model is appropriate for the data at hand. We discuss this use in Chapter 3.

Properties of Fitted Regression Line

The estimated regression line (1.12) fitted by the method of least squares has a number of properties worth noting. These properties of the least squares estimated regression function do not apply to all regression models, as we shall see in Chapter 4.

1. The sum of the residuals is zero:

$$\sum_{i=1}^{n} e_i = 0 \qquad \textbf{(1.17)}$$

Table 1.2, column 4, illustrates this property for the Toluca Company example. Rounding errors may, of course, be present in any particular case, resulting in a sum of the residuals that does not equal zero exactly.

2. The sum of the squared residuals, $\sum e_i^2$, is a minimum. This was the requirement to be satisfied in deriving the least squares estimators of the regression parameters since the

criterion Q in (1.8) to be minimized equals $\sum e_i^2$ when the least squares estimators b_0 and b_1 are used for estimating β_0 and β_1.

3. The sum of the observed values Y_i equals the sum of the fitted values $\hat{Y}_i$:

$$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i \tag{1.18}$$

This property is illustrated in Table 1.2, columns 2 and 3, for the Toluca Company example. It follows that the mean of the fitted values $\hat{Y}_i$ is the same as the mean of the observed values Y_i, namely, $\bar{Y}$.

4. The sum of the weighted residuals is zero when the residual in the ith trial is weighted by the level of the predictor variable in the ith trial:

$$\sum_{i=1}^{n} X_i e_i = 0 \tag{1.19}$$

5. A consequence of properties (1.17) and (1.19) is that the sum of the weighted residuals is zero when the residual in the ith trial is weighted by the fitted value of the response variable for the ith trial:

$$\sum_{i=1}^{n} \hat{Y}_i e_i = 0 \tag{1.20}$$

6. The regression line always goes through the point $(\bar{X}, \bar{Y})$.

Comment

The six properties of the fitted regression line follow directly from the least squares normal equations (1.9). For example, property 1 in (1.17) is proven as follows:

$$\sum e_i = \sum (Y_i - b_0 - b_1 X_i) = \sum Y_i - nb_0 - b_1 \sum X_i$$
$$= 0 \qquad \text{by the first normal equation (1.9a)}$$

Property 6, that the regression line always goes through the point $(\bar{X}, \bar{Y})$, can be demonstrated easily from the alternative form (1.15) of the estimated regression line. When $X = \bar{X}$, we have:

$$\hat{Y} = \bar{Y} + b_1(X - \bar{X}) = \bar{Y} + b_1(\bar{X} - \bar{X}) = \bar{Y}$$ ■

1.7 Estimation of Error Terms Variance σ^2

The variance σ^2 of the error terms ε_i in regression model (1.1) needs to be estimated to obtain an indication of the variability of the probability distributions of Y. In addition, as we shall see in the next chapter, a variety of inferences concerning the regression function and the prediction of Y require an estimate of σ^2.

Point Estimator of σ^2

To lay the basis for developing an estimator of σ^2 for regression model (1.1), we first consider the simpler problem of sampling from a single population.

Single Population. We know that the variance σ^2 of a single population is estimated by the sample variance s^2. In obtaining the sample variance s^2, we consider the deviation of

an observation Y_i from the estimated mean $\bar{Y}$, square it, and then sum all such squared deviations:

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

Such a sum is called a *sum of squares*. The sum of squares is then divided by the degrees of freedom associated with it. This number is $n - 1$ here, because one degree of freedom is lost by using $\bar{Y}$ as an estimate of the unknown population mean μ. The resulting estimator is the usual sample variance:

$$s^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n - 1}$$

which is an unbiased estimator of the variance σ^2 of an infinite population. The sample variance is often called a *mean square,* because a sum of squares has been divided by the appropriate number of degrees of freedom.

Regression Model. The logic of developing an estimator of σ^2 for the regression model is the same as for sampling from a single population. Recall in this connection from (1.4) that the variance of each observation Y_i for regression model (1.1) is σ^2, the same as that of each error term ε_i. We again need to calculate a sum of squared deviations, but must recognize that the Y_i now come from different probability distributions with different means that depend upon the level X_i. Thus, the deviation of an observation Y_i must be calculated around its own estimated mean $\hat{Y}_i$. Hence, the deviations are the residuals:

$$Y_i - \hat{Y}_i = e_i$$

and the appropriate sum of squares, denoted by *SSE*, is:

$$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2 \tag{1.21}$$

where *SSE* stands for *error sum of squares* or *residual sum of squares*.

The sum of squares *SSE* has $n - 2$ degrees of freedom associated with it. Two degrees of freedom are lost because both β_0 and β_1 had to be estimated in obtaining the estimated means $\hat{Y}_i$. Hence, the appropriate mean square, denoted by *MSE* or s^2, is:

$$s^2 = MSE = \frac{SSE}{n-2} = \frac{\sum(Y_i - \hat{Y}_i)^2}{n-2} = \frac{\sum e_i^2}{n-2} \tag{1.22}$$

where *MSE* stands for *error mean square* or *residual mean square*.

It can be shown that *MSE* is an unbiased estimator of σ^2 for regression model (1.1):

$$E\{MSE\} = \sigma^2 \tag{1.23}$$

An estimator of the standard deviation σ is simply $s = \sqrt{MSE}$, the positive square root of *MSE*.

Example

We will calculate *SSE* for the Tolúca Company example by (1.21). The residuals were obtained earlier in Table 1.2, column 4. This table also shows the squared residuals in column 5. From these results, we obtain:

$$SSE = 54{,}825$$

Since $25 - 2 = 23$ degrees of freedom are associated with *SSE*, we find:

$$s^2 = MSE = \frac{54{,}825}{23} = 2{,}384$$

Finally, a point estimate of σ, the standard deviation of the probability distribution of Y for any X, is $s = \sqrt{2{,}384} = 48.8$ hours.

Consider again the case where the lot size is $X = 65$ units. We found earlier that the mean of the probability distribution of Y for this lot size is estimated to be 294.4 hours. Now, we have the additional information that the standard deviation of this distribution is estimated to be 48.8 hours. This estimate is shown in the MINITAB output in Figure 1.11, labeled as s. We see that the variation in work hours from lot to lot for lots of 65 units is quite substantial (49 hours) compared to the mean of the distribution (294 hours).

1.8 Normal Error Regression Model

No matter what may be the form of the distribution of the error terms ε_i (and hence of the Y_i), the least squares method provides unbiased point estimators of β_0 and β_1 that have minimum variance among all unbiased linear estimators. To set up interval estimates and make tests, however, we need to make an assumption about the form of the distribution of the ε_i. The standard assumption is that the error terms ε_i are normally distributed, and we will adopt it here. A normal error term greatly simplifies the theory of regression analysis and, as we shall explain shortly, is justifiable in many real-world situations where regression analysis is applied.

Model

The normal error regression model is as follows:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad \textbf{(1.24)}$$

where:

Y_i is the observed response in the ith trial

X_i is a known constant, the level of the predictor variable in the ith trial

β_0 and β_1 are parameters

ε_i are independent $N(0, \sigma^2)$

$i = 1, \ldots, n$

Comments

1. The symbol $N(0, \sigma^2)$ stands for normally distributed, with mean 0 and variance σ^2.

2. The normal error model (1.24) is the same as regression model (1.1) with unspecified error distribution, except that model (1.24) assumes that the errors ε_i are normally distributed.

3. Because regression model (1.24) assumes that the errors are normally distributed, the assumption of uncorrelatedness of the ε_i in regression model (1.1) becomes one of independence in the normal error model. Hence, the outcome in any one trial has no effect on the error term for any other trial—as to whether it is positive or negative, small or large.

4. Regression model (1.24) implies that the Y_i are independent normal random variables, with mean $E\{Y_i\} = \beta_0 + \beta_1 X_i$ and variance σ^2. Figure 1.6 pictures this normal error model. Each of the probability distributions of Y in Figure 1.6 is normally distributed, with constant variability, and the regression function is linear.

5. The normality assumption for the error terms is justifiable in many situations because the error terms frequently represent the effects of factors omitted from the model that affect the response to some extent and that vary at random without reference to the variable X. For instance, in the Toluca Company example, the effects of such factors as time lapse since the last production run, particular machines used, season of the year, and personnel employed could vary more or less at random from run to run, independent of lot size. Also, there might be random measurement errors in the recording of Y, the hours required. Insofar as these random effects have a degree of mutual independence, the composite error term ε_i representing all these factors would tend to comply with the central limit theorem and the error term distribution would approach normality as the number of factor effects becomes large.

A second reason why the normality assumption of the error terms is frequently justifiable is that the estimation and testing procedures to be discussed in the next chapter are based on the t distribution and are usually only sensitive to large departures from normality. Thus, unless the departures from normality are serious, particularly with respect to skewness, the actual confidence coefficients and risks of errors will be close to the levels for exact normality. ■

Estimation of Parameters by Method of Maximum Likelihood

When the functional form of the probability distribution of the error terms is specified, estimators of the parameters β_0, β_1, and σ^2 can be obtained by the *method of maximum likelihood*. Essentially, the method of maximum likelihood chooses as estimates those values of the parameters that are most consistent with the sample data. We explain the method of maximum likelihood first for the simple case when a single population with one parameter is sampled. Then we explain this method for regression models.

Single Population. Consider a normal population whose standard deviation is known to be $\sigma = 10$ and whose mean is unknown. A random sample of $n = 3$ observations is selected from the population and yields the results $Y_1 = 250$, $Y_2 = 265$, $Y_3 = 259$. We now wish to ascertain which value of μ is most consistent with the sample data. Consider $\mu = 230$. Figure 1.13a shows the normal distribution with $\mu = 230$ and $\sigma = 10$; also shown there are the locations of the three sample observations. Note that the sample observations

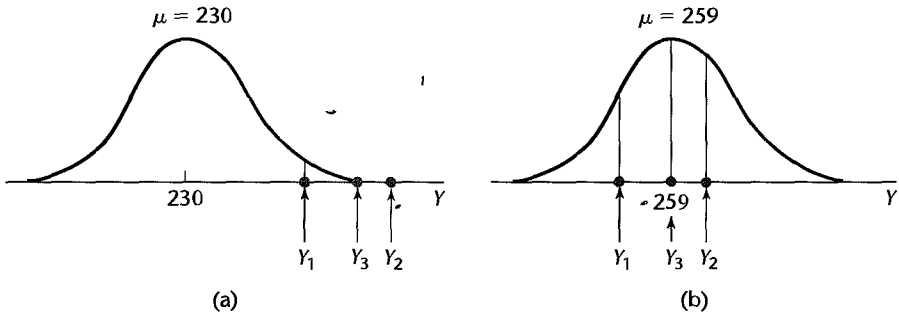


FIGURE 1.13 **Densities for Sample Observations for Two Possible Values of μ: $Y_1 = 250$, $Y_2 = 265$, $Y_3 = 259$.**

would be in the right tail of the distribution if μ were equal to 230. Since these are unlikely occurrences, $\mu = 230$ is not consistent with the sample data.

Figure 1.13b shows the population and the locations of the sample data if μ were equal to 259. Now the observations would be in the center of the distribution and much more likely. Hence, $\mu = 259$ is more consistent with the sample data than $\mu = 230$.

The method of maximum likelihood uses the density of the probability distribution at Y_i (i.e., the height of the curve at Y_i) as a measure of consistency for the observation Y_i. Consider observation Y_1 in our example. If Y_1 is in the tail, as in Figure 1.13a, the height of the curve will be small. If Y_1 is nearer to the center of the distribution, as in Figure 1.13b, the height will be larger. Using the density function for a normal probability distribution in (A.34) in Appendix A, we find the densities for Y_1, denoted by f_1, for the two cases of μ in Figure 1.13 as follows:

$$\mu = 230: \qquad f_1 = \frac{1}{\sqrt{2\pi}(10)} \exp\left[-\frac{1}{2}\left(\frac{250-230}{10}\right)^2\right] = .005399$$

$$\mu = 259: \qquad f_1 = \frac{1}{\sqrt{2\pi}(10)} \exp\left[-\frac{1}{2}\left(\frac{250-259}{10}\right)^2\right] = .026609$$

The densities for all three sample observations for the two cases of μ are as follows:

	$\mu = 230$	$\mu = 259$
f_1	.005399	.026609
f_2	.000087	.033322
f_3	.000595	.039894

The method of maximum likelihood uses the product of the densities (i.e., here, the product of the three heights) as the measure of consistency of the parameter value with the sample data. The product is called the *likelihood value* of the parameter value μ and is denoted by $L(\mu)$. If the value of μ is consistent with the sample data, the densities will be relatively large and so will be the product (i.e., the likelihood value). If the value of μ is not consistent with the data, the densities will be small and the product $L(\mu)$ will be small.

For our simple example, the likelihood values are as follows for the two cases of μ:

$$L(\mu = 230) = .005399(.000087)(.000595) = .279 \times 10^{-9}$$

$$L(\mu = 259) = .026609(.033322)(.039894) = .0000354$$

Since the likelihood value $L(\mu = 230)$ is a very small number, it is shown in scientific notation, which indicates that there are nine zeros after the decimal place before 279. Note that $L(\mu = 230)$ is much smaller than $L(\mu = 259)$, indicating that $\mu = 259$ is much more consistent with the sample data than $\mu = 230$.

The method of maximum likelihood chooses as the maximum likelihood estimate that value of μ for which the likelihood value is largest. Just as for the method of least squares,

there are two methods of finding maximum likelihood estimates: by a systematic numerical search and by use of an analytical solution. For some problems, analytical solutions for the maximum likelihood estimators are available. For others, a computerized numerical search must be conducted.

For our example, an analytical solution is available. It can be shown that for a normal population the maximum likelihood estimator of μ is the sample mean $\bar{Y}$. In our example, $\bar{Y} = 258$ and the maximum likelihood estimate of μ therefore is 258. The likelihood value of $\mu = 258$ is $L(\mu = 258) = .0000359$, which is slightly larger than the likelihood value of .0000354 for $\mu = 259$ that we had calculated earlier.

The product of the densities viewed as a function of the unknown parameters is called the *likelihood function*. For our example, where $\sigma = 10$, the likelihood function is:

$$L(\mu) = \left[\frac{1}{\sqrt{2\pi}(10)}\right]^3 \exp\left[-\frac{1}{2}\left(\frac{250-\mu}{10}\right)^2\right] \exp\left[-\frac{1}{2}\left(\frac{265-\mu}{10}\right)^2\right] \times \exp\left[-\frac{1}{2}\left(\frac{259-\mu}{10}\right)^2\right]$$

Figure 1.14 shows a computer plot of the likelihood function for our example. It is based on the calculation of likelihood values $L(\mu)$ for many values of μ. Note that the likelihood values at $\mu = 230$ and $\mu = 259$ correspond to the ones we determined earlier. Also note that the likelihood function reaches a maximum at $\mu = 258$.

The fact that the likelihood function in Figure 1.14 is relatively peaked in the neighborhood of the maximum likelihood estimate $\bar{Y} = 258$ is of particular interest. Note, for instance, that for $\mu = 250$ or $\mu = 266$, the likelihood value is already only a little more than one-half as large as the likelihood value at $\mu = 258$. This indicates that the maximum likelihood estimate here is relatively precise because values of μ not near the maximum likelihood estimate $\bar{Y} = 258$ are much less consistent with the sample data. When the likelihood function is relatively flat in a fairly wide region around the maximum likelihood

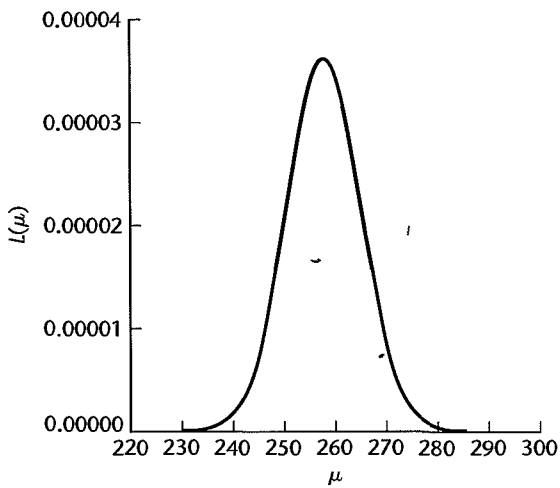


FIGURE 1.14
Likelihood Function for Estimation of Mean of Normal Population: $Y_1 = 250$, $Y_2 = 265$, $Y_3 = 259$.

estimate, many values of the parameter are almost as consistent with the sample data as the maximum likelihood estimate, and the maximum likelihood estimate would therefore be relatively imprecise.

Regression Model. The concepts just presented for maximum likelihood estimation of a population mean carry over directly to the estimation of the parameters of normal error regression model (1.24). For this model, each Y_i observation is normally distributed with mean $\beta_0 + \beta_1 X_i$ and standard deviation σ. To illustrate the method of maximum likelihood estimation here, consider the earlier persistence study example on page 15. For simplicity, let us suppose that we know $\sigma = 2.5$. We wish to determine the likelihood value for the parameter values $\beta_0 = 0$ and $\beta_1 = .5$. For subject 1, $X_1 = 20$ and hence the mean of the probability distribution would be $\beta_0 + \beta_1 X_1 = 0 + .5(20) = 10.0$. Figure 1.15a shows the normal distribution with mean 10.0 and standard deviation 2.5. Note that the observed value $Y_1 = 5$ is in the left tail of the distribution and that the density there is relatively small. For the second subject, $X_2 = 55$ and hence $\beta_0 + \beta_1 X_2 = 27.5$. The normal distribution with mean 27.5 is shown in Figure 1.15b. Note that the observed value $Y_2 = 12$ is most unlikely for this case and that the density there is extremely small. Finally, note that the observed value $Y_3 = 10$ is also in the left tail of its distribution if $\beta_0 = 0$ and $\beta_1 = .5$, as shown in Figure 1.15c, and that the density there is also relatively small.

FIGURE 1.15 **Densities for Sample Observations if $\beta_0 = 0$ and $\beta_1 = 5$—Persistence Study Example.**

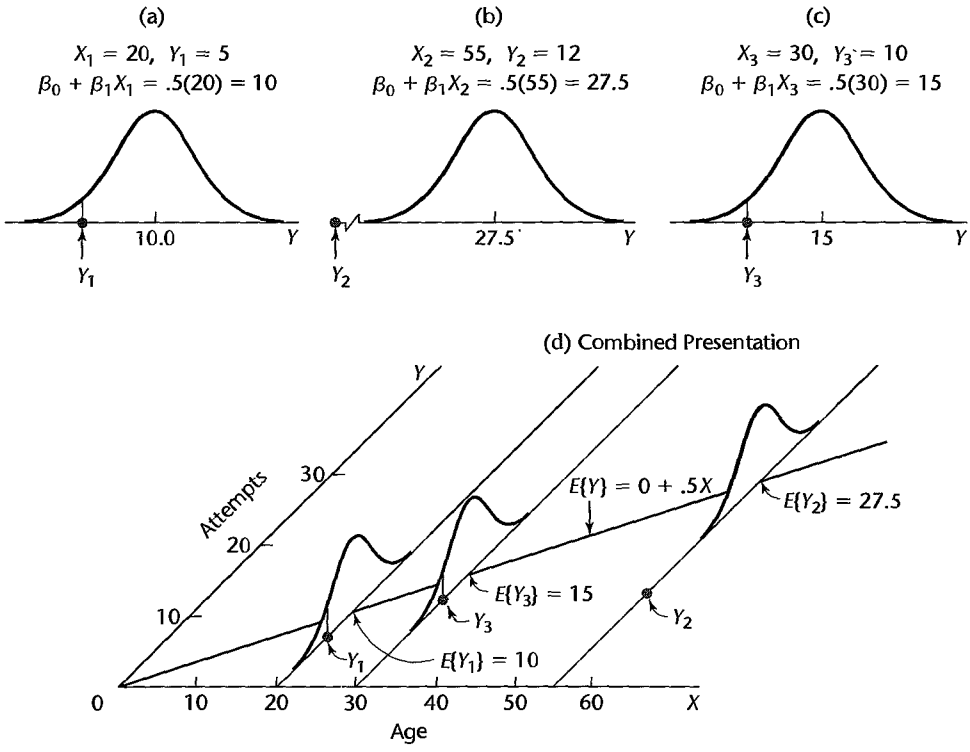


Figure 1.15d combines all of this information, showing the regression function $E\{Y\} = 0 + .5X$, the three sample cases, and the three normal distributions. Note how poorly the regression line fits the three sample cases, as was also indicated by the three small density values. Thus, it appears that $\beta_0 = 0$ and $\beta_1 = .5$ are not consistent with the data.

We calculate the densities (i.e., heights of the curve) in the usual way. For $Y_1 = 5$, $X_1 = 20$, the normal density is as follows when $\beta_0 = 0$ and $\beta_1 = .5$:

$$f_1 = \frac{1}{\sqrt{2\pi}(2.5)} \exp\left[-\frac{1}{2}\left(\frac{5 - 10.0}{2.5}\right)^2 \right] = .021596$$

The other densities are $f_2 = .7175 \times 10^{-9}$ and $f_3 = .021596$, and the likelihood value of $\beta_0 = 0$ and $\beta_1 = .5$ therefore is:

$$L(\beta_0 = 0, \beta_1 = .5) = .021596(.7175 \times 10^{-9})(.021596) = .3346 \times 10^{-12}$$

In general, the density of an observation Y_i for the normal error regression model (1.24) is as follows, utilizing the fact that $E\{Y_i\} = \beta_0 + \beta_1 X_i$ and $\sigma^2\{Y_i\} = \sigma^2$:

$$f_i = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{Y_i - \beta_0 - \beta_1 X_i}{\sigma}\right)^2 \right] \tag{1.25}$$

The likelihood function for n observations $Y_1, Y_2, \ldots, Y_n$ is the product of the individual densities in (1.25). Since the variance σ^2 of the error terms is usually unknown, the likelihood function is a function of three parameters, β_0, β_1, and σ^2:

$$\begin{aligned} L(\beta_0, \beta_1, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left[-\frac{1}{2\sigma^2}(Y_i - \beta_0 - \beta_1 X_i)^2 \right] \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 \right] \end{aligned} \tag{1.26}$$

The values of β_0, β_1, and σ^2 that maximize this likelihood function are the maximum likelihood estimators and are denoted by $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\sigma}^2$, respectively. These estimators can be found analytically, and they are as follows:

Parameter	Maximum Likelihood Estimator	
β_0	$\hat{\beta}_0 = b_0$	same as (1.10b)
β_1	$\hat{\beta}_1 = b_1$	same as (1.10a)
σ^2	$\hat{\sigma}^2 = \dfrac{\sum(Y_i - \hat{Y}_i)^2}{n}$	

(1.27)

Thus, the maximum likelihood estimators of β_0 and β_1 are the same estimators as those provided by the method of least squares. The maximum likelihood estimator $\hat{\sigma}^2$ is biased, and ordinarily the unbiased estimator *MSE* as given in (1.22) is used. Note that the unbiased estimator *MSE* or s^2 differs but slightly from the maximum likelihood estimator $\hat{\sigma}^2$,

especially if n is not small:

$$s^2 = MSE = \frac{n}{n-2}\hat{\sigma}^2 \tag{1.28}$$

Example

For the persistence study example, we know now that the maximum likelihood estimates of β_0 and β_1 are $b_0 = 2.81$ and $b_1 = .177$, the same as the least squares estimates in Figure 1.9b.

Comments

1. Since the maximum likelihood estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are the same as the least squares estimators b_0 and b_1, they have the properties of all least squares estimators:
   a. They are unbiased.
   b. They have minimum variance among all unbiased linear estimators.
   In addition, the maximum likelihood estimators b_0 and b_1 for the normal error regression model (1.24) have other desirable properties:
   c. They are consistent, as defined in (A.52).
   d. They are sufficient, as defined in (A.53).
   e. They are minimum variance unbiased; that is, they have minimum variance in the class of all unbiased estimators (linear or otherwise).
   Thus, for the normal error model, the estimators b_0 and b_1 have many desirable properties.
2. We find the values of β_0, β_1, and σ^2 that maximize the likelihood function L in (1.26) by taking partial derivatives of L with respect to β_0, β_1, and σ^2, equating each of the partials to zero, and solving the system of equations thus obtained. We can work with $\log_e L$, rather than L, because both L and $\log_e L$ are maximized for the same values of β_0, β_1, and σ^2:

$$\log_e L = -\frac{n}{2}\log_e 2\pi - \frac{n}{2}\log_e \sigma^2 - \frac{1}{2\sigma^2}\sum(Y_i - \beta_0 - \beta_1 X_i)^2 \tag{1.29}$$

Partial differentiation of the logarithm of the likelihood function is much easier; it yields:

$$\frac{\partial(\log_e L)}{\partial\beta_0} = \frac{1}{\sigma^2}\sum(Y_i - \beta_0 - \beta_1 X_i)$$

$$\frac{\partial(\log_e L)}{\partial\beta_1} = \frac{1}{\sigma^2}\sum X_i(Y_i - \beta_0 - \beta_1 X_i)$$

$$\frac{\partial(\log_e L)}{\partial\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum(Y_i - \beta_0 - \beta_1 X_i)^2$$

We now set these partial derivatives equal to zero, replacing β_0, β_1, and σ^2 by the estimators $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\sigma}^2$. We obtain, after some simplification:

$$\sum(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0 \tag{1.30a}$$

$$\sum X_i(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0 \tag{1.30b}$$

$$\frac{\sum(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2}{n} = \hat{\sigma}^2 \tag{1.30c}$$

Formulas (1.30a) and (1.30b) are identical to the earlier least squares normal equations (1.9), and formula (1.30c) is the biased estimator of σ^2 given earlier in (1.27). ■

1.1. BMDP New System 2.0. Statistical Solutions, Inc.
1.2. MINITAB Release 13. Minitab Inc.
1.3. SAS/STAT Release 8.2. SAS Institute, Inc.
1.4. SPSS 11.5 for Windows. SPSS Inc.
1.5. SYSTAT 10.2. SYSTAT Software, Inc.
1.6. JMP Version 5. SAS Institute, Inc.
1.7. S-Plus 6 for Windows. Insightful Corporation.
1.8. MATLAB 6.5. The MathWorks, Inc.

1.1. Refer to the sales volume example on page 3. Suppose that the number of units sold is measured accurately, but clerical errors are frequently made in determining the dollar sales. Would the relation between the number of units sold and dollar sales still be a functional one? Discuss.

1.2. The members of a health spa pay annual membership dues of \$300 plus a charge of \$2 for each visit to the spa. Let Y denote the dollar cost for the year for a member and X the number of visits by the member during the year. Express the relation between X and Y mathematically. Is it a functional relation or a statistical relation?

1.3. Experience with a certain type of plastic indicates that a relation exists between the hardness (measured in Brinell units) of items molded from the plastic (Y) and the elapsed time since termination of the molding process (X). It is proposed to study this relation by means of regression analysis. A participant in the discussion objects, pointing out that the hardening of the plastic "is the result of a natural chemical process that doesn't leave anything to chance, so the relation must be mathematical and regression analysis is not appropriate." Evaluate this objection.

1.4. In Table 1.1, the lot size X is the same in production runs 1 and 24 but the work hours Y differ. What feature of regression model (1.1) is illustrated by this?

1.5. When asked to state the simple linear regression model, a student wrote it as follows: $E\{Y_i\} = \beta_0 + \beta_1 X_i + \varepsilon_i$. Do you agree?

1.6. Consider the normal error regression model (1.24). Suppose that the parameter values are $\beta_0 = 200$, $\beta_1 = 5.0$, and $\sigma = 4$.

a. Plot this normal error regression model in the fashion of Figure 1.6. Show the distributions of Y for $X = 10$, 20, and 40.

b. Explain the meaning of the parameters β_0 and β_1. Assume that the scope of the model includes $X = 0$.

1.7. In a simulation exercise, regression model (1.1) applies with $\beta_0 = 100$, $\beta_1 = 20$, and $\sigma^2 = 25$. An observation on Y will be made for $X = 5$.

a. Can you state the exact probability that Y will fall between 195 and 205? Explain.

b. If the normal error regression model (1.24) is applicable, can you now state the exact probability that Y will fall between 195 and 205? If so, state it.

1.8. In Figure 1.6, suppose another Y observation is obtained at $X = 45$. Would $E\{Y\}$ for this new observation still be 104? Would the Y value for this new case again be 108?

1.9. A student in accounting enthusiastically declared: "Regression is a very powerful tool. We can isolate fixed and variable costs by fitting a linear regression model, even when we have no data for small lots." Discuss.

1.10. An analyst in a large corporation studied the relation between current annual salary (Y) and age (X) for the 46 computer programmers presently employed in the company. The analyst concluded that the relation is curvilinear, reaching a maximum at 47 years. Does this imply that the salary for a programmer increases until age 47 and then decreases? Explain.

1.11. The regression function relating production output by an employee after taking a training program (Y) to the production output before the training program (X) is $E\{Y\} = 20 + .95X$, where X ranges from 40 to 100. An observer concludes that the training program does not raise production output on the average because β_1 is not greater than 1.0. Comment.

1.12. In a study of the relationship for senior citizens between physical activity and frequency of colds, participants were asked to monitor their weekly time spent in exercise over a five-year period and the frequency of colds. The study demonstrated that a negative statistical relation exists between time spent in exercise and frequency of colds. The investigator concluded that increasing the time spent in exercise is an effective strategy for reducing the frequency of colds for senior citizens.

a. Were the data obtained in the study observational or experimental data?

b. Comment on the validity of the conclusions reached by the investigator.

c. Identify two or three other explanatory variables that might affect both the time spent in exercise and the frequency of colds for senior citizens simultaneously.

d. How might the study be changed so that a valid conclusion about causal relationship between amount of exercise and frequency of colds can be reached?

1.13. Computer programmers employed by a software developer were asked to participate in a month-long training seminar. During the seminar, each employee was asked to record the number of hours spent in class preparation each week. After completing the seminar, the productivity level of each participant was measured. A positive linear statistical relationship between participants' productivity levels and time spent in class preparation was found. The seminar leader concluded that increases in employee productivity are caused by increased class preparation time.

a. Were the data used by the seminar leader observational or experimental data?

b. Comment on the validity of the conclusion reached by the seminar leader.

c. Identify two or three alternative variables that might cause both the employee productivity scores and the employee class participation times to increase (decrease) simultaneously.

d. How might the study be changed so that a valid conclusion about causal relationship between class preparation time and employee productivity can be reached?

1.14. Refer to Problem 1.3. Four different elapsed times since termination of the molding process (treatments) are to be studied to see how they affect the hardness of a plastic. Sixteen batches (experimental units) are available for the study. Each treatment is to be assigned to four experimental units selected at random. Use a table of random digits or a random number generator to make an appropriate randomization of assignments.

1.15. The effects of five dose levels are to be studied in a completely randomized design, and 20 experimental units are available. Each dose level is to be assigned to four experimental units selected at random. Use a table of random digits or a random number generator to make an appropriate randomization of assignments.

1.16. Evaluate the following statement: "For the least squares method to be fully valid, it is required that the distribution of Y be normal."

1.17. A person states that b_0 and b_1 in the fitted regression function (1.13) can be estimated by the method of least squares. Comment.

1.18. According to (1.17), $\sum e_i = 0$ when regression model (1.1) is fitted to a set of n cases by the method of least squares. Is it also true that $\sum \varepsilon_i = 0$? Comment.

1.19. **Grade point average.** The director of admissions of a small college selected 120 students at random from the new freshman class in a study to determine whether a student's grade point average (GPA) at the end of the freshman year (Y) can be predicted from the ACT test score (X). The results of the study follow. Assume that first-order regression model (1.1) is appropriate.

i:	1	2	3	...	118	119	120
X_i:	21	14	28	...	28	16	28
Y_i:	3.897	3.885	3.778	...	3.914	1.860	2.948

a. Obtain the least squares estimates of β_0 and β_1, and state the estimated regression function.

b. Plot the estimated regression function and the data. Does the estimated regression function appear to fit the data well?

c. Obtain a point estimate of the mean freshman GPA for students with ACT test score $X = 30$.

d. What is the point estimate of the change in the mean response when the entrance test score increases by one point?

*1.20. **Copier maintenance.** The Tri-City Office Equipment Corporation sells an imported copier on a franchise basis and performs preventive maintenance and repair service on this copier. The data below have been collected from 45 recent calls on users to perform routine preventive maintenance service; for each call, X is the number of copiers serviced and Y is the total number of minutes spent by the service person. Assume that first-order regression model (1.1) is appropriate.

i:	1	2	3	...	43	44	45
X_i:	2	4	3	...	2	4	5
Y_i:	20	60	46	...	27	61	77

a. Obtain the estimated regression function.

b. Plot the estimated regression function and the data. How well does the estimated regression function fit the data?

c. Interpret b_0 in your estimated regression function. Does b_0 provide any relevant information here? Explain.

d. Obtain a point estimate of the mean service time when $X = 5$ copiers are serviced.

*1.21. **Airfreight breakage.** A substance used in biological and medical research is shipped by air-freight to users in cartons of 1,000 ampules. The data below, involving 10 shipments, were collected on the number of times the carton was transferred from one aircraft to another over the shipment route (X) and the number of ampules found to be broken upon arrival (Y). Assume that first-order regression model (1.1) is appropriate.

i:	1	2	3	4	5	6	7	8	9	10
X_i:	1	0	2	0	3	1	0	1	2	0
Y_i:	16	9	17	12	22	13	8	15	19	11

a. Obtain the estimated regression function. Plot the estimated regression function and the data. Does a linear regression function appear to give a good fit here?

b. Obtain a point estimate of the expected number of broken ampules when $X = 1$ transfer is made.

c. Estimate the increase in the expected number of ampules broken when there are 2 transfers as compared to 1 transfer.

d. Verify that your fitted regression line goes through the point $(\bar{X}, \bar{Y})$.

1.22. **Plastic hardness.** Refer to Problems 1.3 and 1.14. Sixteen batches of the plastic were made, and from each batch one test item was molded. Each test item was randomly assigned to one of the four predetermined time levels, and the hardness was measured after the assigned elapsed time. The results are shown below; X is the elapsed time in hours, and Y is hardness in Brinell units. Assume that first-order regression model (1.1) is appropriate.

i:	1	2	3	...	14	15	16
X_i:	16	16	16	...	40	40	40
Y_i:	199	205	196	...	248	253	246

a. Obtain the estimated regression function. Plot the estimated regression function and the data. Does a linear regression function appear to give a good fit here?

b. Obtain a point estimate of the mean hardness when $X = 40$ hours.

c. Obtain a point estimate of the change in mean hardness when X increases by 1 hour.

1.23. Refer to **Grade point average** Problem 1.19.

a. Obtain the residuals e_i. Do they sum to zero in accord with (1.17)?

b. Estimate σ^2 and σ. In what units is σ expressed?

*1.24. Refer to **Copier maintenance** Problem 1.20.

a. Obtain the residuals e_i and the sum of the squared residuals $\sum e_i^2$. What is the relation between the sum of the squared residuals here and the quantity Q in (1.8)?

b. Obtain point estimates of σ^2 and σ. In what units is σ expressed?

*1.25. Refer to **Airfreight breakage** Problem 1.21.

a. Obtain the residual for the first case. What is its relation to ε_1?

b. Compute $\sum e_i^2$ and MSE. What is estimated by MSE?

1.26. Refer to **Plastic hardness** Problem 1.22.

a. Obtain the residuals e_i. Do they sum to zero in accord with (1.17)?

b. Estimate σ^2 and σ. In what units is σ expressed?

*1.27. **Muscle mass.** A person's muscle mass is expected to decrease with age. To explore this relationship in women, a nutritionist randomly selected 15 women from each 10-year age group, beginning with age 40 and ending with age 79. The results follow; X is age, and Y is a measure of muscle mass. Assume that first-order regression model (1.1) is appropriate.

i:	1	2	3	...	58	59	60
X_i:	43	41	47	...	76	72	76
Y_i:	106	106	97	...	56	70	74

a. Obtain the estimated regression function. Plot the estimated regression function and the data. Does a linear regression function appear to give a good fit here? Does your plot support the anticipation that muscle mass decreases with age?

b. Obtain the following: (1) a point estimate of the difference in the mean muscle mass for women differing in age by one year, (2) a point estimate of the mean muscle mass for women aged $X = 60$ years, (3) the value of the residual for the eighth case, (4) a point estimate of σ^2.

1.28. **Crime rate.** A criminologist studying the relationship between level of education and crime rate in medium-sized U.S. counties collected the following data for a random sample of 84 counties; X is the percentage of individuals in the county having at least a high-school diploma, and Y is the crime rate (crimes reported per 100,000 residents) last year. Assume that first-order regression model (1.1) is appropriate.

i:	1	2	3	...	82	83	84
X_i:	74	82	81	...	88	83	76
Y_i:	8,487	8,179	8,362	...	8,040	6,981	7,582

a. Obtain the estimated regression function. Plot the estimated regression function and the data. Does the linear regression function appear to give a good fit here? Discuss.

b. Obtain point estimates of the following: (1) the difference in the mean crime rate for two counties whose high-school graduation rates differ by one percentage point, (2) the mean crime rate last year in counties with high school graduation percentage $X = 80$, (3) ε_{10}, (4) σ^2.

Exercises

1.29. Refer to regression model (1.1). Assume that $X = 0$ is within the scope of the model. What is the implication for the regression function if $\beta_0 = 0$ so that the model is $Y_i = \beta_1 X_i + \varepsilon_i$? How would the regression function plot on a graph?

1.30. Refer to regression model (1.1). What is the implication for the regression function if $\beta_1 = 0$ so that the model is $Y_i = \beta_0 + \varepsilon_i$? How would the regression function plot on a graph?

1.31. Refer to **Plastic hardness** Problem 1.22. Suppose one test item was molded from a single batch of plastic and the hardness of this one item was measured at 16 different points in time. Would the error term in the regression model for this case still reflect the same effects as for the experiment initially described? Would you expect the error terms for the different points in time to be uncorrelated? Discuss.

1.32. Derive the expression for b_1 in (1.10a) from the normal equations in (1.9).

1.33. (Calculus needed.) Refer to the regression model $Y_i = \beta_0 + \varepsilon_i$ in Exercise 1.30. Derive the least squares estimator of β_0 for this model.

1.34. Prove that the least squares estimator of β_0 obtained in Exercise 1.33 is unbiased.

1.35. Prove the result in (1.18)—that the sum of the Y observations is the same as the sum of the fitted values.

1.36. Prove the result in (1.20)—that the sum of the residuals weighted by the fitted values is zero.

1.37. Refer to Table 1.1 for the Toluca Company example. When asked to present a point estimate of the expected work hours for lot sizes of 30 pieces, a person gave the estimate 202 because this is the mean number of work hours in the three runs of size 30 in the study. A critic states that this person's approach "throws away" most of the data in the study because cases with lot sizes other than 30 are ignored. Comment.

1.38. In **Airfreight breakage** Problem 1.21, the least squares estimates are $b_0 = 10.20$ and $b_1 = 4.00$, and $\sum e_i^2 = 17.60$. Evaluate the least squares criterion Q in (1.8) for the estimates (1) $b_0 = 9$, $b_1 = 3$; (2) $b_0 = 11$, $b_1 = 5$. Is the criterion Q larger for these estimates than for the least squares estimates?

1.39. Two observations on Y were obtained at each of three X levels, namely, at $X = 5$, $X = 10$, and $X = 15$.

a. Show that the least squares regression line fitted to the *three* points $(5, \bar{Y}_1)$, $(10, \bar{Y}_2)$, and $(15, \bar{Y}_3)$, where $\bar{Y}_1$, $\bar{Y}_2$, and $\bar{Y}_3$ denote the means of the Y observations at the three X levels, is identical to the least squares regression line fitted to the original six cases.

b. In this study, could the error term variance σ^2 be estimated without fitting a regression line? Explain.

1.40. In fitting regression model (1.1), it was found that observation Y_i fell directly on the fitted regression line (i.e., $Y_i = \hat{Y}_i$). If this case were deleted, would the least squares regression line fitted to the remaining $n - 1$ cases be changed? [*Hint*: What is the contribution of case i to the least squares criterion Q in (1.8)?]

1.41. (Calculus needed.) Refer to the regression model $Y_i = \beta_1 X_i + \varepsilon_i, i = 1, \ldots, n$, in Exercise 1.29.

a. Find the least squares estimator of β_1.

b. Assume that the error terms ε_i are independent $N(0, \sigma^2)$ and that σ^2 is known. State the likelihood function for the n sample observations on Y and obtain the maximum likelihood estimator of β_1. Is it the same as the least squares estimator?

c. Show that the maximum likelihood estimator of β_1 is unbiased.

1.42. **Typographical errors.** Shown below are the number of galleys for a manuscript (X) and the dollar cost of correcting typographical errors (Y) in a random sample of recent orders handled by a firm specializing in technical manuscripts. Assume that the regression model $Y_i = \beta_1 X_i + \varepsilon_i$ is appropriate, with normally distributed independent error terms whose variance is $\sigma^2 = 16$.

i:	1	2	3	4	5	6
X_i:	7	12	4	14	25	30
Y_i:	128	213	75	250	446	540

a. State the likelihood function for the six Y observations, for $\sigma^2 = 16$.

b. Evaluate the likelihood function for $\beta_1 = 17$, 18, and 19. For which of these β_1 values is the likelihood function largest?

c. The maximum likelihood estimator is $b_1 = \sum X_i Y_i / \sum X_i^2$. Find the maximum likelihood estimate. Are your results in part (b) consistent with this estimate?

d. Using a computer graphics or statistics package, evaluate the likelihood function for values of β_1 between $\beta_1 = 17$ and $\beta_1 = 19$ and plot the function. Does the point at which the likelihood function is maximized correspond to the maximum likelihood estimate found in part (c)?

Projects

1.43. Refer to the **CDI** data set in Appendix C.2. The number of active physicians in a CDI (Y) is expected to be related to total population, number of hospital beds, and total personal income. Assume that first-order regression model (1.1) is appropriate for each of the three predictor variables.

a. Regress the number of active physicians in turn on each of the three predictor variables. State the estimated regression functions.

b. Plot the three estimated regression functions and data on separate graphs. Does a linear regression relation appear to provide a good fit for each of the three predictor variables?

c. Calculate *MSE* for each of the three predictor variables. Which predictor variable leads to the smallest variability around the fitted regression line?

1.44. Refer to the **CDI** data set in Appendix C.2.

a. For each geographic region, regress per capita income in a CDI (Y) against the percentage of individuals in a county having at least a bachelor's degree (X). Assume that

first-order regression model (1.1) is appropriate for each region. State the estimated regression functions.

b. Are the estimated regression functions similar for the four regions? Discuss.

c. Calculate *MSE* for each region. Is the variability around the fitted regression line approximately the same for the four regions? Discuss.

1.45. Refer to the **SENIC** data set in Appendix C.1. The average length of stay in a hospital (Y) is anticipated to be related to infection risk, available facilities and services, and routine chest X-ray ratio. Assume that first-order regression model (1.1) is appropriate for each of the three predictor variables.

a. Regress average length of stay on each of the three predictor variables. State the estimated regression functions.

b. Plot the three estimated regression functions and data on separate graphs. Does a linear relation appear to provide a good fit for each of the three predictor variables?

c. Calculate *MSE* for each of the three predictor variables. Which predictor variable leads to the smallest variability around the fitted regression line?

1.46. Refer to the **SENIC** data set in Appendix C.1.

a. For each geographic region, regress average length of stay in hospital (Y) against infection risk (X). Assume that first-order regression model (1.1) is appropriate for each region. State the estimated regression functions.

b. Are the estimated regression functions similar for the four regions? Discuss.

c. Calculate *MSE* for each region. Is the variability around the fitted regression line approximately the same for the four regions? Discuss.

1.47. Refer to **Typographical errors** Problem 1.42. Assume that first-order regression model (1.1) is appropriate, with normally distributed independent error terms whose variance is $\sigma^2 = 16$.

a. State the likelihood function for the six observations, for $\sigma^2 = 16$.

b. Obtain the maximum likelihood estimates of β_0 and β_1, using (1.27).

c. Using a computer graphics or statistics package, obtain a three-dimensional plot of the likelihood function for values of β_0 between $\beta_0 = -10$ and $\beta_0 = 10$ and for values of β_1 between $\beta_1 = 17$ and $\beta_1 = 19$. Does the likelihood appear to be maximized by the maximum likelihood estimates found in part (b)?

Chapter

Inferences in Regression and Correlation Analysis

In this chapter, we first take up inferences concerning the regression parameters β_0 and β_1, considering both interval estimation of these parameters and tests about them. We then discuss interval estimation of the mean $E\{Y\}$ of the probability distribution of Y, for given X, prediction intervals for a new observation Y, confidence bands for the regression line, the analysis of variance approach to regression analysis, the general linear test approach, and descriptive measures of association. Finally, we take up the correlation coefficient, a measure of association between X and Y when both X and Y are random variables.

Throughout this chapter (excluding Section 2.11), and in the remainder of Part I unless otherwise stated, we assume that the normal error regression model (1.24) is applicable. This model is:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{2.1}$$

where:

β_0 and β_1 are parameters
X_i are known constants
ε_i are independent $N(0, \sigma^2)$

2.1 Inferences Concerning β_1

Frequently, we are interested in drawing inferences about β_1, the slope of the regression line in model (2.1). For instance, a market research analyst studying the relation between sales (Y) and advertising expenditures (X) may wish to obtain an interval estimate of β_1 because it will provide information as to how many additional sales dollars, on the average, are generated by an additional dollar of advertising expenditure.

At times, tests concerning β_1 are of interest, particularly one of the form:

$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$

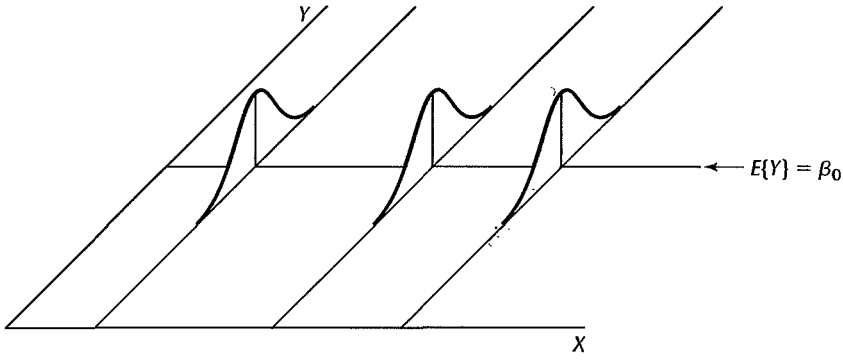


FIGURE 2.1
Regression Model (2.1) when $\beta_1 = 0$.

The reason for interest in testing whether or not $\beta_1 = 0$ is that, when $\beta_1 = 0$, there is no linear association between Y and X. Figure 2.1 illustrates the case when $\beta_1 = 0$. Note that the regression line is horizontal and that the means of the probability distributions of Y are therefore all equal, namely:

$$E\{Y\} = \beta_0 + (0)X = \beta_0$$

For normal error regression model (2.1), the condition $\beta_1 = 0$ implies even more than no linear association between Y and X. Since for this model all probability distributions of Y are normal with constant variance, and since the means are equal when $\beta_1 = 0$, it follows that the probability distributions of Y are identical when $\beta_1 = 0$. This is shown in Figure 2.1. Thus, $\beta_1 = 0$ for the normal error regression model (2.1) implies not only that there is no linear association between Y and X but also that there is no relation of any type between Y and X, since the probability distributions of Y are then identical at all levels of X.

Before discussing inferences concerning β_1 further, we need to consider the sampling distribution of b_1, the point estimator of β_1.

Sampling Distribution of b_1

The point estimator b_1 was given in (1.10a) as follows:

$$b_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} \qquad \textbf{(2.2)}$$

The sampling distribution of b_1 refers to the different values of b_1 that would be obtained with repeated sampling when the levels of the predictor variable X are held constant from sample to sample.

For normal error regression model (2.1), the sampling distribution of b_1 is normal, with mean and variance: **(2.3)**

$$E\{b_1\} = \beta_1 \qquad \textbf{(2.3a)}$$

$$\sigma^2\{b_1\} = \frac{\sigma^2}{\sum(X_i - \bar{X})^2} \qquad \textbf{(2.3b)}$$

To show this, we need to recognize that b_1 is a linear combination of the observations Y_i.

b_1 as Linear Combination of the Y_i. It can be shown that b_1, as defined in (2.2), can be expressed as follows:

$$b_1 = \sum k_i Y_i \tag{2.4}$$

where:

$$k_i = \frac{X_i - \bar{X}}{\sum (X_i - \bar{X})^2} \tag{2.4a}$$

Observe that the k_i are a function of the X_i and therefore are fixed quantities since the X_i are fixed. Hence, b_1 is a linear combination of the Y_i where the coefficients are solely a function of the fixed X_i.

The coefficients k_i have a number of interesting properties that will be used later:

$$\sum k_i = 0 \tag{2.5}$$

$$\sum k_i X_i = 1 \tag{2.6}$$

$$\sum k_i^2 = \frac{1}{\sum (X_i - \bar{X})^2} \tag{2.7}$$

Comments

1. To show that b_1 is a linear combination of the Y_i with coefficients k_i, we first prove:

$$\sum (X_i - \bar{X})(Y_i - \bar{Y}) = \sum (X_i - \bar{X}) Y_i \tag{2.8}$$

This follows since:

$$\sum (X_i - \bar{X})(Y_i - \bar{Y}) = \sum (X_i - \bar{X}) Y_i - \sum (X_i - \bar{X}) \bar{Y}$$

But $\sum (X_i - \bar{X}) \bar{Y} = \bar{Y} \sum (X_i - \bar{X}) = 0$ since $\sum (X_i - \bar{X}) = 0$, Hence, (2.8) holds.

We now express b_1 using (2.8) and (2.4a):

$$b_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \frac{\sum (X_i - \bar{X}) Y_i}{\sum (X_i - \bar{X})^2} = \sum k_i Y_i$$

2. The proofs of the properties of the k_i are direct. For example, property (2.5) follows because:

$$\sum k_i = \sum \left[\frac{X_i - \bar{X}}{\sum (X_i - \bar{X})^2} \right] = \frac{1}{\sum (X_i - \bar{X})^2} \sum (X_i - \bar{X}) = \frac{0}{\sum (X_i - \bar{X})^2} = 0$$

Similarly, property (2.7) follows because:

$$\sum k_i^2 = \sum \left[\frac{X_i - \bar{X}}{\sum (X_i - \bar{X})^2} \right]^2 = \frac{1}{\left[\sum (X_i - \bar{X})^2 \right]^2} \sum (X_i - \bar{X})^2 = \frac{1}{\sum (X_i - \bar{X})^2}$$

■

Normality. We return now to the sampling distribution of b_1 for the normal error regression model (2.1). The normality of the sampling distribution of b_1 follows at once from the fact that b_1 is a linear combination of the Y_i. The Y_i are independently, normally distributed

according to model (2.1), and (A.40) in Appendix A states that a linear combination of independent normal random variables is normally distributed.

Mean. The unbiasedness of the point estimator b_1, stated earlier in the Gauss-Markov theorem (1.11), is easy to show:

$$E\{b_1\} = E\left\{\sum k_i Y_i\right\} = \sum k_i E\{Y_i\} = \sum k_i(\beta_0 + \beta_1 X_i)$$
$$= \beta_0 \sum k_i + \beta_1 \sum k_i X_i$$

By (2.5) and (2.6), we then obtain $E\{b_1\} = \beta_1$.

Variance. The variance of b_1 can be derived readily. We need only remember that the Y_i are independent random variables, each with variance σ^2, and that the k_i are constants. Hence, we obtain by (A.31):

$$\sigma^2\{b_1\} = \sigma^2\left\{\sum k_i Y_i\right\} = \sum k_i^2 \sigma^2\{Y_i\}$$
$$= \sum k_i^2 \sigma^2 = \sigma^2 \sum k_i^2$$
$$= \sigma^2 \frac{1}{\sum (X_i - \bar{X})^2}$$

The last step follows from (2.7).

Estimated Variance. We can estimate the variance of the sampling distribution of b_1:

$$\sigma^2\{b_1\} = \frac{\sigma^2}{\sum (X_i - \bar{X})^2}$$

by replacing the parameter σ^2 with *MSE*, the unbiased estimator of σ^2:

$$s^2\{b_1\} = \frac{MSE}{\sum (X_i - \bar{X})^2} \tag{2.9}$$

The point estimator $s^2\{b_1\}$ is an unbiased estimator of $\sigma^2\{b_1\}$. Taking the positive square root, we obtain $s\{b_1\}$, the point estimator of $\sigma\{b_1\}$.

Comment

We stated in theorem (1.11) that b_1 has minimum variance among all unbiased linear estimators of the form:

$$\hat{\beta}_1 = \sum c_i Y_i$$

where the c_i are arbitrary constants. We now prove this. Since $\hat{\beta}_1$ is required to be unbiased, the following must hold:

$$E\{\hat{\beta}_1\} = E\left\{\sum c_i Y_i\right\} = \sum c_i E\{Y_i\} = \beta_1$$

Now $E\{Y_i\} = \beta_0 + \beta_1 X_i$ by (1.2), so the above condition becomes:

$$E\{\hat{\beta}_1\} = \sum c_i(\beta_0 + \beta_1 X_i) = \beta_0 \sum c_i + \beta_1 \sum c_i X_i = \beta_1$$

For the unbiasedness condition to hold, the c_i must follow the restrictions:

$$\sum c_i = 0 \qquad \sum c_i X_i = 1$$

Now the variance of $\hat{\beta}_1$ is, by (A.31):

$$\sigma^2\{\hat{\beta}_1\} = \sum c_i^2 \sigma^2\{Y_i\} = \sigma^2 \sum c_i^2$$

Let us define $c_i = k_i + d_i$, where the k_i are the least squares constants in (2.4a) and the d_i are arbitrary constants. We can then write:

$$\sigma^2\{\hat{\beta}_1\} = \sigma^2 \sum c_i^2 = \sigma^2 \sum (k_i + d_i)^2 = \sigma^2 \left(\sum k_i^2 + \sum d_i^2 + 2 \sum k_i d_i \right)$$

We know that $\sigma^2 \sum k_i^2 = \sigma^2\{b_1\}$ from our proof above. Further, $\sum k_i d_i = 0$ because of the restrictions on the k_i and c_i above:

$$\begin{aligned} \sum k_i d_i &= \sum k_i (c_i - k_i) \\ &= \sum c_i k_i - \sum k_i^2 \\ &= \sum c_i \left[\frac{X_i - \bar{X}}{\sum (X_i - \bar{X})^2} \right] - \frac{1}{\sum (X_i - \bar{X})^2} \\ &= \frac{\sum c_i X_i - \bar{X} \sum c_i}{\sum (X_i - \bar{X})^2} - \frac{1}{\sum (X_i - \bar{X})^2} = 0 \end{aligned}$$

Hence, we have:

$$\sigma^2\{\hat{\beta}_1\} = \sigma^2\{b_1\} + \sigma^2 \sum d_i^2$$

Note that the smallest value of $\sum d_i^2$ is zero. Hence, the variance of $\hat{\beta}_1$ is at a minimum when $\sum d_i^2 = 0$. But this can only occur if all $d_i = 0$, which implies $c_i \equiv k_i$. Thus, the least squares estimator b_1 has minimum variance among all unbiased linear estimators. ■

Sampling Distribution of $(b_1 - \beta_1)/s\{b_1\}$

Since b_1 is normally distributed, we know that the standardized statistic $(b_1 - \beta_1)/\sigma\{b_1\}$ is a standard normal variable. Ordinarily, of course, we need to estimate $\sigma\{b_1\}$ by $s\{b_1\}$, and hence are interested in the distribution of the statistic $(b_1 - \beta_1)/s\{b_1\}$. When a statistic is standardized but the denominator is an estimated standard deviation rather than the true standard deviation, it is called a *studentized statistic*. An important theorem in statistics states the following about the studentized statistic $(b_1 - \beta_1)/s\{b_1\}$:

$$\frac{b_1 - \beta_1}{s\{b_1\}} \text{ is distributed as } t(n-2) \text{ for regression model (2.1)} \qquad \textbf{(2.10)}$$

Intuitively, this result should not be unexpected. We know that if the observations Y_i come from the same normal population, $(\bar{Y} - \mu)/s\{\bar{Y}\}$ follows the t distribution with $n - 1$ degrees of freedom. The estimator b_1, like $\bar{Y}$, is a linear combination of the observations Y_i. The reason for the difference in the degrees of freedom is that two parameters (β_0 and β_1) need to be estimated for the regression model; hence, two degrees of freedom are lost here.

Comment

We can show that the studentized statistic $(b_1 - \beta_1)/s\{b_1\}$ is distributed as t with $n-2$ degrees of freedom by relying on the following theorem:

> For regression model (2.1), SSE/σ^2 is distributed as χ^2 with $n-2$ degrees of freedom and is independent of b_0 and b_1. **(2.11)**

First, let us rewrite $(b_1 - \beta_1)/s\{b_1\}$ as follows:

$$\frac{b_1 - \beta_1}{\sigma\{b_1\}} \div \frac{s\{b_1\}}{\sigma\{b_1\}}$$

The numerator is a standard normal variable z. The nature of the denominator can be seen by first considering:

$$\frac{s^2\{b_1\}}{\sigma^2\{b_1\}} = \frac{\dfrac{MSE}{\sum(X_i - \bar{X})^2}}{\dfrac{\sigma^2}{\sum(X_i - \bar{X})^2}} = \frac{MSE}{\sigma^2} = \frac{\dfrac{SSE}{n-2}}{\sigma^2}$$

$$= \frac{SSE}{\sigma^2(n-2)} \sim \frac{\chi^2(n-2)}{n-2}$$

where the symbol $\sim$ stands for "is distributed as." The last step follows from (2.11). Hence, we have:

$$\frac{b_1 - \beta_1}{s\{b_1\}} \sim \frac{z}{\sqrt{\dfrac{\chi^2(n-2)}{n-2}}}$$

But by theorem (2.11), z and χ^2 are independent since z is a function of b_1 and b_1 is independent of $SSE/\sigma^2 \sim \chi^2$. Hence, by (A.44), it follows that:

$$\frac{b_1 - \beta_1}{s\{b_1\}} \sim t(n-2)$$

This result places us in a position to readily make inferences concerning β_1. ■

Confidence Interval for β_1

Since $(b_1 - \beta_1)/s\{b_1\}$ follows a t distribution, we can make the following probability statement:

$$P\{t(\alpha/2; n-2) \le (b_1 - \beta_1)/s\{b_1\} \le t(1-\alpha/2; n-2)\} = 1 - \alpha \qquad \textbf{(2.12)}$$

Here, $t(\alpha/2; n-2)$ denotes the $(\alpha/2)100$ percentile of the t distribution with $n-2$ degrees of freedom. Because of the symmetry of the t distribution around its mean 0, it follows that:

$$t(\alpha/2; n-2) = -t(1-\alpha/2; n-2) \qquad \textbf{(2.13)}$$

Rearranging the inequalities in (2.12) and using (2.13), we obtain:

$$P\{b_1 - t(1-\alpha/2; n-2)s\{b_1\} \le \beta_1 \le b_1 + t(1-\alpha/2; n-2)s\{b_1\}\} = 1 - \alpha \qquad \textbf{(2.14)}$$

Since (2.14) holds for all possible values of β_1, the $1-\alpha$ confidence limits for β_1 are:

$$b_1 \pm t(1-\alpha/2; n-2)s\{b_1\} \qquad \textbf{(2.15)}$$

Example

Consider the Toluca Company example of Chapter 1. Management wishes an estimate of β_1 with 95 percent confidence coefficient. We summarize in Table 2.1 the needed results obtained earlier. First, we need to obtain $s\{b_1\}$:

$$s^2\{b_1\} = \frac{MSE}{\sum(X_i - \bar{X})^2} = \frac{2{,}384}{19{,}800} = .12040$$

$$s\{b_1\} = .3470$$

This estimated standard deviation is shown in the MINITAB output in Figure 2.2 in the column labeled Stdev corresponding to the row labeled X. Figure 2.2 repeats the MINITAB output presented earlier in Chapter 1 and contains some additional results that we will utilize shortly.

For a 95 percent confidence coefficient, we require $t(.975; 23)$. From Table B.2 in Appendix B, we find $t(.975; 23) = 2.069$. The 95 percent confidence interval, by (2.15), then is:

$$3.5702 - 2.069(.3470) \le \beta_1 \le 3.5702 + 2.069(.3470)$$

$$2.85 \le \beta_1 \le 4.29$$

Thus, with confidence coefficient .95, we estimate that the mean number of work hours increases by somewhere between 2.85 and 4.29 hours for each additional unit in the lot.

Comment

In Chapter 1, we noted that the scope of a regression model is restricted ordinarily to some range of values of the predictor variable. This is particularly important to keep in mind in using estimates of the slope β_1. In our Toluca Company example, a linear regression model appeared appropriate for lot sizes between 20 and 120, the range of the predictor variable in the recent past. It may not be

TABLE 2.1 Results for Toluca Company Example Obtained in Chapter 1.

$n = 25$	$\bar{X} = 70.00$
$b_0 = 62.37$	$b_1 = 3.5702$
$\hat{Y} = 62.37 + 3.5702X$	$SSE = 54{,}825$
$\sum(X_i - \bar{X})^2 = 19{,}800$	$MSE = 2{,}384$
$\sum(Y_i - \bar{Y})^2 = 307{,}203$	

FIGURE 2.2 Portion of MINITAB Regression Output—Toluca Company Example.

```
The regression equation is
Y = 62.4 + 3.57 X

Predictor       Coef      Stdev     t-ratio        p
Constant       62.37      26.18        2.38    0.026
X             3.5702     0.3470       10.29    0.000

s = 48.82       R-sq = 82.2%      R-sq(adj) = 81.4%

Analysis of Variance

SOURCE       DF          SS          MS          F        p
Regression    1      252378      252378     105.88    0.000
Error        23       54825        2384
Total        24      307203
```

reasonable to use the estimate of the slope to infer the effect of lot size on number of work hours far outside this range since the regression relation may not be linear there. ■

Tests Concerning β_1

Since $(b_1 - \beta_1)/s\{b_1\}$ is distributed as t with $n - 2$ degrees of freedom, tests concerning β_1 can be set up in ordinary fashion using the t distribution.

Example 1

Two-Sided Test A cost analyst in the Toluca Company is interested in testing, using regression model (2.1), whether or not there is a linear association between work hours and lot size, i.e., whether or not $\beta_1 = 0$. The two alternatives then are:

$$\begin{aligned} H_0&: \beta_1 = 0 \\ H_a&: \beta_1 \neq 0 \end{aligned} \tag{2.16}$$

The analyst wishes to control the risk of a Type I error at $\alpha = .05$. The conclusion H_a could be reached at once by referring to the 95 percent confidence interval for β_1 constructed earlier, since this interval does not include 0.

An explicit test of the alternatives (2.16) is based on the test statistic:

$$t^* = \frac{b_1}{s\{b_1\}} \tag{2.17}$$

The decision rule with this test statistic for controlling the level of significance at α is:

$$\begin{aligned} &\text{If } |t^*| \leq t(1 - \alpha/2; n - 2), \text{ conclude } H_0 \\ &\text{If } |t^*| > t(1 - \alpha/2; n - 2), \text{ conclude } H_a \end{aligned} \tag{2.18}$$

For the Toluca Company example, where $\alpha = .05$, $b_1 = 3.5702$, and $s\{b_1\} = .3470$, we require $t(.975; 23) = 2.069$. Thus, the decision rule for testing alternatives (2.16) is:

$$\begin{aligned} &\text{If } |t^*| \leq 2.069, \text{ conclude } H_0 \\ &\text{If } |t^*| > 2.069, \text{ conclude } H_a \end{aligned}$$

Since $|t^*| = |3.5702/.3470| = 10.29 > 2.069$, we conclude H_a, that $\beta_1 \neq 0$ or that there is a linear association between work hours and lot size. The value of the test statistic, $t^* = 10.29$, is shown in the MINITAB output in Figure 2.2 in the column labeled t-ratio and the row labeled X.

The two-sided P-value for the sample outcome is obtained by first finding the one-sided P-value, $P\{t(23) > t^* = 10.29\}$. We see from Table B.2 that this probability is less than .0005. Many statistical calculators and computer packages will provide the actual probability; it is almost 0, denoted by 0+. Thus, the two-sided P-value is $2(0+) = 0+$. Since the two-sided P-value is less than the specified level of significance $\alpha = .05$, we could conclude H_a directly. The MINITAB output in Figure 2.2 shows the P-value in the column labeled p, corresponding to the row labeled X. It is shown as 0.000.

Comment

When the test of whether or not $\beta_1 = 0$ leads to the conclusion that $\beta_1 \neq 0$, the association between Y and X is sometimes described to be a linear statistical association. ■

Example 2

One-Sided Test Suppose the analyst had wished to test whether or not β_1 is positive, controlling the level of significance at $\alpha = .05$. The alternatives then would be:

$$\begin{aligned} H_0&: \beta_1 \leq 0 \\ H_a&: \beta_1 > 0 \end{aligned}$$

and the decision rule based on test statistic (2.17) would be:

$$\text{If } t^* \leq t(1-\alpha; n-2), \text{ conclude } H_0$$
$$\text{If } t^* > t(1-\alpha; n-2), \text{ conclude } H_a$$

For $\alpha = .05$, we require $t(.95; 23) = 1.714$. Since $t^* = 10.29 > 1.714$, we would conclude H_a, that β_1 is positive.

This same conclusion could be reached directly from the one-sided P-value, which was noted in Example 1 to be 0+. Since this P-value is less than .05, we would conclude H_a.

Comments

1. The P-value is sometimes called the observed level of significance.

2. Many scientific publications commonly report the P-value together with the value of the test statistic. In this way, one can conduct a test at any desired level of significance α by comparing the P-value with the specified level α.

3. Users of statistical calculators and computer packages need to be careful to ascertain whether one-sided or two-sided P-values are reported. Many commonly used labels, such as PROB or P, do not reveal whether the P-value is one- or two-sided.

4. Occasionally, it is desired to test whether or not β_1 equals some specified nonzero value β_{10}, which may be a historical norm, the value for a comparable process, or an engineering specification. The alternatives now are:

$$H_0: \beta_1 = \beta_{10}$$
$$H_a: \beta_1 \neq \beta_{10} \tag{2.19}$$

and the appropriate test statistic is:

$$t^* = \frac{b_1 - \beta_{10}}{s\{b_1\}} \tag{2.20}$$

The decision rule to be employed here still is (2.18), but it is now based on t^* defined in (2.20).

Note that test statistic (2.20) simplifies to test statistic (2.17) when the test involves $H_0: \beta_1 = \beta_{10} = 0$. ■

2.2 Inferences Concerning β_0

As noted in Chapter 1, there are only infrequent occasions when we wish to make inferences concerning β_0, the intercept of the regression line. These occur when the scope of the model includes $X = 0$.

Sampling Distribution of b_0

The point estimator b_0 was given in (1.10b) as follows:

$$b_0 = \bar{Y} - b_1\bar{X} \tag{2.21}$$

The sampling distribution of b_0 refers to the different values of b_0 that would be obtained with repeated sampling when the levels of the predictor variable X are held constant from

sample to sample.

For regression model (2.1), the sampling distribution of b_0 is normal, with mean and variance: **(2.22)**

$$E\{b_0\} = \beta_0 \quad \textbf{(2.22a)}$$

$$\sigma^2\{b_0\} = \sigma^2\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2}\right] \quad \textbf{(2.22b)}$$

The normality of the sampling distribution of b_0 follows because b_0, like b_1, is a linear combination of the observations Y_i. The results for the mean and variance of the sampling distribution of b_0 can be obtained in similar fashion as those for b_1.

An estimator of $\sigma^2\{b_0\}$ is obtained by replacing σ^2 by its point estimator MSE:

$$s^2\{b_0\} = MSE\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2}\right] \quad \textbf{(2.23)}$$

The positive square root, $s\{b_0\}$, is an estimator of $\sigma\{b_0\}$.

Sampling Distribution of $(b_0 - \beta_0)/s\{b_0\}$

Analogous to theorem (2.10) for b_1, a theorem for b_0 states:

$$\frac{b_0 - \beta_0}{s\{b_0\}} \text{ is distributed as } t(n-2) \text{ for regression model (2.1)} \quad \textbf{(2.24)}$$

Hence, confidence intervals for β_0 and tests concerning β_0 can be set up in ordinary fashion, using the t distribution.

Confidence Interval for β_0

The $1 - \alpha$ confidence limits for β_0 are obtained in the same manner as those for β_1 derived earlier. They are:

$$b_0 \pm t(1 - \alpha/2; n-2)s\{b_0\} \quad \textbf{(2.25)}$$

Example

As noted earlier, the scope of the model for the Toluca Company example does not extend to lot sizes of $X = 0$. Hence, the regression parameter β_0 may not have intrinsic meaning here. If, nevertheless, a 90 percent confidence interval for β_0 were desired, we would proceed by finding $t(.95; 23)$ and $s\{b_0\}$. From Table B.2, we find $t(.95; 23) = 1.714$. Using the earlier results summarized in Table 2.1, we obtain by (2.23):

$$s^2\{b_0\} = MSE\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2}\right] = 2{,}384\left[\frac{1}{25} + \frac{(70.00)^2}{19{,}800}\right] = 685.34$$

or:

$$s\{b_0\} = 26.18$$

The MINITAB output in Figure 2.2 shows this estimated standard deviation in the column labeled Stdev and the row labeled Constant.

The 90 percent confidence interval for β_0 is:

$$62.37 - 1.714(26.18) \le \beta_0 \le 62.37 + 1.714(26.18)$$

$$17.5 \le \beta_0 \le 107.2$$

We caution again that this confidence interval does not necessarily provide meaningful information. For instance, it does not necessarily provide information about the "setup" cost (the cost incurred in setting up the production process for the part) since we are not certain whether a linear regression model is appropriate when the scope of the model is extended to $X = 0$.

2.3 Some Considerations on Making Inferences Concerning β_0 and β_1

Effects of Departures from Normality

If the probability distributions of Y are not exactly normal but do not depart seriously, the sampling distributions of b_0 and b_1 will be approximately normal, and the use of the t distribution will provide approximately the specified confidence coefficient or level of significance. Even if the distributions of Y are far from normal, the estimators b_0 and b_1 generally have the property of *asymptotic normality*—their distributions approach normality under very general conditions as the sample size increases. Thus, with sufficiently large samples, the confidence intervals and decision rules given earlier still apply even if the probability distributions of Y depart far from normality. For large samples, the t value is, of course, replaced by the z value for the standard normal distribution.

Interpretation of Confidence Coefficient and Risks of Errors

Since regression model (2.1) assumes that the X_i are known constants, the confidence coefficient and risks of errors are interpreted with respect to taking repeated samples in which the X observations are kept at the same levels as in the observed sample. For instance, we constructed a confidence interval for β_1 with confidence coefficient .95 in the Toluca Company example. This coefficient is interpreted to mean that if many independent samples are taken where the levels of X (the lot sizes) are the same as in the data set and a 95 percent confidence interval is constructed for each sample, 95 percent of the intervals will contain the true value of β_1.

Spacing of the X Levels

Inspection of formulas (2.3b) and (2.22b) for the variances of b_1 and b_0, respectively, indicates that for given n and σ^2 these variances are affected by the spacing of the X levels in the observed data. For example, the greater is the spread in the X levels, the larger is the quantity $\sum(X_i - \bar{X})^2$ and the smaller is the variance of b_1. We discuss in Chapter 4 how the X observations should be spaced in experiments where spacing can be controlled.

Power of Tests

The power of tests on β_0 and β_1 can be obtained from Appendix Table B.5. Consider, for example, the general test concerning β_1 in (2.19):

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

for which test statistic (2.20) is employed:

$$t^* = \frac{b_1 - \beta_{10}}{s\{b_1\}}$$

and the decision rule for level of significance α is given in (2.18):

$$\text{If } |t^*| \le t(1-\alpha/2; n-2), \text{ conclude } H_0$$
$$\text{If } |t^*| > t(1-\alpha/2; n-2), \text{ conclude } H_a$$

The power of this test is the probability that the decision rule will lead to conclusion H_a when H_a in fact holds. Specifically, the power is given by:

$$\text{Power} = P\{|t^*| > t(1-\alpha/2; n-2) \mid \delta\} \qquad \textbf{(2.26)}$$

where δ is the *noncentrality measure*—i.e., a measure of how far the true value of β_1 is from β_{10}:

$$\delta = \frac{|\beta_1 - \beta_{10}|}{\sigma\{b_1\}} \qquad \textbf{(2.27)}$$

Table B.5 presents the power of the two-sided t test for $\alpha = .05$ and $\alpha = .01$, for various degrees of freedom df. To illustrate the use of this table, let us return to the Toluca Company example where we tested:

$$H_0: \beta_1 = \beta_{10} = 0$$
$$H_a: \beta_1 \ne \beta_{10} = 0$$

Suppose we wish to know the power of the test when $\beta_1 = 1.5$. To ascertain this, we need to know σ^2, the variance of the error terms. Assume, based on prior information or pilot data, that a reasonable planning value for the unknown variance is $\sigma^2 = 2{,}500$, so $\sigma^2\{b_1\}$ for our example would be:

$$\sigma^2\{b_1\} = \frac{\sigma^2}{\sum(X_i - \bar{X})^2} = \frac{2{,}500}{19{,}800} = .1263$$

or $\sigma\{b_1\} = .3553$. Then $\delta = |1.5 - 0| \div .3553 = 4.22$. We enter Table B.5 for $\alpha = .05$ (the level of significance used in the test) and 23 degrees of freedom and interpolate linearly between $\delta = 4.00$ and $\delta = 5.00$. We obtain:

$$.97 + \frac{4.22 - 4.00}{5.00 - 4.00}(1.00 - .97) = .9766$$

Thus, if $\beta_1 = 1.5$, the probability would be about .98 that we would be led to conclude H_a ($\beta_1 \ne 0$). In other words, if $\beta_1 = 1.5$, we would be almost certain to conclude that there is a linear relation between work hours and lot size.

The power of tests concerning β_0 can be obtained from Table B.5 in completely analogous fashion. For one-sided tests, Table B.5 should be entered so that one-half the level of significance shown there is the level of significance of the one-sided test.

2.4 Interval Estimation of $E\{Y_h\}$

A common objective in regression analysis is to estimate the mean for one or more probability distributions of Y. Consider, for example, a study of the relation between level of piecework pay (X) and worker productivity (Y). The mean productivity at high and medium levels of piecework pay may be of particular interest for purposes of analyzing the benefits obtained from an increase in the pay. As another example, the Toluca Company was interested in the mean response (mean number of work hours) for a range of lot sizes for purposes of finding the optimum lot size.

Let X_h denote the level of X for which we wish to estimate the mean response. X_h may be a value which occurred in the sample, or it may be some other value of the predictor variable within the scope of the model. The mean response when $X = X_h$ is denoted by $E\{Y_h\}$. Formula (1.12) gives us the point estimator $\hat{Y}_h$ of $E\{Y_h\}$:

$$\hat{Y}_h = b_0 + b_1 X_h \tag{2.28}$$

We consider now the sampling distribution of $\hat{Y}_h$.

Sampling Distribution of $\hat{Y}_h$

The sampling distribution of $\hat{Y}_h$, like the earlier sampling distributions discussed, refers to the different values of $\hat{Y}_h$ that would be obtained if repeated samples were selected, each holding the levels of the predictor variable X constant, and calculating $\hat{Y}_h$ for each sample.

For normal error regression model (2.1), the sampling distribution of $\hat{Y}_h$ is normal, with mean and variance: **(2.29)**

$$E\{\hat{Y}_h\} = E\{Y_h\} \tag{2.29a}$$

$$\sigma^2\{\hat{Y}_h\} = \sigma^2\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right] \tag{2.29b}$$

Normality. The normality of the sampling distribution of $\hat{Y}_h$ follows directly from the fact that $\hat{Y}_h$, like b_0 and b_1, is a linear combination of the observations Y_i.

Mean. Note from (2.29a) that $\hat{Y}_h$ is an unbiased estimator of $E\{Y_h\}$. To prove this, we proceed as follows:

$$E\{\hat{Y}_h\} = E\{b_0 + b_1 X_h\} = E\{b_0\} + X_h E\{b_1\} = \beta_0 + \beta_1 X_h$$

by (2.3a) and (2.22a).

Variance. Note from (2.29b) that the variability of the sampling distribution of $\hat{Y}_h$ is affected by how far X_h is from $\bar{X}$, through the term $(X_h - \bar{X})^2$. The further from $\bar{X}$ is X_h, the greater is the quantity $(X_h - \bar{X})^2$ and the larger is the variance of $\hat{Y}_h$. An intuitive explanation of this effect is found in Figure 2.3. Shown there are two sample regression lines, based on two samples for the same set of X values. The two regression lines are assumed to go through the same $(\bar{X}, \bar{Y})$ point to isolate the effect of interest, namely, the effect of variation in the estimated slope b_1 from sample to sample. Note that at X_1, near $\bar{X}$, the fitted values $\hat{Y}_1$ for the two sample regression lines are close to each other. At X_2, which is far from $\bar{X}$, the situation is different. Here, the fitted values $\hat{Y}_2$ differ substantially.

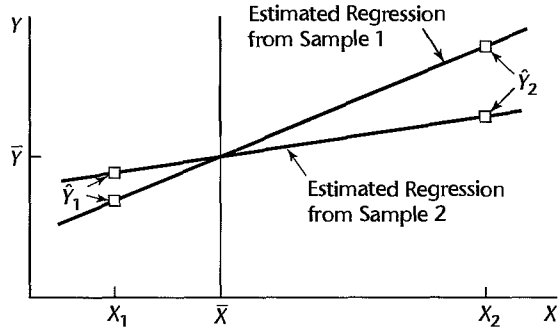


FIGURE 2.3
Effect on $\hat{Y}_h$ of Variation in b_1 from Sample to Sample in Two Samples with Same Means $\bar{Y}$ and $\bar{X}$.

Thus, variation in the slope b_1 from sample to sample has a much more pronounced effect on $\hat{Y}_h$ for X levels far from the mean $\bar{X}$ than for X levels near $\bar{X}$. Hence, the variation in the $\hat{Y}_h$ values from sample to sample will be greater when X_h is far from the mean than when X_h is near the mean.

When *MSE* is substituted for σ^2 in (2.29b), we obtain $s^2\{\hat{Y}_h\}$, the estimated variance of $\hat{Y}_h$:

$$s^2\{\hat{Y}_h\} = MSE\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right] \tag{2.30}$$

The estimated standard deviation of $\hat{Y}_h$ is then $s\{\hat{Y}_h\}$, the positive square root of $s^2\{\hat{Y}_h\}$.

Comments

1. When $X_h = 0$, the variance of $\hat{Y}_h$ in (2.29b) reduces to the variance of b_0 in (2.22b). Similarly, $s^2\{\hat{Y}_h\}$ in (2.30) reduces to $s^2\{b_0\}$ in (2.23). The reason is that $\hat{Y}_h = b_0$ when $X_h = 0$ since $\hat{Y}_h = b_0 + b_1 X_h$.

2. To derive $\sigma^2\{\hat{Y}_h\}$, we first show that b_1 and $\bar{Y}$ are uncorrelated and, hence, for regression model (2.1), independent:

$$\sigma\{\bar{Y}, b_1\} = 0 \tag{2.31}$$

where $\sigma\{\bar{Y}, b_1\}$ denotes the covariance between $\bar{Y}$ and b_1. We begin with the definitions:

$$\bar{Y} = \sum\left(\frac{1}{n}\right)Y_i \qquad b_1 = \sum k_i Y_i$$

where k_i is as defined in (2.4a). We now use (A.32), with $a_i = 1/n$ and $c_i = k_i$; remember that the Y_i are independent random variables:

$$\sigma\{\bar{Y}, b_1\} = \sum\left(\frac{1}{n}\right)k_i\sigma^2\{Y_i\} = \frac{\sigma^2}{n}\sum k_i$$

But we know from (2.5) that $\sum k_i = 0$. Hence, the covariance is 0.

Now we are ready to find the variance of $\hat{Y}_h$. We shall use the estimator in the alternative form (1.15):

$$\sigma^2\{\hat{Y}_h\} = \sigma^2\{\bar{Y} + b_1(X_h - \bar{X})\}$$

Since $\bar{Y}$ and b_1 are independent and X_h and $\bar{X}$ are constants, we obtain:

$$\sigma^2\{\hat{Y}_h\} = \sigma^2\{\bar{Y}\} + (X_h - \bar{X})^2\sigma^2\{b_1\}$$

Now $\sigma^2\{b_1\}$ is given in (2.3b), and:

$$\sigma^2\{\bar{Y}\} = \frac{\sigma^2\{Y_i\}}{n} = \frac{\sigma^2}{n}$$

Hence:

$$\sigma^2\{\hat{Y}_h\} = \frac{\sigma^2}{n} + (X_h - \bar{X})^2\frac{\sigma^2}{\sum(X_i - \bar{X})^2}$$

which, upon a slight rearrangement of terms, yields (2.29b). ■

Sampling Distribution of $(\hat{Y}_h - E\{Y_h\})/s\{\hat{Y}_h\}$

Since we have encountered the t distribution in each type of inference for regression model (2.1) up to this point, it should not be surprising that:

$$\frac{\hat{Y}_h - E\{Y_h\}}{s\{\hat{Y}_h\}} \text{ is distributed as } t(n-2) \text{ for regression model (2.1)} \qquad \textbf{(2.32)}$$

Hence, all inferences concerning $E\{Y_h\}$ are carried out in the usual fashion with the t distribution. We illustrate the construction of confidence intervals, since in practice these are used more frequently than tests.

Confidence Interval for $E\{Y_h\}$

A confidence interval for $E\{Y_h\}$ is constructed in the standard fashion, making use of the t distribution as indicated by theorem (2.32). The $1 - \alpha$ confidence limits are:

$$\hat{Y}_h \pm t(1 - \alpha/2; n-2)s\{\hat{Y}_h\} \qquad \textbf{(2.33)}$$

Example 1

Returning to the Toluca Company example, let us find a 90 percent confidence interval for $E\{Y_h\}$ when the lot size is $X_h = 65$ units. Using the earlier results in Table 2.1, we find the point estimate $\hat{Y}_h$:

$$\hat{Y}_h = 62.37 + 3.5702(65) = 294.4$$

Next, we need to find the estimated standard deviation $s\{\hat{Y}_h\}$. We obtain, using (2.30):

$$s^2\{\hat{Y}_h\} = 2{,}384\left[\frac{1}{25} + \frac{(65 - 70.00)^2}{19{,}800}\right] = 98.37$$

$$s\{\hat{Y}_h\} = 9.918$$

For a 90 percent confidence coefficient, we require $t(.95; 23) = 1.714$. Hence, our confidence interval with confidence coefficient .90 is by (2.33):

$$294.4 - 1.714(9.918) \le E\{Y_h\} \le 294.4 + 1.714(9.918)$$

$$277.4 \le E\{Y_h\} \le 311.4$$

We conclude with confidence coefficient .90 that the mean number of work hours required when lots of 65 units are produced is somewhere between 277.4 and 311.4 hours. We see that our estimate of the mean number of work hours is moderately precise.

Example 2 Suppose the Toluca Company wishes to estimate $E\{Y_h\}$ for lots with $X_h = 100$ units with a 90 percent confidence interval. We require:

$$\hat{Y}_h = 62.37 + 3.5702(100) = 419.4$$

$$s^2\{\hat{Y}_h\} = 2{,}384\left[\frac{1}{25} + \frac{(100 - 70.00)^2}{19{,}800}\right] = 203.72$$

$$s\{\hat{Y}_h\} = 14.27$$

$$t(.95; 23) = 1.714$$

Hence, the 90 percent confidence interval is:

$$419.4 - 1.714(14.27) \leq E\{Y_h\} \leq 419.4 + 1.714(14.27)$$

$$394.9 \leq E\{Y_h\} \leq 443.9$$

Note that this confidence interval is somewhat wider than that for Example 1, since the X_h level here ($X_h = 100$) is substantially farther from the mean $\bar{X} = 70.0$ than the X_h level for Example 1 ($X_h = 65$).

Comments

1. Since the X_i are known constants in regression model (2.1), the interpretation of confidence intervals and risks of errors in inferences on the mean response is in terms of taking repeated samples in which the X observations are at the same levels as in the actual study. We noted this same point in connection with inferences on β_0 and β_1.
2. We see from formula (2.29b) that, for given sample results, the variance of $\hat{Y}_h$ is smallest when $X_h = \bar{X}$. Thus, in an experiment to estimate the mean response at a particular level X_h of the predictor variable, the precision of the estimate will be greatest if (everything else remaining equal) the observations on X are spaced so that $\bar{X} = X_h$.
3. The usual relationship between confidence intervals and tests applies in inferences concerning the mean response. Thus, the two-sided confidence limits (2.33) can be utilized for two-sided tests concerning the mean response at X_h. Alternatively, a regular decision rule can be set up.
4. The confidence limits (2.33) for a mean response $E\{Y_h\}$ are not sensitive to moderate departures from the assumption that the error terms are normally distributed. Indeed, the limits are not sensitive to substantial departures from normality if the sample size is large. This robustness in estimating the mean response is related to the robustness of the confidence limits for β_0 and β_1, noted earlier.
5. Confidence limits (2.33) apply when a single mean response is to be estimated from the study. We discuss in Chapter 4 how to proceed when several mean responses are to be estimated from the same data. ■

2.5 Prediction of New Observation

We consider now the prediction of a new observation Y corresponding to a given level X of the predictor variable. Three illustrations where prediction of a new observation is needed follow.

1. In the Toluca Company example, the next lot to be produced consists of 100 units and management wishes to predict the number of work hours for this particular lot.

2. An economist has estimated the regression relation between company sales and number of persons 16 or more years old from data for the past 10 years. Using a reliable demographic projection of the number of persons 16 or more years old for next year, the economist wishes to predict next year's company sales.
3. An admissions officer at a university has estimated the regression relation between the high school grade point average (GPA) of admitted students and the first-year college GPA. The officer wishes to predict the first-year college GPA for an applicant whose high school GPA is 3.5 as part of the information on which an admissions decision will be based.

The new observation on Y to be predicted is viewed as the result of a new trial, independent of the trials on which the regression analysis is based. We denote the level of X for the new trial as X_h and the new observation on Y as $Y_{h(\text{new})}$. Of course, we assume that the underlying regression model applicable for the basic sample data continues to be appropriate for the new observation.

The distinction between estimation of the mean response $E\{Y_h\}$, discussed in the preceding section, and prediction of a new response $Y_{h(\text{new})}$, discussed now, is basic. In the former case, we estimate the *mean* of the distribution of Y. In the present case, we predict an *individual outcome* drawn from the distribution of Y. Of course, the great majority of individual outcomes deviate from the mean response, and this must be taken into account by the procedure for predicting $Y_{h(\text{new})}$.

Prediction Interval for $Y_{h(\text{new})}$ when Parameters Known

To illustrate the nature of a *prediction interval* for a new observation $Y_{h(\text{new})}$ in as simple a fashion as possible, we shall first assume that all regression parameters are known. Later we drop this assumption and make appropriate modifications.

Suppose that in the college admissions example the relevant parameters of the regression model are known to be:

$$\beta_0 = .10 \qquad \beta_1 = .95$$
$$E\{Y\} = .10 + .95X$$
$$\sigma = .12$$

The admissions officer is considering an applicant whose high school GPA is $X_h = 3.5$. The mean college GPA for students whose high school average is 3.5 is:

$$E\{Y_h\} = .10 + .95(3.5) = 3.425$$

Figure 2.4 shows the probability distribution of Y for $X_h = 3.5$. Its mean is $E\{Y_h\} = 3.425$, and its standard deviation is $\sigma = .12$. Further, the distribution is normal in accord with regression model (2.1).

Suppose we were to predict that the college GPA of the applicant whose high school GPA is $X_h = 3.5$ will be between:

$$E\{Y_h\} \pm 3\sigma$$
$$3.425 \pm 3(.12)$$

so that the prediction interval would be:

$$3.065 \leq Y_{h(\text{new})} \leq 3.785$$

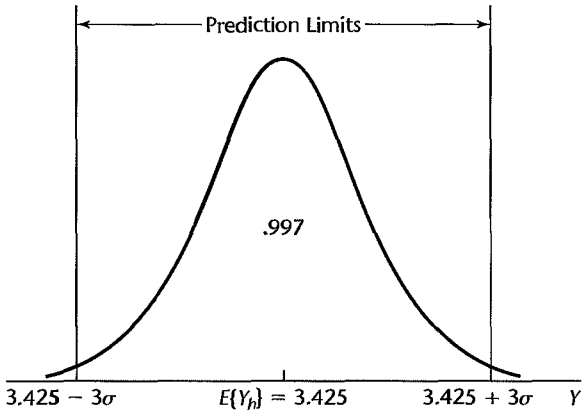


FIGURE 2.4 Prediction of $Y_{h(new)}$ when Parameters Known.

Since 99.7 percent of the area in a normal probability distribution falls within three standard deviations from the mean, the probability is .997 that this prediction interval will give a correct prediction for the applicant with high school GPA of 3.5. While the prediction limits here are rather wide, so that the prediction is not too precise, the prediction interval does indicate to the admissions officer that the applicant is expected to attain at least a 3.0 GPA in the first year of college.

The basic idea of a prediction interval is thus to choose a range in the distribution of Y wherein most of the observations will fall, and then to declare that the next observation will fall in this range. The usefulness of the prediction interval depends, as always, on the width of the interval and the needs for precision by the user.

In general, when the regression parameters of normal error regression model (2.1) are known, the $1 - \alpha$ prediction limits for $Y_{h(new)}$ are:

$$E\{Y_h\} \pm z(1 - \alpha/2)\sigma \tag{2.34}$$

In centering the limits around $E\{Y_h\}$, we obtain the narrowest interval consistent with the specified probability of a correct prediction.

Prediction Interval for $Y_{h(new)}$ when Parameters Unknown

When the regression parameters are unknown, they must be estimated. The mean of the distribution of Y is estimated by $\hat{Y}_h$, as usual, and the variance of the distribution of Y is estimated by *MSE*. We cannot, however, simply use the prediction limits (2.34) with the parameters replaced by the corresponding point estimators. The reason is illustrated intuitively in Figure 2.5. Shown there are two probability distributions of Y, corresponding to the upper and lower limits of a confidence interval for $E\{Y_h\}$. In other words, the distribution of Y could be located as far left as the one shown, as far right as the other one shown, or anywhere in between. Since we do not know the mean $E\{Y_h\}$ and only estimate it by a confidence interval, we cannot be certain of the location of the distribution of Y.

Figure 2.5 also shows the prediction limits for each of the two probability distributions of Y presented there. Since we cannot be certain of the location of the distribution

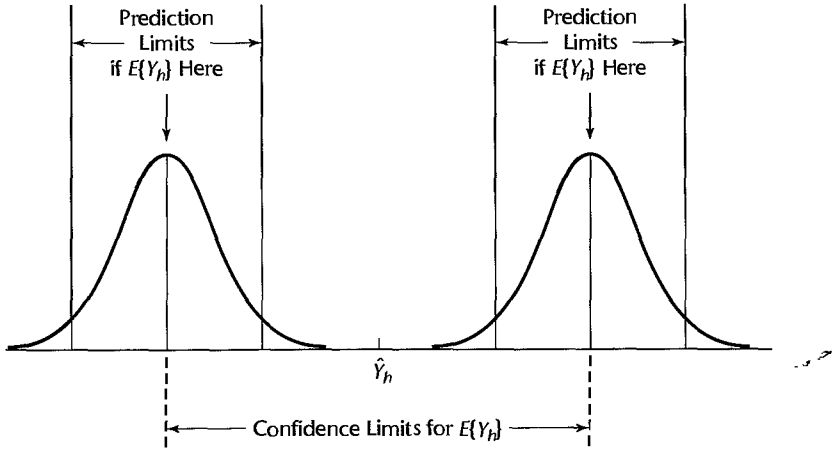


FIGURE 2.5
Prediction of $Y_{h(\text{new})}$ when Parameters Unknown.

of Y, prediction limits for $Y_{h(\text{new})}$ clearly must take account of two elements, as shown in Figure 2.5:

1. Variation in possible location of the distribution of Y.
2. Variation within the probability distribution of Y.

Prediction limits for a new observation $Y_{h(\text{new})}$ at a given level X_h are obtained by means of the following theorem:

$$\frac{Y_{h(\text{new})} - \hat{Y}_h}{s\{\text{pred}\}} \text{ is distributed as } t(n-2) \text{ for normal error regression model (2.1)} \quad \textbf{(2.35)}$$

Note that the studentized statistic (2.35) uses the point estimator $\hat{Y}_h$ in the numerator rather than the true mean $E\{Y_h\}$ because the true mean is unknown and cannot be used in making a prediction. The estimated standard deviation of the prediction, $s\{\text{pred}\}$, in the denominator of the studentized statistic will be defined shortly.

From theorem (2.35), it follows in the usual fashion that the $1-\alpha$ prediction limits for a new observation $Y_{h(\text{new})}$ are (for instance, compare (2.35) to (2.10) and relate $\hat{Y}_h$ to b_1 and $Y_{h(\text{new})}$ to β_1):

$$\hat{Y}_h \pm t(1-\alpha/2; n-2)s\{\text{pred}\} \quad \textbf{(2.36)}$$

Note that the numerator of the studentized statistic (2.35) represents how far the new observation $Y_{h(\text{new})}$ will deviate from the estimated mean $\hat{Y}_h$ based on the original n cases in the study. This difference may be viewed as the prediction error, with $\hat{Y}_h$ serving as the best point estimate of the value of the new observation $Y_{h(\text{new})}$. The variance of this prediction error can be readily obtained by utilizing the independence of the new observation $Y_{h(\text{new})}$ and the original n sample cases on which $\hat{Y}_h$ is based. We denote the variance of the prediction error by $\sigma^2\{\text{pred}\}$, and we obtain by (A.31b):

$$\sigma^2\{\text{pred}\} = \sigma^2\{Y_{h(\text{new})} - \hat{Y}_h\} = \sigma^2\{Y_{h(\text{new})}\} + \sigma^2\{\hat{Y}_h\} = \sigma^2 + \sigma^2\{\hat{Y}_h\} \quad \textbf{(2.37)}$$

Note that $\sigma^2\{\text{pred}\}$ has two components:

1. The variance of the distribution of Y at $X = X_h$, namely σ^2.
2. The variance of the sampling distribution of $\hat{Y}_h$, namely $\sigma^2\{\hat{Y}_h\}$.

An unbiased estimator of $\sigma^2\{\text{pred}\}$ is:

$$s^2\{\text{pred}\} = MSE + s^2\{\hat{Y}_h\} \tag{2.38}$$

which can be expressed as follows, using (2.30):

$$s^2\{\text{pred}\} = MSE\left[1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right] \tag{2.38a}$$

Example

The Toluca Company studied the relationship between lot size and work hours primarily to obtain information on the mean work hours required for different lot sizes for use in determining the optimum lot size. The company was also interested, however, to see whether the regression relationship is useful for predicting the required work hours for individual lots. Suppose that the next lot to be produced consists of $X_h = 100$ units and that a 90 percent prediction interval is desired. We require $t(.95; 23) = 1.714$. From earlier work, we have:

$$\hat{Y}_h = 419.4 \qquad s^2\{\hat{Y}_h\} = 203.72 \qquad MSE = 2{,}384$$

Using (2.38), we obtain:

$$s^2\{\text{pred}\} = 2{,}384 + 203.72 = 2{,}587.72$$
$$s\{\text{pred}\} = 50.87$$

Hence, the 90 percent prediction interval for $Y_{h(\text{new})}$ is by (2.36):

$$419.4 - 1.714(50.87) \le Y_{h(\text{new})} \le 419.4 + 1.714(50.87)$$
$$332.2 \le Y_{h(\text{new})} \le 506.6$$

With confidence coefficient .90, we predict that the number of work hours for the next production run of 100 units will be somewhere between 332 and 507 hours.

This prediction interval is rather wide and may not be too useful for planning worker requirements for the next lot. The interval can still be useful for control purposes, though. For instance, suppose that the actual work hours on the next lot of 100 units were 550 hours. Since the actual work hours fall outside the prediction limits, management would have an indication that a change in the production process may have occurred and would be alerted to the possible need for remedial action.

Note that the primary reason for the wide prediction interval is the large lot-to-lot variability in work hours for any given lot size; $MSE = 2{,}384$ accounts for 92 percent of the estimated prediction variance $s^2\{\text{pred}\} = 2{,}587.72$. It may be that the large lot-to-lot variability reflects other factors that affect the required number of work hours besides lot size, such as the amount of experience of employees assigned to the lot production. If so, a multiple regression model incorporating these other factors might lead to much more precise predictions. Alternatively, a designed experiment could be conducted to determine the main factors leading to the large lot-to-lot variation. A quality improvement program would then use these findings to achieve more uniform performance, for example, by additional training of employees if inadequate training accounted for much of the variability.

Comments

1. The 90 percent prediction interval for $Y_{h(\text{new})}$ obtained in the Toluca Company example is wider than the 90 percent confidence interval for $E\{Y_h\}$ obtained in Example 2 on page 55. The reason is that when predicting the work hours required for a new lot, we encounter both the variability in $\hat{Y}_h$ from sample to sample as well as the lot-to-lot variation within the probability distribution of Y.

2. Formula (2.38a) indicates that the prediction interval is wider the further X_h is from $\bar{X}$. The reason for this is that the estimate of the mean $\hat{Y}_h$, as noted earlier, is less precise as X_h is located farther away from $\bar{X}$.

3. The prediction limits (2.36), unlike the confidence limits (2.33) for a mean response $E\{Y_h\}$, are sensitive to departures from normality of the error terms distribution. In Chapter 3, we discuss diagnostic procedures for examining the nature of the probability distribution of the error terms, and we describe remedial measures if the departure from normality is serious.

4. The confidence coefficient for the prediction limits (2.36) refers to the taking of repeated samples based on the same set of X values, and calculating prediction limits for $Y_{h(\text{new})}$ for each sample.

5. Prediction limits (2.36) apply for a single prediction based on the sample data. Next, we discuss how to predict the mean of several new observations at a given X_h, and in Chapter 4 we take up how to make several predictions at different X_h levels.

6. Prediction intervals resemble confidence intervals. However, they differ conceptually. A confidence interval represents an inference on a parameter and is an interval that is intended to cover the value of the parameter. A prediction interval, on the other hand, is a statement about the value to be taken by a random variable, the new observation $Y_{h(\text{new})}$. ■

Prediction of Mean of *m* New Observations for Given X_h

Occasionally, one would like to predict the mean of m new observations on Y for a given level of the predictor variable. Suppose the Toluca Company has been asked to bid on a contract that calls for $m = 3$ production runs of $X_h = 100$ units during the next few months. Management would like to predict the mean work hours per lot for these three runs and then convert this into a prediction of the total work hours required to fill the contract.

We denote the mean of the new Y observations to be predicted as $\bar{Y}_{h(\text{new})}$. It can be shown that the appropriate $1 - \alpha$ prediction limits are, assuming that the new Y observations are independent:

$$\hat{Y}_h \pm t(1 - \alpha/2; n - 2)s\{\text{predmean}\} \tag{2.39}$$

where:

$$s^2\{\text{predmean}\} = \frac{MSE}{m} + s^2\{\hat{Y}_h\} \tag{2.39a}$$

or equivalently:

$$s^2\{\text{predmean}\} = MSE\left[\frac{1}{m} + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right] \tag{2.39b}$$

Note from (2.39a) that the variance $s^2\{\text{predmean}\}$ has two components:

1. The variance of the mean of m observations from the probability distribution of Y at $X = X_h$.
2. The variance of the sampling distribution of $\hat{Y}_h$.

Example

In the Toluca Company example, let us find the 90 percent prediction interval for the mean number of work hours $\bar{Y}_{h(\text{new})}$ in three new production runs, each for $X_h = 100$ units. From previous work, we have:

$$\hat{Y}_h = 419.4 \qquad s^2\{\hat{Y}_h\} = 203.72$$
$$MSE = 2{,}384 \qquad t(.95; 23) = 1.714$$

Hence, we obtain:

$$s^2\{\text{predmean}\} = \frac{2{,}384}{3} + 203.72 = 998.4$$
$$s\{\text{predmean}\} = 31.60$$

The prediction interval for the mean work hours per lot then is:

$$419.4 - 1.714(31.60) \le \bar{Y}_{h(\text{new})} \le 419.4 + 1.714(31.60)$$
$$365.2 \le \bar{Y}_{h(\text{new})} \le 473.6$$

Note that these prediction limits are narrower than those for predicting the work hours for a single lot of 100 units because they involve a prediction of the mean work hours for three lots.

We obtain the prediction interval for the total number of work hours for the three lots by multiplying the prediction limits for $\bar{Y}_{h(\text{new})}$ by 3:

$$1{,}095.6 = 3(365.2) \le \text{Total work hours} \le 3(473.6) = 1{,}420.8$$

Thus, it can be predicted with 90 percent confidence that between 1,096 and 1,421 work hours will be needed to fill the contract for three lots of 100 units each.

Comment

The 90 percent prediction interval for $\bar{Y}_{h(\text{new})}$, obtained for the Toluca Company example above, is narrower than that obtained for $Y_{h(\text{new})}$ on page 59, as expected. Furthermore, both of the prediction intervals are wider than the 90 percent confidence interval for $E\{Y_h\}$ obtained in Example 2 on page 55—also as expected. ■

2.6 Confidence Band for Regression Line

At times we would like to obtain a confidence band for the entire regression line $E\{Y\} = \beta_0 + \beta_1 X$. This band enables us to see the region in which the entire regression line lies. It is particularly useful for determining the appropriateness of a fitted regression function, as we explain in Chapter 3.

The Working-Hotelling $1 - \alpha$ confidence band for the regression line for regression model (2.1) has the following two boundary values at any level X_h:

$$\hat{Y}_h \pm Ws\{\hat{Y}_h\} \tag{2.40}$$

where:

$$W^2 = 2F(1 - \alpha; 2, n - 2) \tag{2.40a}$$

and $\hat{Y}_h$ and $s\{\hat{Y}_h\}$ are defined in (2.28) and (2.30), respectively. Note that the formula for the boundary values is of exactly the same form as formula (2.33) for the confidence limits for the mean response at X_h, except that the t multiple has been replaced by the W

multiple. Consequently, the boundary points of the confidence band for the regression line are wider apart the further X_h is from the mean $\bar{X}$ of the X observations. The W multiple will be larger than the t multiple in (2.33) because the confidence band must encompass the entire regression line, whereas the confidence limits for $E\{Y_h\}$ at X_h apply only at the single level X_h.

Example

We wish to determine how precisely we have been able to estimate the regression function for the Toluca Company example by obtaining the 90 percent confidence band for the regression line. We illustrate the calculations of the boundary values of the confidence band when $X_h = 100$. We found earlier for this case:

$$\hat{Y}_h = 419.4 \qquad s\{\hat{Y}_h\} = 14.27$$

We now require:

$$W^2 = 2F(1-\alpha; 2, n-2) = 2F(.90; 2, 23) = 2(2.549) = 5.098$$
$$W = 2.258$$

Hence, the boundary values of the confidence band for the regression line at $X_h = 100$ are $419.4 \pm 2.258(14.27)$, and the confidence band there is:

$$387.2 \leq \beta_0 + \beta_1 X_h \leq 451.6 \qquad \text{for } X_h = 100$$

In similar fashion, we can calculate the boundary values for other values of X_h by obtaining $\hat{Y}_h$ and $s\{\hat{Y}_h\}$ for each X_h level from (2.28) and (2.30) and then finding the boundary values by means of (2.40). Figure 2.6 contains a plot of the confidence band for the regression line. Note that at $X_h = 100$, the boundary values are 387.2 and 451.6, as we calculated earlier.

We see from Figure 2.6 that the regression line for the Toluca Company example has been estimated fairly precisely. The slope of the regression line is clearly positive, and the levels of the regression line at different levels of X are estimated fairly precisely except for small and large lot sizes.

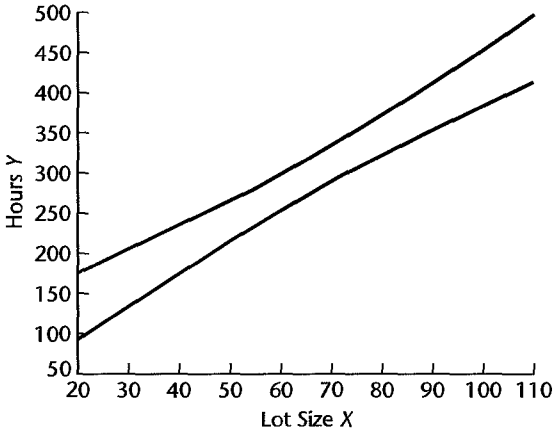


FIGURE 2.6 **Confidence Band for Regression Line—Toluca Company Example.**

Comments

1. The boundary values of the confidence band for the regression line in (2.40) define a hyperbola, as may be seen by replacing $\hat{Y}_h$ and $s\{\hat{Y}_h\}$ by their definitions in (2.28) and (2.30), respectively:

$$b_0 + b_1 X \pm W\sqrt{MSE}\left[\frac{1}{n} + \frac{(X - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right]^{1/2} \tag{2.41}$$

2. The boundary values of the confidence band for the regression line at any value X_h often are not substantially wider than the confidence limits for the mean response at that single X_h level. In the Toluca Company example, the t multiple for estimating the mean response at $X_h = 100$ with a 90 percent confidence interval was $t(.95; 23) = 1.714$. This compares with the W multiple for the 90 percent confidence band for the entire regression line of $W = 2.258$. With the somewhat wider limits for the entire regression line, one is able to draw conclusions about any and all mean responses for the entire regression line and not just about the mean response at a given X level. Some uses of this broader base for inference will be explained in the next two chapters.

3. The confidence band (2.40) applies to the entire regression line over all real-numbered values of X from $-\infty$ to ∞. The confidence coefficient indicates the proportion of time that the estimating procedure will yield a band that covers the entire line, in a long series of samples in which the X observations are kept at the same level as in the actual study.

In applications, the confidence band is ignored for that part of the regression line which is not of interest in the problem at hand. In the Toluca Company example, for instance, negative lot sizes would be ignored. The confidence coefficient for a limited segment of the band of interest is somewhat higher than $1 - \alpha$, so $1 - \alpha$ serves then as a lower bound to the confidence coefficient.

4. Some alternative procedures for developing confidence bands for the regression line have been developed. The simplicity of the Working-Hotelling confidence band (2.40) arises from the fact that it is a direct extension of the confidence limits for a single mean response in (2.33). ■

2.7 Analysis of Variance Approach to Regression Analysis

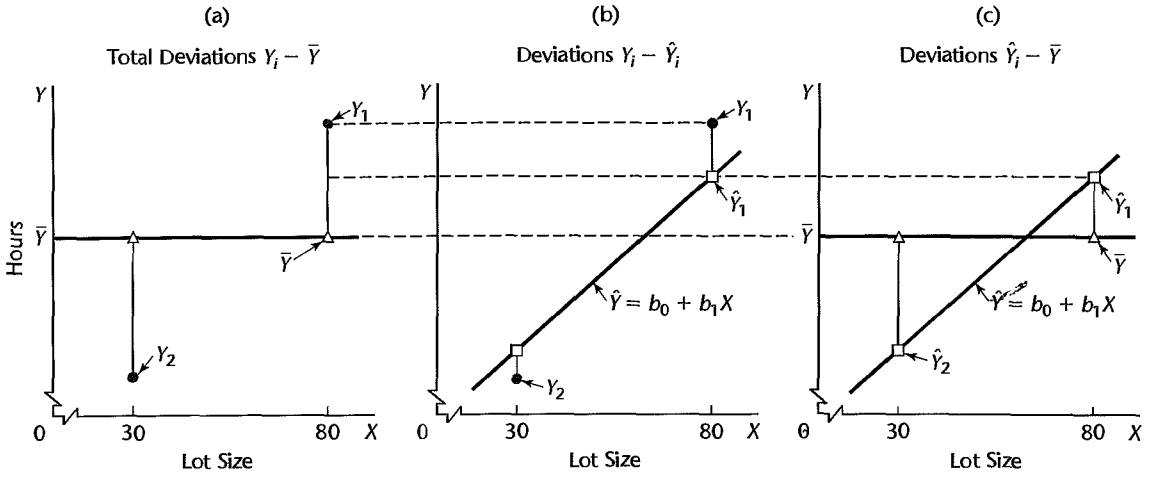
We now have developed the basic regression model and demonstrated its major uses. At this point, we consider the regression analysis from the perspective of analysis of variance. This new perspective will not enable us to do anything new, but the analysis of variance approach will come into its own when we take up multiple regression models and other types of linear statistical models.

Partitioning of Total Sum of Squares

Basic Notions. The analysis of variance approach is based on the partitioning of sums of squares and degrees of freedom associated with the response variable Y. To explain the motivation of this approach, consider again the Toluca Company example. Figure 2.7a shows the observations Y_i for the first two production runs presented in Table 1.1. Disregarding the lot sizes, we see that there is variation in the number of work hours Y_i, as in all statistical data. This variation is conventionally measured in terms of the deviations of the Y_i around their mean $\bar{Y}$:

$$Y_i - \bar{Y} \tag{2.42}$$

FIGURE 2.7 Illustration of Partitioning of Total Deviations $Y_i - \bar{Y}$—Toluca Company Example (not drawn to scale; only observations Y_1 and Y_2 are shown).



These deviations are shown by the vertical lines in Figure 2.7a. The measure of total variation, denoted by *SSTO*, is the sum of the squared deviations (2.42):

$$SSTO = \sum (Y_i - \bar{Y})^2 \tag{2.43}$$

Here *SSTO* stands for *total sum of squares*. If all Y_i observations are the same, $SSTO = 0$. The greater the variation among the Y_i observations, the larger is *SSTO*. Thus, *SSTO* for our example is a measure of the uncertainty pertaining to the work hours required for a lot, when the lot size is not taken into account.

When we utilize the predictor variable X, the variation reflecting the uncertainty concerning the variable Y is that of the Y_i observations around the fitted regression line:

$$Y_i - \hat{Y}_i \tag{2.44}$$

These deviations are shown by the vertical lines in Figure 2.7b. The measure of variation in the Y_i observations that is present when the predictor variable X is taken into account is the sum of the squared deviations (2.44), which is the familiar *SSE* of (1.21):

$$SSE = \sum (Y_i - \hat{Y}_i)^2 \tag{2.45}$$

Again, *SSE* denotes *error sum of squares*. If all Y_i observations fall on the fitted regression line, $SSE = 0$. The greater the variation of the Y_i observations around the fitted regression line, the larger is *SSE*.

For the Toluca Company example, we know from earlier work (Table 2.1) that:

$$SSTO = 307{,}203 \qquad SSE = 54{,}825$$

What accounts for the substantial difference between these two sums of squares? The difference, as we show shortly, is another sum of squares:

$$SSR = \sum (\hat{Y}_i - \bar{Y})^2 \tag{2.46}$$

where *SSR* stands for *regression sum of squares*. Note that *SSR* is a sum of squared deviations, the deviations being:

$$\hat{Y}_i - \bar{Y} \tag{2.47}$$

These deviations are shown by the vertical lines in Figure 2.7c. Each deviation is simply the difference between the fitted value on the regression line and the mean of the fitted values $\bar{Y}$. (Recall from (1.18) that the mean of the fitted values $\hat{Y}_i$ is $\bar{Y}$.) If the regression line is horizontal so that $\hat{Y}_i - \bar{Y} \equiv 0$, then $SSR = 0$. Otherwise, *SSR* is positive.

SSR may be considered a measure of that part of the variability of the Y_i which is associated with the regression line. The larger *SSR* is in relation to *SSTO*, the greater is the effect of the regression relation in accounting for the total variation in the Y_i observations.

For the Toluca Company example, we have:

$$SSR = SSTO - SSE = 307{,}203 - 54{,}825 = 252{,}378$$

which indicates that most of the total variability in work hours is accounted for by the relation between lot size and work hours.

Formal Development of Partitioning. The total deviation $Y_i - \bar{Y}$, used in the measure of the total variation of the observations Y_i without taking the predictor variable into account, can be decomposed into two components:

$$\underbrace{Y_i - \bar{Y}}_{\substack{\text{Total} \\ \text{deviation}}} = \underbrace{\hat{Y}_i - \bar{Y}}_{\substack{\text{Deviation of fitted} \\ \text{regression value} \\ \text{around mean}}} + \underbrace{Y_i - \hat{Y}_i}_{\substack{\text{Deviation} \\ \text{around} \\ \text{fitted} \\ \text{regression} \\ \text{line}}} \tag{2.48}$$

The two components are:

1. The deviation of the fitted value $\hat{Y}_i$ around the mean $\bar{Y}$.
2. The deviation of the observation Y_i around the fitted regression line.

Figure 2.7 shows this decomposition for observation Y_1 by the broken lines.

It is a remarkable property that the sums of these squared deviations have the same relationship:

$$\sum (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 + \sum (Y_i - \hat{Y}_i)^2 \tag{2.49}$$

or, using the notation in (2.43), (2.45), and (2.46):

$$SSTO = SSR + SSE \tag{2.50}$$

To prove this basic result in the analysis of variance, we proceed as follows:

$$\begin{aligned} \sum (Y_i - \bar{Y})^2 &= \sum [(\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i)]^2 \\ &= \sum [(\hat{Y}_i - \bar{Y})^2 + (Y_i - \hat{Y}_i)^2 + 2(\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i)] \\ &= \sum (\hat{Y}_i - \bar{Y})^2 + \sum (Y_i - \hat{Y}_i)^2 + 2 \sum (\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i) \end{aligned}$$

The last term on the right equals zero, as we can see by expanding it:

$$2\sum(\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i) = 2\sum \hat{Y}_i(Y_i - \hat{Y}_i) - 2\bar{Y}\sum(Y_i - \hat{Y}_i)$$

The first summation on the right equals zero by (1.20), and the second equals zero by (1.17). Hence, (2.49) follows.

Comment

The formulas for *SSTO*, *SSR*, and *SSE* given in (2.43), (2.45), and (2.46) are best for computational accuracy. Alternative formulas that are algebraically equivalent are available. One that is useful for deriving analytical results is:

$$SSR = b_1^2\sum(X_i - \bar{X})^2 \qquad (2.51)$$

■

Breakdown of Degrees of Freedom

Corresponding to the partitioning of the total sum of squares *SSTO*, there is a partitioning of the associated degrees of freedom (abbreviated *df*). We have $n - 1$ degrees of freedom associated with *SSTO*. One degree of freedom is lost because the deviations $Y_i - \bar{Y}$ are subject to one constraint: they must sum to zero. Equivalently, one degree of freedom is lost because the sample mean $\bar{Y}$ is used to estimate the population mean.

SSE, as noted earlier, has $n - 2$ degrees of freedom associated with it. Two degrees of freedom are lost because the two parameters β_0 and β_1 are estimated in obtaining the fitted values $\hat{Y}_i$.

SSR has one degree of freedom associated with it. Although there are n deviations $\hat{Y}_i - \bar{Y}$, all fitted values $\hat{Y}_i$ are calculated from the same estimated regression line. Two degrees of freedom are associated with a regression line, corresponding to the intercept and the slope of the line. One of the two degrees of freedom is lost because the deviations $\hat{Y}_i - \bar{Y}$ are subject to a constraint: they must sum to zero.

Note that the degrees of freedom are additive:

$$n - 1 = 1 + (n - 2)$$

For the Toluca Company example, these degrees of freedom are:

$$24 = 1 + 23$$

Mean Squares

A sum of squares divided by its associated degrees of freedom is called a *mean square* (abbreviated *MS*). For instance, an ordinary sample variance is a mean square since a sum of squares, $\sum(Y_i - \bar{Y})^2$, is divided by its associated degrees of freedom, $n - 1$. We are interested here in the *regression mean square*, denoted by *MSR*:

$$MSR = \frac{SSR}{1} = SSR \qquad (2.52)$$

and in the *error mean square*, *MSE*, defined earlier in (1.22):

$$MSE = \frac{SSE}{n-2} \qquad (2.53)$$

For the Toluca Company example, we have $SSR = 252{,}378$ and $SSE = 54{,}825$. Hence:

$$MSR = \frac{252{,}378}{1} = 252{,}378$$

Also, we obtained earlier:

$$MSE = \frac{54{,}825}{23} = 2{,}384$$

Comment

The two mean squares *MSR* and *MSE* do not add to

$$\frac{SSTO}{(n-1)} = \frac{307{,}203}{24} = 12{,}800$$

Thus, mean squares are not additive. ■

Analysis of Variance Table

Basic Table. The breakdowns of the total sum of squares and associated degrees of freedom are displayed in the form of an analysis of variance table (ANOVA table) in Table 2.2. Mean squares of interest also are shown. In addition, the ANOVA table contains a column of expected mean squares that will be utilized shortly. The ANOVA table for the Toluca Company example is shown in Figure 2.2. The columns for degrees of freedom and sums of squares are reversed in the MINITAB output.

Modified Table. Sometimes an ANOVA table showing one additional element of decomposition is utilized. This modified table is based on the fact that the total sum of squares can be decomposed into two parts, as follows:

$$SSTO = \sum(Y_i - \bar{Y})^2 = \sum Y_i^2 - n\bar{Y}^2$$

In the modified ANOVA table, the *total uncorrected sum of squares,* denoted by *SSTOU*, is defined as:

$$SSTOU = \sum Y_i^2 \tag{2.54}$$

and the *correction for the mean sum of squares*, denoted by *SS*(correction for mean), is defined as:

$$SS(\text{correction for mean}) = n\bar{Y}^2 \tag{2.55}$$

Table 2.3 shows the general format of this modified ANOVA table. While both types of ANOVA tables are widely used, we shall usually utilize the basic type of table.

TABLE 2.2 ANOVA Table for Simple Linear Regression.

Source of Variation	*SS*	*df*	*MS*	*E{MS}*
Regression	$SSR = \sum(\hat{Y}_i - \bar{Y})^2$	1	$MSR = \frac{SSR}{1}$	$\sigma^2 + \beta_1^2 \sum(X_i - \bar{X})^2$
Error	$SSE = \sum(Y_i - \hat{Y}_i)^2$	$n-2$	$MSE = \frac{SSE}{n-2}$	σ^2
Total	$SSTO = \sum(Y_i - \bar{Y})^2$	$n-1$		

TABLE 2.3 Modified ANOVA Table for Simple Linear Regression.

Source of Variation	*SS*	*df*	*MS*
Regression	$SSR = \sum(\hat{Y}_i - \bar{Y})^2$	1	$MSR = \frac{SSR}{1}$
Error	$SSE = \sum(Y_i - \hat{Y}_i)^2$	$n-2$	$MSE = \frac{SSE}{n-2}$
Total	$SSTO = \sum(Y_i - \bar{Y})^2$	$n-1$	
Correction for mean	$SS(\text{correction for mean}) = n\bar{Y}^2$	1	
Total, uncorrected	$SSTOU = \sum Y_i^2$	n	

Expected Mean Squares

In order to make inferences based on the analysis of variance approach, we need to know the expected value of each of the mean squares. The expected value of a mean square is the mean of its sampling distribution and tells us what is being estimated by the mean square. Statistical theory provides the following results:

$$E\{MSE\} = \sigma^2 \tag{2.56}$$

$$E\{MSR\} = \sigma^2 + \beta_1^2 \sum(X_i - \bar{X})^2 \tag{2.57}$$

The expected mean squares in (2.56) and (2.57) are shown in the analysis of variance table in Table 2.2. Note that result (2.56) is in accord with our earlier statement that *MSE* is an unbiased estimator of σ^2.

Two important implications of the expected mean squares in (2.56) and (2.57) are the following:

1. The mean of the sampling distribution of *MSE* is σ^2 whether or not X and Y are linearly related, i.e., whether or not $\beta_1 = 0$.
2. The mean of the sampling distribution of *MSR* is also σ^2 when $\beta_1 = 0$. Hence, when $\beta_1 = 0$, the sampling distributions of *MSR* and *MSE* are located identically and *MSR* and *MSE* will tend to be of the same order of magnitude.

   On the other hand, when $\beta_1 \neq 0$, the mean of the sampling distribution of *MSR* is greater than σ^2 since the term $\beta_1^2 \sum(X_i - \bar{X})^2$ in (2.57) then must be positive. Thus, when $\beta_1 \neq 0$, the mean of the sampling distribution of *MSR* is located to the right of that of *MSE* and, hence, *MSR* will tend to be larger than *MSE*.

This suggests that a comparison of *MSR* and *MSE* is useful for testing whether or not $\beta_1 = 0$. If *MSR* and *MSE* are of the same order of magnitude, this would suggest that $\beta_1 = 0$. On the other hand, if *MSR* is substantially greater than *MSE*, this would suggest that $\beta_1 \neq 0$. This indeed is the basic idea underlying the analysis of variance test to be discussed next.

Comment

The derivation of (2.56) follows from theorem (2.11), which states that $SSE/\sigma^2 \sim \chi^2(n-2)$ for regression model (2.1). Hence, it follows from property (A.42) of the chi-square distribution

that:

$$E\left\{\frac{SSE}{\sigma^2}\right\} = n - 2$$

or that:

$$E\left\{\frac{SSE}{n-2}\right\} = E\{MSE\} = \sigma^2$$

To find the expected value of *MSR*, we begin with (2.51):

$$SSR = b_1^2 \sum (X_i - \bar{X})^2$$

Now by (A.15a), we have:

$$\sigma^2\{b_1\} = E\{b_1^2\} - (E\{b_1\})^2 \tag{2.58}$$

We know from (2.3a) that $E\{b_1\} = \beta_1$ and from (2.3b) that:

$$\sigma^2\{b_1\} = \frac{\sigma^2}{\sum (X_i - \bar{X})^2}$$

Hence, substituting into (2.58), we obtain:

$$E\{b_1^2\} = \frac{\sigma^2}{\sum (X_i - \bar{X})^2} + \beta_1^2$$

It now follows that:

$$E\{SSR\} = E\{b_1^2\} \sum (X_i - \bar{X})^2 = \sigma^2 + \beta_1^2 \sum (X_i - \bar{X})^2$$

Finally, $E\{MSR\}$ is:

$$E\{MSR\} = E\left\{\frac{SSR}{1}\right\} = \sigma^2 + \beta_1^2 \sum (X_i - \bar{X})^2$$

■

F Test of $\beta_1 = 0$ versus $\beta_1 \neq 0$

The analysis of variance approach provides us with a battery of highly useful tests for regression models (and other linear statistical models). For the simple linear regression case considered here, the analysis of variance provides us with a test for:

$$\begin{aligned} H_0&: \beta_1 = 0 \\ H_a&: \beta_1 \neq 0 \end{aligned} \tag{2.59}$$

Test Statistic. The test statistic for the analysis of variance approach is denoted by F^*. As just mentioned, it compares *MSR* and *MSE* in the following fashion:

$$F^* = \frac{MSR}{MSE} \tag{2.60}$$

The earlier motivation, based on the expected mean squares in Table 2.2, suggests that large values of F^* support H_a and values of F^* near 1 support H_0. In other words, the appropriate test is an upper-tail one.

Sampling Distribution of F^*. In order to be able to construct a statistical decision rule and examine its properties, we need to know the sampling distribution of F^*. We begin by considering the sampling distribution of F^* when H_0 ($\beta_1 = 0$) holds. *Cochran's theorem*

will be most helpful in this connection. For our purposes, this theorem can be stated as follows:

> If all n observations Y_i come from the same normal distribution with mean μ and variance σ^2, and *SSTO* is decomposed into k sums of squares SS_r, each with degrees of freedom df_r, then the SS_r/σ^2 terms are independent χ^2 variables with df_r degrees of freedom if: **(2.61)**

$$\sum_{r=1}^{k} df_r = n - 1$$

Note from Table 2.2 that we have decomposed *SSTO* into the two sums of squares *SSR* and *SSE* and that their degrees of freedom are additive. Hence:

> If $\beta_1 = 0$ so that all Y_i have the same mean $\mu = \beta_0$ and the same variance σ^2, SSE/σ^2 and SSR/σ^2 are independent χ^2 variables.

Now consider test statistic F^*, which we can write as follows:

$$F^* = \frac{\frac{SSR}{\sigma^2}}{1} \div \frac{\frac{SSE}{\sigma^2}}{n-2} = \frac{MSR}{MSE}$$

But by Cochran's theorem, we have when H_0 holds:

$$F^* \sim \frac{\chi^2(1)}{1} \div \frac{\chi^2(n-2)}{n-2} \qquad \text{when } H_0 \text{ holds}$$

where the χ^2 variables are independent. Thus, when H_0 holds, F^* is the ratio of two independent χ^2 variables, each divided by its degrees of freedom. But this is the definition of an F random variable in (A.47).

We have thus established that if H_0 holds, F^* follows the F distribution, specifically the $F(1, n-2)$ distribution.

When H_a holds, it can be shown that F^* follows the noncentral F distribution, a complex distribution that we need not consider further at this time.

Comment

Even if $\beta_1 \neq 0$, *SSR* and *SSE* are independent and $SSE/\sigma^2 \sim \chi^2$. However, the condition that both SSR/σ^2 and SSE/σ^2 are χ^2 random variables requires $\beta_1 = 0$. ■

Construction of Decision Rule. Since the test is upper-tail and F^* is distributed as $F(1, n-2)$ when H_0 holds, the decision rule is as follows when the risk of a Type I error is to be controlled at α:

$$\begin{aligned} &\text{If } F^* \leq F(1-\alpha; 1, n-2), \text{ conclude } H_0 \\ &\text{If } F^* > F(1-\alpha; 1, n-2), \text{ conclude } H_a \end{aligned} \qquad \textbf{(2.62)}$$

where $F(1-\alpha; 1, n-2)$ is the $(1-\alpha)100$ percentile of the appropriate F distribution.

Example

For the Toluca Company example, we shall repeat the earlier test on β_1, this time using the F test. The alternative conclusions are:

$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$

As before, let $\alpha = .05$. Since $n = 25$, we require $F(.95; 1, 23) = 4.28$. The decision rule is:

$$\text{If } F^* \leq 4.28, \text{ conclude } H_0$$
$$\text{If } F^* > 4.28, \text{ conclude } H_a$$

We have from earlier that $MSR = 252{,}378$ and $MSE = 2{,}384$. Hence, F^* is:

$$F^* = \frac{252{,}378}{2{,}384} = 105.9$$

Since $F^* = 105.9 > 4.28$, we conclude H_a, that $\beta_1 \neq 0$, or that there is a linear association between work hours and lot size. This is the same result as when the t test was employed, as it must be according to our discussion below.

The MINITAB output in Figure 2.2 on page 46 shows the F^* statistic in the column labeled F. Next to it is shown the P-value, $P\{F(1, 23) > 105.9\}$, namely, 0+, indicating that the data are not consistent with $\beta_1 = 0$.

Equivalence of F Test and t Test. For a given α level, the F test of $\beta_1 = 0$ versus $\beta_1 \neq 0$ is equivalent algebraically to the two-tailed t test. To see this, recall from (2.51) that:

$$SSR = b_1^2 \sum (X_i - \bar{X})^2$$

Thus, we can write:

$$F^* = \frac{SSR \div 1}{SSE \div (n-2)} = \frac{b_1^2 \sum (X_i - \bar{X})^2}{MSE}$$

But since $s^2\{b_1\} = MSE / \sum (X_i - \bar{X})^2$, we obtain:

$$F^* = \frac{b_1^2}{s^2\{b_1\}} = \left(\frac{b_1}{s\{b_1\}}\right)^2 = (t^*)^2 \tag{2.63}$$

The last step follows because the t^* statistic for testing whether or not $\beta_1 = 0$ is by (2.17):

$$t^* = \frac{b_1}{s\{b_1\}}$$

In the Toluca Company example, we just calculated that $F^* = 105.9$. From earlier work, we have $t^* = 10.29$ (see Figure 2.2). We thus see that $(10.29)^2 = 105.9$.

Corresponding to the relation between t^* and F^*, we have the following relation between the required percentiles of the t and F distributions for the tests: $[t(1 - \alpha/2; n - 2)]^2 = F(1 - \alpha; 1, n - 2)$. In our tests on β_1, these percentiles were $[t(.975; 23)]^2 = (2.069)^2 = 4.28 = F(.95; 1, 23)$. Remember that the t test is two-tailed whereas the F test is one-tailed.

Thus, at any given α level, we can use either the t test or the F test for testing $\beta_1 = 0$ versus $\beta_1 \neq 0$. Whenever one test leads to H_0, so will the other, and correspondingly for H_a. The t test, however, is more flexible since it can be used for one-sided alternatives involving $\beta_1 (\leq \geq) 0$ versus $\beta_1 (> <) 0$, while the F test cannot.

2.8 General Linear Test Approach

The analysis of variance test of $\beta_1 = 0$ versus $\beta_1 \neq 0$ is an example of the general test for a linear statistical model. We now explain this general test approach in terms of the simple linear regression model. We do so at this time because of the generality of the approach and the wide use we shall make of it, and because of the simplicity of understanding the approach in terms of simple linear regression.

The general linear test approach involves three basic steps, which we now describe in turn.

Full Model

We begin with the model considered to be appropriate for the data, which in this context is called the *full* or *unrestricted model.* For the simple linear regression case, the full model is the normal error regression model (2.1):

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad \text{Full model} \tag{2.64}$$

We fit this full model, either by the method of least squares or by the method of maximum likelihood, and obtain the error sum of squares. The error sum of squares is the sum of the squared deviations of each observation Y_i around its estimated expected value. In this context, we shall denote this sum of squares by $SSE(F)$ to indicate that it is the error sum of squares for the full model. Here, we have:

$$SSE(F) = \sum [Y_i - (b_0 + b_1 X_i)]^2 = \sum (Y_i - \hat{Y}_i)^2 = SSE \tag{2.65}$$

Thus, for the full model (2.64), the error sum of squares is simply SSE, which measures the variability of the Y_i observations around the fitted regression line.

Reduced Model

Next, we consider H_0. In this instance, we have:

$$\begin{aligned} H_0&: \beta_1 = 0 \\ H_a&: \beta_1 \neq 0 \end{aligned} \tag{2.66}$$

The model when H_0 holds is called the *reduced* or *restricted model.* When $\beta_1 = 0$, model (2.64) reduces to:

$$Y_i = \beta_0 + \varepsilon_i \qquad \text{Reduced model} \tag{2.67}$$

We fit this reduced model, by either the method of least squares or the method of maximum likelihood, and obtain the error sum of squares for this reduced model, denoted by $SSE(R)$. When we fit the particular reduced model (2.67), it can be shown that the least squares and maximum likelihood estimator of β_0 is $\bar{Y}$. Hence, the estimated expected value for each observation is $b_0 = \bar{Y}$, and the error sum of squares for this reduced model is:

$$SSE(R) = \sum (Y_i - b_0)^2 = \sum (Y_i - \bar{Y})^2 = SSTO \tag{2.68}$$

Test Statistic

The logic now is to compare the two error sums of squares $SSE(F)$ and $SSE(R)$. It can be shown that $SSE(F)$ never is greater than $SSE(R)$:

$$SSE(F) \leq SSE(R) \tag{2.69}$$

The reason is that the more parameters are in the model, the better one can fit the data and the smaller are the deviations around the fitted regression function. When $SSE(F)$ is not much less than $SSE(R)$, using the full model does not account for much more of the variability of the Y_i than does the reduced model, in which case the data suggest that the reduced model is adequate (i.e., that H_0 holds). To put this another way, when $SSE(F)$ is close to $SSE(R)$, the variation of the observations around the fitted regression function for the full model is almost as great as the variation around the fitted regression function for the reduced model. In this case, the added parameters in the full model really do not help to reduce the variation in the Y_i about the fitted regression function. Thus, a small difference $SSE(R) - SSE(F)$ suggests that H_0 holds. On the other hand, a large difference suggests that H_a holds because the additional parameters in the model do help to reduce substantially the variation of the observations Y_i around the fitted regression function.

The actual test statistic is a function of $SSE(R) - SSE(F)$, namely:

$$F^* = \frac{SSE(R) - SSE(F)}{df_R - df_F} \div \frac{SSE(F)}{df_F} \tag{2.70}$$

which follows the F distribution when H_0 holds. The degrees of freedom df_R and df_F are those associated with the reduced and full model error sums of squares, respectively. Large values of F^* lead to H_a because a large difference $SSE(R) - SSE(F)$ suggests that H_a holds. The decision rule therefore is:

$$\begin{aligned} &\text{If } F^* \leq F(1-\alpha; df_R - df_F,\ df_F), \text{ conclude } H_0 \\ &\text{If } F^* > F(1-\alpha; df_R - df_F,\ df_F), \text{ conclude } H_a \end{aligned} \tag{2.71}$$

For testing whether or not $\beta_1 = 0$, we therefore have:

$$\begin{aligned} SSE(R) &= SSTO & SSE(F) &= SSE \\ df_R &= n-1 & df_F &= n-2 \end{aligned}$$

so that we obtain when substituting into (2.70):

$$F^* = \frac{SSTO - SSE}{(n-1)-(n-2)} \div \frac{SSE}{n-2} = \frac{SSR}{1} \div \frac{SSE}{n-2} = \frac{MSR}{MSE}$$

which is identical to the analysis of variance test statistic (2.60).

Summary

The general linear test approach can be used for highly complex tests of linear statistical models, as well as for simple tests. The basic steps in summary form are:

1. Fit the full model and obtain the error sum of squares $SSE(F)$.
2. Fit the reduced model under H_0 and obtain the error sum of squares $SSE(R)$.
3. Use test statistic (2.70) and decision rule (2.71).

2.9 Descriptive Measures of Linear Association between X and Y

We have discussed the major uses of regression analysis—estimation of parameters and means and prediction of new observations—without mentioning the "degree of linear association" between X and Y, or similar terms. The reason is that the usefulness of estimates or predictions depends upon the width of the interval and the user's needs for precision, which vary from one application to another. Hence, no single descriptive measure of the "degree of linear association" can capture the essential information as to whether a given regression relation is useful in any particular application.

Nevertheless, there are times when the degree of linear association is of interest in its own right. We shall now briefly discuss two descriptive measures that are frequently used in practice to describe the degree of linear association between X and Y.

Coefficient of Determination

We saw earlier that *SSTO* measures the variation in the observations Y_i, or the uncertainty in predicting Y, when no account of the predictor variable X is taken. Thus, *SSTO* is a measure of the uncertainty in predicting Y when X is not considered. Similarly, *SSE* measures the variation in the Y_i when a regression model utilizing the predictor variable X is employed. A natural measure of the effect of X in reducing the variation in Y, i.e., in reducing the uncertainty in predicting Y, is to express the reduction in variation ($SSTO - SSE = SSR$) as a proportion of the total variation:

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO} \tag{2.72}$$

The measure R^2 is called the *coefficient of determination*. Since $0 \leq SSE \leq SSTO$, it follows that:

$$0 \leq R^2 \leq 1 \tag{2.72a}$$

We may interpret R^2 as the proportionate reduction of total variation associated with the use of the predictor variable X. Thus, the larger R^2 is, the more the total variation of Y is reduced by introducing the predictor variable X. The limiting values of R^2 occur as follows:

1. When all observations fall on the fitted regression line, then $SSE = 0$ and $R^2 = 1$. This case is shown in Figure 2.8a. Here, the predictor variable X accounts for all variation in the observations Y_i.
2. When the fitted regression line is horizontal so that $b_1 = 0$ and $\hat{Y}_i \equiv \bar{Y}$, then $SSE = SSTO$ and $R^2 = 0$. This case is shown in Figure 2.8b. Here, there is no linear association between X and Y in the sample data, and the predictor variable X is of no help in reducing the variation in the observations Y_i with linear regression.

In practice, R^2 is not likely to be 0 or 1 but somewhere between these limits. The closer it is to 1, the greater is said to be the degree of linear association between X and Y.

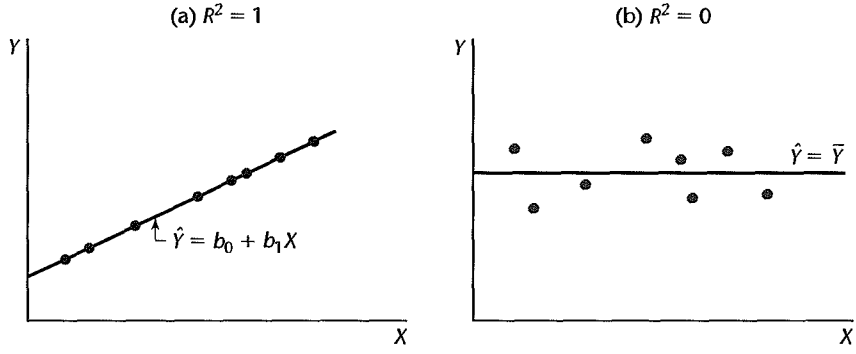


FIGURE 2.8
Scatter Plots when $R^2 = 1$ and $R^2 = 0$.

Example

For the Toluca Company example, we obtained $SSTO = 307{,}203$ and $SSR = 252{,}378$. Hence:

$$R^2 = \frac{252{,}378}{307{,}203} = .822$$

Thus, the variation in work hours is reduced by 82.2 percent when lot size is considered.

The MINITAB output in Figure 2.2 shows the coefficient of determination R^2 labeled as R–sq in percent form. The output also shows the coefficient R–sq(adj), which will be explained in Chapter 6.

Limitations of R^2

We noted that no single measure will be adequate for describing the usefulness of a regression model for different applications. Still, the coefficient of determination is widely used. Unfortunately, it is subject to serious misunderstandings. We consider now three common misunderstandings:

Misunderstanding 1. A high coefficient of determination indicates that useful predictions can be made. This is not necessarily correct. In the Toluca Company example, we saw that the coefficient of determination was high ($R^2 = .82$). Yet the 90 percent prediction interval for the next lot, consisting of 100 units, was wide (332 to 507 hours) and not precise enough to permit management to schedule workers effectively.

Misunderstanding 2. A high coefficient of determination indicates that the estimated regression line is a good fit. Again, this is not necessarily correct. Figure 2.9a shows a scatter plot where the coefficient of determination is high ($R^2 = .69$). Yet a linear regression function would not be a good fit since the regression relation is curvilinear.

Misunderstanding 3. A coefficient of determination near zero indicates that X and Y are not related. This also is not necessarily correct. Figure 2.9b shows a scatter plot where the coefficient of determination between X and Y is $R^2 = .02$. Yet X and Y are strongly related; however, the relationship between the two variables is curvilinear.

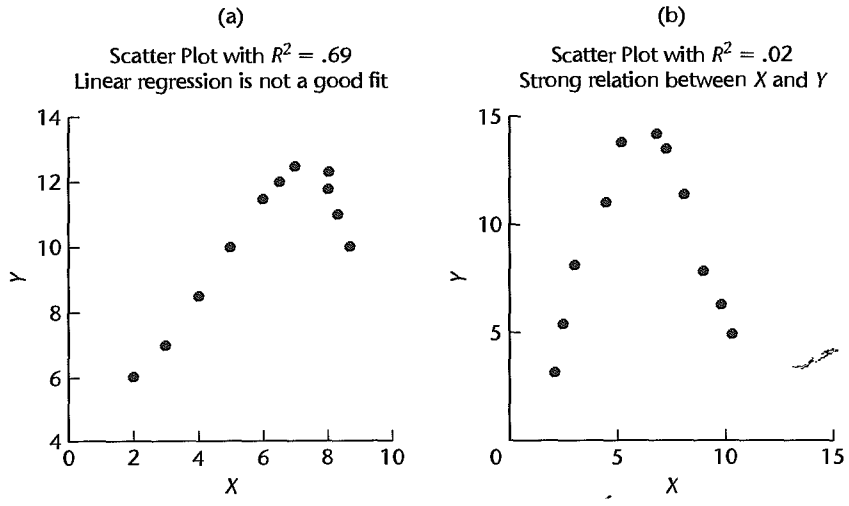


FIGURE 2.9 Illustrations of Two Misunderstandings about Coefficient of Determination.

Misunderstanding 1 arises because R^2 measures only a relative reduction from *SSTO* and provides no information about absolute precision for estimating a mean response or predicting a new observation. Misunderstandings 2 and 3 arise because R^2 measures the degree of *linear* association between X and Y, whereas the actual regression relation may be curvilinear.

Coefficient of Correlation

A measure of linear association between Y and X when both Y and X are random is the *coefficient of correlation*. This measure is the signed square root of R^2:

$$r = \pm\sqrt{R^2} \tag{2.73}$$

A plus or minus sign is attached to this measure according to whether the slope of the fitted regression line is positive or negative. Thus, the range of r is: $-1 \leq r \leq 1$.

Example

For the Toluca Company example, we obtained $R^2 = .822$. Treating X as a random variable, the correlation coefficient here is:

$$r = +\sqrt{.822} = .907$$

The plus sign is affixed since b_1 is positive. We take up the topic of correlation analysis in more detail in Section 2.11.

Comments

1. The value taken by R^2 in a given sample tends to be affected by the spacing of the X observations. This is implied in (2.72). *SSE* is not affected systematically by the spacing of the X_i since, for regression model (2.1), $\sigma^2\{Y_i\} = \sigma^2$ at all X levels. However, the wider the spacing of the X_i in the sample when $b_1 \neq 0$, the greater will tend to be the spread of the observed Y_i around $\bar{Y}$ and hence the greater *SSTO* will be. Consequently, the wider the X_i are spaced, the higher R^2 will tend to be.

2. The regression sum of squares *SSR* is often called the "explained variation" in Y, and the residual sum of squares *SSE* is called the "unexplained variation." The coefficient R^2 then is interpreted in terms of the proportion of the total variation in Y (*SSTO*) which has been "explained" by X. Unfortunately,

this terminology frequently is taken literally and, hence, misunderstood. Remember that in a regression model there is no implication that Y necessarily depends on X in a causal or explanatory sense.

3. Regression models do not contain a parameter to be estimated by R^2 or r. These are simply descriptive measures of the degree of linear association between X and Y in the sample observations that may, or may not, be useful in any instance. ■

2.10 Considerations in Applying Regression Analysis

We have now discussed the major uses of regression analysis—to make inferences about the regression parameters, to estimate the mean response for a given X, and to predict a new observation Y for a given X. It remains to make a few cautionary remarks about implementing applications of regression analysis.

1. Frequently, regression analysis is used to make inferences for the future. For instance, for planning staffing requirements, a school board may wish to predict future enrollments by using a regression model containing several demographic variables as predictor variables. In applications of this type, it is important to remember that the validity of the regression application depends upon whether basic causal conditions in the period ahead will be similar to those in existence during the period upon which the regression analysis is based. This caution applies whether mean responses are to be estimated, new observations predicted, or regression parameters estimated.

2. In predicting new observations on Y, the predictor variable X itself often has to be predicted. For instance, we mentioned earlier the prediction of company sales for next year from the demographic projection of the number of persons 16 years of age or older next year. A prediction of company sales under these circumstances is a conditional prediction, dependent upon the correctness of the population projection. It is easy to forget the conditional nature of this type of prediction.

3. Another caution deals with inferences pertaining to levels of the predictor variable that fall outside the range of observations. Unfortunately, this situation frequently occurs in practice. A company that predicts its sales from a regression relation of company sales to disposable personal income will often find the level of disposable personal income of interest (e.g., for the year ahead) to fall beyond the range of past data. If the X level does not fall far beyond this range, one may have reasonable confidence in the application of the regression analysis. On the other hand, if the X level falls far beyond the range of past data, extreme caution should be exercised since one cannot be sure that the regression function that fits the past data is appropriate over the wider range of the predictor variable.

4. A statistical test that leads to the conclusion that $\beta_1 \neq 0$ does not establish a cause-and-effect relation between the predictor and response variables. As we noted in Chapter 1, with nonexperimental data both the X and Y variables may be simultaneously influenced by other variables not in the regression model. On the other hand, the existence of a regression relation in controlled experiments is often good evidence of a cause-and-effect relation.

5. We should note again that frequently we wish to estimate several mean responses or predict several new observations for different levels of the predictor variable, and that special problems arise in this case. The confidence coefficients for the limits (2.33) for estimating a mean response and for the prediction limits (2.36) for a new observation apply

only for a single level of X for a given sample. In Chapter 4, we discuss how to make multiple inferences from a given sample.

6. Finally, when observations on the predictor variable X are subject to measurement errors, the resulting parameter estimates are generally no longer unbiased. In Chapter 4, we discuss several ways to handle this situation.

2.11 Normal Correlation Models

Distinction between Regression and Correlation Model

The normal error regression model (2.1), which has been used throughout this chapter and which will continue to be used, assumes that the X values are known constants. As a consequence of this, the confidence coefficients and risks of errors refer to repeated sampling when the X values are kept the same from sample to sample.

Frequently, it may not be appropriate to consider the X values as known constants. For instance, consider regressing daily bathing suit sales by a department store on mean daily temperature. Surely, the department store cannot control daily temperatures, so it would not be meaningful to think of repeated sampling where the temperature levels are the same from sample to sample. As a second example, an analyst may use a correlation model for the two variables "height of person" and "weight of person" in a study of a sample of persons, each variable being taken as random. The analyst might wish to study the relation between the two variables or might be interested in making inferences about weight of a person on the basis of the person's height, in making inferences about height on the basis of weight, or in both.

Other examples where a correlation model, rather than a regression model, may be appropriate are:

1. To study the relation between service station sales of gasoline, and sales of auxiliary products.
2. To study the relation between company net income determined by generally accepted accounting principles and net income according to tax regulations.
3. To study the relation between blood pressure and age in human subjects.

The correlation model most widely employed is the normal correlation model. We discuss it here for the case of two variables.

Bivariate Normal Distribution

The normal correlation model for the case of two variables is based on the *bivariate normal distribution*. Let us denote the two variables as Y_1 and Y_2. (We do not use the notation X and Y here because both variables play a symmetrical role in correlation analysis.) We say that Y_1 and Y_2 are *jointly normally distributed* if the density function of their joint distribution is that of the bivariate normal distribution.

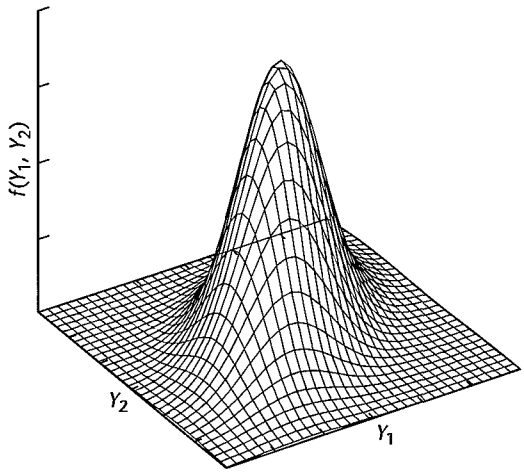


FIGURE 2.10
Example of Bivariate Normal Distribution.

Density Function. The density function of the bivariate normal distribution is as follows:

$$f(Y_1, Y_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho_{12}^2}} \exp\left\{ -\frac{1}{2(1-\rho_{12}^2)}\left[\left(\frac{Y_1-\mu_1}{\sigma_1}\right)^2 - 2\rho_{12}\left(\frac{Y_1-\mu_1}{\sigma_1}\right)\left(\frac{Y_2-\mu_2}{\sigma_2}\right) + \left(\frac{Y_2-\mu_2}{\sigma_2}\right)^2\right]\right\} \quad (2.74)$$

Note that this density function involves five parameters: $\mu_1, \mu_2, \sigma_1, \sigma_2, \rho_{12}$. We shall explain the meaning of these parameters shortly. First, let us consider a graphic representation of the bivariate normal distribution.

Figure 2.10 contains a SYSTAT three-dimensional plot of a bivariate normal probability distribution. The probability distribution is a surface in three-dimensional space. For every pair of (Y_1, Y_2) values, the density $f(Y_1, Y_2)$ represents the height of the surface at that point. The surface is continuous, and probability corresponds to volume under the surface.

Marginal Distributions. If Y_1 and Y_2 are jointly normally distributed, it can be shown that their marginal distributions have the following characteristics:

The marginal distribution of Y_1 is normal with mean μ_1 and standard deviation σ_1: **(2.75a)**

$$f_1(Y_1) = \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left[-\frac{1}{2}\left(\frac{Y_1-\mu_1}{\sigma_1}\right)^2\right]$$

The marginal distribution of Y_2 is normal with mean μ_2 and standard deviation σ_2: **(2.75b)**

$$f_2(Y_2) = \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left[-\frac{1}{2}\left(\frac{Y_2-\mu_2}{\sigma_2}\right)^2\right]$$

Thus, when Y_1 and Y_2 are jointly normally distributed, each of the two variables by itself is normally distributed. The converse, however, is not generally true; if Y_1 and Y_2 are each normally distributed, they need not be jointly normally distributed in accord with (2.74).

Meaning of Parameters. The five parameters of the bivariate normal density function (2.74) have the following meaning:

1. μ_1 and σ_1 are the mean and standard deviation of the marginal distribution of Y_1.
2. μ_2 and σ_2 are the mean and standard deviation of the marginal distribution of Y_2.
3. ρ_{12} is the *coefficient of correlation* between the random variables Y_1 and Y_2. This coefficient is denoted by $\rho\{Y_1, Y_2\}$ in Appendix A, using the correlation operator notation, and defined in (A.25a):

$$\rho_{12} = \rho\{Y_1, Y_2\} = \frac{\sigma_{12}}{\sigma_1\sigma_2} \tag{2.76}$$

Here, σ_1 and σ_2, as just mentioned, denote the standard deviations of Y_1 and Y_2, and σ_{12} denotes the covariance $\sigma\{Y_1, Y_2\}$ between Y_1 and Y_2 as defined in (A.21):

$$\sigma_{12} = \sigma\{Y_1, Y_2\} = E\{(Y_1 - \mu_1)(Y_2 - \mu_2)\} \tag{2.77}$$

Note that $\sigma_{12} \equiv \sigma_{21}$ and $\rho_{12} \equiv \rho_{21}$.

If Y_1 and Y_2 are independent, $\sigma_{12} = 0$ according to (A.28) so that $\rho_{12} = 0$. If Y_1 and Y_2 are positively related—that is, Y_1 tends to be large when Y_2 is large, or small when Y_2 is small—σ_{12} is positive and so is ρ_{12}. On the other hand, if Y_1 and Y_2 are negatively related—that is, Y_1 tends to be large when Y_2 is small, or vice versa—σ_{12} is negative and so is ρ_{12}. The coefficient of correlation ρ_{12} can take on any value between -1 and 1 inclusive. It assumes 1 if the linear relation between Y_1 and Y_2 is perfectly positive (direct) and -1 if it is perfectly negative (inverse).

Conditional Inferences

As noted, one principal use of a bivariate correlation model is to make conditional inferences regarding one variable, given the other variable. Suppose Y_1 represents a service station's gasoline sales and Y_2 its sales of auxiliary products. We may then wish to predict a service station's sales of auxiliary products Y_2, given that its gasoline sales are $Y_1 = \$5{,}500$.

Such conditional inferences require the use of conditional probability distributions, which we discuss next.

Conditional Probability Distribution of Y_1. The density function of the conditional probability distribution of Y_1 for any given value of Y_2 is denoted by $f(Y_1|Y_2)$ and defined as follows:

$$f(Y_1|Y_2) = \frac{f(Y_1, Y_2)}{f_2(Y_2)} \tag{2.78}$$

where $f(Y_1, Y_2)$ is the joint density function of Y_1 and Y_2, and $f_2(Y_2)$ is the marginal density function of Y_2. When Y_1 and Y_2 are jointly normally distributed according to (2.74) so that the marginal density function $f_2(Y_2)$ is given by (2.75b), it can be shown that:

The conditional probability distribution of Y_1 for any given value of Y_2 is normal with mean $\alpha_{1|2} + \beta_{12}Y_2$ and standard deviation $\sigma_{1|2}$ and its density function is: **(2.79)**

$$f(Y_1|Y_2) = \frac{1}{\sqrt{2\pi}\sigma_{1|2}} \exp\left[-\frac{1}{2}\left(\frac{Y_1 - \alpha_{1|2} - \beta_{12}Y_2}{\sigma_{1|2}}\right)^2\right]$$

The parameters $\alpha_{1|2}$, β_{12}, and $\sigma_{1|2}$ of the conditional probability distributions of Y_1 are functions of the parameters of the joint probability distribution (2.74), as follows:

$$\alpha_{1|2} = \mu_1 - \mu_2\rho_{12}\frac{\sigma_1}{\sigma_2} \tag{2.80a}$$

$$\beta_{12} = \rho_{12}\frac{\sigma_1}{\sigma_2} \tag{2.80b}$$

$$\sigma^2_{1|2} = \sigma_1^2(1 - \rho^2_{12}) \tag{2.80c}$$

The parameter $\alpha_{1|2}$ is the intercept of the line of regression of Y_1 on Y_2, and the parameter β_{12} is the slope of this line. Thus we find that the conditional distribution of Y_1, given Y_2, is equivalent to the normal error regression model (1.24).

Conditional Probability Distributions of Y_2. The random variables Y_1 and Y_2 play symmetrical roles in the bivariate normal probability distribution (2.74). Hence, it follows:

The conditional probability distribution of Y_2 for any given value of Y_1 is normal with mean $\alpha_{2|1} + \beta_{21}Y_1$ and standard deviation $\sigma_{2|1}$ and its density function is: **(2.81)**

$$f(Y_2|Y_1) = \frac{1}{\sqrt{2\pi}\sigma_{2|1}} \exp\left[-\frac{1}{2}\left(\frac{Y_2 - \alpha_{2|1} - \beta_{21}Y_1}{\sigma_{2|1}}\right)^2\right]$$

The parameters $\alpha_{2|1}$, β_{21}, and $\sigma_{2|1}$ of the conditional probability distributions of Y_2 are functions of the parameters of the joint probability distribution (2.74), as follows:

$$\alpha_{2|1} = \mu_2 - \mu_1\rho_{12}\frac{\sigma_2}{\sigma_1} \tag{2.82a}$$

$$\beta_{21} = \rho_{12}\frac{\sigma_2}{\sigma_1} \tag{2.82b}$$

$$\sigma^2_{2|1} = \sigma_2^2(1 - \rho^2_{12}) \tag{2.82c}$$

Important Characteristics of Conditional Distributions. Three important characteristics of the conditional probability distributions of Y_1 are normality, linear regression, and constant variance. We take up each of these in turn.

1. The conditional probability distribution of Y_1 for any given value of Y_2 is normal. Imagine that we slice a bivariate normal distribution vertically at a given value of Y_2, say, at Y_{h2}. That is, we slice it parallel to the Y_1 axis. This slicing is shown in Figure 2.11. The exposed cross section has the shape of a normal distribution, and after being scaled so that its area is 1, it portrays the conditional probability distribution of Y_1, given that $Y_2 = Y_{h2}$.

This property of normality holds no matter what the value Y_{h2} is. Thus, whenever we slice the bivariate normal distribution parallel to the Y_1 axis, we obtain (after proper scaling) a normal conditional probability distribution.

2. The means of the conditional probability distributions of Y_1 fall on a straight line, and hence are a linear function of Y_2:

$$E\{Y_1|Y_2\} = \alpha_{1|2} + \beta_{12}Y_2 \tag{2.83}$$

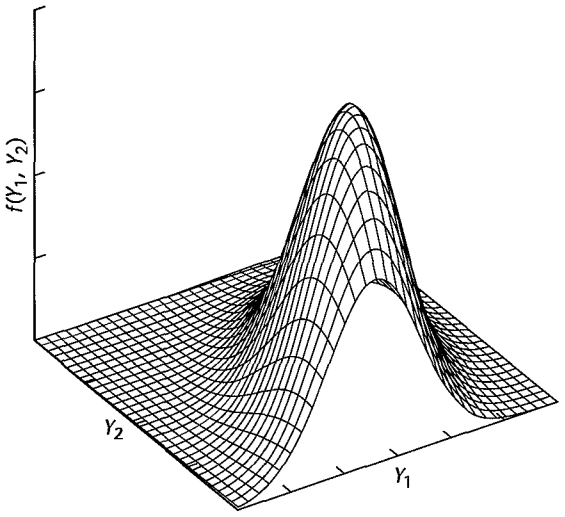


FIGURE 2.11
Cross Section of Bivariate Normal Distribution at Y_{h2}.

Here, $\alpha_{1|2}$ is the intercept parameter and β_{12} the slope parameter. Thus, the relation between the conditional means and Y_2 is given by a linear regression function.

3. All conditional probability distributions of Y_1 have the same standard deviation $\sigma_{1|2}$. Thus, no matter where we slice the bivariate normal distribution parallel to the Y_1 axis, the resulting conditional probability distribution (after scaling to have an area of 1) has the same standard deviation. Hence, constant variances characterize the conditional probability distributions of Y_1.

Equivalence to Normal Error Regression Model. Suppose that we select a random sample of observations (Y_1, Y_2) from a bivariate normal population and wish to make conditional inferences about Y_1, given Y_2. The preceding discussion makes it clear that the normal error regression model (1.24) is entirely applicable because:

1. The Y_1 observations are independent.
2. The Y_1 observations when Y_2 is considered given or fixed are normally distributed with mean $E\{Y_1|Y_2\} = \alpha_{1|2} + \beta_{12}Y_2$ and constant variance $\sigma^2_{1|2}$.

Use of Regression Analysis. In view of the equivalence of each of the conditional bivariate normal correlation models (2.81) and (2.79) with the normal error regression model (1.24), all conditional inferences with these correlation models can be made by means of the usual regression methods. For instance, if a researcher has data that can be appropriately described as having been generated from a bivariate normal distribution and wishes to make inferences about Y_2, given a particular value of Y_1, the ordinary regression techniques will be applicable. Thus, the regression function of Y_2 on Y_1 can be estimated by means of (1.12), the slope of the regression line can be estimated by means of the interval estimate (2.15), a new observation Y_2, given the value of Y_1, can be predicted by means of (2.36), and so on. Computer regression packages can be used in the usual manner. To avoid notational problems, it may be helpful to relabel the variables according to regression usage: $Y = Y_2$, $X = Y_1$. Of course, if conditional inferences on Y_1 for given values of Y_2 are desired, the notation correspondences would be: $Y = Y_1$, $X = Y_2$.

Can we still use regression model (2.1) if Y_1 and Y_2 are not bivariate normal? It can be shown that all results on estimation, testing, and prediction obtained from regression model (2.1) apply if $Y_1 = Y$ and $Y_2 = X$ are random variables, and if the following conditions hold:

1. The conditional distributions of the Y_i, given X_i, are normal and independent, with conditional means $\beta_0 + \beta_1 X_i$ and conditional variance σ^2.
2. The X_i are independent random variables whose probability distribution $g(X_i)$ does not involve the parameters β_0, β_1, σ^2.

These conditions require only that regression model (2.1) is appropriate for each *conditional* distribution of Y_i, and that the probability distribution of the X_i does not involve the regression parameters. If these conditions are met, all earlier results on estimation, testing, and prediction still hold even though the X_i are now random variables. The major modification occurs in the interpretation of confidence coefficients and specified risks of error. When X is random, these refer to repeated sampling of pairs of (X_i, Y_i) values, where the X_i values as well as the Y_i values change from sample to sample. Thus, in our bathing suit sales illustration, a confidence coefficient would refer to the proportion of correct interval estimates if repeated samples of n days' sales and temperatures were obtained and the confidence interval calculated for each sample. Another modification occurs in the test's power, which is different when X is a random variable.

Comments

1. The notation for the parameters of the conditional correlation models departs somewhat from our previous notation for regression models. The symbol α is now used to denote the regression intercept. The subscript $1|2$ to α indicates that Y_1 is regressed on Y_2. Similarly, the subscript $2|1$ to α indicates that Y_2 is regressed on Y_1. The symbol β_{12} indicates that it is the slope in the regression of Y_1 on Y_2, while β_{21} is the slope in the regression of Y_2 on Y_1. Finally, $\sigma_{2|1}$ is the standard deviation of the conditional probability distributions of Y_2 for any given Y_1, while $\sigma_{1|2}$ is the standard deviation of the conditional probability distributions of Y_1 for any given Y_2.

2. Two distinct regressions are involved in a bivariate normal model, that of Y_1 on Y_2 when Y_2 is fixed and that of Y_2 on Y_1 when Y_1 is fixed. In general, the two regression lines are not the same. For instance, the two slopes β_{12} and β_{21} are the same only if $\sigma_1 = \sigma_2$, as can be seen from (2.80b) and (2.82b).

3. When interval estimates for the conditional correlation models are obtained, the confidence coefficient refers to repeated samples where pairs of observations (Y_1, Y_2) are obtained from the bivariate normal distribution. ■

Inferences on Correlation Coefficients

A principal use of the bivariate normal correlation model is to study the relationship between two variables. In a bivariate normal model, the parameter ρ_{12} provides information about the degree of the linear relationship between the two variables Y_1 and Y_2.

Point Estimator of ρ_{12}. The maximum likelihood estimator of ρ_{12}, denoted by r_{12}, is given by:

$$r_{12} = \frac{\sum(Y_{i1} - \bar{Y}_1)(Y_{i2} - \bar{Y}_2)}{\left[\sum(Y_{i1} - \bar{Y}_1)^2 \sum(Y_{i2} - \bar{Y}_2)^2\right]^{1/2}} \tag{2.84}$$

This estimator is often called the *Pearson product-moment correlation coefficient.* It is a biased estimator of ρ_{12} (unless $\rho_{12} = 0$ or 1), but the bias is small when n is large.

It can be shown that the range of r_{12} is:

$$-1 \leq r_{12} \leq 1 \tag{2.85}$$

Generally, values of r_{12} near 1 indicate a strong positive (direct) linear association between Y_1 and Y_2 whereas values of r_{12} near -1 indicate a strong negative (indirect) linear association. Values of r_{12} near 0 indicate little or no linear association between Y_1 and Y_2.

Test whether $\rho_{12} = 0$. When the population is bivariate normal, it is frequently desired to test whether the coefficient of correlation is zero:

$$\begin{aligned} H_0&: \rho_{12} = 0 \\ H_a&: \rho_{12} \neq 0 \end{aligned} \tag{2.86}$$

The reason for interest in this test is that in the case where Y_1 and Y_2 are jointly normally distributed, $\rho_{12} = 0$ implies that Y_1 and Y_2 are independent.

We can use regression procedures for the test since (2.80b) implies that the following alternatives are equivalent to those in (2.86):

$$\begin{aligned} H_0&: \beta_{12} = 0 \\ H_a&: \beta_{12} \neq 0 \end{aligned} \tag{2.86a}$$

and (2.82b) implies that the following alternatives are also equivalent to the ones in (2.86):

$$\begin{aligned} H_0&: \beta_{21} = 0 \\ H_a&: \beta_{21} \neq 0 \end{aligned} \tag{2.86b}$$

It can be shown that the test statistics for testing either (2.86a) or (2.86b) are the same and can be expressed directly in terms of r_{12}:

$$t^* = \frac{r_{12}\sqrt{n-2}}{\sqrt{1-r_{12}^2}} \tag{2.87}$$

If H_0 holds, t^* follows the $t(n-2)$ distribution. The appropriate decision rule to control the Type I error at α is:

$$\begin{aligned} &\text{If } |t^*| \leq t(1-\alpha/2; n-2), \text{ conclude } H_0 \\ &\text{If } |t^*| > t(1-\alpha/2; n-2), \text{ conclude } H_a \end{aligned} \tag{2.88}$$

Test statistic (2.87) is identical to the regression t^* test statistic (2.17).

Example

A national oil company was interested in the relationship between its service station gasoline sales and its sales of auxiliary products. A company analyst obtained a random sample of 23 of its service stations and obtained average monthly sales data on gasoline sales (Y_1) and comparable sales of its auxiliary products and services (Y_2). These data (not shown) resulted in an estimated correlation coefficient $r_{12} = .52$. Suppose the analyst wished to test whether or not the association was positive, controlling the level of significance at $\alpha = .05$. The alternatives would then be:

$$\begin{aligned} H_0&: \rho_{12} \leq 0 \\ H_a&: \rho_{12} > 0 \end{aligned}$$

and the decision rule based on test statistic (2.87) would be:

$$\text{If } t^* \leq t(1-\alpha; n-2), \text{ conclude } H_0$$
$$\text{If } t^* > t(1-\alpha; n-2), \text{ conclude } H_a$$

For $\alpha = .05$, we require $t(.95; 21) = 1.721$. Since:

$$t^* = \frac{.52\sqrt{21}}{\sqrt{1-(.52)^2}} = 2.79$$

is greater than 1.721, we would conclude H_a, that $\rho_{12} > 0$. The P-value for this test is .006.

Interval Estimation of ρ_{12} Using the z' Transformation. Because the sampling distribution of r_{12} is complicated when $\rho_{12} \neq 0$, interval estimation of ρ_{12} is usually carried out by means of an approximate procedure based on a transformation. This transformation, known as the *Fisher z transformation,* is as follows:

$$z' = \frac{1}{2}\log_e\left(\frac{1+r_{12}}{1-r_{12}}\right) \tag{2.89}$$

When n is large (25 or more is a useful rule of thumb), the distribution of z' is approximately normal with approximate mean and variance:

$$E\{z'\} = \zeta = \frac{1}{2}\log_e\left(\frac{1+\rho_{12}}{1-\rho_{12}}\right) \tag{2.90}$$

$$\sigma^2\{z'\} = \frac{1}{n-3} \tag{2.91}$$

Note that the transformation from r_{12} to z' in (2.89) is the same as the relation in (2.90) between ρ_{12} and $E\{z'\} = \zeta$. Also note that the approximate variance of z' is a known constant, depending only on the sample size n.

Table B.8 gives paired values for the left and right sides of (2.89) and (2.90), thus eliminating the need for calculations. For instance, if r_{12} or ρ_{12} equals .25, Table B.8 indicates that z' or ζ equals .2554, and vice versa. The values on the two sides of the transformation always have the same sign. Thus, if r_{12} or ρ_{12} is negative, a minus sign is attached to the value in Table B.8. For instance, if $r_{12} = -.25$, $z' = -.2554$.

Interval Estimate. When the sample size is large ($n \geq 25$), the standardized statistic:

$$\frac{z'-\zeta}{\sigma\{z'\}} \tag{2.92}$$

is approximately a standard normal variable. Therefore, approximate $1-\alpha$ confidence limits for ζ are:

$$z' \pm z(1-\alpha/2)\sigma\{z'\} \tag{2.93}$$

where $z(1-\alpha/2)$ is the $(1-\alpha/2)100$ percentile of the standard normal distribution. The $1-\alpha$ confidence limits for ρ_{12} are then obtained by transforming the limits on ζ by means of (2.90).

Example

An economist investigated food purchasing patterns by households in a midwestern city. Two hundred households with family incomes between $40,000 and $60,000 were selected to ascertain, among other things, the proportions of the food budget expended for beef and poultry, respectively. The economist expected these to be negatively related, and wished to estimate the coefficient of correlation with a 95 percent confidence interval. Some supporting evidence suggested that the joint distribution of the two variables does not depart markedly from a bivariate normal one.

The point estimate of ρ_{12} was $r_{12} = -.61$ (data and calculations not shown). To obtain an approximate 95 percent confidence interval estimate, we require:

$$z' = -.7089 \quad \text{when } r_{12} = -.61 \qquad \text{(from Table B.8)}$$

$$\sigma\{z'\} = \frac{1}{\sqrt{200-3}} = .07125$$

$$z(.975) = 1.960$$

Hence, the confidence limits for ζ, by (2.93), are $-.7089 \pm 1.960(.07125)$, and the approximate 95 percent confidence interval is:

$$-.849 \leq \zeta \leq -.569$$

Using Table B.8 to transform back to ρ_{12}, we obtain:

$$-.69 \leq \rho_{12} \leq -.51$$

This confidence interval was sufficiently precise to be useful to the economist, confirming the negative relation and indicating that the degree of linear association is moderately high.

Comments

1. As usual, a confidence interval for ρ_{12} can be employed to test whether or not ρ_{12} has a specified value—say, .5—by noting whether or not the specified value falls within the confidence limits.

2. It can be shown that the square of the coefficient of correlation, namely ρ_{12}^2, measures the relative reduction in the variability of Y_2 associated with the use of variable Y_1. To see this, we noted earlier in (2.80c) and (2.82c) that:

$$\sigma_{1|2}^2 = \sigma_1^2(1 - \rho_{12}^2) \tag{2.94a}$$

$$\sigma_{2|1}^2 = \sigma_2^2(1 - \rho_{12}^2) \tag{2.94b}$$

We can rewrite these expressions as follows:

$$\rho_{12}^2 = \frac{\sigma_1^2 - \sigma_{1|2}^2}{\sigma_1^2} \tag{2.95a}$$

$$\rho_{12}^2 = \frac{\sigma_2^2 - \sigma_{2|1}^2}{\sigma_2^2} \tag{2.95b}$$

The meaning of ρ_{12}^2 is now clear. Consider first (2.95a). ρ_{12}^2 measures how much smaller relatively is the variability in the conditional distributions of Y_1, for any given level of Y_2, than is the variability in the marginal distribution of Y_1. Thus, ρ_{12}^2 measures the relative reduction in the variability of Y_1 associated with the use of variable Y_2. Correspondingly, (2.95b) shows that ρ_{12}^2 also measures the relative reduction in the variability of Y_2 associated with the use of variable Y_1.

It can be shown that:

$$0 \leq \rho_{12}^2 \leq 1 \tag{2.96}$$

The limiting value $\rho_{12}^2 = 0$ occurs when Y_1 and Y_2 are independent, so that the variances of each variable in the conditional probability distributions are then no smaller than the variance in the marginal distribution. The limiting value $\rho_{12}^2 = 1$ occurs when there is no variability in the conditional probability distributions for each variable, so perfect predictions of either variable can be made from the other.

3. The interpretation of ρ_{12}^2 as measuring the relative reduction in the conditional variances as compared with the marginal variance is valid for the case of a bivariate normal population, but not for many other bivariate populations. Of course, the interpretation implies nothing in a causal sense.

4. Confidence limits for ρ_{12}^2 can be obtained by squaring the respective confidence limits for ρ_{12}, provided the latter limits do not differ in sign. ■

Spearman Rank Correlation Coefficient

At times the joint distribution of two random variables Y_1 and Y_2 differs considerably from the bivariate normal distribution (2.74). In those cases, transformations of the variables Y_1 and Y_2 may be sought to make the joint distribution of the transformed variables approximately bivariate normal and thus permit the use of the inference procedures about ρ_{12} described earlier.

When no appropriate transformations can be found, a nonparametric *rank correlation* procedure may be useful for making inferences about the association between Y_1 and Y_2. The *Spearman rank correlation coefficient* is widely used for this purpose. First, the observations on Y_1 (i.e., $Y_{11}, \ldots, Y_{n1}$) are expressed in ranks from 1 to n. We denote the rank of Y_{i1} by R_{i1}. Similarly, the observations on Y_2 (i.e., $Y_{12}, \ldots, Y_{n2}$) are ranked, with the rank of Y_{i2} denoted by R_{i2}. The Spearman rank correlation coefficient, to be denoted by r_S, is then defined as the ordinary Pearson product-moment correlation coefficient in (2.84) based on the rank data:

$$r_S = \frac{\sum(R_{i1} - \bar{R}_1)(R_{i2} - \bar{R}_2)}{\left[\sum(R_{i1} - \bar{R}_1)^2 \sum(R_{i2} - \bar{R}_2)^2\right]^{1/2}} \tag{2.97}$$

Here $\bar{R}_1$ is the mean of the ranks R_{i1} and $\bar{R}_2$ is the mean of the ranks R_{i2}. Of course, since the ranks R_{i1} and R_{i2} are the integers $1, \ldots, n$, it follows that $\bar{R}_1 = \bar{R}_2 = (n+1)/2$.

Like an ordinary correlation coefficient, the Spearman rank correlation coefficient takes on values between -1 and 1 inclusive:

$$-1 \leq r_S \leq 1 \tag{2.98}$$

The coefficient r_S equals 1 when the ranks for Y_1 are identical to those for Y_2, that is, when the case with rank 1 for Y_1 also has rank 1 for Y_2, and so on. In that case, there is perfect association between the ranks for the two variables. The coefficient r_S equals -1 when the case with rank 1 for Y_1 has rank n for Y_2, the case with rank 2 for Y_1 has rank $n-1$ for Y_2, and so on. In that event, there is perfect inverse association between the ranks for the two variables. When there is little, if any, association between the ranks of Y_1 and Y_2, the Spearman rank correlation coefficient tends to have a value near zero.

The Spearman rank correlation coefficient can be used to test the alternatives:

$$H_0: \text{There is no association between } Y_1 \text{ and } Y_2$$
$$H_a: \text{There is an association between } Y_1 \text{ and } Y_2 \quad (2.99)$$

A two-sided test is conducted here since H_a includes either positive or negative association. When the alternative H_a is:

$$H_a: \text{There is positive (negative) association between } Y_1 \text{ and } Y_2 \quad (2.100)$$

an upper-tail (lower-tail) one-sided test is conducted.

The probability distribution of r_S under H_0 is not difficult to obtain. It is based on the condition that, for any ranking of Y_1, all rankings of Y_2 are equally likely when there is no association between Y_1 and Y_2. Tables have been prepared and are presented in specialized texts such as Reference 2.1. Computer packages generally do not present the probability distribution of r_S under H_0 but give only the two-sided P-value. When the sample size n exceeds 10, the test can be carried out approximately by using test statistic (2.87):

$$t^* = \frac{r_S\sqrt{n-2}}{\sqrt{1-r_S^2}} \quad (2.101)$$

based on the t distribution with $n-2$ degrees of freedom.

Example

A market researcher wished to examine whether an association exists between population size (Y_1) and per capita expenditures for a new food product (Y_2). The data for a random sample of 12 test markets are given in Table 2.4, columns 1 and 2. Because the distributions of the variables do not appear to be approximately normal, a nonparametric test of association is desired. The ranks for the variables are given in Table 2.4, columns 3 and 4. A computer package found that the coefficient of simple correlation between the ranked data in columns 3 and 4 is $r_S = .895$. The alternatives of interest are the two-sided ones in (2.99). Since n

TABLE 2.4 Data on Population and Expenditures and Their Ranks—Sales Marketing Example.

	(1)	(2)	(3)	(4)
Test Market i	Population (in thousands) Y_{i1}	Per Capita Expenditure (dollars) Y_{i2}	R_{i1}	R_{i2}
1	29	127	1	2
2	435	214	8	11
3	86	133	3	4
4	1,090	208	11	10
5	219	153	7	6
6	503	184	9	8
7	47	130	2	3
8	3,524	217	12	12
9	185	141	6	5
10	98	154	5	7
11	952	194	10	9
12	89	103	4	1

exceeds 10 here, we use test statistic (2.101):

$$t^* = \frac{.895\sqrt{12-2}}{\sqrt{1-(.895)^2}} = 6.34$$

For $\alpha = .01$, we require $t(.995; 10) = 3.169$. Since $|t^*| = 6.34 > 3.169$, we conclude H_a, that there is an association between population size and per capita expenditures for the food product. The two-sided P-value of the test is .00008.

Comments

1. In case of ties among some data values, each of the tied values is given the average of the ranks involved.

2. It is interesting to note that had the data in Table 2.4 been analyzed by assuming the bivariate normal distribution assumption (2.74) and test statistic (2.87), then the strength of the association would have been somewhat weaker. In particular, the Pearson product-moment correlation coefficient is $r_{12} = .674$, with $t^* = .674\sqrt{10}/\sqrt{1-(.674)^2} = 2.885$. Our conclusion would have been to conclude H_0, that there is no association between population size and per capita expenditures for the food product. The two-sided P-value of the test is .016.

3. Another nonparametric rank procedure similar to Spearman's r_S is *Kendall's* τ. This statistic also measures how far the rankings of Y_1 and Y_2 differ from each other, but in a somewhat different way than the Spearman rank correlation coefficient. A discussion of Kendall's τ may be found in Reference 2.2. ■

Cited References

2.1. Gibbons, J. D. *Nonparametric Methods for Quantitative Analysis*. 2nd ed. Columbus, Ohio: American Sciences Press, 1985.

2.2. Kendall, M. G., and J. D. Gibbons. *Rank Correlation Methods*. 5th ed. London: Oxford University Press, 1990.

Problems

2.1. A student working on a summer internship in the economic research department of a large corporation studied the relation between sales of a product (Y, in million dollars) and population (X, in million persons) in the firm's 50 marketing districts. The normal error regression model (2.1) was employed. The student first wished to test whether or not a linear association between Y and X existed. The student accessed a simple linear regression program and obtained the following information on the regression coefficients:

Parameter	Estimated Value	95 Percent Confidence Limits	
Intercept	7.43119	−1.18518	16.0476
Slope	.755048	.452886	1.05721

a. The student concluded from these results that there is a linear association between Y and X. Is the conclusion warranted? What is the implied level of significance?

b. Someone questioned the negative lower confidence limit for the intercept, pointing out that dollar sales cannot be negative even if the population in a district is zero. Discuss.

2.2. In a test of the alternatives H_0: $\beta_1 \leq 0$ versus H_a: $\beta_1 > 0$, an analyst concluded H_0. Does this conclusion imply that there is no linear association between X and Y? Explain.

2.3. A member of a student team playing an interactive marketing game received the following computer output when studying the relation between advertising expenditures (X) and sales (Y) for one of the team's products:

Estimated regression equation: $\hat{Y} = 350.7 - .18X$

Two-sided P-value for estimated slope: .91

The student stated: "The message I get here is that the more we spend on advertising this product, the fewer units we sell!" Comment.

2.4. Refer to **Grade point average** Problem 1.19.

a. Obtain a 99 percent confidence interval for β_1. Interpret your confidence interval. Does it include zero? Why might the director of admissions be interested in whether the confidence interval includes zero?

b. Test, using the test statistic t^*, whether or not a linear association exists between student's ACT score (X) and GPA at the end of the freshman year (Y). Use a level of significance of .01. State the alternatives, decision rule, and conclusion.

c. What is the P-value of your test in part (b)? How does it support the conclusion reached in part (b)?

*2.5. Refer to **Copier maintenance** Problem 1.20.

a. Estimate the change in the mean service time when the number of copiers serviced increases by one. Use a 90 percent confidence interval. Interpret your confidence interval.

b. Conduct a t test to determine whether or not there is a linear association between X and Y here; control the α risk at .10. State the alternatives, decision rule, and conclusion. What is the P-value of your test?

c. Are your results in parts (a) and (b) consistent? Explain.

d. The manufacturer has suggested that the mean required time should not increase by more than 14 minutes for each additional copier that is serviced on a service call. Conduct a test to decide whether this standard is being satisfied by Tri-City. Control the risk of a Type I error at .05. State the alternatives, decision rule, and conclusion. What is the P-value of the test?

e. Does b_0 give any relevant information here about the "start-up" time on calls—i.e., about the time required before service work is begun on the copiers at a customer location?

*2.6. Refer to **Airfreight breakage** Problem 1.21.

a. Estimate β_1 with a 95 percent confidence interval. Interpret your interval estimate.

b. Conduct a t test to decide whether or not there is a linear association between number of times a carton is transferred (X) and number of broken ampules (Y). Use a level of significance of .05. State the alternatives, decision rule, and conclusion. What is the P-value of the test?

c. β_0 represents here the mean number of ampules broken when no transfers of the shipment are made—i.e., when $X = 0$. Obtain a 95 percent confidence interval for β_0 and interpret it.

d. A consultant has suggested, on the basis of previous experience, that the mean number of broken ampules should not exceed 9.0 when no transfers are made. Conduct an appropriate test, using $\alpha = .025$. State the alternatives, decision rule, and conclusion. What is the P-value of the test?

e. Obtain the power of your test in part (b) if actually $\beta_1 = 2.0$. Assume $\sigma\{b_1\} = .50$. Also obtain the power of your test in part (d) if actually $\beta_0 = 11$. Assume $\sigma\{b_0\} = .75$.

2.7 Refer to **Plastic hardness** Problem 1.22.

a. Estimate the change in the mean hardness when the elapsed time increases by one hour. Use a 99 percent confidence interval. Interpret your interval estimate.

b. The plastic manufacturer has stated that the mean hardness should increase by 2 Brinell units per hour. Conduct a two-sided test to decide whether this standard is being satisfied; use $\alpha = .01$. State the alternatives, decision rule, and conclusion. What is the P-value of the test?

c. Obtain the power of your test in part (b) if the standard actually is being exceeded by .3 Brinell units per hour. Assume $\sigma\{b_1\} = .1$.

2.8. Refer to Figure 2.2 for the Toluca Company example. A consultant has advised that an increase of one unit in lot size should require an increase of 3.0 in the expected number of work hours for the given production item.

a. Conduct a test to decide whether or not the increase in the expected number of work hours in the Toluca Company equals this standard. Use $\alpha = .05$. State the alternatives, decision rule, and conclusion.

b. Obtain the power of your test in part (a) if the consultant's standard actually is being exceeded by .5 hour. Assume $\sigma\{b_1\} = .35$.

c. Why is $F^* = 105.88$, given in the printout, not relevant for the test in part (a)?

2.9. Refer to Figure 2.2. A student, noting that $s\{b_1\}$ is furnished in the printout, asks why $s\{\hat{Y}_h\}$ is not also given. Discuss.

2.10. For each of the following questions, explain whether a confidence interval for a mean response or a prediction interval for a new observation is appropriate.

a. What will be the humidity level in this greenhouse tomorrow when we set the temperature level at 31°C?

b. How much do families whose disposable income is $23,500 spend, on the average, for meals away from home?

c. How many kilowatt-hours of electricity will be consumed next month by commercial and industrial users in the Twin Cities service area, given that the index of business activity for the area remains at its present level?

2.11. A person asks if there is a difference between the "mean response at $X = X_h$" and the "mean of m new observations at $X = X_h$." Reply.

2.12. Can $\sigma^2\{\text{pred}\}$ in (2.37) be brought increasingly close to 0 as n becomes large? Is this also the case for $\sigma^2\{\hat{Y}_h\}$ in (2.29b)? What is the implication of this difference?

2.13. Refer to **Grade point average** Problem 1.19.

a. Obtain a 95 percent interval estimate of the mean freshman GPA for students whose ACT test score is 28. Interpret your confidence interval.

b. Mary Jones obtained a score of 28 on the entrance test. Predict her freshman GPA using a 95 percent prediction interval. Interpret your prediction interval.

c. Is the prediction interval in part (b) wider than the confidence interval in part (a)? Should it be?

d. Determine the boundary values of the 95 percent confidence band for the regression line when $X_h = 28$. Is your confidence band wider at this point than the confidence interval in part (a)? Should it be?

*2.14. Refer to **Copier maintenance** Problem 1.20.

a. Obtain a 90 percent confidence interval for the mean service time on calls in which six copiers are serviced. Interpret your confidence interval.

b. Obtain a 90 percent prediction interval for the service time on the next call in which six copiers are serviced. Is your prediction interval wider than the corresponding confidence interval in part (a)? Should it be?

c. Management wishes to estimate the expected service time *per copier* on calls in which six copiers are serviced. Obtain an appropriate 90 percent confidence interval by converting the interval obtained in part (a). Interpret the converted confidence interval.

d. Determine the boundary values of the 90 percent confidence band for the regression line when $X_h = 6$. Is your confidence band wider at this point than the confidence interval in part (a)? Should it be?

*2.15. Refer to **Airfreight breakage** Problem 1.21.

a. Because of changes in airline routes, shipments may have to be transferred more frequently than in the past. Estimate the mean breakage for the following numbers of transfers: $X = 2$, 4. Use separate 99 percent confidence intervals. Interpret your results.

b. The next shipment will entail two transfers. Obtain a 99 percent prediction interval for the number of broken ampules for this shipment. Interpret your prediction interval.

c. In the next several days, three independent shipments will be made, each entailing two transfers. Obtain a 99 percent prediction interval for the mean number of ampules broken in the three shipments. Convert this interval into a 99 percent prediction interval for the total number of ampules broken in the three shipments.

d. Determine the boundary values of the 99 percent confidence band for the regression line when $X_h = 2$ and when $X_h = 4$. Is your confidence band wider at these two points than the corresponding confidence intervals in part (a)? Should it be?

2.16. Refer to **Plastic hardness** Problem 1.22.

a. Obtain a 98 percent confidence interval for the mean hardness of molded items with an elapsed time of 30 hours. Interpret your confidence interval.

b. Obtain a 98 percent prediction interval for the hardness of a newly molded test item with an elapsed time of 30 hours.

c. Obtain a 98 percent prediction interval for the mean hardness of 10 newly molded test items, each with an elapsed time of 30 hours.

d. Is the prediction interval in part (c) narrower than the one in part (b)? Should it be?

e. Determine the boundary values of the 98 percent confidence band for the regression line when $X_h = 30$. Is your confidence band wider at this point than the confidence interval in part (a)? Should it be?

2.17. An analyst fitted normal error regression model (2.1) and conducted an F test of $\beta_1 = 0$ versus $\beta_1 \neq 0$. The P-value of the test was .033, and the analyst concluded H_a: $\beta_1 \neq 0$. Was the α level used by the analyst greater than or smaller than .033? If the α level had been .01, what would have been the appropriate conclusion?

2.18. For conducting statistical tests concerning the parameter β_1, why is the t test more versatile than the F test?

2.19. When testing whether or not $\beta_1 = 0$, why is the F test a one-sided test even though H_a includes both $\beta_1 < 0$ and $\beta_1 > 0$? [*Hint*: Refer to (2.57).]

2.20. A student asks whether R^2 is a point estimator of any parameter in the normal error regression model (2.1). Respond.

2.21. A value of R^2 near 1 is sometimes interpreted to imply that the relation between Y and X is sufficiently close so that suitably precise predictions of Y can be made from knowledge of X. Is this implication a necessary consequence of the definition of R^2?

2.22. Using the normal error regression model (2.1) in an engineering safety experiment, a researcher found for the first 10 cases that R^2 was zero. Is it possible that for the complete set of 30 cases R^2 will not be zero? Could R^2 not be zero for the first 10 cases, yet equal zero for all 30 cases? Explain.

2.23. Refer to **Grade point average** Problem 1.19.

a. Set up the ANOVA table.

b. What is estimated by *MSR* in your ANOVA table? by *MSE*? Under what condition do *MSR* and *MSE* estimate the same quantity?

c. Conduct an F test of whether or not $\beta_1 = 0$. Control the α risk at .01. State the alternatives, decision rule, and conclusion.

d. What is the absolute magnitude of the reduction in the variation of Y when X is introduced into the regression model? What is the relative reduction? What is the name of the latter measure?

e. Obtain r and attach the appropriate sign.

f. Which measure, R^2 or r, has the more clear-cut operational interpretation? Explain.

*2.24. Refer to **Copier maintenance** Problem 1.20.

a. Set up the basic ANOVA table in the format of Table 2.2. Which elements of your table are additive? Also set up the ANOVA table in the format of Table 2.3. How do the two tables differ?

b. Conduct an F test to determine whether or not there is a linear association between time spent and number of copiers serviced; use $\alpha = .10$. State the alternatives, decision rule, and conclusion.

c. By how much, relatively, is the total variation in number of minutes spent on a call reduced when the number of copiers serviced is introduced into the analysis? Is this a relatively small or large reduction? What is the name of this measure?

d. Calculate r and attach the appropriate sign.

e. Which measure, r or R^2, has the more clear-cut operational interpretation?

*2.25. Refer to **Airfreight breakage** Problem 1.21.

a. Set up the ANOVA table. Which elements are additive?

b. Conduct an F test to decide whether or not there is a linear association between the number of times a carton is transferred and the number of broken ampules; control the α risk at .05. State the alternatives, decision rule, and conclusion.

c. Obtain the t^* statistic for the test in part (b) and demonstrate numerically its equivalence to the F^* statistic obtained in part (b).

d. Calculate R^2 and r. What proportion of the variation in Y is accounted for by introducing X into the regression model?

2.26. Refer to **Plastic hardness** Problem 1.22.

a. Set up the ANOVA table.

b. Test by means of an F test whether or not there is a linear association between the hardness of the plastic and the elapsed time. Use $\alpha = .01$. State the alternatives, decision rule, and conclusion.

c. Plot the deviations $Y_i - \hat{Y}_i$ against X_i on a graph. Plot the deviations $\hat{Y}_i - \bar{Y}$ against X_i on another graph, using the same scales as for the first graph. From your two graphs, does *SSE* or *SSR* appear to be the larger component of *SSTO*? What does this imply about the magnitude of R^2?

d. Calculate R^2 and r.

*2.27. Refer to **Muscle mass** Problem 1.27.

a. Conduct a test to decide whether or not there is a negative linear association between amount of muscle mass and age. Control the risk of Type I error at .05. State the alternatives, decision rule, and conclusion. What is the P-value of the test?

b. The two-sided P-value for the test whether $\beta_0 = 0$ is 0+. Can it now be concluded that b_0 provides relevant information on the amount of muscle mass at birth for a female child?

c. Estimate with a 95 percent confidence interval the difference in expected muscle mass for women whose ages differ by one year. Why is it not necessary to know the specific ages to make this estimate?

*2.28. Refer to **Muscle mass** Problem 1.27.

a. Obtain a 95 percent confidence interval for the mean muscle mass for women of age 60. Interpret your confidence interval.

b. Obtain a 95 percent prediction interval for the muscle mass of a woman whose age is 60. Is the prediction interval relatively precise?

c. Determine the boundary values of the 95 percent confidence band for the regression line when $X_h = 60$. Is your confidence band wider at this point than the confidence interval in part (a)? Should it be?

*2.29. Refer to **Muscle mass** Problem 1.27.

a. Plot the deviations $Y_i - \hat{Y}_i$ against X_i on one graph. Plot the deviations $\hat{Y}_i - \bar{Y}$ against X_i on another graph, using the same scales as in the first graph. From your two graphs, does *SSE* or *SSR* appear to be the larger component of *SSTO*? What does this imply about the magnitude of R^2?

b. Set up the ANOVA table.

c. Test whether or not $\beta_1 = 0$ using an F test with $\alpha = .05$. State the alternatives, decision rule, and conclusion.

d. What proportion of the total variation in muscle mass remains "unexplained" when age is introduced into the analysis? Is this proportion relatively small or large?

e. Obtain R^2 and r.

2.30. Refer to **Crime rate** Problem 1.28.

a. Test whether or not there is a linear association between crime rate and percentage of high school graduates, using a t test with $\alpha = .01$. State the alternatives, decision rule, and conclusion. What is the P-value of the test?

b. Estimate β_1 with a 99 percent confidence interval. Interpret your interval estimate.

2.31. Refer to **Crime rate** Problem 1.28

a. Set up the ANOVA table.

b. Carry out the test in Problem 2.30a by means of the F test. Show the numerical equivalence of the two test statistics and decision rules. Is the P-value for the F test the same as that for the t test?

c. By how much is the total variation in crime rate reduced when percentage of high school graduates is introduced into the analysis? Is this a relatively large or small reduction?

d. Obtain r.

2.32. Refer to **Crime rate** Problems 1.28 and 2.30. Suppose that the test in Problem 2.30a is to be carried out by means of a general linear test.

a. State the full and reduced models.

b. Obtain (1) $SSE(F)$, (2) $SSE(R)$, (3) df_F, (4) df_R, (5) test statistic F^* for the general linear test, (6) decision rule.

c. Are the test statistic F^* and the decision rule for the general linear test numerically equivalent to those in Problem 2.30a?

2.33. In developing empirically a cost function from observed data on a complex chemical experiment, an analyst employed normal error regression model (2.1). β_0 was interpreted here as the cost of setting up the experiment. The analyst hypothesized that this cost should be \$7.5 thousand and wished to test the hypothesis by means of a general linear test.

a. Indicate the alternative conclusions for the test.

b. Specify the full and reduced models.

c. Without additional information, can you tell what the quantity $df_R - df_F$ in test statistic (2.70) will equal in the analyst's test? Explain.

2.34. Refer to **Grade point average** Problem 1.19.

a. Would it be more reasonable to consider the X_i as known constants or as random variables here? Explain.

b. If the X_i were considered to be random variables, would this have any effect on prediction intervals for new applicants? Explain.

2.35. Refer to **Copier maintenance** Problems 1.20 and 2.5. How would the meaning of the confidence coefficient in Problem 2.5a change if the predictor variable were considered a random variable and the conditions on page 83 were applicable?

2.36. A management trainee in a production department wished to study the relation between weight of rough casting and machining time to produce the finished block. The trainee selected castings so that the weights would be spaced equally apart in the sample and then observed the corresponding machining times. Would you recommend that a regression or a correlation model be used? Explain.

2.37. A social scientist stated: "The conditions for the bivariate normal distribution are so rarely met in my experience that I feel much safer using a regression model." Comment.

2.38. A student was investigating from a large sample whether variables Y_1 and Y_2 follow a bivariate normal distribution. The student obtained the residuals when regressing Y_1 on Y_2, and also obtained the residuals when regressing Y_2 on Y_1, and then prepared a normal probability plot for each set of residuals. Do these two normal probability plots provide sufficient information for determining whether the two variables follow a bivariate normal distribution? Explain.

2.39. For the bivariate normal distribution with parameters $\mu_1 = 50$, $\mu_2 = 100$, $\sigma_1 = 3$, $\sigma_2 = 4$, and $\rho_{12} = .80$.

a. State the characteristics of the marginal distribution of Y_1.

b. State the characteristics of the conditional distribution of Y_2 when $Y_1 = 55$.

c. State the characteristics of the conditional distribution of Y_1 when $Y_2 = 95$.

2.40. Explain whether any of the following would be affected if the bivariate normal model (2.74) were employed instead of the normal error regression model (2.1) with fixed levels of the predictor variable: (1) point estimates of the regression coefficients, (2) confidence limits for the regression coefficients, (3) interpretation of the confidence coefficient.

2.41. Refer to **Plastic hardness** Problem 1.22. A student was analyzing these data and received the following standard query from the interactive regression and correlation computer package: CALCULATE CONFIDENCE INTERVAL FOR POPULATION CORRELATION COEFFICIENT RHO? ANSWER Y OR N. Would a "yes" response lead to meaningful information here? Explain.

*2.42. **Property assessments.** The data that follow show assessed value for property tax purposes (Y_1, in thousand dollars) and sales price (Y_2, in thousand dollars) for a sample of 15 parcels of land for industrial development sold recently in "arm's length" transactions in a tax district. Assume that bivariate normal model (2.74) is appropriate here.

i:	1	2	3	...	13	14	15
Y_{i1}:	13.9	16.0	10.3	...	14.9	12.9	15.8
Y_{i2}:	28.6	34.7	21.0	...	35.1	30.0	36.2

a. Plot the data in a scatter diagram. Does the bivariate normal model appear to be appropriate here? Discuss.

b. Calculate r_{12}. What parameter is estimated by r_{12}? What is the interpretation of this parameter?

c. Test whether or not Y_1 and Y_2 are statistically independent in the population, using test statistic (2.87) and level of significance .01. State the alternatives, decision rule, and conclusion.

d. To test $\rho_{12} = .6$ versus $\rho_{12} \neq .6$, would it be appropriate to use test statistic (2.87)?

2.43. **Contract profitability.** A cost analyst for a drilling and blasting contractor examined 84 contracts handled in the last two years and found that the coefficient of correlation between value of contract (Y_1) and profit contribution generated by the contract (Y_2) is $r_{12} = .61$. Assume that bivariate normal model (2.74) applies.

a. Test whether or not Y_1 and Y_2 are statistically independent in the population; use $\alpha = .05$. State the alternatives, decision rule, and conclusion.

b. Estimate ρ_{12} with a 95 percent confidence interval.

c. Convert the confidence interval in part (b) to a 95 percent confidence interval for ρ_{12}^2. Interpret this interval estimate.

*2.44. **Bid preparation.** A building construction consultant studied the relationship between cost of bid preparation (Y_1) and amount of bid (Y_2) for the consulting firm's clients. In a sample of 103 bids prepared by clients, $r_{12} = .87$. Assume that bivariate normal model (2.74) applies.

a. Test whether or not $\rho_{12} = 0$; control the risk of Type I error at .10. State the alternatives, decision rule, and conclusion. What would be the implication if $\rho_{12} = 0$?

b. Obtain a 90 percent confidence interval for ρ_{12}. Interpret this interval estimate.

c. Convert the confidence interval in part (b) to a 90 percent confidence interval for ρ_{12}^2.

2.45. **Water flow.** An engineer, desiring to estimate the coefficient of correlation ρ_{12} between rate of water flow at point A in a stream (Y_1) and concurrent rate of flow at point B (Y_2), obtained $r_{12} = .83$ in a sample of 147 cases. Assume that bivariate normal model (2.74) is appropriate.

a. Obtain a 99 percent confidence interval for ρ_{12}.

b. Convert the confidence interval in part (a) to a 99 percent confidence interval for ρ_{12}^2.

2.46. Refer to **Property assessments** Problem 2.42. There is some question as to whether or not bivariate model (2.74) is appropriate.

a. Obtain the Spearman rank correlation coefficient r_S.

b. Test by means of the Spearman rank correlation coefficient whether an association exists between property assessments and sales prices using test statistic (2.101) with $\alpha = .01$. State the alternatives, decision rule, and conclusion.

c. How do your estimates and conclusions in parts (a) and (b) compare to those obtained in Problem 2.42?

*2.47. Refer to **Muscle mass** Problem 1.27. Assume that the normal bivariate model (2.74) is appropriate.

a. Compute the Pearson product-moment correlation coefficient r_{12}.

b. Test whether muscle mass and age are statistically independent in the population; use $\alpha = .05$. State the alternatives, decision rule, and conclusion.

c. The bivariate normal model (2.74) assumption is possibly inappropriate here. Compute the Spearman rank correlation coefficient, r_S.

d. Repeat part (b), this time basing the test of independence on the Spearman rank correlation computed in part (c) and test statistic (2.101). Use $\alpha = .05$. State the alternatives, decision rule, and conclusion.

e. How do your estimates and conclusions in parts (a) and (b) compare to those obtained in parts (c) and (d)?

2.48. Refer to **Crime rate** Problems 1.28, 2.30, and 2.31. Assume that the normal bivariate model (2.74) is appropriate.

a. Compute the Pearson product-moment correlation coefficient r_{12}.

b. Test whether crime rate and percentage of high school graduates are statistically independent in the population; use $\alpha = .01$. State the alternatives, decision rule, and conclusion.

c. How do your estimates and conclusions in parts (a) and (b) compare to those obtained in 2.31b and 2.30a, respectively?

2.49. Refer to **Crime rate** Problems 1.28 and 2.48. The bivariate normal model (2.74) assumption is possibly inappropriate here.

a. Compute the Spearman rank correlation coefficient r_S.

b. Test by means of the Spearman rank correlation coefficient whether an association exists between crime rate and percentage of high school graduates using test statistic (2.101) and a level of significance .01. State the alternatives, decision rule, and conclusion.

c. How do your estimates and conclusions in parts (a) and (b) compare to those obtained in Problems 2.48a and 2.48b, respectively?

Exercises

2.50. Derive the property in (2.6) for the k_i.

2.51. Show that b_0 as defined in (2.21) is an unbiased estimator of β_0.

2.52. Derive the expression in (2.22b) for the variance of b_0, making use of (2.31). Also explain how variance (2.22b) is a special case of variance (2.29b).

2.53. (Calculus needed.)

a. Obtain the likelihood function for the sample observations $Y_1, \ldots, Y_n$ given $X_1, \ldots, X_n$, if the conditions on page 83 apply.

b. Obtain the maximum likelihood estimators of β_0, β_1, and σ^2. Are the estimators of β_0 and β_1 the same as those in (1.27) when the X_i are fixed?

2.54. Suppose that normal error regression model (2.1) is applicable except that the error variance is not constant; rather the variance is larger, the larger is X. Does $\beta_1 = 0$ still imply that there is no linear association between X and Y? That there is no association between X and Y? Explain.

2.55. Derive the expression for SSR in (2.51).

2.56. In a small-scale regression study, five observations on Y were obtained corresponding to $X = 1$, 4, 10, 11, and 14. Assume that $\sigma = .6$, $\beta_0 = 5$, and $\beta_1 = 3$.

a. What are the expected values of MSR and MSE here?

b. For determining whether or not a regression relation exists, would it have been better or worse to have made the five observations at $X = 6$, 7, 8, 9, and 10? Why? Would the same answer apply if the principal purpose were to estimate the mean response for $X = 8$? Discuss.

2.57. The normal error regression model (2.1) is assumed to be applicable.

a. When testing H_0: $\beta_1 = 5$ versus H_a: $\beta_1 \neq 5$ by means of a general linear test, what is the reduced model? What are the degrees of freedom df_R?

b. When testing H_0: $\beta_0 = 2$, $\beta_1 = 5$ versus H_a: not both $\beta_0 = 2$ and $\beta_1 = 5$ by means of a general linear test, what is the reduced model? What are the degrees of freedom df_R?

2.58. The random variables Y_1 and Y_2 follow the bivariate normal distribution in (2.74). Show that if $\rho_{12} = 0$, Y_1 and Y_2 are independent random variables.

2.59. (Calculus needed.)

a. Obtain the maximum likelihood estimators of the parameters of the bivariate normal distribution in (2.74).

b. Using the results in part (a), obtain the maximum likelihood estimators of the parameters of the conditional probability distribution of Y_1 for any value of Y_2 in (2.80).

c. Show that the maximum likelihood estimators of $\alpha_{1|2}$ and β_{12} obtained in part (b) are the same as the least squares estimators (1.10) for the regression coefficients in the simple linear regression model.

2.60. Show that test statistics (2.17) and (2.87) are equivalent.

2.61. Show that the ratio *SSR/SSTO* is the same whether Y_1 is regressed on Y_2 or Y_2 is regressed on Y_1. [*Hint*: Use (1.10a) and (2.51).]

Projects

2.62. Refer to the **CDI** data set in Appendix C.2 and Project 1.43. Using R^2 as the criterion, which predictor variable accounts for the largest reduction in the variability in the number of active physicians?

2.63. Refer to the **CDI** data set in Appendix C.2 and Project 1.44. Obtain a separate interval estimate of β_1 for each region. Use a 90 percent confidence coefficient in each case. Do the regression lines for the different regions appear to have similar slopes?

2.64. Refer to the **SENIC** data set in Appendix C.1 and Project 1.45. Using R^2 as the criterion, which predictor variable accounts for the largest reduction in the variability of the average length of stay?

2.65. Refer to the **SENIC** data set in Appendix C.1 and Project 1.46. Obtain a separate interval estimate of β_1 for each region. Use a 95 percent confidence coefficient in each case. Do the regression lines for the different regions appear to have similar slopes?

2.66. Five observations on Y are to be taken when $X = 4$, 8, 12, 16, and 20, respectively. The true regression function is $E\{Y\} = 20 + 4X$, and the ε_i are independent $N(0, 25)$.

a. Generate five normal random numbers, with mean 0 and variance 25. Consider these random numbers as the error terms for the five Y observations at $X = 4, 8, 12, 16$, and 20 and calculate Y_1, Y_2, Y_3, Y_4, and Y_5. Obtain the least squares estimates b_0 and b_1 when fitting a straight line to the five cases. Also calculate $\hat{Y}_h$ when $X_h = 10$ and obtain a 95 percent confidence interval for $E\{Y_h\}$ when $X_h = 10$.

b. Repeat part (a) 200 times, generating new random numbers each time.

c. Make a frequency distribution of the 200 estimates b_1. Calculate the mean and standard deviation of the 200 estimates b_1. Are the results consistent with theoretical expectations?

d. What proportion of the 200 confidence intervals for $E\{Y_h\}$ when $X_h = 10$ include $E\{Y_h\}$? Is this result consistent with theoretical expectations?

2.67. Refer to **Grade point average** Problem 1.19.

a. Plot the data, with the least squares regression line for ACT scores between 20 and 30 superimposed.

b. On the plot in part (a), superimpose a plot of the 95 percent confidence band for the true regression line for ACT scores between 20 and 30. Does the confidence band suggest that the true regression relation has been precisely estimated? Discuss.

2.68. Refer to **Copier maintenance** Problem 1.20.

a. Plot the data, with the least squares regression line for numbers of copiers serviced between 1 and 8 superimposed.

b. On the plot in part (a), superimpose a plot of the 90 percent confidence band for the true regression line for numbers of copiers serviced between 1 and 8. Does the confidence band suggest that the true regression relation has been precisely estimated? Discuss.

Chapter 3

Diagnostics and Remedial Measures

When a regression model, such as the simple linear regression model (2.1), is considered for an application, we can usually not be certain in advance that the model is appropriate for that application. Any one, or several, of the features of the model, such as linearity of the regression function or normality of the error terms, may not be appropriate for the particular data at hand. Hence, it is important to examine the aptness of the model for the data before inferences based on that model are undertaken. In this chapter, we discuss some simple graphic methods for studying the appropriateness of a model, as well as some formal statistical tests for doing so. We also consider some remedial techniques that can be helpful when the data are not in accordance with the conditions of regression model (2.1). We conclude the chapter with a case example that brings together the concepts and methods presented in this and the earlier chapters.

While the discussion in this chapter is in terms of the appropriateness of the simple linear regression model (2.1), the basic principles apply to all statistical models discussed in this book. In later chapters, additional methods useful for examining the appropriateness of statistical models and other remedial measures will be presented, as well as methods for validating the statistical model.

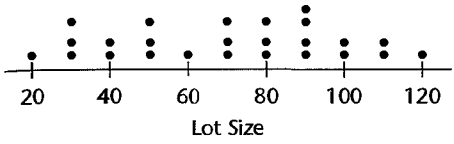
3.1 Diagnostics for Predictor Variable

We begin by considering some graphic diagnostics for the predictor variable. We need diagnostic information about the predictor variable to see if there are any outlying X values that could influence the appropriateness of the fitted regression function. We discuss the role of influential cases in detail in Chapter 10. Diagnostic information about the range and concentration of the X levels in the study is also useful for ascertaining the range of validity for the regression analysis.

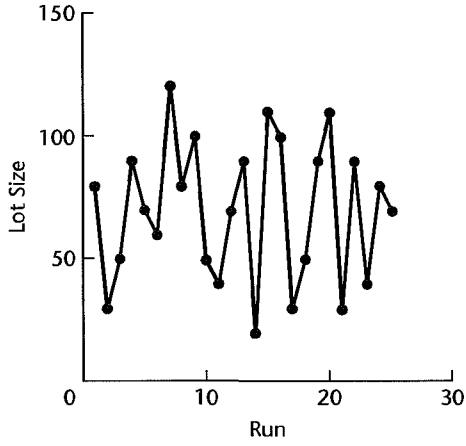
Figure 3.1a contains a simple *dot plot* for the lot sizes in the Toluca Company example in Figure 1.10. A dot plot is helpful when the number of observations in the data set is not large. The dot plot in Figure 3.1a shows that the minimum and maximum lot sizes are 20 and 120, respectively, that the lot size levels are spread throughout this interval, and that

FIGURE 3.1 MINITAB and SYGRAPH Diagnostic Plots for Predictor Variable—Toluca Company Example.

(a) Dot Plot



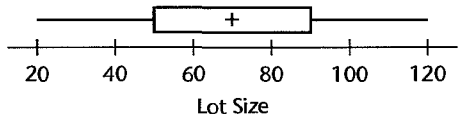
(b) Sequence Plot



(c) Stem-and-Leaf Plot

```
 2   0
 3   000
 4   00
5H   000
 6   0
7M   000
 8   000
 9H  0000
10   00
11   00
12   0
```

(d) Box Plot



there are no lot sizes that are far outlying. The dot plot also shows that in a number of cases several runs were made for the same lot size.

A second useful diagnostic for the predictor variable is a *sequence plot.* Figure 3.1b contains a time sequence plot of the lot sizes for the Toluca Company example. Lot size is here plotted against production run (i.e., against time sequence). The points in the plot are connected to show more effectively the time sequence. Sequence plots should be utilized whenever data are obtained in a sequence, such as over time or for adjacent geographic areas. The sequence plot in Figure 3.1b contains no special pattern. If, say, the plot had shown that smaller lot sizes had been utilized early on and larger lot sizes later on, this information could be very helpful for subsequent diagnostic studies of the aptness of the fitted regression model.

Figures 3.1c and 3.1d contain two other diagnostic plots that present information similar to the dot plot in Figure 3.1a. The *stem-and-leaf plot* in Figure 3.1c provides information similar to a frequency histogram. By displaying the last digits, this plot also indicates here that all lot sizes in the Toluca Company example were multiples of 10. The letter M in the

SYGRAPH output denotes the stem where the median is located, and the letter H denotes the stems where the first and third quartiles (hinges) are located.

The *box plot* in Figure 3.1d shows the minimum and maximum lot sizes, the first and third quartiles, and the median lot size. We see that the middle half of the lot sizes range from 50 to 90, and that they are fairly symmetrically distributed because the median is located in the middle of the central box. A box plot is particularly helpful when there are many observations in the data set.

3.2 Residuals

Direct diagnostic plots for the response variable Y are ordinarily not too useful in regression analysis because the values of the observations on the response variable are a function of the level of the predictor variable. Instead, diagnostics for the response variable are usually carried out indirectly through an examination of the residuals.

The residual e_i, as defined in (1.16), is the difference between the observed value Y_i and the fitted value $\hat{Y}_i$:

$$e_i = Y_i - \hat{Y}_i \tag{3.1}$$

The residual may be regarded as the observed error, in distinction to the unknown true error ε_i in the regression model:

$$\varepsilon_i = Y_i - E\{Y_i\} \tag{3.2}$$

For regression model (2.1), the error terms ε_i are assumed to be independent normal random variables, with mean 0 and constant variance σ^2. If the model is appropriate for the data at hand, the observed residuals e_i should then reflect the properties assumed for the ε_i. This is the basic idea underlying *residual analysis*, a highly useful means of examining the aptness of a statistical model.

Properties of Residuals

Mean. The mean of the n residuals e_i for the simple linear regression model (2.1) is, by (1.17):

$$\bar{e} = \frac{\sum e_i}{n} = 0 \tag{3.3}$$

where $\bar{e}$ denotes the mean of the residuals. Thus, since $\bar{e}$ is always 0, it provides no information as to whether the true errors ε_i have expected value $E\{\varepsilon_i\} = 0$.

Variance. The variance of the n residuals e_i is defined as follows for regression model (2.1):

$$s^2 = \frac{\sum (e_i - \bar{e})^2}{n-2} = \frac{\sum e_i^2}{n-2} = \frac{SSE}{n-2} = MSE \tag{3.4}$$

If the model is appropriate, MSE is, as noted earlier, an unbiased estimator of the variance of the error terms σ^2.

Nonindependence. The residuals e_i are not independent random variables because they involve the fitted values $\hat{Y}_i$ which are based on the same fitted regression function. As

a result, the residuals for regression model (2.1) are subject to two constraints. These are constraint (1.17)—that the sum of the e_i must be 0—and constraint (1.19)—that the products $X_i e_i$ must sum to 0.

When the sample size is large in comparison to the number of parameters in the regression model, the dependency effect among the residuals e_i is relatively unimportant and can be ignored for most purposes.

Semistudentized Residuals

At times, it is helpful to standardize the residuals for residual analysis. Since the standard deviation of the error terms ε_i is σ, which is estimated by $\sqrt{MSE}$, it is natural to consider the following form of standardization:

$$e_i^* = \frac{e_i - \bar{e}}{\sqrt{MSE}} = \frac{e_i}{\sqrt{MSE}} \tag{3.5}$$

If $\sqrt{MSE}$ were an estimate of the standard deviation of the residual e_i, we would call e_i^* a studentized residual. However, the standard deviation of e_i is complex and varies for the different residuals e_i, and $\sqrt{MSE}$ is only an approximation of the standard deviation of e_i. Hence, we call the statistic e_i^* in (3.5) a *semistudentized residual*. We shall take up studentized residuals in Chapter 10. Both semistudentized residuals and studentized residuals can be very helpful in identifying outlying observations.

Departures from Model to Be Studied by Residuals

We shall consider the use of residuals for examining six important types of departures from the simple linear regression model (2.1) with normal errors:

1. The regression function is not linear.
2. The error terms do not have constant variance.
3. The error terms are not independent.
4. The model fits all but one or a few outlier observations.
5. The error terms are not normally distributed.
6. One or several important predictor variables have been omitted from the model.

3.3 Diagnostics for Residuals

We take up now some informal diagnostic plots of residuals to provide information on whether any of the six types of departures from the simple linear regression model (2.1) just mentioned are present. The following plots of residuals (or semistudentized residuals) will be utilized here for this purpose:

1. Plot of residuals against predictor variable.
2. Plot of absolute or squared residuals against predictor variable.
3. Plot of residuals against fitted values.
4. Plot of residuals against time or other sequence.
5. Plots of residuals against omitted predictor variables.
6. Box plot of residuals.
7. Normal probability plot of residuals.

FIGURE 3.2 MINITAB and SYGRAPH Diagnostic Residual Plots—Toluca Company Example.

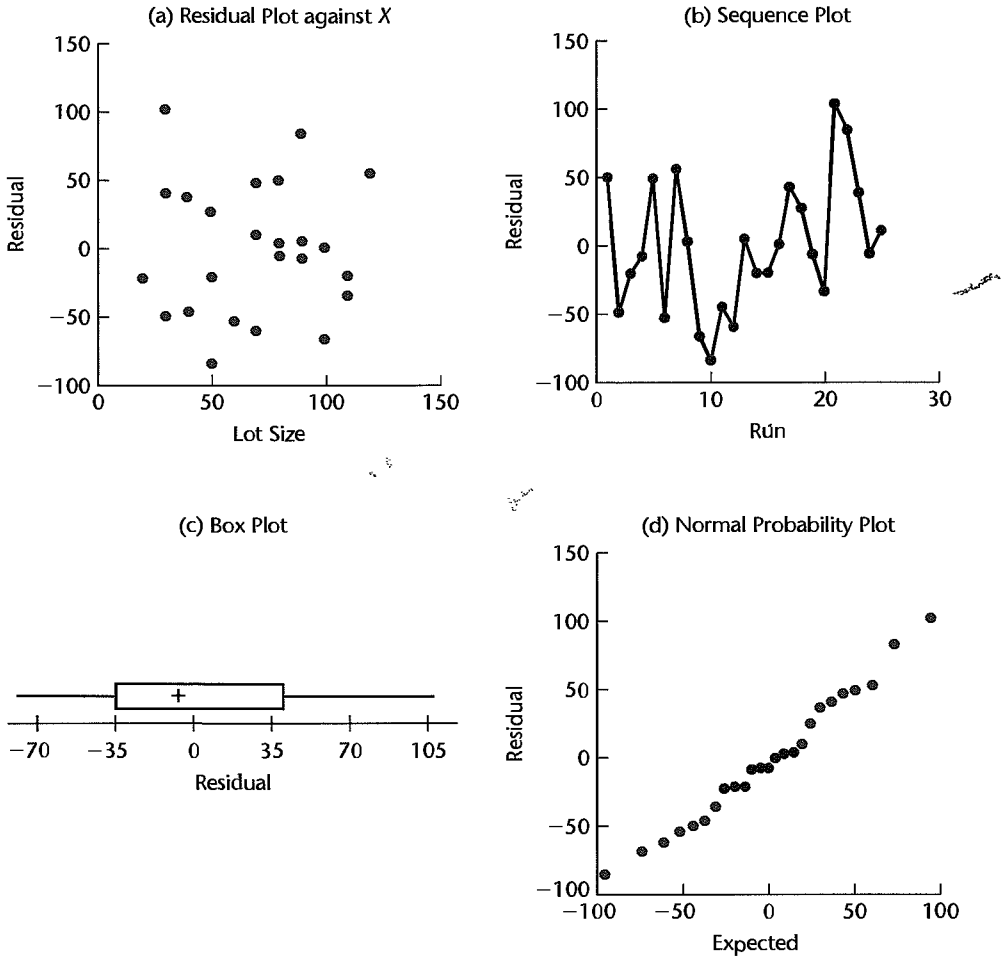


Figure 3.2 contains, for the Toluca Company example, MINITAB and SYGRAPH plots of the residuals in Table 1.2 against the predictor variable and against time, a box plot, and a normal probability plot. All of these plots, as we shall see, support the appropriateness of regression model (2.1) for the data.

We turn now to consider how residual analysis can be helpful in studying each of the six departures from regression model (2.1).

Nonlinearity of Regression Function

Whether a linear regression function is appropriate for the data being analyzed can be studied from a *residual plot against the predictor variable* or, equivalently, from a *residual plot against the fitted values.* Nonlinearity of the regression function can also be studied from a *scatter plot,* but this plot is not always as effective as a residual plot. Figure 3.3a

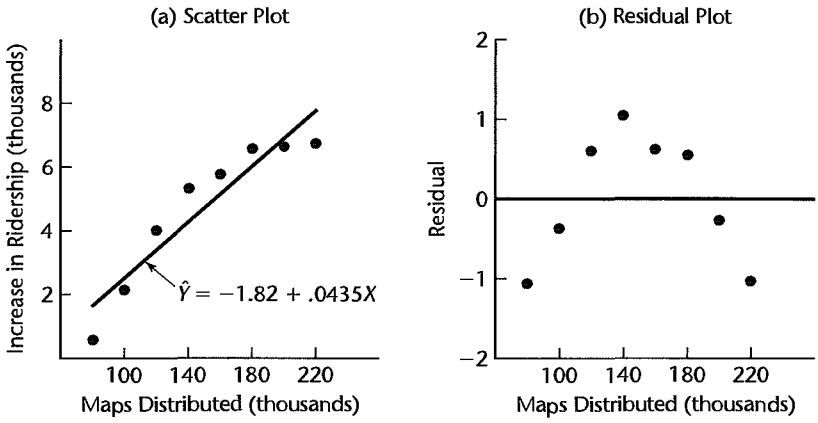


FIGURE 3.3 Scatter Plot and Residual Plot Illustrating Nonlinear Regression Function—Transit Example.

TABLE 3.1 Number of Maps Distributed and Increase in Ridership—Transit Example.

City i	(1) Increase in Ridership (thousands) Y_i	(2) Maps Distributed (thousands) X_i	(3) Fitted Value $\hat{Y}_i$	(4) Residual $Y_i - \hat{Y}_i = e_i$
1	.60	80	1.66	−1.06
2	6.70	220	7.75	−1.05
3	5.30	140	4.27	1.03
4	4.00	120	3.40	.60
5	6.55	180	6.01	.54
6	2.15	100	2.53	−.38
7	6.60	200	6.88	−.28
8	5.75	160	5.14	.61

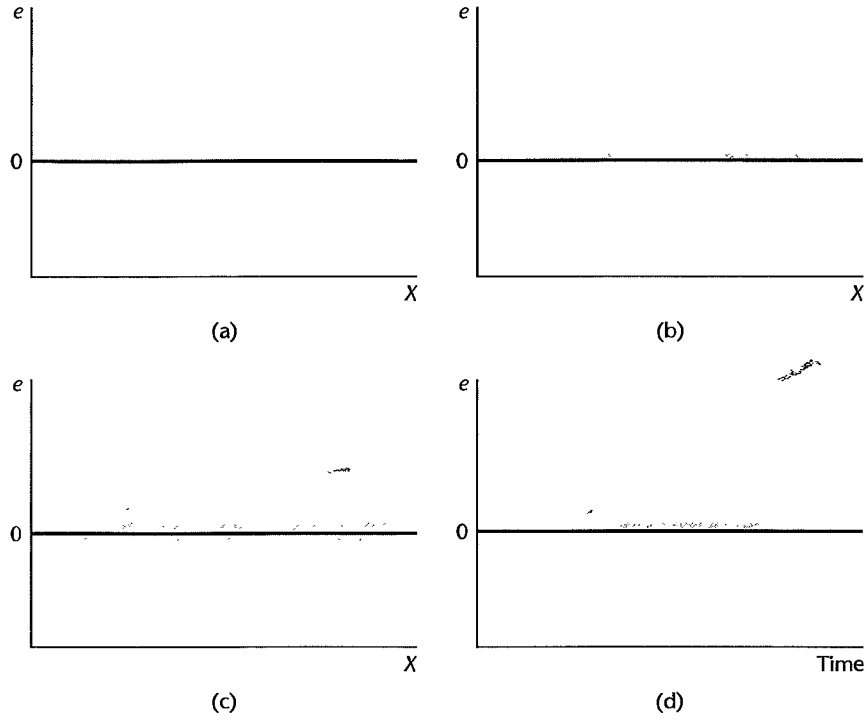
$\hat{Y} = -1.82 + .0435X$

contains a scatter plot of the data and the fitted regression line for a study of the relation between maps distributed and bus ridership in eight test cities. Here, X is the number of bus transit maps distributed free to residents of the city at the beginning of the test period and Y is the increase during the test period in average daily bus ridership during nonpeak hours. The original data and fitted values are given in Table 3.1, columns 1, 2, and 3. The plot suggests strongly that a linear regression function is not appropriate.

Figure 3.3b presents a plot of the residuals, shown in Table 3.1, column 4, against the predictor variable X. The lack of fit of the linear regression function is even more strongly suggested by the residual plot against X in Figure 3.3b than by the scatter plot. Note that the residuals depart from 0 in a systematic fashion; they are negative for smaller X values, positive for medium-size X values, and negative again for large X values.

In this case, both Figures 3.3a and 3.3b point out the lack of linearity of the regression function. In general, however, the residual plot is to be preferred, because it has some important advantages over the scatter plot. First, the residual plot can easily be used for examining other facets of the aptness of the model. Second, there are occasions when the

FIGURE 3.4
Prototype Residual Plots.



scaling of the scatter plot places the Y_i observations close to the fitted values $\hat{Y}_i$, for instance, when there is a steep slope. It then becomes more difficult to study the appropriateness of a linear regression function from the scatter plot. A residual plot, on the other hand, can clearly show any systematic pattern in the deviations around the fitted regression line under these conditions.

Figure 3.4a shows a prototype situation of the residual plot against X when a linear regression model is appropriate. The residuals then fall within a horizontal band centered around 0, displaying no systematic tendencies to be positive and negative. This is the case in Figure 3.2a for the Toluca Company example.

Figure 3.4b shows a prototype situation of a departure from the linear regression model that indicates the need for a curvilinear regression function. Here the residuals tend to vary in a systematic fashion between being positive and negative. This is the case in Figure 3.3b for the transit example. A different type of departure from linearity would, of course, lead to a picture different from the prototype pattern in Figure 3.4b.

Comment

A plot of residuals against the fitted values $\hat{Y}$ provides equivalent information as a plot of residuals against X for the simple linear regression model, and thus is not needed in addition to the residual plot against X. The two plots provide the same information because the fitted values $\hat{Y}_i$ are a linear function of the values X_i for the predictor variable. Thus, only the X scale values, not the basic pattern of the plotted points, are affected by whether the residual plot is against the X_i or the $\hat{Y}_i$. For curvilinear regression and multiple regression, on the other hand, separate plots of the residuals against the fitted values and against the predictor variable(s) are usually helpful. ■

Nonconstancy of Error Variance

Plots of the residuals against the predictor variable or against the fitted values are not only helpful to study whether a linear regression function is appropriate but also to examine whether the variance of the error terms is constant. Figure 3.5a shows a residual plot against age for a study of the relation between diastolic blood pressure of healthy, adult women (Y) and their age (X). The plot suggests that the older the woman is, the more spread out the residuals are. Since the relation between blood pressure and age is positive, this suggests that the error variance is larger for older women than for younger ones.

The prototype plot in Figure 3.4a exemplifies residual plots when the error term variance is constant. The residual plot in Figure 3.2a for the Toluca Company example is of this type, suggesting that the error terms have constant variance here.

Figure 3.4c shows a prototype picture of residual plots when the error variance increases with X. In many business, social science, and biological science applications, departures from constancy of the error variance tend to be of the "megaphone" type shown in Figure 3.4c, as in the blood pressure example in Figure 3.5a. One can also encounter error variances decreasing with increasing levels of the predictor variable and occasionally varying in some more complex fashion.

Plots of the absolute values of the residuals or of the squared residuals against the predictor variable X or against the fitted values $\hat{Y}$ are also useful for diagnosing nonconstancy of the error variance since the signs of the residuals are not meaningful for examining the constancy of the error variance. These plots are especially useful when there are not many cases in the data set because plotting of either the absolute or squared residuals places all of the information on changing magnitudes of the residuals above the horizontal zero line so that one can more readily see whether the magnitude of the residuals (irrespective of sign) is changing with the level of X or $\hat{Y}$.

Figure 3.5b contains a plot of the absolute residuals against age for the blood pressure example. This plot shows more clearly that the residuals tend to be larger in absolute magnitude for older-aged women.

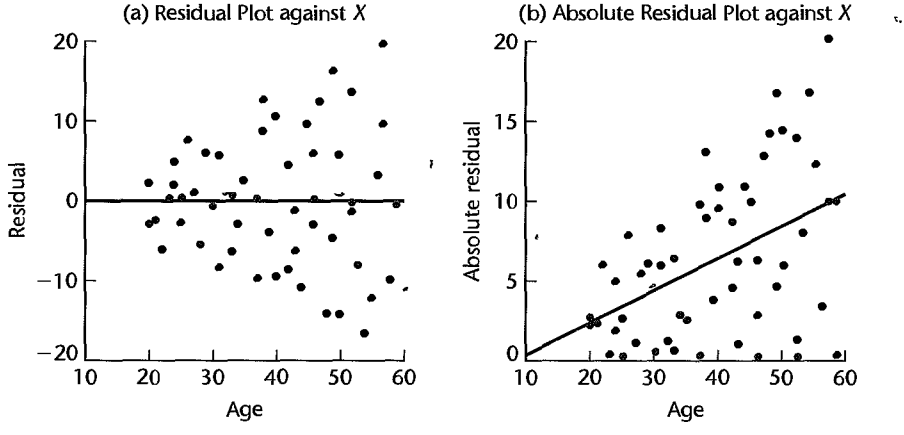


FIGURE 3.5
Residual Plots Illustrating Nonconstant Error Variance.

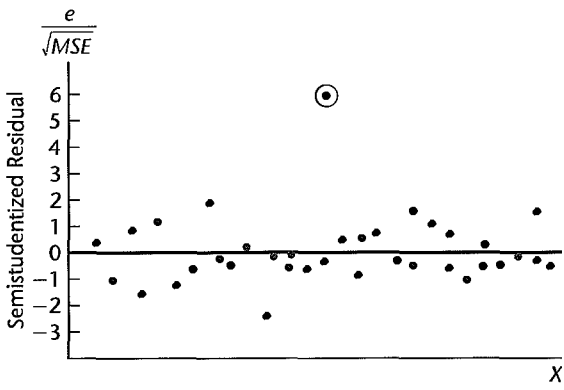


FIGURE 3.6
Residual Plot with Outlier.

Presence of Outliers

Outliers are extreme observations. Residual outliers can be identified from *residual plots against X or $\hat{Y}$*, as well as from *box plots, stem-and-leaf plots,* and *dot plots* of the residuals. Plotting of semistudentized residuals is particularly helpful for distinguishing outlying observations, since it then becomes easy to identify residuals that lie many standard deviations from zero. A rough rule of thumb when the number of cases is large is to consider semistudentized residuals with absolute value of four or more to be outliers. We shall take up more refined procedures for identifying outliers in Chapter 10.

The residual plot in Figure 3.6 presents semistudentized residuals and contains one outlier, which is circled. Note that this residual represents an observation almost six standard deviations from the fitted value.

Outliers can create great difficulty. When we encounter one, our first suspicion is that the observation resulted from a mistake or other extraneous effect, and hence should be discarded. A major reason for discarding it is that under the least squares method, a fitted line may be pulled disproportionately toward an outlying observation because the sum of the *squared* deviations is minimized. This could cause a misleading fit if indeed the outlying observation resulted from a mistake or other extraneous cause. On the other hand, outliers may convey significant information, as when an outlier occurs because of an interaction with another predictor variable omitted from the model. A safe rule frequently suggested is to discard an outlier only if there is direct evidence that it represents an error in recording, a miscalculation, a malfunctioning of equipment, or a similar type of circumstance.

Comment

When a linear regression model is fitted to a data set with a small number of cases and an outlier is present, the fitted regression can be so distorted by the outlier that the residual plot may improperly suggest a lack of fit of the linear regression model, in addition to flagging the outlier. Figure 3.7 illustrates this situation. The scatter plot in Figure 3.7a presents a situation where all observations except the outlier fall around a straight-line statistical relationship. When a linear regression function is fitted to these data, the outlier causes such a shift in the fitted regression line as to lead to a systematic pattern of deviations from the fitted line for the other observations, suggesting a lack of fit of the linear regression function. This is shown by the residual plot in Figure 3.7b. ■

Nonindependence of Error Terms

Whenever data are obtained in a time sequence or some other type of sequence, such as for adjacent geographic areas, it is a good idea to prepare a *sequence plot of the residuals.*

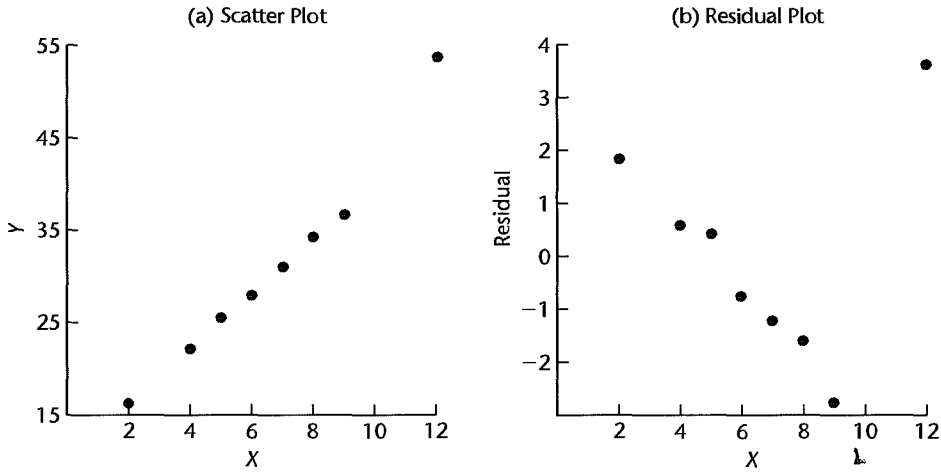
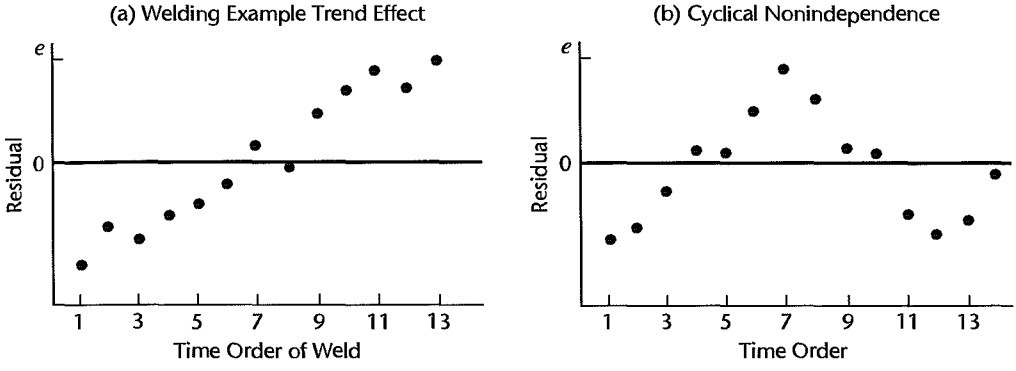


FIGURE 3.7 Distorting Effect on Residuals Caused by an Outlier When Remaining Data Follow Linear Regression.

FIGURE 3.8 Residual Time Sequence Plots Illustrating Nonindependence of Error Terms.



The purpose of plotting the residuals against time or in some other type of sequence is to see if there is any correlation between error terms that are near each other in the sequence. Figure 3.8a contains a time sequence plot of the residuals in an experiment to study the relation between the diameter of a weld (X) and the shear strength of the weld (Y): An evident correlation between the error terms stands out. Negative residuals are associated mainly with the early trials, and positive residuals with the later trials. Apparently, some effect connected with time was present, such as learning by the welder or a gradual change in the welding equipment, so the shear strength tended to be greater in the later welds because of this effect.

A prototype residual plot showing a time-related trend effect is presented in Figure 3.4d, which portrays a linear time-related trend effect, as in the welding example. It is sometimes useful to view the problem of nonindependence of the error terms as one in which an important variable (in this case, time) has been omitted from the model. We shall discuss this type of problem shortly.

Another type of nonindependence of the error terms is illustrated in Figure 3.8b. Here the adjacent error terms are also related, but the resulting pattern is a cyclical one with no trend effect present.

When the error terms are independent, we expect the residuals in a sequence plot to fluctuate in a more or less random pattern around the base line 0, such as the scattering shown in Figure 3.2b for the Toluca Company example. Lack of randomness can take the form of too much or too little alternation of points around the zero line. In practice, there is little concern with the former because it does not arise frequently. Too little alternation, in contrast, frequently occurs, as in the welding example in Figure 3.8a.

Comment

When the residuals are plotted against X, as in Figure 3.3b for the transit example, the scatter may not appear to be random. For this plot, however, the basic problem is probably not lack of independence of the error terms but a poorly fitting regression function. This, indeed, is the situation portrayed in the scatter plot in Figure 3.3a. ■

Nonnormality of Error Terms

As we noted earlier, small departures from normality do not create any serious problems. Major departures, on the other hand, should be of concern. The normality of the error terms can be studied informally by examining the residuals in a variety of graphic ways.

Distribution Plots. A *box plot* of the residuals is helpful for obtaining summary information about the symmetry of the residuals and about possible outliers. Figure 3.2c contains a box plot of the residuals in the Toluca Company example. No serious departures from symmetry are suggested by this plot. A *histogram, dot plot,* or *stem-and-leaf plot* of the residuals can also be helpful for detecting gross departures from normality. However, the number of cases in the regression study must be reasonably large for any of these plots to convey reliable information about the shape of the distribution of the error terms.

Comparison of Frequencies. Another possibility when the number of cases is reasonably large is to compare actual frequencies of the residuals against expected frequencies under normality. For example, one can determine whether, say, about 68 percent of the residuals e_i fall between $\pm\sqrt{MSE}$ or about 90 percent fall between $\pm 1.645\sqrt{MSE}$. When the sample size is moderately large, corresponding t values may be used for the comparison.

To illustrate this procedure, we again consider the Toluca Company example of Chapter 1. Table 3.2, column 1, repeats the residuals from Table 1.2. We see from Figure 2.2 that $\sqrt{MSE} = 48.82$. Using the t distribution, we expect under normality about 90 percent of the residuals to fall between $\pm t(.95; 23)\sqrt{MSE} = \pm 1.714(48.82)$, or between -83.68 and 83.68. Actually, 22 residuals, or 88 percent, fall within these limits. Similarly, under normality, we expect about 60 percent of the residuals to fall between -41.89 and 41.89. The actual percentage here is 52 percent. Thus, the actual frequencies here are reasonably consistent with those expected under normality.

Normal Probability Plot. Still another possibility is to prepare a *normal probability plot of the residuals.* Here each residual is plotted against its expected value under normality. A plot that is nearly linear suggests agreement with normality, whereas a plot that departs substantially from linearity suggests that the error distribution is not normal.

Table 3.2, column 1, contains the residuals for the Toluca Company example. To find the expected values of the ordered residuals under normality, we utilize the facts that (1)

TABLE 3.2 Residuals and Expected Values under Normality—Toluca Company Example.

Run i	(1) Residual e_i	(2) Rank k	(3) Expected Value under Normality
1	51.02	22	51.95
2	−48.47	5	−44.10
3	−19.88	10	−14.76
...	...	...	...
23	38.83	19	31.05
24	−5.98	13	0
25	10.72	17	19.93

the expected value of the error terms for regression model (2.1) is zero and (2) the standard deviation of the error terms is estimated by $\sqrt{MSE}$. Statistical theory has shown that for a normal random variable with mean 0 and estimated standard deviation $\sqrt{MSE}$, a good approximation of the expected value of the kth smallest observation in a random sample of n is:

$$\sqrt{MSE}\left[z\left(\frac{k-.375}{n+.25}\right)\right] \tag{3.6}$$

where $z(A)$ as usual denotes the $(A)100$ percentile of the standard normal distribution.

Using this approximation, let us calculate the expected values of the residuals under normality for the Toluca Company example. Column 2 of Table 3.2 shows the ranks of the residuals, with the smallest residual being assigned rank 1. We see that the rank of the residual for run 1, $e_1 = 51.02$, is 22, which indicates that this residual is the 22nd smallest among the 25 residuals. Hence, for this residual $k = 22$. We found earlier (Table 2.1) that $MSE = 2{,}384$. Hence:

$$\frac{k-.375}{n+.25} = \frac{22-.375}{25+.25} = \frac{21.625}{25.25} = .8564$$

so that the expected value of this residual under normality is:

$$\sqrt{2{,}384}[z(.8564)] = \sqrt{2{,}384}(1.064) = 51.95$$

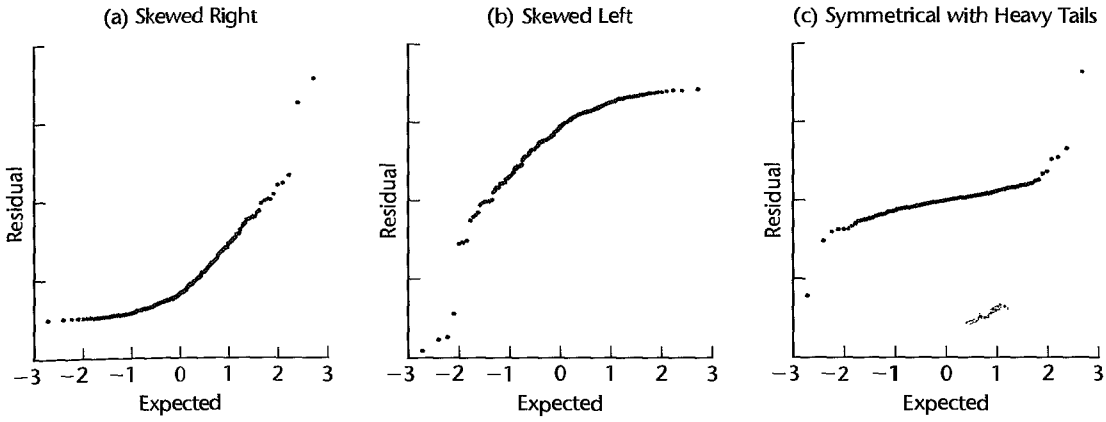
Similarly, the expected value of the residual for run 2, $e_2 = -48.47$, is obtained by noting that the rank of this residual is $k = 5$; in other words, this residual is the fifth smallest one among the 25 residuals. Hence, we require $(k-.375)/(n+.25) = (5-.375)/(25+.25) = .1832$, so that the expected value of this residual under normality is:

$$\sqrt{2{,}384}[z(.1832)] = \sqrt{2{,}384}(-.9032) = -44.10$$

Table 3.2, column 3, contains the expected values under the assumption of normality for a portion of the 25 residuals. Figure 3.2d presents a plot of the residuals against their expected values under normality. Note that the points in Figure 3.2d fall reasonably close to a straight line, suggesting that the distribution of the error terms does not depart substantially from a normal distribution.

Figure 3.9 shows three normal probability plots when the distribution of the error terms departs substantially from normality. Figure 3.9a shows a normal probability plot when the error term distribution is highly skewed to the right. Note the concave-upward shape

FIGURE 3.9 Normal Probability Plots when Error Term Distribution Is Not Normal.



of the plot. Figure 3.9b shows a normal probability plot when the error term distribution is highly skewed to the left. Here, the pattern is concave downward. Finally, Figure 3.9c shows a normal probability plot when the distribution of the error terms is symmetrical but has heavy tails; in other words, the distribution has higher probabilities in the tails than a normal distribution. Note the concave-downward curvature in the plot at the left end, corresponding to the plot for a left-skewed distribution, and the concave-upward plot at the right end, corresponding to a right-skewed distribution.

Comments

1. Many computer packages will prepare normal probability plots, either automatically or at the option of the user. Some of these plots utilize semistudentized residuals, others omit the factor $\sqrt{MSE}$ in (3.6), but neither of these variations affect the nature of the plot.

2. For continuous data, ties among the residuals should occur only rarely. If two residuals do have the same value, a simple procedure is to use the average rank for the tied residuals for calculating the corresponding expected values. ■

Difficulties in Assessing Normality. The analysis for model departures with respect to normality is, in many respects, more difficult than that for other types of departures. In the first place, random variation can be particularly mischievous when studying the nature of a probability distribution unless the sample size is quite large. Even worse, other types of departures can and do affect the distribution of the residuals. For instance, residuals may appear to be not normally distributed because an inappropriate regression function is used or because the error variance is not constant. Hence, it is usually a good strategy to investigate these other types of departures first, before concerning oneself with the normality of the error terms.

Omission of Important Predictor Variables

Residuals should also be plotted against variables omitted from the model that might have important effects on the response. The time variable cited earlier in the welding example is

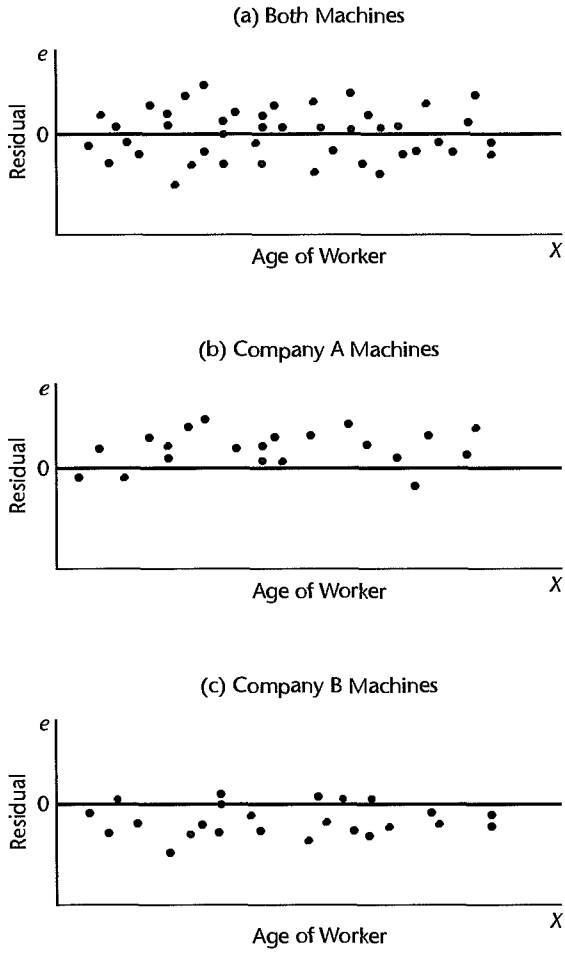


FIGURE 3.10 **Residual Plots for Possible Omission of Important Predictor Variable—Productivity Example.**

an illustration. The purpose of this additional analysis is to determine whether there are any other key variables that could provide important additional descriptive and predictive power to the model.

As another example, in a study to predict output by piece-rate workers in an assembling operation, the relation between output (Y) and age (X) of worker was studied for a sample of employees. The plot of the residuals against X, shown in Figure 3.10a, indicates no ground for suspecting the appropriateness of the linearity of the regression function or the constancy of the error variance. Since machines produced by two companies (A and B) are used in the assembling operation and could have an effect on output, residual plots against X by type of machine were undertaken and are shown in Figures 3.10b and 3.10c. Note that the residuals for Company A machines tend to be positive, while those for Company B machines tend to be negative. Thus, type of machine appears to have a definite effect on productivity, and output predictions may turn out to be far superior when this variable is added to the model.

While this second example dealt with a qualitative variable (type of machine), the residual analysis for an additional quantitative variable is analogous. The residuals are plotted against the additional predictor variable to see whether or not the residuals tend to vary systematically with the level of the additional predictor variable.

Comment

We do not say that the original model is "wrong" when it can be improved materially by adding one or more predictor variables. Only a few of the factors operating on any response variable Y in real-world situations can be included explicitly in a regression model. The chief purpose of residual analysis in identifying other important predictor variables is therefore to test the adequacy of the model and see whether it could be improved materially by adding one or more predictor variables. ■

Some Final Comments

1. We discussed model departures one at a time. In actuality, several types of departures may occur together. For instance, a linear regression function may be a poor fit and the variance of the error terms may not be constant. In these cases, the prototype patterns of Figure 3.4 can still be useful, but they would need to be combined into composite patterns.

2. Although graphic analysis of residuals is only an informal method of analysis, in many cases it suffices for examining the aptness of a model.

3. The basic approach to residual analysis explained here applies not only to simple linear regression but also to more complex regression and other types of statistical models.

4. Several types of departures from the simple linear regression model have been identified by diagnostic tests of the residuals. Model misspecification due to either nonlinearity or the omission of important predictor variables tends to be serious, leading to biased estimates of the regression parameters and error variance. These problems are discussed further in Section 3.9 and Chapter 10. Nonconstancy of error variance tends to be less serious, leading to less efficient estimates and invalid error variance estimates. The problem is discussed in depth in Section 11.1. The presence of outliers can be serious for smaller data sets when their influence is large. Influential outliers are discussed further in Section 10.4. Finally, the nonindependence of error terms results in estimators that are unbiased but whose variances are seriously biased. Alternative estimation methods for correlated errors are discussed in Chapter 12.

3.4 Overview of Tests Involving Residuals

Graphic analysis of residuals is inherently subjective. Nevertheless, subjective analysis of a variety of interrelated residual plots will frequently reveal difficulties with the model more clearly than particular formal tests. There are occasions, however, when one wishes to put specific questions to a test. We now briefly review some of the relevant tests.

Most statistical tests require independent observations. As we have seen, however, the residuals are dependent. Fortunately, the dependencies become quite small for large samples, so that one can usually then ignore them.

Tests for Randomness

A runs test is frequently used to test for lack of randomness in the residuals arranged in time order. Another test, specifically designed for lack of randomness in least squares residuals, is the Durbin-Watson test. This test is discussed in Chapter 12.

Tests for Constancy of Variance

When a residual plot gives the impression that the variance may be increasing or decreasing in a systematic manner related to X or $E\{Y\}$, a simple test is based on the rank correlation between the absolute values of the residuals and the corresponding values of the predictor variable. Two other simple tests for constancy of the error variance—the Brown-Forsythe test and the Breusch-Pagan test—are discussed in Section 3.6.

Tests for Outliers

A simple test for identifying an outlier observation involves fitting a new regression line to the other $n - 1$ observations. The suspect observation, which was not used in fitting the new line, can now be regarded as a new observation. One can calculate the probability that in n observations, a deviation from the fitted line as great as that of the outlier will be obtained by chance. If this probability is sufficiently small, the outlier can be rejected as not having come from the same population as the other $n - 1$ observations. Otherwise, the outlier is retained. We discuss this approach in detail in Chapter 10.

Many other tests to aid in evaluating outliers have been developed. These are discussed in specialized references, such as Reference 3.1.

Tests for Normality

Goodness of fit tests can be used for examining the normality of the error terms. For instance, the chi-square test or the Kolmogorov-Smirnov test and its modification, the Lilliefors test, can be employed for testing the normality of the error terms by analyzing the residuals. A simple test based on the normal probability plot of the residuals will be taken up in Section 3.5.

Comment

The runs test, rank correlation, and goodness of fit tests are commonly used statistical procedures and are discussed in many basic statistics texts. ■

3.5 Correlation Test for Normality

In addition to visually assessing the approximate linearity of the points plotted in a normal probability plot, a formal test for normality of the error terms can be conducted by calculating the coefficient of correlation (2.74) between the residuals e_i and their expected values under normality. A high value of the correlation coefficient is indicative of normality. Table B.6, prepared by Looney and Gulledge (Ref. 3.2), contains critical values (percentiles) for various sample sizes for the distribution of the coefficient of correlation between the ordered residuals and their expected values under normality when the error terms are normally distributed. If the observed coefficient of correlation is at least as large as the tabled value, for a given α level, one can conclude that the error terms are reasonably normally distributed.

Example

For the Toluca Company example in Table 3.2, the coefficient of correlation between the ordered residuals and their expected values under normality is .991. Controlling the α risk at .05, we find from Table B.6 that the critical value for $n = 25$ is .959. Since the observed coefficient exceeds this level, we have support for our earlier conclusion that the distribution of the error terms does not depart substantially from a normal distribution.

Comment

The correlation test for normality presented here is simpler than the Shapiro-Wilk test (Ref. 3.3), which can be viewed as being based approximately also on the coefficient of correlation between the ordered residuals and their expected values under normality. ■

3.6 Tests for Constancy of Error Variance

We present two formal tests for ascertaining whether the error terms have constant variance: the Brown-Forsythe test and the Breusch-Pagan test.

Brown-Forsythe Test

The Brown-Forsythe test, a modification of the Levene test (Ref. 3.4), does not depend on normality of the error terms. Indeed, this test is robust against serious departures from normality, in the sense that the nominal significance level remains approximately correct when the error terms have equal variances even if the distribution of the error terms is far from normal. Yet the test is still relatively efficient when the error terms are normally distributed. The Brown-Forsythe test as described is applicable to simple linear regression when the variance of the error terms either increases or decreases with X, as illustrated in the prototype megaphone plot in Figure 3.4c. The sample size needs to be large enough so that the dependencies among the residuals can be ignored.

The test is based on the variability of the residuals. The larger the error variance, the larger the variability of the residuals will tend to be. To conduct the Brown-Forsythe test, we divide the data set into two groups, according to the level of X, so that one group consists of cases where the X level is comparatively low and the other group consists of cases where the X level is comparatively high. If the error variance is either increasing or decreasing with X, the residuals in one group will tend to be more variable than those in the other group. Equivalently, the absolute deviations of the residuals around their group mean will tend to be larger for one group than for the other group. In order to make the test more robust, we utilize the absolute deviations of the residuals around the median for the group (Ref. 3.5). The Brown-Forsythe test then consists simply of the two-sample t test based on test statistic (A.67) to determine whether the mean of the absolute deviations for one group differs significantly from the mean absolute deviation for the second group.

Although the distribution of the absolute deviations of the residuals is usually not normal, it has been shown that the t^* test statistic still follows approximately the t distribution when the variance of the error terms is constant and the sample sizes of the two groups are not extremely small.

We shall now use e_{i1} to denote the ith residual for group 1 and e_{i2} to denote the ith residual for group 2. Also we shall use n_1 and n_2 to denote the sample sizes of the two groups, where:

$$n = n_1 + n_2 \tag{3.7}$$

Further, we shall use $\tilde{e}_1$ and $\tilde{e}_2$ to denote the medians of the residuals in the two groups. The Brown-Forsythe test uses the absolute deviations of the residuals around their group median, to be denoted by d_{i1} and d_{i2}:

$$d_{i1} = |e_{i1} - \tilde{e}_1| \qquad d_{i2} = |e_{i2} - \tilde{e}_2| \tag{3.8}$$

With this notation, the two-sample t test statistic (A.67) becomes:

$$t_{BF}^* = \frac{\bar{d}_1 - \bar{d}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{3.9}$$

where $\bar{d}_1$ and $\bar{d}_2$ are the sample means of the d_{i1} and d_{i2}, respectively, and the pooled variance s^2 in (A.63) becomes:

$$s^2 = \frac{\sum(d_{i1} - \bar{d}_1)^2 + \sum(d_{i2} - \bar{d}_2)^2}{n - 2} \tag{3.9a}$$

We denote the test statistic for the Brown-Forsythe test by t_{BF}^*.

If the error terms have constant variance and n_1 and n_2 are not extremely small, t_{BF}^* follows approximately the t distribution with $n - 2$ degrees of freedom. Large absolute values of t_{BF}^* indicate that the error terms do not have constant variance.

Example

We wish to use the Brown-Forsythe test for the Toluca Company example to determine whether or not the error term variance varies with the level of X. Since the X levels are spread fairly uniformly (see Figure 3.1a), we divide the 25 cases into two groups with approximately equal X ranges. The first group consists of the 13 runs with lot sizes from 20 to 70. The second group consists of the 12 runs with lot sizes from 80 to 120. Table 3.3

TABLE 3.3 Calculations for Brown-Forsythe Test for Constancy of Error Variance—Toluca Company Example.

Group 1

i	Run	(1) Lot Size	(2) Residual e_{i1}	(3) d_{i1}	(4) $(d_{i1} - \bar{d}_1)^2$
1	14	20	−20.77	.89	1,929.41
2	2	30	−48.47	28.59	263.25
...	...	...	...	...	...
12	12	70	−60.28	40.40	19.49
13	25	70	10.72	30.60	202.07
	Total			582.60	12,566.6

$\tilde{e}_1 = -19.88 \qquad \bar{d}_1 = 44.815$

Group 2

i	Run	(1) Lot Size	(2) Residual e_{i2}	(3) d_{i2}	(4) $(d_{i2} - \bar{d}_2)^2$
1	1	80	51.02	53.70	637.56
2	8	80	4.02	6.70	473.06
...	...	...	...	...	...
11	20	110	−34.09	31.41	8.76
12	7	120	55.21	57.89	866.71
	Total			341.40	9,610.2

$\tilde{e}_2 = -2.68 \qquad \bar{d}_2 = 28.450$

presents a portion of the data for each group. In columns 1 and 2 are repeated the lot sizes and residuals from Table 1.2. We see from Table 3.3 that the median residual is $\tilde{e}_1 = -19.88$ for group 1 and $\tilde{e}_2 = -2.68$ for group 2. Column 3 contains the absolute deviations of the residuals around their respective group medians. For instance, we obtain:

$$d_{11} = |e_{11} - \tilde{e}_1| = |-20.77 - (-19.88)| = .89$$
$$d_{12} = |e_{12} - \tilde{e}_2| = |51.02 - (-2.68)| = 53.70$$

The means of the absolute deviations are obtained in the usual fashion:

$$\bar{d}_1 = \frac{582.60}{13} = 44.815 \qquad \bar{d}_2 = \frac{341.40}{12} = 28.450$$

Finally, column 4 contains the squares of the deviations of the d_{i1} and d_{i2} around their respective group means. For instance, we have:

$$(d_{11} - \bar{d}_1)^2 = (.89 - 44.815)^2 = 1{,}929.41$$
$$(d_{12} - \bar{d}_2)^2 = (53.70 - 28.450)^2 = 637.56$$

We are now ready to calculate test statistic (3.9):

$$s^2 = \frac{12{,}566.6 + 9{,}610.2}{25 - 2} = 964.21$$
$$s = 31.05$$
$$t_{BF}^* = \frac{44.815 - 28.450}{31.05\sqrt{\frac{1}{13} + \frac{1}{12}}} = 1.32$$

To control the α risk at .05, we require $t(.975; 23) = 2.069$. The decision rule therefore is:

If $|t_{BF}^*| \leq 2.069$, conclude the error variance is constant

If $|t_{BF}^*| > 2.069$, conclude the error variance is not constant

Since $|t_{BF}^*| = 1.32 \leq 2.069$, we conclude that the error variance is constant and does not vary with the level of X. The two-sided P-value of this test is .20.

Comments

1. If the data set contains many cases, the two-sample t test for constancy of error variance can be conducted after dividing the cases into three or four groups, according to the level of X, and using the two extreme groups.

2. A robust test for constancy of the error variance is desirable because nonnormality and lack of constant variance often go hand in hand. For example, the distribution of the error terms may become increasingly skewed and hence more variable with increasing levels of X. ■

Breusch-Pagan Test

A second test for the constancy of the error variance is the Breusch-Pagan test (Ref. 3.6). This test, a large-sample test, assumes that the error terms are independent and normally distributed and that the variance of the error term ε_i, denoted by σ_i^2, is related to the level

of X in the following way:

$$\log_e \sigma_i^2 = \gamma_0 + \gamma_1 X_i \tag{3.10}$$

Note that (3.10) implies that σ_i^2 either increases or decreases with the level of X, depending on the sign of γ_1. Constancy of error variance corresponds to $\gamma_1 = 0$. The test of H_0: $\gamma_1 = 0$ versus H_a: $\gamma_1 \neq 0$ is carried out by means of regressing the squared residuals e_i^2 against X_i in the usual manner and obtaining the regression sum of squares, to be denoted by SSR^*. The test statistic X_{BP}^2 is as follows:

$$X_{BP}^2 = \frac{SSR^*}{2} \div \left(\frac{SSE}{n}\right)^2 \tag{3.11}$$

where SSR^* is the regression sum of squares when regressing e^2 on X and SSE is the error sum of squares when regressing Y on X. If H_0: $\gamma_1 = 0$ holds and n is reasonably large, X_{BP}^2 follows approximately the chi-square distribution with one degree of freedom. Large values of X_{BP}^2 lead to conclusion H_a, that the error variance is not constant.

Example

To conduct the Breusch-Pagan test for the Toluca Company example, we regress the squared residuals in Table 1.2, column 5, against X and obtain $SSR^* = 7{,}896{,}128$. We know from Figure 2.2 that $SSE = 54{,}825$. Hence, test statistic (3.11) is:

$$X_{BP}^2 = \frac{7{,}896{,}128}{2} \div \left(\frac{54{,}825}{25}\right)^2 = .821$$

To control the α risk at .05, we require $\chi^2(.95; 1) = 3.84$. Since $X_{BP}^2 = .821 \leq 3.84$, we conclude H_0, that the error variance is constant. The P-value of this test is .64 so that the data are quite consistent with constancy of the error variance.

Comments

1. The Breusch-Pagan test can be modified to allow for different relationships between the error variance and the level of X than the one in (3.10).
2. Test statistic (3.11) was developed independently by Cook and Weisberg (Ref. 3.7), and the test is sometimes referred to as the Cook-Weisberg test. ■

3.7 *F* Test for Lack of Fit

We next take up a formal test for determining whether a specific type of regression function adequately fits the data. We illustrate this test for ascertaining whether a linear regression function is a good fit for the data.

Assumptions

The lack of fit test assumes that the observations Y for given X are (1) independent and (2) normally distributed, and that (3) the distributions of Y have the same variance σ^2.

The lack of fit test requires repeat observations at one or more X levels. In nonexperimental data, these may occur fortuitously, as when in a productivity study relating workers' output and age, several workers of the same age happen to be included in the study. In an experiment, one can assure by design that there are repeat observations. For instance, in an

experiment on the effect of size of salesperson bonus on sales, three salespersons can be offered a particular size of bonus, for each of six bonus sizes, and their sales then observed.

Repeat trials for the same level of the predictor variable, of the type described, are called *replications*. The resulting observations are called *replicates*.

Example

In an experiment involving 12 similar but scattered suburban branch offices of a commercial bank, holders of checking accounts at the offices were offered gifts for setting up money market accounts. Minimum initial deposits in the new money market account were specified to qualify for the gift. The value of the gift was directly proportional to the specified minimum deposit. Various levels of minimum deposit and related gift values were used in the experiment in order to ascertain the relation between the specified minimum deposit and gift value, on the one hand, and number of accounts opened at the office, on the other. Altogether, six levels of minimum deposit and proportional gift value were used, with two of the branch offices assigned at random to each level. One branch office had a fire during the period and was dropped from the study. Table 3.4a contains the results, where X is the amount of minimum deposit and Y is the number of new money market accounts that were opened and qualified for the gift during the test period.

A linear regression function was fitted in the usual fashion; it is:

$$\hat{Y} = 50.72251 + .48670X$$

The analysis of variance table also was obtained and is shown in Table 3.4b. A scatter plot, together with the fitted regression line, is shown in Figure 3.11. The indications are strong that a linear regression function is inappropriate. To test this formally, we shall use the general linear test approach described in Section 2.8.

TABLE 3.4 Data and Analysis of Variance Table—Bank Example.

(a) Data

Branch i	Size of Minimum Deposit (dollars) X_i	Number of New Accounts Y_i	Branch i	Size of Minimum Deposit (dollars) X_i	Number of New Accounts Y_i
1	125	160	7	75	42
2	100	112	8	175	124
3	200	124	9	125	150
4	75	28	10	200	104
5	150	152	11	100	136
6	175	156			

(b) ANOVA Table

Source of Variation	*SS*	*df*	*MS*
Regression	5,141.3	1	5,141.3
Error	14,741.6	9	1,638.0
Total	19,882.9	10	

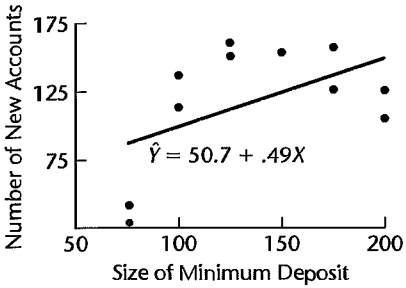


FIGURE 3.11 Scatter Plot and Fitted Regression Line—Bank Example.

TABLE 3.5 Data Arranged by Replicate Number and Minimum Deposit—Bank Example.

	Size of Minimum Deposit (dollars)					
Replicate	$j=1$ $X_1=75$	$j=2$ $X_2=100$	$j=3$ $X_3=125$	$j=4$ $X_4=150$	$j=5$ $X_5=175$	$j=6$ $X_6=200$
$i=1$	28	112	160	152	156	124
$i=2$	42	136	150		124	104
Mean $\bar{Y}_j$	35	124	155	152	140	114

Notation

First, we need to modify our notation to recognize the existence of replications at some levels of X. Table 3.5 presents the same data as Table 3.4a, but in an arrangement that recognizes the replicates. We shall denote the different X levels in the study, whether or not replicated observations are present, as $X_1, \ldots, X_c$. For the bank example, $c = 6$ since there are six minimum deposit size levels in the study, for five of which there are two observations and for one there is a single observation. We shall let $X_1 = 75$ (the smallest minimum deposit level), $X_2 = 100, \ldots, X_6 = 200$. Further, we shall denote the number of replicates for the jth level of X as n_j; for our example, $n_1 = n_2 = n_3 = n_5 = n_6 = 2$ and $n_4 = 1$. Thus, the total number of observations n is given by:

$$n = \sum_{j=1}^{c} n_j \tag{3.12}$$

We shall denote the observed value of the response variable for the ith replicate for the jth level of X by Y_{ij}, where $i = 1, \ldots, n_j$, $j = 1, \ldots, c$. For the bank example (Table 3.5), $Y_{11} = 28$, $Y_{21} = 42$, $Y_{12} = 112$, and so on. Finally, we shall denote the mean of the Y observations at the level $X = X_j$ by $\bar{Y}_j$. Thus, $\bar{Y}_1 = (28 + 42)/2 = 35$ and $\bar{Y}_4 = 152/1 = 152$.

Full Model

The general linear test approach begins with the specification of the full model. The full model used for the lack of fit test makes the same assumptions as the simple linear regression model (2.1) except for assuming a linear regression relation, the subject of the test. This full model is:

$$Y_{ij} = \mu_j + \varepsilon_{ij} \qquad \text{Full model} \tag{3.13}$$

where:

μ_j are parameters $j = 1, \ldots, c$
ε_{ij} are independent $N(0, \sigma^2)$

Since the error terms have expectation zero, it follows that:

$$E\{Y_{ij}\} = \mu_j \tag{3.14}$$

Thus, the parameter μ_j ($j = 1, \ldots, c$) is the mean response when $X = X_j$.

The full model (3.13) is like the regression model (2.1) in stating that each response Y is made up of two components: the mean response when $X = X_j$ and a random error term. The difference between the two models is that in the full model (3.13) there are no restrictions on the means μ_j, whereas in the regression model (2.1) the mean responses are linearly related to X (i.e., $E\{Y\} = \beta_0 + \beta_1 X$).

To fit the full model to the data, we require the least squares or maximum likelihood estimators for the parameters μ_j. It can be shown that these estimators of μ_j are simply the sample means $\bar{Y}_j$:

$$\hat{\mu}_j = \bar{Y}_j \tag{3.15}$$

Thus, the estimated expected value for observation Y_{ij} is $\bar{Y}_j$, and the error sum of squares for the full model therefore is:

$$SSE(F) = \sum_j \sum_i (Y_{ij} - \bar{Y}_j)^2 = SSPE \tag{3.16}$$

In the context of the test for lack of fit, the full model error sum of squares (3.16) is called the *pure error sum of squares* and is denoted by *SSPE*.

Note that *SSPE* is made up of the sums of squared deviations at each X level. At level $X = X_j$, this sum of squared deviations is:

$$\sum_i (Y_{ij} - \bar{Y}_j)^2 \tag{3.17}$$

These sums of squares are then added over all of the X levels ($j = 1, \ldots, c$). For the bank example, we have:

$$\begin{aligned} SSPE &= (28 - 35)^2 + (42 - 35)^2 + (112 - 124)^2 + (136 - 124)^2 + (160 - 155)^2 \\ &\quad + (150 - 155)^2 + (152 - 152)^2 + (156 - 140)^2 + (124 - 140)^2 \\ &\quad + (124 - 114)^2 + (104 - 114)^2 \\ &= 1{,}148 \end{aligned}$$

Note that any X level with no replications makes no contribution to *SSPE* because $\bar{Y}_j = Y_{1j}$ then. Thus, $(152 - 152)^2 = 0$ for $j = 4$ in the bank example.

The degrees of freedom associated with *SSPE* can be obtained by recognizing that the sum of squared deviations (3.17) at a given level of X is like an ordinary total sum of squares based on n observations, which has $n - 1$ degrees of freedom associated with it. Here, there are n_j observations when $X = X_j$; hence the degrees of freedom are $n_j - 1$. Just as *SSPE* is the sum of the sums of squares (3.17), so the number of degrees of freedom associated

with *SSPE* is the sum of the component degrees of freedom:

$$df_F = \sum_j (n_j - 1) = \sum_j n_j - c = n - c \tag{3.18}$$

For the bank example, we have $df_F = 11 - 6 = 5$. Note that any X level with no replications makes no contribution to df_F because $n_j - 1 = 1 - 1 = 0$ then, just as such an X level makes no contribution to *SSPE*.

Reduced Model

The general linear test approach next requires consideration of the reduced model under H_0. For testing the appropriateness of a linear regression relation, the alternatives are:

$$\begin{aligned} H_0&: E\{Y\} = \beta_0 + \beta_1 X \\ H_a&: E\{Y\} \neq \beta_0 + \beta_1 X \end{aligned} \tag{3.19}$$

Thus, H_0 postulates that μ_j in the full model (3.13) is linearly related to X_j:

$$\mu_j = \beta_0 + \beta_1 X_j$$

The reduced model under H_0 therefore is:

$$Y_{ij} = \beta_0 + \beta_1 X_j + \varepsilon_{ij} \qquad \text{Reduced model} \tag{3.20}$$

Note that the reduced model is the ordinary simple linear regression model (2.1), with the subscripts modified to recognize the existence of replications. We know that the estimated expected value for observation Y_{ij} with regression model (2.1) is the fitted value $\hat{Y}_{ij}$:

$$\hat{Y}_{ij} = b_0 + b_1 X_j \tag{3.21}$$

Hence, the error sum of squares for the reduced model is the usual error sum of squares *SSE*:

$$\begin{aligned} SSE(R) &= \sum\sum [Y_{ij} - (b_0 + b_1 X_j)]^2 \\ &= \sum\sum (Y_{ij} - \hat{Y}_{ij})^2 = SSE \end{aligned} \tag{3.22}$$

We also know that the degrees of freedom associated with $SSE(R)$ are:

$$df_R = n - 2$$

For the bank example, we have from Table 3.4b:

$$SSE(R) = SSE = 14{,}741.6$$
$$df_R = 9$$

Test Statistic

The general linear test statistic (2.70):

$$F^* = \frac{SSE(R) - SSE(F)}{df_R - df_F} \div \frac{SSE(F)}{df_F}$$

here becomes:

$$F^* = \frac{SSE - SSPE}{(n-2) - (n-c)} \div \frac{SSPE}{n-c} \tag{3.23}$$

The difference between the two error sums of squares is called the *lack of fit sum of squares* here and is denoted by *SSLF*:

$$SSLF = SSE - SSPE \tag{3.24}$$

We can then express the test statistic as follows:

$$F^* = \frac{SSLF}{c-2} \div \frac{SSPE}{n-c} = \frac{MSLF}{MSPE} \tag{3.25}$$

where *MSLF* denotes the *lack of fit mean square* and *MSPE* denotes the *pure error mean square*.

We know that large values of F^* lead to conclusion H_a in the general linear test. Decision rule (2.71) here becomes:

$$\begin{aligned} &\text{If } F^* \le F(1-\alpha; c-2, n-c), \text{ conclude } H_0 \\ &\text{If } F^* > F(1-\alpha; c-2, n-c), \text{ conclude } H_a \end{aligned} \tag{3.26}$$

For the bank example, the test statistic can be constructed easily from our earlier results:

$$\begin{aligned} SSPE &= 1{,}148.0 & n-c &= 11-6=5 \\ SSE &= 14{,}741.6 \\ SSLF &= 14{,}741.6 - 1{,}148.0 = 13{,}593.6 & c-2 &= 6-2=4 \end{aligned}$$

$$F^* = \frac{13{,}593.6}{4} \div \frac{1{,}148.0}{5} = \frac{3{,}398.4}{229.6} = 14.80$$

If the level of significance is to be $\alpha = .01$, we require $F(.99; 4, 5) = 11.4$. Since $F^* = 14.80 > 11.4$, we conclude H_a, that the regression function is not linear. This, of course, accords with our visual impression from Figure 3.11. The *P*-value for the test is .006.

ANOVA Table

The definition of the lack of fit sum of squares *SSLF* in (3.24) indicates that we have, in fact, decomposed the error sum of squares *SSE* into two components:

$$SSE = SSPE + SSLF \tag{3.27}$$

This decomposition follows from the identity:

$$\underbrace{Y_{ij} - \hat{Y}_{ij}}_{\substack{\text{Error} \\ \text{deviation}}} = \underbrace{Y_{ij} - \bar{Y}_j}_{\substack{\text{Pure error} \\ \text{deviation}}} + \underbrace{\bar{Y}_j - \hat{Y}_{ij}}_{\substack{\text{Lack of fit} \\ \text{deviation}}} \tag{3.28}$$

This identity shows that the error deviations in *SSE* are made up of a pure error component and a lack of fit component. Figure 3.12 illustrates this partitioning for the case $Y_{13} = 160$, $X_3 = 125$ in the bank example.

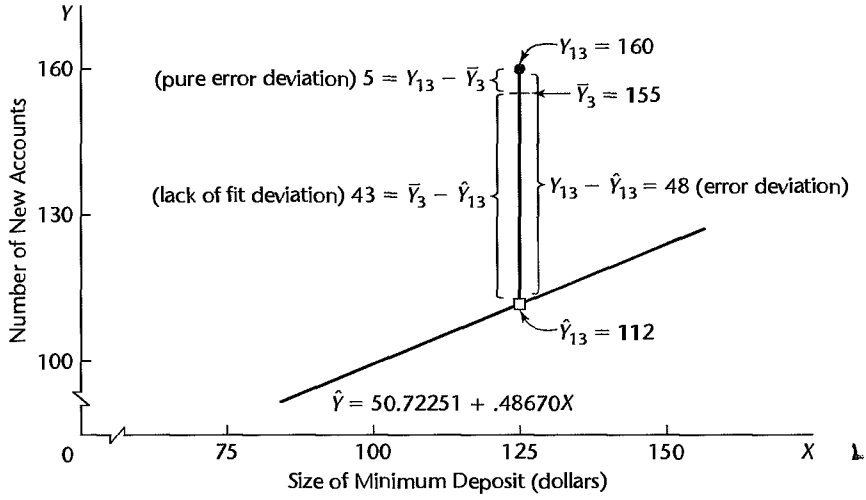


FIGURE 3.12 **Illustration of Decomposition of Error Deviation $Y_{ij} - \hat{Y}_{ij}$—Bank Example.**

When (3.28) is squared and summed over all observations, we obtain (3.27) since the cross-product sum equals zero:

$$\underset{SSE}{\sum\sum(Y_{ij} - \hat{Y}_{ij})^2} = \underset{SSPE}{\sum\sum(Y_{ij} - \bar{Y}_j)^2} + \underset{SSLF}{\sum\sum(\bar{Y}_j - \hat{Y}_{ij})^2} \qquad (3.29)$$

Note from (3.29) that we can define the lack of fit sum of squares directly as follows:

$$SSLF = \sum\sum(\bar{Y}_j - \hat{Y}_{ij})^2 \qquad (3.30)$$

Since all Y_{ij} observations at the level X_j have the same fitted value, which we can denote by $\hat{Y}_j$, we can express (3.30) equivalently as:

$$SSLF = \sum_j n_j(\bar{Y}_j - \hat{Y}_j)^2 \qquad (3.30a)$$

Formula (3.30a) indicates clearly why *SSLF* measures lack of fit. If the linear regression function is appropriate, then the means $\bar{Y}_j$ will be near the fitted values $\hat{Y}_j$ calculated from the estimated linear regression function and *SSLF* will be small. On the other hand, if the linear regression function is not appropriate, the means $\bar{Y}_j$ will not be near the fitted values calculated from the estimated linear regression function, as in Figure 3.11 for the bank example, and *SSLF* will be large.

Formula (3.30a) also indicates why $c - 2$ degrees of freedom are associated with *SSLF*. There are c means $\bar{Y}_j$ in the sum of squares, and two degrees of freedom are lost in estimating the parameters β_0 and β_1 of the linear regression function to obtain the fitted values $\hat{Y}_j$.

An ANOVA table can be constructed for the decomposition of *SSE*. Table 3.6a contains the general ANOVA table, including the decomposition of *SSE* just explained and the mean squares of interest, and Table 3.6b contains the ANOVA decomposition for the bank example.

TABLE 3.6 General ANOVA Table for Testing Lack of Fit of Simple Linear Regression Function and ANOVA Table—Bank Example.

(a) General

Source of Variation	*SS*	*df*	*MS*
Regression	$SSR = \sum\sum(\hat{Y}_{ij} - \bar{Y})^2$	1	$MSR = \frac{SSR}{1}$
Error	$SSE = \sum\sum(Y_{ij} - \hat{Y}_{ij})^2$	$n-2$	$MSE = \frac{SSE}{n-2}$
Lack of fit	$SSLF = \sum\sum(\bar{Y}_j - \hat{Y}_{ij})^2$	$c-2$	$MSLF = \frac{SSLF}{c-2}$
Pure error	$SSPE = \sum\sum(Y_{ij} - \bar{Y}_j)^2$	$n-c$	$MSPE = \frac{SSPE}{n-c}$
Total	$SSTO = \sum\sum(Y_{ij} - \bar{Y})^2$	$n-1$	

(b) Bank Example

Source of Variation	*SS*	*df*	*MS*
Regression	5,141.3	1	5,141.3
Error	14,741.6	9	1,638.0
Lack of fit	13,593.6	4	3,398.4
Pure error	1,148.0	5	229.6
Total	19,882.9	10	

Comments

1. As shown by the bank example, not all levels of X need have repeat observations for the F test for lack of fit to be applicable. Repeat observations at only one or some levels of X are sufficient.

2. It can be shown that the mean squares *MSPE* and *MSLF* have the following expectations when testing whether the regression function is linear:

$$E\{MSPE\} = \sigma^2 \tag{3.31}$$

$$E\{MSLF\} = \sigma^2 + \frac{\sum n_j[\mu_j - (\beta_0 + \beta_1 X_j)]^2}{c-2} \tag{3.32}$$

The reason for the term "pure error" is that *MSPE* is always an unbiased estimator of the error term variance σ^2, no matter what is the true regression function. The expected value of *MSLF* also is σ^2 if the regression function is linear, because $\mu_j = \beta_0 + \beta_1 X_j$ then and the second term in (3.32) becomes zero. On the other hand, if the regression function is not linear, $\mu_j \neq \beta_0 + \beta_1 X_j$ and $E\{MSLF\}$ will be greater than σ^2. Hence, a value of F^* near 1 accords with a linear regression function; large values of F^* indicate that the regression function is not linear.

3. The terminology "error sum of squares" and "error mean square" is not precise when the regression function under test in H_0 is not the true function since the error sum of squares and error mean square then reflect the effects of both the lack of fit and the variability of the error terms. We continue to use the terminology for consistency and now use the term "pure error" to identify the variability associated with the error term only.

4. Suppose that prior to any analysis of the appropriateness of the model, we had fitted a linear regression model and wished to test whether or not $\beta_1 = 0$ for the bank example (Table 3.4b). Test statistic (2.60) would be:

$$F^* = \frac{MSR}{MSE} = \frac{5{,}141.3}{1{,}638.0} = 3.14$$

For $\alpha = .10$, $F(.90; 1, 9) = 3.36$, and we would conclude H_0, that $\beta_1 = 0$ or that there is no *linear association* between minimum deposit size (and value of gift) and number of new accounts. A conclusion that there is no *relation* between these variables would be improper, however. Such an inference requires that regression model (2.1) be appropriate. Here, there is a definite relationship, but the regression function is not linear. This illustrates the importance of always examining the appropriateness of a model before any inferences are drawn.

5. The general linear test approach just explained can be used to test the appropriateness of other regression functions. Only the degrees of freedom for *SSLF* will need be modified. In general, $c - p$ degrees of freedom are associated with *SSLF*, where p is the number of parameters in the regression function. For the test of a simple linear regression function, $p = 2$ because there are two parameters, β_0 and β_1, in the regression function.

6. The alternative H_a in (3.19) includes all regression functions other than a linear one. For instance, it includes a quadratic regression function or a logarithmic one. If H_a is concluded, a study of residuals can be helpful in identifying an appropriate function.

7. When we conclude that the employed model in H_0 is appropriate, the usual practice is to use the error mean square *MSE* as an estimator of σ^2 in preference to the pure error mean square *MSPE*, since the former contains more degrees of freedom.

8. Observations at the same level of X are genuine repeats only if they involve independent trials with respect to the error term. Suppose that in a regression analysis of the relation between hardness (Y) and amount of carbon (X) in specimens of an alloy, the error term in the model covers, among other things, random errors in the measurement of hardness by the analyst and effects of uncontrolled production factors, which vary at random from specimen to specimen and affect hardness. If the analyst takes two readings on the hardness of a specimen, this will not provide a genuine replication because the effects of random variation in the production factors are fixed in any given specimen. For genuine replications, different specimens with the same carbon content (X) would have to be measured by the analyst so that *all* the effects covered in the error term could vary at random from one repeated observation to the next.

9. When no replications are present in a data set, an approximate test for lack of fit can be conducted if there are some cases at adjacent X levels for which the mean responses are quite close to each other. Such adjacent cases are grouped together and treated as pseudoreplicates, and the test for lack of fit is then carried out using these groupings of adjacent cases. A useful summary of this and related procedures for conducting a test for lack of fit when no replicates are present may be found in Reference 3.8. ■

3.8 Overview of Remedial Measures

If the simple linear regression model (2.1) is not appropriate for a data set, there are two basic choices:

1. Abandon regression model (2.1) and develop and use a more appropriate model.
2. Employ some transformation on the data so that regression model (2.1) is appropriate for the transformed data.

Each approach has advantages and disadvantages. The first approach may entail a more complex model that could yield better insights, but may also lead to more complex procedures for estimating the parameters. Successful use of transformations, on the other hand, leads to relatively simple methods of estimation and may involve fewer parameters than a complex model, an advantage when the sample size is small. Yet transformations may obscure the fundamental interconnections between the variables, though at other times they may illuminate them.

We consider the use of transformations in this chapter and the use of more complex models in later chapters. First, we provide a brief overview of remedial measures.

Nonlinearity of Regression Function

When the regression function is not linear, a direct approach is to modify regression model (2.1) by altering the nature of the regression function. For instance, a quadratic regression function might be used:

$$E\{Y\} = \beta_0 + \beta_1 X + \beta_2 X^2$$

or an exponential regression function:

$$E\{Y\} = \beta_0 \beta_1^X$$

In Chapter 7, we discuss polynomial regression functions, and in Part III we take up nonlinear regression functions, such as an exponential regression function.

The transformation approach employs a transformation to linearize, at least approximately, a nonlinear regression function. We discuss the use of transformations to linearize regression functions in Section 3.9.

When the nature of the regression function is not known, exploratory analysis that does not require specifying a particular type of function is often useful. We discuss exploratory regression analysis in Section 3.10.

Nonconstancy of Error Variance

When the error variance is not constant but varies in a systematic fashion, a direct approach is to modify the model to allow for this and use the method of *weighted least squares* to obtain the estimators of the parameters. We discuss the use of weighted least squares for this purpose in Chapter 11.

Transformations can also be effective in stabilizing the variance. Some of these are discussed in Section 3.9.

Nonindependence of Error Terms

When the error terms are correlated, a direct remedial measure is to work with a model that calls for correlated error terms. We discuss such a model in Chapter 12. A simple remedial transformation that is often helpful is to work with first differences, a topic also discussed in Chapter 12.

Nonnormality of Error Terms

Lack of normality and nonconstant error variances frequently go hand in hand. Fortunately, it is often the case that the same transformation that helps stabilize the variance is also helpful in approximately normalizing the error terms. It is therefore desirable that the transformation

for stabilizing the error variance be utilized first, and then the residuals studied to see if serious departures from normality are still present. We discuss transformations to achieve approximate normality in Section 3.9.

Omission of Important Predictor Variables

When residual analysis indicates that an important predictor variable has been omitted from the model, the solution is to modify the model. In Chapter 6 and later chapters, we discuss multiple regression analysis in which two or more predictor variables are utilized.

Outlying Observations

When outlying observations are present, as in Figure 3.7a, use of the least squares and maximum likelihood estimators (1.10) for regression model (2.1) may lead to serious distortions in the estimated regression function. When the outlying observations do not represent recording errors and should not be discarded, it may be desirable to use an estimation procedure that places less emphasis on such outlying observations. We discuss one such robust estimation procedure in Chapter 11.

3.9 Transformations

We now consider in more detail the use of transformations of one or both of the original variables before carrying out the regression analysis. Simple transformations of either the response variable Y or the predictor variable X, or of both, are often sufficient to make the simple linear regression model appropriate for the transformed data.

Transformations for Nonlinear Relation Only

We first consider transformations for linearizing a nonlinear regression relation when the distribution of the error terms is reasonably close to a normal distribution and the error terms have approximately constant variance. In this situation, transformations on X should be attempted. The reason why transformations on Y may not be desirable here is that a transformation on Y, such as $Y' = \sqrt{Y}$, may materially change the shape of the distribution of the error terms from the normal distribution and may also lead to substantially differing error term variances.

Figure 3.13 contains some prototype nonlinear regression relations with constant error variance and also presents some simple transformations on X that may be helpful to linearize the regression relationship without affecting the distributions of Y. Several alternative transformations may be tried. Scatter plots and residual plots based on each transformation should then be prepared and analyzed to decide which transformation is most effective.

Example

Data from an experiment on the effect of number of days of training received (X) on performance (Y) in a battery of simulated sales situations are presented in Table 3.7, columns 1 and 2, for the 10 participants in the study. A scatter plot of these data is shown in Figure 3.14a. Clearly the regression relation appears to be curvilinear, so the simple linear regression model (2.1) does not seem to be appropriate. Since the variability at the different X levels appears to be fairly constant, we shall consider a transformation on X. Based on the prototype plot in Figure 3.13a, we shall consider initially the square root transformation $X' = \sqrt{X}$. The transformed values are shown in column 3 of Table 3.7.

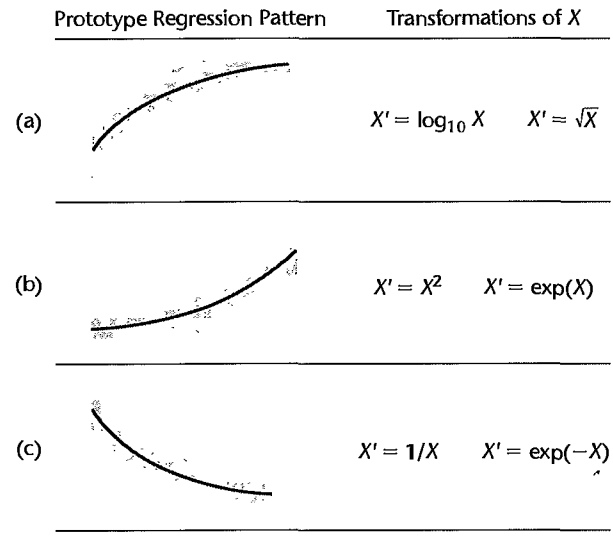


FIGURE 3.13 **Prototype Nonlinear Regression Patterns with Constant Error Variance and Simple Transformations of X.**

TABLE 3.7 **Use of Square Root Transformation of X to Linearize Regression Relation—Sales Training Example.**

Sales Trainee i	(1) Days of Training X_i	(2) Performance Score Y_i	(3) $X_i' = \sqrt{X_i}$
1	.5	42.5	.70711
2	.5	50.6	.70711
3	1.0	68.5	1.00000
4	1.0	80.7	1.00000
5	1.5	89.0	1.22474
6	1.5	99.6	1.22474
7	2.0	105.3	1.41421
8	2.0	111.8	1.41421
9	2.5	112.3	1.58114
10	2.5	125.7	1.58114

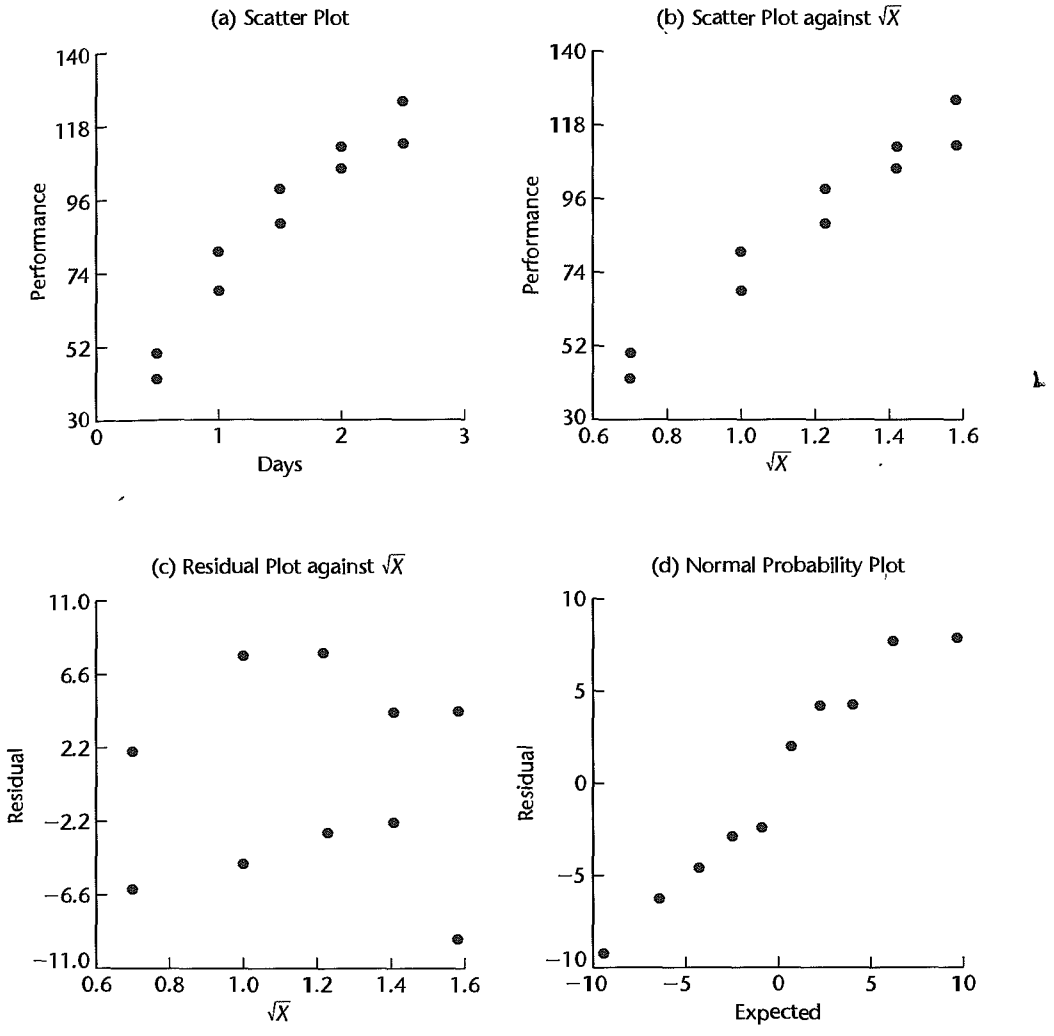
In Figure 3.14b, the same data are plotted with the predictor variable transformed to $X' = \sqrt{X}$. Note that the scatter plot now shows a reasonably linear relation. The variability of the scatter at the different X levels is the same as before, since we did not make a transformation on Y.

To examine further whether the simple linear regression model (2.1) is appropriate now, we fit it to the transformed X data. The regression calculations with the transformed X data are carried out in the usual fashion, except that the predictor variable now is X'. We obtain the following fitted regression function:

$$\hat{Y} = -10.33 + 83.45X'$$

Figure 3.14c contains a plot of the residuals against X'. There is no evidence of lack of fit or of strongly unequal error variances. Figure 3.14d contains a normal probability plot of

FIGURE 3.14 Scatter Plots and Residual Plots—Sales Training Example.



the residuals. No strong indications of substantial departures from normality are indicated by this plot. This conclusion is supported by the high coefficient of correlation between the ordered residuals and their expected values under normality, .979. For $\alpha = .01$, Table B.6 shows that the critical value is .879, so the observed coefficient is substantially larger and supports the reasonableness of normal error terms. Thus, the simple linear regression model (2.1) appears to be appropriate here for the transformed data.

The fitted regression function in the original units of X can easily be obtained, if desired:

$$\hat{Y} = -10.33 + 83.45\sqrt{X}$$

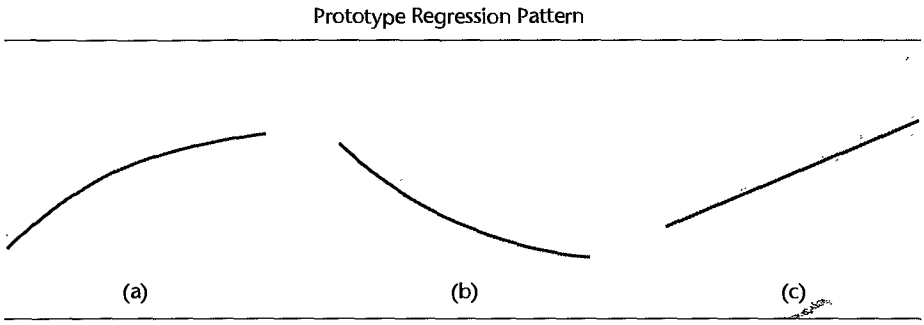


FIGURE 3.15
Prototype Regression Patterns with Unequal Error Variances and Simple Transformations of *Y*.

Transformations on Y

$Y' = \sqrt{Y}$

$Y' = \log_{10} Y$

$Y' = 1/Y$

Note: A simultaneous transformation on X may also be helpful or necessary.

Comment

At times, it may be helpful to introduce a constant into the transformation. For example, if some of the X data are near zero and the reciprocal transformation is desired, we can shift the origin by using the transformation $X' = 1/(X + k)$, where k is an appropriately chosen constant. ■

Transformations for Nonnormality and Unequal Error Variances

Unequal error variances and nonnormality of the error terms frequently appear together. To remedy these departures from the simple linear regression model (2.1), we need a transformation on Y, since the shapes and spreads of the distributions of Y need to be changed. Such a transformation on Y may also at the same time help to linearize a curvilinear regression relation. At other times, a simultaneous transformation on X may be needed to obtain or maintain a linear regression relation.

Frequently, the nonnormality and unequal variances departures from regression model (2.1) take the form of increasing skewness and increasing variability of the distributions of the error terms as the mean response $E\{Y\}$ increases. For example, in a regression of yearly household expenditures for vacations (Y) on household income (X), there will tend to be more variation and greater positive skewness (i.e., some very high yearly vacation expenditures) for high-income households than for low-income households, who tend to consistently spend much less for vacations. Figure 3.15 contains some prototype regression relations where the skewness and the error variance increase with the mean response $E\{Y\}$. This figure also presents some simple transformations on Y that may be helpful for these cases. Several alternative transformations on Y may be tried, as well as some simultaneous transformations on X. Scatter plots and residual plots should be prepared to determine the most effective transformation(s).

Example

Data on age (X) and plasma level of a polyamine (Y) for a portion of the 25 healthy children in a study are presented in columns 1 and 2 of Table 3.8. These data are plotted in Figure 3.16a as a scatter plot. Note the distinct curvilinear regression relationship, as well as the greater variability for younger children than for older ones.

TABLE 3.8 Use of Logarithmic Transformation of Y to Linearize Regression Relation and Stabilize Error Variance—Plasma Levels Example.

Child i	(1) Age X_i	(2) Plasma Level Y_i	(3) $Y_i' = \log_{10} Y_i$
1	0 (newborn)	13.44	1.1284
2	0 (newborn)	12.84	1.1086
3	0 (newborn)	11.91	1.0759
4	0 (newborn)	20.09	1.3030
5	0 (newborn)	15.60	1.1931
6	1.0	10.11	1.0048
7	1.0	11.38	1.0561
...	...	...	...
19	3.0	6.90	.8388
20	3.0	6.77	.8306
21	4.0	4.86	.6866
22	4.0	5.10	.7076
23	4.0	5.67	.7536
24	4.0	5.75	.7597
25	4.0	6.23	.7945

On the basis of the prototype regression pattern in Figure 3.15b, we shall first try the logarithmic transformation $Y' = \log_{10} Y$. The transformed Y values are shown in column 3 of Table 3.8. Figure 3.16b contains the scatter plot with this transformation. Note that the transformation not only has led to a reasonably linear regression relation, but the variability at the different levels of X also has become reasonably constant.

To further examine the reasonableness of the transformation $Y' = \log_{10} Y$, we fitted the simple linear regression model (2.1) to the transformed Y data and obtained:

$$\hat{Y}' = 1.135 - .1023X$$

A plot of the residuals against X is shown in Figure 3.16c, and a normal probability plot of the residuals is shown in Figure 3.16d. The coefficient of correlation between the ordered residuals and their expected values under normality is .981. For $\alpha = .05$, Table B.6 indicates that the critical value is .959 so that the observed coefficient supports the assumption of normality of the error terms. All of this evidence supports the appropriateness of regression model (2.1) for the transformed Y data.

Comments

1. At times it may be desirable to introduce a constant into a transformation of Y, such as when Y may be negative. For instance, the logarithmic transformation to shift the origin in Y and make all Y observations positive would be $Y' = \log_{10}(Y + k)$, where k is an appropriately chosen constant.

2. When unequal error variances are present but the regression relation is linear, a transformation on Y may not be sufficient. While such a transformation may stabilize the error variance, it will also change the linear relationship to a curvilinear one. A transformation on X may therefore also be required. This case can also be handled by using weighted least squares, a procedure explained in Chapter 11. ■

The difference between the two error sums of squares is called the *lack of fit sum of squares* here and is denoted by *SSLF*:

$$SSLF = SSE - SSPE \tag{3.24}$$

We can then express the test statistic as follows:

$$\begin{aligned} F^* &= \frac{SSLF}{c-2} \div \frac{SSPE}{n-c} \\ &= \frac{MSLF}{MSPE} \end{aligned} \tag{3.25}$$

where *MSLF* denotes the *lack of fit mean square* and *MSPE* denotes the *pure error mean square.*

We know that large values of F^* lead to conclusion H_a in the general linear test. Decision rule (2.71) here becomes:

$$\begin{aligned} &\text{If } F^* \leq F(1-\alpha; c-2, n-c), \text{ conclude } H_0 \\ &\text{If } F^* > F(1-\alpha; c-2, n-c), \text{ conclude } H_a \end{aligned} \tag{3.26}$$

For the bank example, the test statistic can be constructed easily from our earlier results:

$$SSPE = 1{,}148.0 \qquad n-c = 11-6 = 5$$

$$SSE = 14{,}741.6$$

$$SSLF = 14{,}741.6 - 1{,}148.0 = 13{,}593.6 \qquad c-2 = 6-2 = 4$$

$$F^* = \frac{13{,}593.6}{4} \div \frac{1{,}148.0}{5}$$

$$= \frac{3{,}398.4}{229.6} = 14.80$$

If the level of significance is to be $\alpha = .01$, we require $F(.99; 4, 5) = 11.4$. Since $F^* = 14.80 > 11.4$, we conclude H_a, that the regression function is not linear. This, of course, accords with our visual impression from Figure 3.11. The P-value for the test is .006.

ANOVA Table

The definition of the lack of fit sum of squares *SSLF* in (3.24) indicates that we have, in fact, decomposed the error sum of squares *SSE* into two components:

$$SSE = SSPE + SSLF \tag{3.27}$$

This decomposition follows from the identity:

$$\underbrace{Y_{ij} - \hat{Y}_{ij}}_{\substack{\text{Error} \\ \text{deviation}}} = \underbrace{Y_{ij} - \bar{Y}_j}_{\substack{\text{Pure error} \\ \text{deviation}}} + \underbrace{\bar{Y}_j - \hat{Y}_{ij}}_{\substack{\text{Lack of fit} \\ \text{deviation}}} \tag{3.28}$$

This identity shows that the error deviations in *SSE* are made up of a pure error component and a lack of fit component. Figure 3.12 illustrates this partitioning for the case $Y_{13} = 160$, $X_3 = 125$ in the bank example.

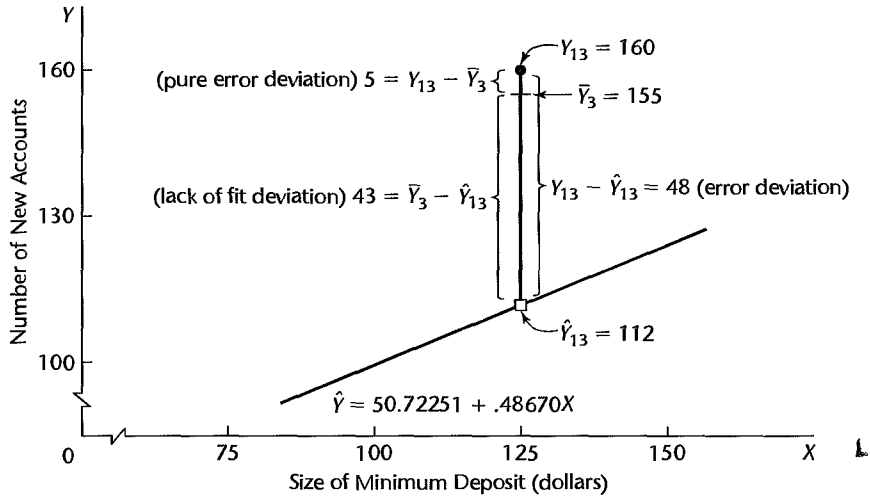


FIGURE 3.12 Illustration of Decomposition of Error Deviation $Y_{ij} - \hat{Y}_{ij}$—Bank Example.

When (3.28) is squared and summed over all observations, we obtain (3.27) since the cross-product sum equals zero:

$$\underset{SSE}{\sum\sum(Y_{ij} - \hat{Y}_{ij})^2} = \underset{SSPE}{\sum\sum(Y_{ij} - \bar{Y}_j)^2} + \underset{SSLF}{\sum\sum(\bar{Y}_j - \hat{Y}_{ij})^2} \qquad (3.29)$$

Note from (3.29) that we can define the lack of fit sum of squares directly as follows:

$$SSLF = \sum\sum(\bar{Y}_j - \hat{Y}_{ij})^2 \qquad (3.30)$$

Since all Y_{ij} observations at the level X_j have the same fitted value, which we can denote by $\hat{Y}_j$, we can express (3.30) equivalently as:

$$SSLF = \sum_j n_j(\bar{Y}_j - \hat{Y}_j)^2 \qquad (3.30a)$$

Formula (3.30a) indicates clearly why *SSLF* measures lack of fit. If the linear regression function is appropriate, then the means $\bar{Y}_j$ will be near the fitted values $\hat{Y}_j$ calculated from the estimated linear regression function and *SSLF* will be small. On the other hand, if the linear regression function is not appropriate, the means $\bar{Y}_j$ will not be near the fitted values calculated from the estimated linear regression function, as in Figure 3.11 for the bank example, and *SSLF* will be large.

Formula (3.30a) also indicates why $c - 2$ degrees of freedom are associated with *SSLF*. There are c means $\bar{Y}_j$ in the sum of squares, and two degrees of freedom are lost in estimating the parameters β_0 and β_1 of the linear regression function to obtain the fitted values $\hat{Y}_j$.

An ANOVA table can be constructed for the decomposition of *SSE*. Table 3.6a contains the general ANOVA table, including the decomposition of *SSE* just explained and the mean squares of interest, and Table 3.6b contains the ANOVA decomposition for the bank example.

TABLE 3.6 General ANOVA Table for Testing Lack of Fit of Simple Linear Regression Function and ANOVA Table—Bank Example.

(a) General

Source of Variation	*SS*	*df*	*MS*
Regression	$SSR = \sum\sum(\hat{Y}_{ij} - \bar{Y})^2$	1	$MSR = \frac{SSR}{1}$
Error	$SSE = \sum\sum(Y_{ij} - \hat{Y}_{ij})^2$	$n-2$	$MSE = \frac{SSE}{n-2}$
Lack of fit	$SSLF = \sum\sum(\bar{Y}_j - \hat{Y}_{ij})^2$	$c-2$	$MSLF = \frac{SSLF}{c-2}$
Pure error	$SSPE = \sum\sum(Y_{ij} - \bar{Y}_j)^2$	$n-c$	$MSPE = \frac{SSPE}{n-c}$
Total	$SSTO = \sum\sum(Y_{ij} - \bar{Y})^2$	$n-1$	

(b) Bank Example

Source of Variation	*SS*	*df*	*MS*
Regression	5,141.3	1	5,141.3
Error	14,741.6	9	1,638.0
Lack of fit	13,593.6	4	3,398.4
Pure error	1,148.0	5	229.6
Total	19,882.9	10	

Comments

1. As shown by the bank example, not all levels of X need have repeat observations for the F test for lack of fit to be applicable. Repeat observations at only one or some levels of X are sufficient.

2. It can be shown that the mean squares *MSPE* and *MSLF* have the following expectations when testing whether the regression function is linear:

$$E\{MSPE\} = \sigma^2 \tag{3.31}$$

$$E\{MSLF\} = \sigma^2 + \frac{\sum n_j[\mu_j - (\beta_0 + \beta_1 X_j)]^2}{c-2} \tag{3.32}$$

The reason for the term "pure error" is that *MSPE* is always an unbiased estimator of the error term variance σ^2, no matter what is the true regression function. The expected value of *MSLF* also is σ^2 if the regression function is linear, because $\mu_j = \beta_0 + \beta_1 X_j$ then and the second term in (3.32) becomes zero. On the other hand, if the regression function is not linear, $\mu_j \neq \beta_0 + \beta_1 X_j$ and $E\{MSLF\}$ will be greater than σ^2. Hence, a value of F^* near 1 accords with a linear regression function; large values of F^* indicate that the regression function is not linear.

3. The terminology "error sum of squares" and "error mean square" is not precise when the regression function under test in H_0 is not the true function since the error sum of squares and error mean square then reflect the effects of both the lack of fit and the variability of the error terms. We continue to use the terminology for consistency and now use the term "pure error" to identify the variability associated with the error term only.

4. Suppose that prior to any analysis of the appropriateness of the model, we had fitted a linear regression model and wished to test whether or not $\beta_1 = 0$ for the bank example (Table 3.4b). Test statistic (2.60) would be:

$$F^* = \frac{MSR}{MSE} = \frac{5{,}141.3}{1{,}638.0} = 3.14$$

For $\alpha = .10$, $F(.90; 1, 9) = 3.36$, and we would conclude H_0, that $\beta_1 = 0$ or that there is no *linear association* between minimum deposit size (and value of gift) and number of new accounts. A conclusion that there is no *relation* between these variables would be improper, however. Such an inference requires that regression model (2.1) be appropriate. Here, there is a definite relationship, but the regression function is not linear. This illustrates the importance of always examining the appropriateness of a model before any inferences are drawn.

5. The general linear test approach just explained can be used to test the appropriateness of other regression functions. Only the degrees of freedom for *SSLF* will need be modified. In general, $c - p$ degrees of freedom are associated with *SSLF*, where p is the number of parameters in the regression function. For the test of a simple linear regression function, $p = 2$ because there are two parameters, β_0 and β_1, in the regression function.

6. The alternative H_a in (3.19) includes all regression functions other than a linear one. For instance, it includes a quadratic regression function or a logarithmic one. If H_a is concluded, a study of residuals can be helpful in identifying an appropriate function.

7. When we conclude that the employed model in H_0 is appropriate, the usual practice is to use the error mean square *MSE* as an estimator of σ^2 in preference to the pure error mean square *MSPE*, since the former contains more degrees of freedom.

8. Observations at the same level of X are genuine repeats only if they involve independent trials with respect to the error term. Suppose that in a regression analysis of the relation between hardness (Y) and amount of carbon (X) in specimens of an alloy, the error term in the model covers, among other things, random errors in the measurement of hardness by the analyst and effects of uncontrolled production factors, which vary at random from specimen to specimen and affect hardness. If the analyst takes two readings on the hardness of a specimen, this will not provide a genuine replication because the effects of random variation in the production factors are fixed in any given specimen. For genuine replications, different specimens with the same carbon content (X) would have to be measured by the analyst so that *all* the effects covered in the error term could vary at random from one repeated observation to the next.

9. When no replications are present in a data set, an approximate test for lack of fit can be conducted if there are some cases at adjacent X levels for which the mean responses are quite close to each other. Such adjacent cases are grouped together and treated as pseudoreplicates, and the test for lack of fit is then carried out using these groupings of adjacent cases. A useful summary of this and related procedures for conducting a test for lack of fit when no replicates are present may be found in Reference 3.8. ■

3.8 Overview of Remedial Measures

If the simple linear regression model (2.1) is not appropriate for a data set, there are two basic choices:

1. Abandon regression model (2.1) and develop and use a more appropriate model.
2. Employ some transformation on the data so that regression model (2.1) is appropriate for the transformed data.

Each approach has advantages and disadvantages. The first approach may entail a more complex model that could yield better insights, but may also lead to more complex procedures for estimating the parameters. Successful use of transformations, on the other hand, leads to relatively simple methods of estimation and may involve fewer parameters than a complex model, an advantage when the sample size is small. Yet transformations may obscure the fundamental interconnections between the variables, though at other times they may illuminate them.

We consider the use of transformations in this chapter and the use of more complex models in later chapters. First, we provide a brief overview of remedial measures.

Nonlinearity of Regression Function

When the regression function is not linear, a direct approach is to modify regression model (2.1) by altering the nature of the regression function. For instance, a quadratic regression function might be used:

$$E\{Y\} = \beta_0 + \beta_1 X + \beta_2 X^2$$

or an exponential regression function:

$$E\{Y\} = \beta_0 \beta_1^X$$

In Chapter 7, we discuss polynomial regression functions, and in Part III we take up nonlinear regression functions, such as an exponential regression function.

The transformation approach employs a transformation to linearize, at least approximately, a nonlinear regression function. We discuss the use of transformations to linearize regression functions in Section 3.9.

When the nature of the regression function is not known, exploratory analysis that does not require specifying a particular type of function is often useful. We discuss exploratory regression analysis in Section 3.10.

Nonconstancy of Error Variance

When the error variance is not constant but varies in a systematic fashion, a direct approach is to modify the model to allow for this and use the method of *weighted least squares* to obtain the estimators of the parameters. We discuss the use of weighted least squares for this purpose in Chapter 11.

Transformations can also be effective in stabilizing the variance. Some of these are discussed in Section 3.9.

Nonindependence of Error Terms

When the error terms are correlated, a direct remedial measure is to work with a model that calls for correlated error terms. We discuss such a model in Chapter 12. A simple remedial transformation that is often helpful is to work with first differences, a topic also discussed in Chapter 12.

Nonnormality of Error Terms

Lack of normality and nonconstant error variances frequently go hand in hand. Fortunately, it is often the case that the same transformation that helps stabilize the variance is also helpful in approximately normalizing the error terms. It is therefore desirable that the transformation

for stabilizing the error variance be utilized first, and then the residuals studied to see if serious departures from normality are still present. We discuss transformations to achieve approximate normality in Section 3.9.

Omission of Important Predictor Variables

When residual analysis indicates that an important predictor variable has been omitted from the model, the solution is to modify the model. In Chapter 6 and later chapters, we discuss multiple regression analysis in which two or more predictor variables are utilized.

Outlying Observations

When outlying observations are present, as in Figure 3.7a, use of the least squares and maximum likelihood estimators (1.10) for regression model (2.1) may lead to serious distortions in the estimated regression function. When the outlying observations do not represent recording errors and should not be discarded, it may be desirable to use an estimation procedure that places less emphasis on such outlying observations. We discuss one such robust estimation procedure in Chapter 11.

3.9 Transformations

We now consider in more detail the use of transformations of one or both of the original variables before carrying out the regression analysis. Simple transformations of either the response variable Y or the predictor variable X, or of both, are often sufficient to make the simple linear regression model appropriate for the transformed data.

Transformations for Nonlinear Relation Only

We first consider transformations for linearizing a nonlinear regression relation when the distribution of the error terms is reasonably close to a normal distribution and the error terms have approximately constant variance. In this situation, transformations on X should be attempted. The reason why transformations on Y may not be desirable here is that a transformation on Y, such as $Y' = \sqrt{Y}$, may materially change the shape of the distribution of the error terms from the normal distribution and may also lead to substantially differing error term variances.

Figure 3.13 contains some prototype nonlinear regression relations with constant error variance and also presents some simple transformations on X that may be helpful to linearize the regression relationship without affecting the distributions of Y. Several alternative transformations may be tried. Scatter plots and residual plots based on each transformation should then be prepared and analyzed to decide which transformation is most effective.

Example

Data from an experiment on the effect of number of days of training received (X) on performance (Y) in a battery of simulated sales situations are presented in Table 3.7, columns 1 and 2, for the 10 participants in the study. A scatter plot of these data is shown in Figure 3.14a. Clearly the regression relation appears to be curvilinear, so the simple linear regression model (2.1) does not seem to be appropriate. Since the variability at the different X levels appears to be fairly constant, we shall consider a transformation on X. Based on the prototype plot in Figure 3.13a, we shall consider initially the square root transformation $X' = \sqrt{X}$. The transformed values are shown in column 3 of Table 3.7.

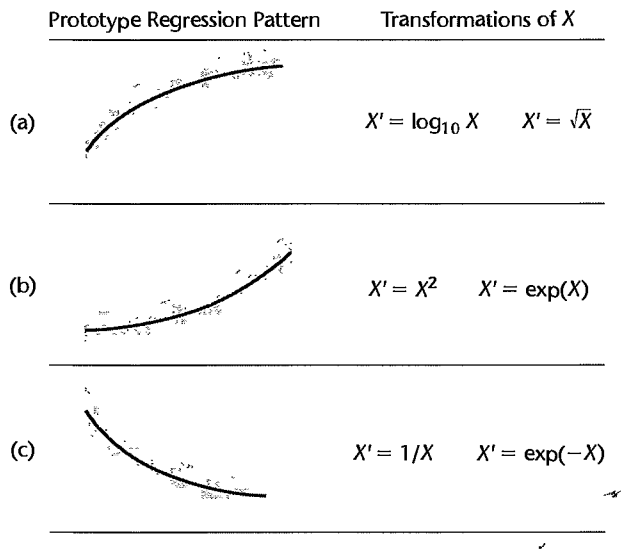


FIGURE 3.13 **Prototype Nonlinear Regression Patterns with Constant Error Variance and Simple Transformations of X.**

TABLE 3.7 **Use of Square Root Transformation of X to Linearize Regression Relation—Sales Training Example.**

Sales Trainee i	(1) Days of Training X_i	(2) Performance Score Y_i	(3) $X_i' = \sqrt{X_i}$
1	.5	42.5	.70711
2	.5	50.6	.70711
3	1.0	68.5	1.00000
4	1.0	80.7	1.00000
5	1.5	89.0	1.22474
6	1.5	99.6	1.22474
7	2.0	105.3	1.41421
8	2.0	111.8	1.41421
9	2.5	112.3	1.58114
10	2.5	125.7	1.58114

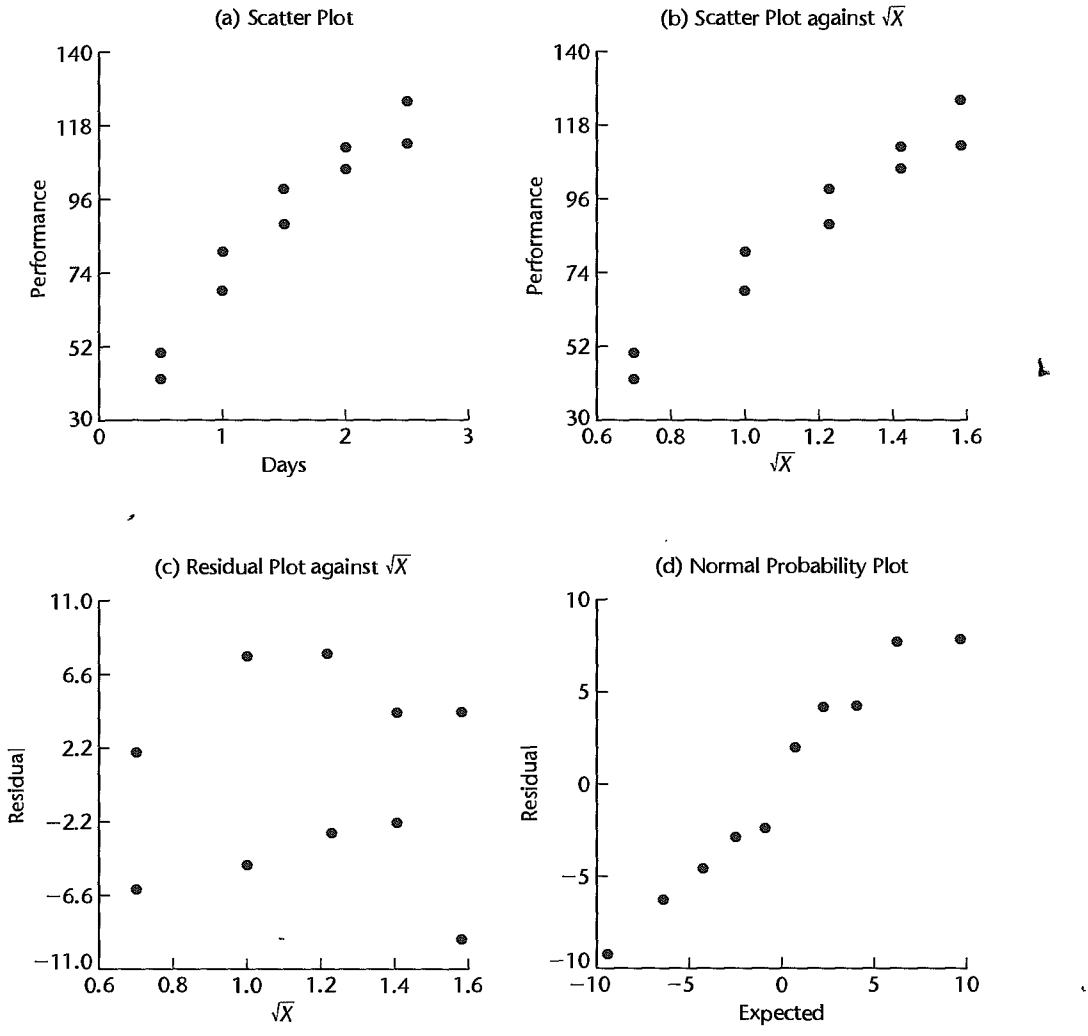
In Figure 3.14b, the same data are plotted with the predictor variable transformed to $X' = \sqrt{X}$. Note that the scatter plot now shows a reasonably linear relation. The variability of the scatter at the different X levels is the same as before, since we did not make a transformation on Y.

To examine further whether the simple linear regression model (2.1) is appropriate now, we fit it to the transformed X data. The regression calculations with the transformed X data are carried out in the usual fashion, except that the predictor variable now is X'. We obtain the following fitted regression function:

$$\hat{Y} = -10.33 + 83.45X'$$

Figure 3.14c contains a plot of the residuals against X'. There is no evidence of lack of fit or of strongly unequal error variances. Figure 3.14d contains a normal probability plot of

FIGURE 3.14 Scatter Plots and Residual Plots—Sales Training Example.



the residuals. No strong indications of substantial departures from normality are indicated by this plot. This conclusion is supported by the high coefficient of correlation between the ordered residuals and their expected values under normality, .979. For $\alpha = .01$, Table B.6 shows that the critical value is .879, so the observed coefficient is substantially larger and supports the reasonableness of normal error terms. Thus, the simple linear regression model (2.1) appears to be appropriate here for the transformed data.

The fitted regression function in the original units of X can easily be obtained, if desired:

$$\hat{Y} = -10.33 + 83.45\sqrt{X}$$

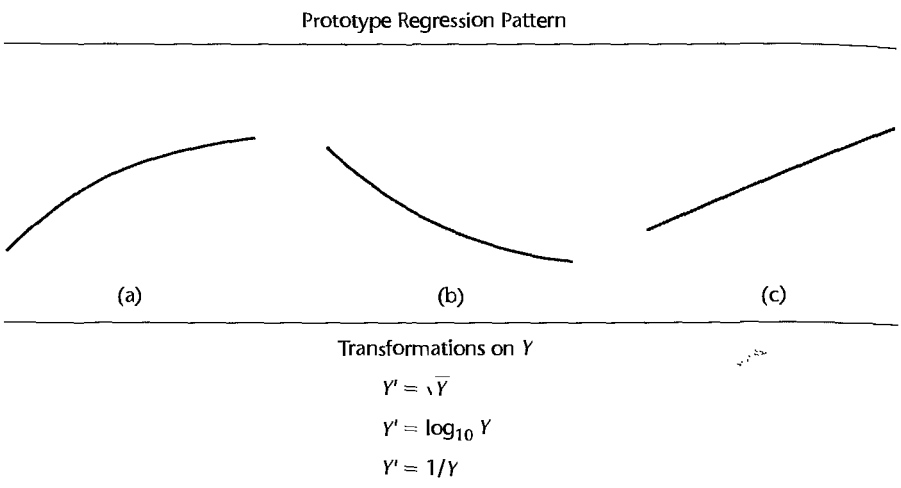


FIGURE 3.15 Prototype Regression Patterns with Unequal Error Variances and Simple Transformations of *Y*.

Note: A simultaneous transformation on X may also be helpful or necessary.

Comment

At times, it may be helpful to introduce a constant into the transformation. For example, if some of the X data are near zero and the reciprocal transformation is desired, we can shift the origin by using the transformation $X' = 1/(X + k)$, where k is an appropriately chosen constant. ■

Transformations for Nonnormality and Unequal Error Variances

Unequal error variances and nonnormality of the error terms frequently appear together. To remedy these departures from the simple linear regression model (2.1), we need a transformation on Y, since the shapes and spreads of the distributions of Y need to be changed. Such a transformation on Y may also at the same time help to linearize a curvilinear regression relation. At other times, a simultaneous transformation on X may be needed to obtain or maintain a linear regression relation.

Frequently, the nonnormality and unequal variances departures from regression model (2.1) take the form of increasing skewness and increasing variability of the distributions of the error terms as the mean response $E\{Y\}$ increases. For example, in a regression of yearly household expenditures for vacations (Y) on household income (X), there will tend to be more variation and greater positive skewness (i.e., some very high yearly vacation expenditures) for high-income households than for low-income households, who tend to consistently spend much less for vacations. Figure 3.15 contains some prototype regression relations where the skewness and the error variance increase with the mean response $E\{Y\}$. This figure also presents some simple transformations on Y that may be helpful for these cases. Several alternative transformations on Y may be tried, as well as some simultaneous transformations on X. Scatter plots and residual plots should be prepared to determine the most effective transformation(s).

Example

Data on age (X) and plasma level of a polyamine (Y) for a portion of the 25 healthy children in a study are presented in columns 1 and 2 of Table 3.8. These data are plotted in Figure 3.16a as a scatter plot. Note the distinct curvilinear regression relationship, as well as the greater variability for younger children than for older ones.

TABLE 3.8 Use of Logarithmic Transformation of Y to Linearize Regression Relation and Stabilize Error Variance—Plasma Levels Example.

Child i	(1) Age X_i	(2) Plasma Level Y_i	(3) $Y_i' = \log_{10} Y_i$
1	0 (newborn)	13.44	1.1284
2	0 (newborn)	12.84	1.1086
3	0 (newborn)	11.91	1.0759
4	0 (newborn)	20.09	1.3030
5	0 (newborn)	15.60	1.1931
6	1.0	10.11	1.0048
7	1.0	11.38	1.0561
...	...	...	...
19	3.0	6.90	.8388
20	3.0	6.77	.8306
21	4.0	4.86	.6866
22	4.0	5.10	.7076
23	4.0	5.67	.7536
24	4.0	5.75	.7597
25	4.0	6.23	.7945

On the basis of the prototype regression pattern in Figure 3.15b, we shall first try the logarithmic transformation $Y' = \log_{10} Y$. The transformed Y values are shown in column 3 of Table 3.8. Figure 3.16b contains the scatter plot with this transformation. Note that the transformation not only has led to a reasonably linear regression relation, but the variability at the different levels of X also has become reasonably constant.

To further examine the reasonableness of the transformation $Y' = \log_{10} Y$, we fitted the simple linear regression model (2.1) to the transformed Y data and obtained:

$$\hat{Y}' = 1.135 - .1023X$$

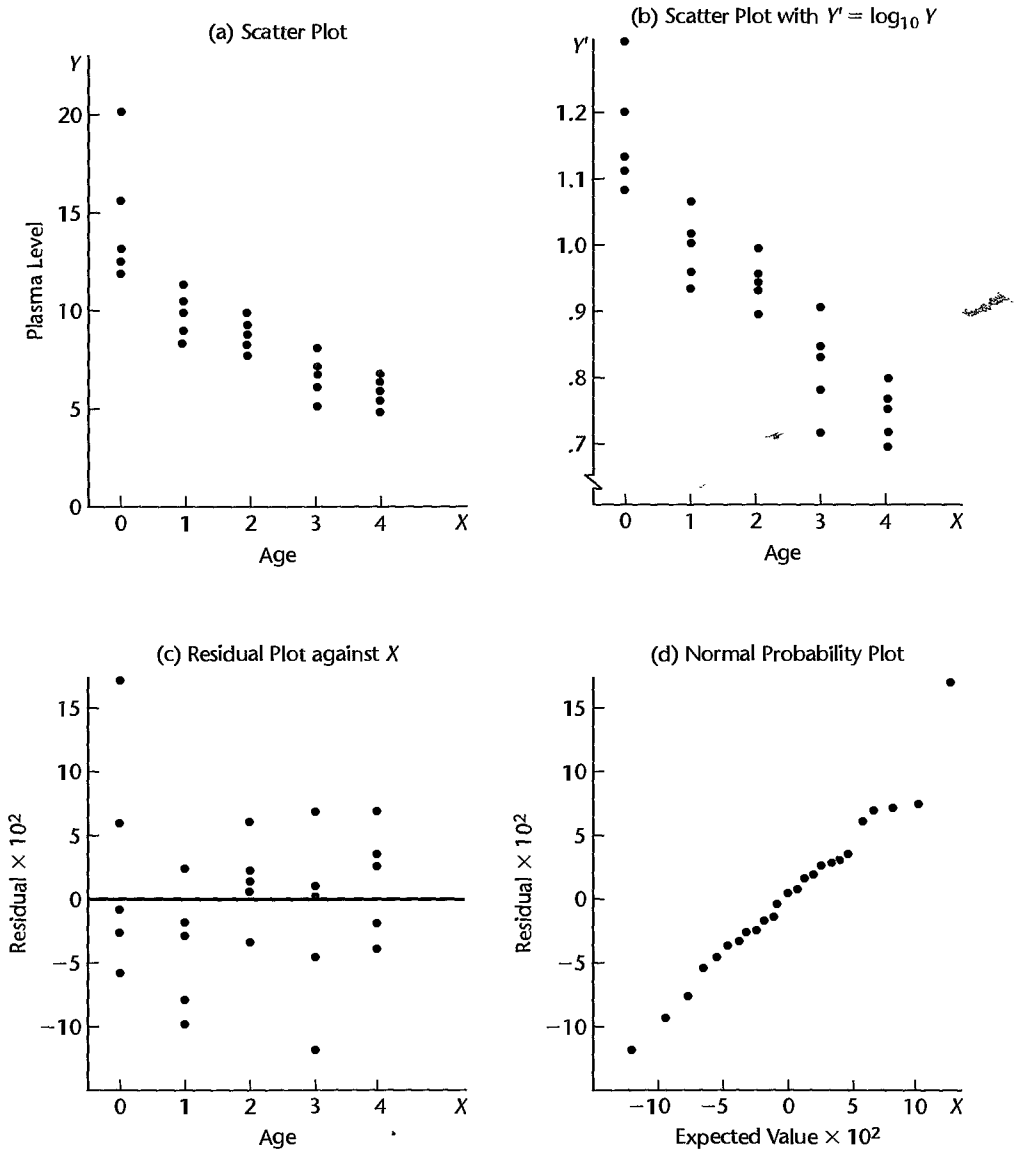
A plot of the residuals against X is shown in Figure 3.16c, and a normal probability plot of the residuals is shown in Figure 3.16d. The coefficient of correlation between the ordered residuals and their expected values under normality is .981. For $\alpha = .05$, Table B.6 indicates that the critical value is .959 so that the observed coefficient supports the assumption of normality of the error terms. All of this evidence supports the appropriateness of regression model (2.1) for the transformed Y data.

Comments

1. At times it may be desirable to introduce a constant into a transformation of Y, such as when Y may be negative. For instance, the logarithmic transformation to shift the origin in Y and make all Y observations positive would be $Y' = \log_{10}(Y + k)$, where k is an appropriately chosen constant.

2. When unequal error variances are present but the regression relation is linear, a transformation on Y may not be sufficient. While such a transformation may stabilize the error variance, it will also change the linear relationship to a curvilinear one. A transformation on X may therefore also be required. This case can also be handled by using weighted least squares, a procedure explained in Chapter 11. ■

FIGURE 3.16 Scatter Plots and Residual Plots—Plasma Levels Example.



Box-Cox Transformations

It is often difficult to determine from diagnostic plots, such as the one in Figure 3.16a for the plasma levels example, which transformation of Y is most appropriate for correcting skewness of the distributions of error terms, unequal error variances, and nonlinearity of the regression function. The Box-Cox procedure (Ref. 3.9) automatically identifies a transformation from the family of power transformations on Y. The family of power transformations

is of the form:

$$Y' = Y^{\lambda} \tag{3.33}$$

where λ is a parameter to be determined from the data. Note that this family encompasses the following simple transformations:

$$\begin{array}{lll} \lambda = 2 & Y' = Y^2 & \\ \lambda = .5 & Y' = \sqrt{Y} & \\ \lambda = 0 & Y' = \log_e Y & \text{(by definition)} \\ \lambda = -.5 & Y' = \dfrac{1}{\sqrt{Y}} & \\ \lambda = -1.0 & Y' = \dfrac{1}{Y} & \end{array} \tag{3.34}$$

The normal error regression model with the response variable a member of the family of power transformations in (3.33) becomes:

$$Y_i^{\lambda} = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{3.35}$$

Note that regression model (3.35) includes an additional parameter, λ, which needs to be estimated. The Box-Cox procedure uses the method of maximum likelihood to estimate λ, as well as the other parameters β_0, β_1, and σ^2. In this way, the Box-Cox procedure identifies $\hat{\lambda}$, the maximum likelihood estimate of λ to use in the power transformation.

Since some statistical software packages do not automatically provide the Box-Cox maximum likelihood estimate $\hat{\lambda}$ for the power transformation, a simple procedure for obtaining $\hat{\lambda}$ using standard regression software can be employed instead. This procedure involves a numerical search in a range of potential λ values; for example, $\lambda = -2$, $\lambda = -1.75, \ldots,$ $\lambda = 1.75$, $\lambda = 2$. For each λ value, the Y_i^{λ} observations are first standardized so that the magnitude of the error sum of squares does not depend on the value of λ:

$$W_i = \begin{cases} K_1(Y_i^{\lambda} - 1) & \lambda \neq 0 \\ K_2(\log_e Y_i) & \lambda = 0 \end{cases} \tag{3.36}$$

where:

$$K_2 = \left(\prod_{i=1}^{n} Y_i\right)^{1/n} \tag{3.36a}$$

$$K_1 = \frac{1}{\lambda K_2^{\lambda - 1}} \tag{3.36b}$$

Note that K_2 is the geometric mean of the Y_i observations.

Once the standardized observations W_i have been obtained for a given λ value, they are regressed on the predictor variable X and the error sum of squares SSE is obtained. It can be shown that the maximum likelihood estimate $\hat{\lambda}$ is that value of λ for which SSE is a minimum.

If desired, a finer search can be conducted in the neighborhood of the λ value that minimizes SSE. However, the Box-Cox procedure ordinarily is used only to provide a guide for selecting a transformation, so overly precise results are not needed. In any case, scatter

and residual plots should be utilized to examine the appropriateness of the transformation identified by the Box-Cox procedure.

Example

Table 3.9 contains the Box-Cox results for the plasma levels example. Selected values of λ, ranging from -1.0 to 1.0, were chosen, and for each chosen λ the transformation (3.36) was made and the linear regression of W on X was fitted. For instance, for $\lambda = .5$, the transformation $W_i = K_1(\sqrt{Y_i} - 1)$ was made and the linear regression of W on X was fitted. For this fitted linear regression, the error sum of squares is $SSE = 48.4$. The transformation that leads to the smallest value of SSE corresponds to $\lambda = -.5$, for which $SSE = 30.6$.

Figure 3.17 contains the SAS-JMP Box-Cox results for this example. It consists of a plot of SSE as a function of λ. From the plot, it is clear that a power value near $\lambda = -.50$ is indicated. However, SSE as a function of λ is fairly stable in the range from near 0 to -1.0, so the earlier choice of the logarithmic transformation $Y' = \log_{10} Y$ for the plasma levels example, corresponding to $\lambda = 0$, is not unreasonable according to the Box-Cox approach. One reason the logarithmic transformation was chosen here is because of the ease of interpreting it. The use of logarithms to base 10 rather than natural logarithms does not, of course, affect the appropriateness of the logarithmic transformation.

Comments

1. At times, theoretical or a priori considerations can be utilized to help in choosing an appropriate transformation. For example, when the shape of the scatter in a study of the relation between price of a commodity (X) and quantity demanded (Y) is that in Figure 3.15b, economists may prefer logarithmic transformations of both Y and X because the slope of the regression line for the transformed variables then measures the price elasticity of demand. The slope is then commonly interpreted as showing the percent change in quantity demanded per 1 percent change in price, where it is understood that the changes are in opposite directions.

TABLE 3.9 Box-Cox Results—Plasma Levels Example.

λ	SSE	λ	SSE
1.0	78.0	−.1	33.1
.9	70.4	−.3	31.2
.7	57.8	−.4	30.7
.5	48.4	−.5	30.6
.3	41.4	−.6	30.7
.1	36.4	−.7	31.1
0	34.5	−.9	32.7
		−1.0	33.9

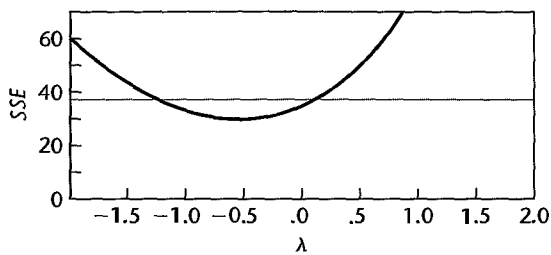


FIGURE 3.17 SAS-JMP Box-Cox Results—Plasma Levels Example.

Similarly, scientists may prefer logarithmic transformations of both Y and X when studying the relation between radioactive decay (Y) of a substance and time (X) for a curvilinear relation of the type illustrated in Figure 3.15b because the slope of the regression line for the transformed variables then measures the decay rate.

2. After a transformation has been tentatively selected, residual plots and other analyses described earlier need to be employed to ascertain that the simple linear regression model (2.1) is appropriate for the transformed data.

3. When transformed models are employed, the estimators b_0 and b_1 obtained by least squares have the least squares properties with respect to the transformed observations, not the original ones.

4. The maximum likelihood estimate of λ with the Box-Cox procedure is subject to sampling variability. In addition, the error sum of squares *SSE* is often fairly stable in a neighborhood around the estimate. It is therefore often reasonable to use a nearby λ value for which the power transformation is easy to understand. For example, use of $\lambda = 0$ instead of the maximum likelihood estimate $\hat{\lambda} = .13$ or use of $\lambda = -.5$ instead of $\hat{\lambda} = -.79$ may facilitate understanding without sacrificing much in terms of the effectiveness of the transformation. To determine the reasonableness of using an easier-to-understand value of λ, one should examine the flatness of the likelihood function in the neighborhood of $\hat{\lambda}$, as we did in the plasma levels example. Alternatively, one may construct an approximate confidence interval for λ; the procedure for constructing such an interval is discussed in Reference 3.10.

5. When the Box-Cox procedure leads to a λ value near 1, no transformation of Y may be needed.

■

3.10 Exploration of Shape of Regression Function

Scatter plots often indicate readily the nature of the regression function. For instance, Figure 1.3 clearly shows the curvilinear nature of the regression relationship between steroid level and age. At other times, however, the scatter plot is complex and it becomes difficult to see the nature of the regression relationship, if any, from the plot. In these cases, it is helpful to explore the nature of the regression relationship by fitting a smoothed curve without any constraints on the regression function. These smoothed curves are also called *nonparametric regression curves*. They are useful not only for exploring regression relationships but also for confirming the nature of the regression function when the scatter plot visually suggests the nature of the regression relationship.

Many smoothing methods have been developed for obtaining smoothed curves for time series data, where the X_i denote time periods that are equally spaced apart. The *method of moving averages* uses the mean of the Y observations for adjacent time periods to obtain smoothed values. For example, the mean of the Y values for the first three time periods in the time series might constitute the first smoothed value corresponding to the middle of the three time periods, in other words, corresponding to time period 2. Then the mean of the Y values for the second, third, and fourth time periods would constitute the second smoothed value, corresponding to the middle of these three time periods, in other words, corresponding to time period 3, and so on. Special procedures are required for obtaining smoothed values at the two ends of the time series. The larger the successive neighborhoods used for obtaining the smoothed values, the smoother the curve will be.

The *method of running medians* is similar to the method of moving averages, except that the median is used as the average measure in order to reduce the influence of outlying

observations. With this method, as well as with the moving average method, successive smoothing of the smoothed values and other refinements may be undertaken to provide a suitable smoothed curve for the time series. Reference 3.11 provides a good introduction to the running median smoothing method.

Many smoothing methods have also been developed for regression data when the X values are not equally spaced apart. A simple smoothing method, *band regression,* divides the data set into a number of groups or "bands" consisting of adjacent cases according to their X levels. For each band, the median X value and the median Y value are calculated, and the points defined by the pairs of these median values are then connected by straight lines. For example, consider the following simple data set divided into three groups:

X	Y	Median X	Median Y
2.0	13.1	2.7	14.4
3.4	15.7		
3.7	14.9		
4.5	16.8	4.5	16.8
5.0	17.1		
5.2	16.9	5.55	17.35
5.9	17.8		

The three pairs of medians are then plotted on the scatter plot of the data and connected by straight lines as a simple smoothed nonparametric regression curve.

Lowess Method

The *lowess method,* developed by Cleveland (Ref. 3.12), is a more refined nonparametric method than band regression. It obtains a smoothed curve by fitting successive linear regression functions in local neighborhoods. The name lowess stands for *locally weighted regression scatter plot smoothing.* The method is similar to the moving average and running median methods in that it uses a neighborhood around each X value to obtain a smoothed Y value corresponding to that X value. It obtains the smoothed Y value at a given X by fitting a linear regression to the data in the neighborhood of the X value and then using the fitted value at X as the smoothed value. To illustrate this concretely, let (X_1, Y_1) denote the sample case with the smallest X value, (X_2, Y_2) denote the sample case with the second smallest X value, and so on. If neighborhoods of three X values are used with the lowess method, then a linear regression would be fitted to the data:

$$(X_1, Y_1) \quad (X_2, Y_2) \quad (X_3, Y_3)$$

The fitted value at X_2 would constitute the smoothed value corresponding to X_2. Another linear regression would be fitted to the data:

$$(X_2, Y_2) \quad (X_3, Y_3) \quad (X_4, Y_4)$$

and the fitted value at X_3 would constitute the smoothed value corresponding to X_3. Smoothed values at each end of the X range are also obtained by the lowess procedure.

The lowess method uses a number of refinements in obtaining the final smoothed values to improve the smoothing and to make the procedure robust to outlying observations.

1. The linear regression is weighted to give cases further from the middle X level in each neighborhood smaller weights.
2. To make the procedure robust to outlying observations, the linear regression fitting is repeated, with the weights revised so that cases that had large residuals in the first fitting receive smaller weights in the second fitting.
3. To improve the robustness of the procedure further, step 2 is repeated one or more times by revising the weights according to the size of the residuals in the latest fitting.

To implement the lowess procedure, one must choose the size of the successive neighborhoods to be used when fitting each linear regression. One must also choose the weight function that gives less weight to neighborhood cases with X values far from each center X level and another weight function that gives less weight to cases with large residuals. Finally, the number of iterations to make the procedure robust must be chosen.

In practice, two iterations appear to be sufficient to provide robustness. Also, the weight functions suggested by Cleveland appear to be adequate for many circumstances. Hence, the primary choice to be made for a particular application is the size of the successive neighborhoods. The larger the size, the smoother the function but the greater the danger that the smoothing will lose essential features of the regression relationship. It may require some experimentation with different neighborhood sizes in order to find the size that best brings out the regression relationship. We explain the lowess method in detail in Chapter 11 in the context of multiple regression. Specific choices of weight functions and neighborhood sizes are discussed there.

Example

Figure 3.18a contains a scatter plot based on a study of research quality at 24 research laboratories. The response variable is a measure of the quality of the research done at the laboratory, and the explanatory variable is a measure of the volume of research performed at the laboratory. Note that it is very difficult to tell from this scatter plot whether or not a relationship exists between research quality and quantity. Figure 3.18b repeats the scatter plot and also shows the lowess smoothed curve. The curve suggests that there might be somewhat higher research quality for medium-sized laboratories. However, the scatter is great so that this suggested relationship should be considered only as a possibility. Also, because any particular measures of research quality and quantity are so limited, other measures should be considered to see if these corroborate the relationship suggested in Figure 3.18b.

Use of Smoothed Curves to Confirm Fitted Regression Function

Smoothed curves are useful not only in the exploratory stages when a regression model is selected but they are also helpful in confirming the regression function chosen. The procedure for confirmation is simple: The smoothed curve is plotted together with the confidence band for the fitted regression function. If the smoothed curve falls within the confidence band, we have supporting evidence of the appropriateness of the fitted regression function.

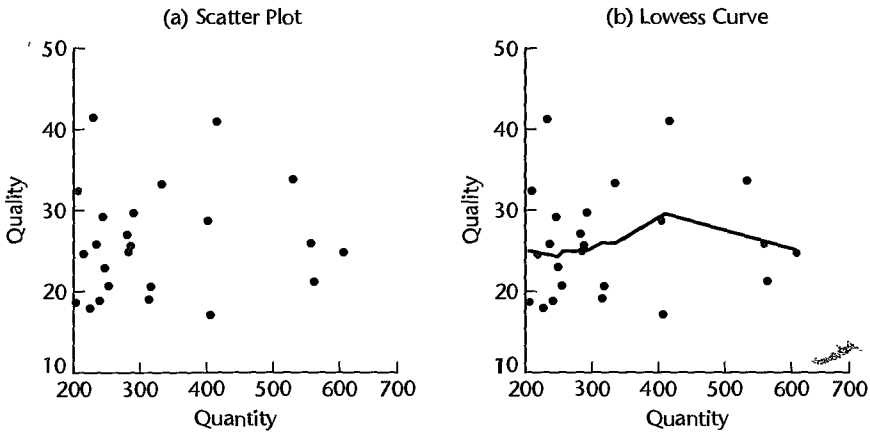


FIGURE 3.18
MINITAB Scatter Plot and Lowess Smoothed Curve—Research Laboratories Example.

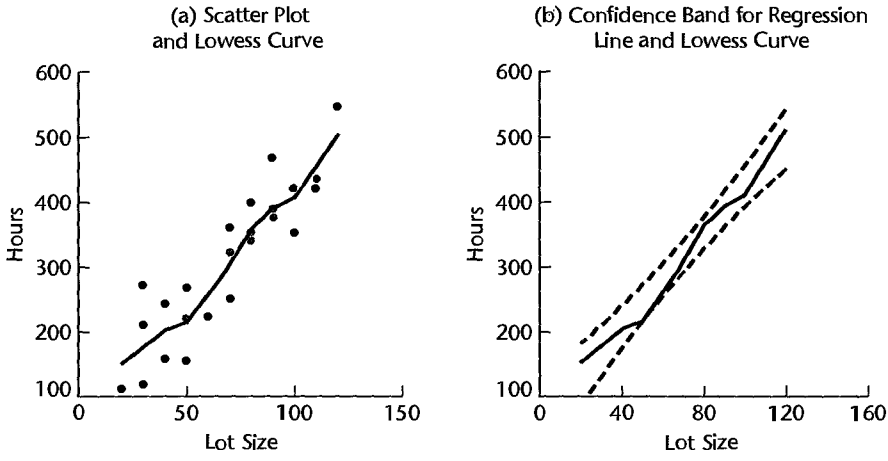


FIGURE 3.19
MINITAB Lowess Curve and Confidence Band for Regression Line—Toluca Company Example.

Example

Figure 3.19a repeats the scatter plot for the Toluca Company example from Figure 1.10a and shows the lowess smoothed curve. It appears that the regression relation is linear or possibly slightly curved. Figure 3.19b repeats the confidence band for the regression line from Figure 2.6 and shows the lowess smoothed curve. We see that the smoothed curve falls within the confidence band for the regression line and thereby supports the appropriateness of a linear regression function.

Comments

1. Smoothed curves, such as the lowess curve, do not provide an analytical expression for the functional form of the regression relationship. They only suggest the shape of the regression curve.

2. The lowess procedure is not restricted to fitting linear regression functions in each neighborhood. Higher-degree polynomials can also be utilized with this method.

3. Smoothed curves are also useful when examining residual plots to ascertain whether the residuals (or the absolute or squared residuals) follow some relationship with X or $\hat{Y}$.

4. References 3.13 and 3.14 provide good introductions to other nonparametric methods in regression analysis. ■

3.11 Case Example—Plutonium Measurement

Some environmental cleanup work requires that nuclear materials, such as plutonium 238, be located and completely removed from a restoration site. When plutonium has become mixed with other materials in very small amounts, detecting its presence can be a difficult task. Even very small amounts can be traced, however, because plutonium emits subatomic particles—alpha particles—that can be detected. Devices that are used to detect plutonium record the intensity of alpha particle strikes in counts per second (#/sec). The regression relationship between alpha counts per second (the response variable) and plutonium activity (the explanatory variable) is then used to estimate the activity of plutonium in the material under study. This use of a regression relationship involves inverse prediction [i.e., predicting plutonium activity (X) from the observed alpha count (Y)], a procedure discussed in Chapter 4.

The task here is to estimate the regression relationship between alpha counts per second and plutonium activity. This relationship varies for each measurement device and must be established precisely each time a different measurement device is used. It is reasonable to assume here that the level of alpha counts increases with plutonium activity, but the exact nature of the relationship is generally unknown.

In a study to establish the regression relationship for a particular measurement device, four plutonium *standards* were used. These standards are aluminum/plutonium rods containing a fixed, known level of plutonium activity. The levels of plutonium activity in the four standards were 0.0, 5.0, 10.0, and 20.0 picocuries per gram (pCi/g). Each standard was exposed to the detection device from 4 to 10 times, and the rate of alpha strikes, measured as counts per second, was observed for each replication. A portion of the data is shown in Table 3.10, and the data are plotted as a scatter plot in Figure 3.20a. Notice that, as expected, the strike rate tends to increase with the activity level of plutonium. Notice also that nonzero strike rates are recorded for the standard containing no plutonium. This results from background radiation and indicates that a regression model with an intercept term is required here.

TABLE 3.10 **Basic Data—Plutonium Measurement Example.**

Case	Plutonium Activity (pCi/g)	Alpha Count Rate (#/sec)
1	20	.150
2	0	.004
3	10	.069
...	...	...
22	0	.002
23	5	.049
24	0	.106

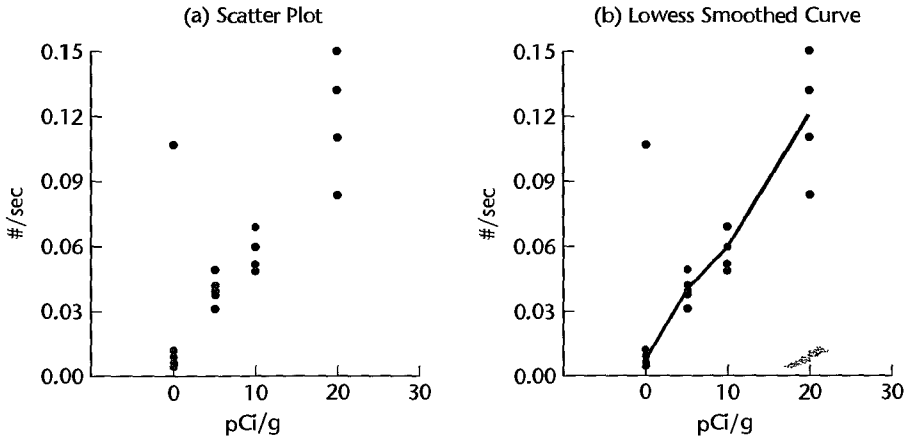


FIGURE 3.20
SAS-JMP Scatter Plot and Lowess Smoothed Curve—Plutonium Measurement Example.

As an initial step to examine the nature of the regression relationship, a lowess smoothed curve was obtained; this curve is shown in Figure 3.20b. We see that the regression relationship may be linear or slightly curvilinear in the range of the plutonium activity levels included in the study. We also see that one of the readings taken at 0.0 pCi/g (case 24) does not appear to fit with the rest of the observations. An examination of laboratory records revealed that the experimental conditions were not properly maintained for the last case, and it was therefore decided that case 24 should be discarded. Note, incidentally, how robust the lowess smoothing process was here by assigning very little weight to the outlying observation.

A linear regression function was fitted next, based on the remaining 23 cases. The SAS-JMP regression output is shown in Figure 3.21a, a plot of the residuals against the fitted values is shown in Figure 3.21b, and a normal probability plot is shown in Figure 3.21c. The JMP output uses the label Model to denote the regression component of the analysis of variance; the label C Total stands for corrected total. We see from the regression output that the slope of the regression line is not zero ($F^* = 228.9984$, P-value $= .0000$) so that a regression relationship exists. We also see from the flared, megaphone shape of the residual plot that the error variance appears to be increasing with the level of plutonium activity. The normal probability plot suggests nonnormality (heavy tails), but the nonlinearity of the plot is likely to be related (at least in part) to the unequal error variances. The existence of nonconstant variance is confirmed by the Breusch-Pagan test statistic (3.11):

$$X^2_{BP} = 23.29 > \chi^2(.95; 1) = 3.84$$

The presence of nonconstant variance clearly requires remediation. A number of approaches could be followed, including the use of weighted least squares discussed in Chapter 11. Often with count data, the error variance can be stabilized through the use of a square root transformation of the response variable. Since this is just one in a range of power transformations that might be useful, we shall use the Box-Cox procedure to suggest an appropriate power transformation. Using the standardized variable (3.36), we find the maximum likelihood estimate of λ to be $\hat{\lambda} = .65$. Because the likelihood function is fairly flat in the neighborhood of $\hat{\lambda} = .65$, the Box-Cox procedure supports the use of the square root transformation (i.e., use of $\lambda = .5$). The results of fitting a linear regression function when the response variable is $Y' = \sqrt{Y}$ are shown in Figure 3.22a.

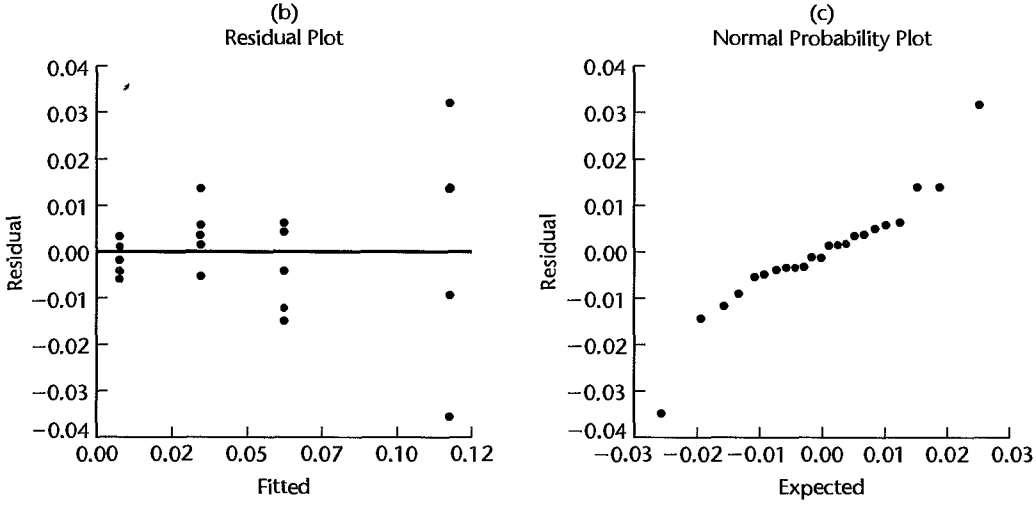
FIGURE 3.21 SAS-JMP Regression Output and Diagnostic Plots for Untransformed Data—Plutonium Measurement Example.

(a) Regression Output

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.0070331	0.0036	1.95	0.0641
Plutonium	0.005537	0.00037	15.13	0.0000

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	0.03619042	0.036190	228.9984
Error	21	0.00331880	0.000158	Prob>F
C Total	22	0.03950922		0.0000

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack of Fit	2	0.00016811	0.000084	0.5069
Pure Error	19	0.00315069	0.000166	Prob>F
Total Error	21	0.00331880		0.6103



At this point a new problem has arisen. Although the residual plot in Figure 3.22b shows that the error variance appears to be more stable and the points in the normal probability plot in Figure 3.22c fall roughly on a straight line, the residual plot now suggests that Y' is nonlinearly related to X. This concern is confirmed by the lack of fit test statistic (3.25) ($F^* = 10.1364$, P-value $= .0010$). Of course, this result is not completely unexpected, since Y was linearly related to X.

To restore a linear relation with the transformed Y variable, we shall see if a square root transformation of X will lead to a satisfactory linear fit. The regression results when regressing $Y' = \sqrt{Y}$ on $X' = \sqrt{X}$ are presented in Figure 3.23. Notice from the residual plot in Figure 3.23b that the square root transformation of the predictor variable has eliminated the lack of fit. Also, the normal probability plot of the residuals in Figure 3.23c appears to be satisfactory, and the correlation test ($r = .986$) supports the assumption of normally distributed error terms (the interpolated critical value in Table B.6 for $\alpha = .05$ and $n = 23$ is .9555). However, the residual plot suggests that some nonconstancy of the error variance

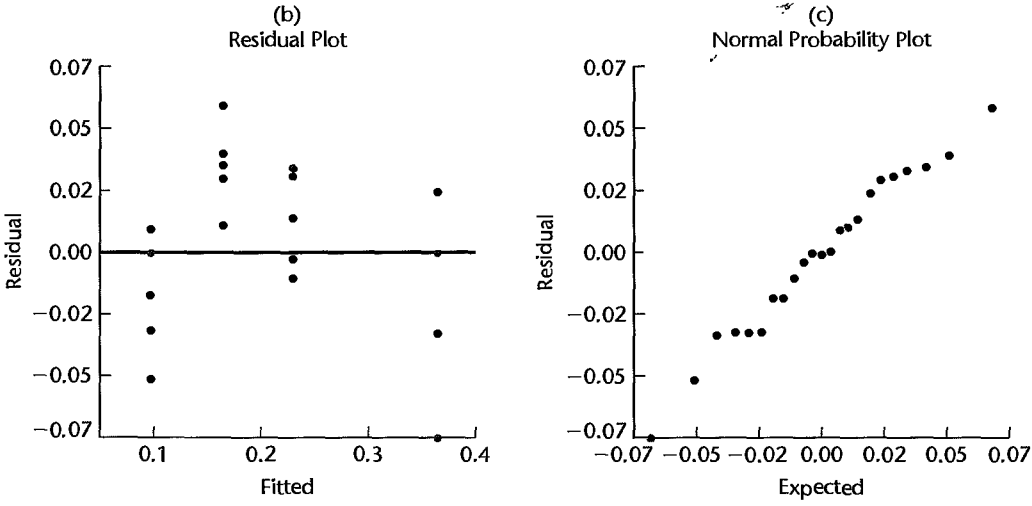
FIGURE 3.22 SAS-JMP Regression Output and Diagnostic Plots for Transformed Response Variable—Plutonium Measurement Example.

(a) Regression Output

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.0947596	0.00957	9.91	0.0000
Plutonium	0.0133648	0.00097	13.74	0.0000

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	0.21084655	0.210847	188.7960
Error	21	0.02345271	0.001117	Prob>F
C Total	22	0.23429926		0.0000

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack of Fit	2	0.01210640	0.006053	10.1364
Pure Error	19	0.01134631	0.000597	Prob>F
Total Error	21	0.02345271		0.0010



may still remain; but if so, it does not appear to be substantial. The Breusch-Pagan test statistic (3.11) is $X^2_{BP} = 3.85$, which corresponds to a P-value of .05, supporting the conclusion from the residual plot that the nonconstancy of the error variance is not substantial.

Figure 3.23d contains a SYSTAT plot of the confidence band (2.40) for the fitted regression line:

$$\hat{Y}' = .0730 + .0573X'$$

We see that the regression line has been estimated fairly precisely. Also plotted in this figure is the lowess smoothed curve. This smoothed curve falls entirely within the confidence band, supporting the reasonableness of a linear regression relation between Y' and X'. The lack of fit test statistic (3.25) now is $F^* = 1.2868$ (P-value = .2992), also supporting the linearity of the regression relating $Y' = \sqrt{Y}$ to $X' = \sqrt{X}$.

FIGURE 3.23 SAS-JMP Regression Output and Diagnostic Plots for Transformed Response and Predictor Variables—Plutonium Measurement Example.

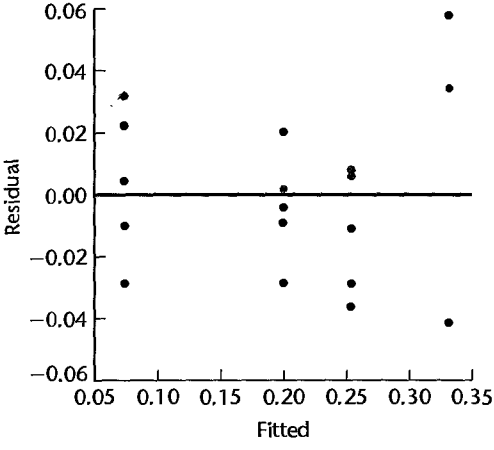
(a) Regression Output

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.0730056	0.00783	9.32	0.0000
Sqrt Plutonium	0.0573055	0.00302	19.00	0.0000

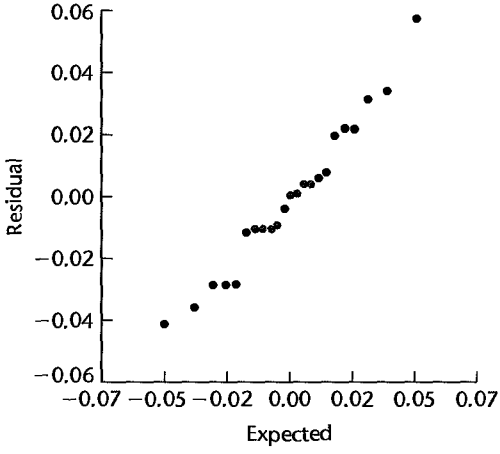
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	0.22141612	0.221416	360.9166
Error	21	0.01288314	0.000613	Prob>F
C Total	22	0.23429926		0.0000

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack of Fit	2	0.00153683	0.000768	1.2868
Pure Error	19	0.01134631	0.000597	Prob>F
Total Error	21	0.01288314		0.2992

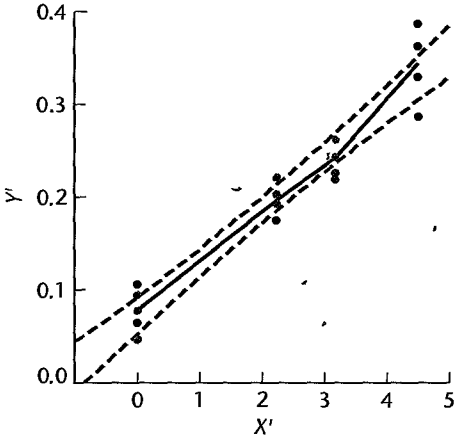
(b)
Residual Plot



(c)
Normal Probability Plot



(d)
Confidence Band for Regression Line and Lowess Curve



Cited References

3.1. Barnett, V., and T. Lewis. *Outliers in Statistical Data.* 3rd ed. New York: John Wiley & Sons, 1994.

3.2. Looney, S. W., and T. R. Gulledge, Jr. "Use of the Correlation Coefficient with Normal Probability Plots," *The American Statistician* 39 (1985), pp. 75–79.

3.3. Shapiro, S. S., and M. B. Wilk. "An Analysis of Variance Test for Normality (Complete Samples)," *Biometrika* 52 (1965), pp. 591–611.

3.4. Levene, H. "Robust Tests for Equality of Variances," in *Contributions to Probability and Statistics,* ed. I. Olkin. Palo Alto, Calif.: Stanford University Press, 1960, pp. 278–92.

3.5. Brown, M. B., and A. B. Forsythe. "Robust Tests for Equality of Variances," *Journal of the American Statistical Association* 69 (1974), pp. 364–67.

3.6. Breusch, T. S., and A. R. Pagan. "A Simple Test for Heteroscedasticity and Random Coefficient Variation," *Econometrica* 47 (1979), pp. 1287–94.

3.7. Cook, R. D., and S. Weisberg. "Diagnostics for Heteroscedasticity in Regression," *Biometrika* 70 (1983), pp. 1–10.

3.8. Joglekar, G., J. H. Schuenemeyer, and V. LaRiccia. "Lack-of-Fit Testing When Replicates Are Not Available," *The American Statistician* 43 (1989), pp. 135–43.

3.9. Box, G. E. P., and D. R. Cox. "An Analysis of Transformations," *Journal of the Royal Statistical Society B* 26 (1964), pp. 211–43.

3.10. Draper, N. R., and H. Smith. *Applied Regression Analysis.* 3rd ed. New York: John Wiley & Sons, 1998.

3.11. Velleman, P. F., and D. C. Hoaglin. *Applications, Basics, and Computing of Exploratory Data Analysis.* Boston: Duxbury Press, 1981.

3.12. Cleveland, W. S. "Robust Locally Weighted Regression and Smoothing Scatterplots," *Journal of the American Statistical Association* 74 (1979), pp. 829–36.

3.13. Altman, N. S. "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression," *The American Statistician* 46 (1992), pp. 175–85.

3.14. Haerdle, W. *Applied Nonparametric Regression.* Cambridge: Cambridge University Press, 1990.

Problems

3.1. Distinguish between (1) residual and semistudentized residual, (2) $E\{\varepsilon_i\} = 0$ and $\bar{e} = 0$, (3) error term and residual.

3.2. Prepare a prototype residual plot for each of the following cases: (1) error variance decreases with X; (2) true regression function is U shaped, but a linear regression function is fitted.

3.3. Refer to **Grade point average** Problem 1.19.

a. Prepare a box plot for the ACT scores X_i. Are there any noteworthy features in this plot?

b. Prepare a dot plot of the residuals. What information does this plot provide?

c. Plot the residual e_i against the fitted values $\hat{Y}_i$. What departures from regression model (2.1) can be studied from this plot? What are your findings?

d. Prepare a normal probability plot of the residuals. Also obtain the coefficient of correlation between the ordered residuals and their expected values under normality. Test the reasonableness of the normality assumption here using Table B.6 and $\alpha = .05$. What do you conclude?

e. Conduct the Brown-Forsythe test to determine whether or not the error variance varies with the level of X. Divide the data into the two groups, $X < 26$, $X \geq 26$, and use $\alpha = .01$. State the decision rule and conclusion. Does your conclusion support your preliminary findings in part (c)?

f. Information is given below for each student on two variables not included in the model, namely, intelligence test score (X_2) and high school class rank percentile (X_3). (Note that larger class rank percentiles indicate higher standing in the class, e.g., 1% is near the bottom of the class and 99% is near the top of the class.) Plot the residuals against X_2 and X_3 on separate graphs to ascertain whether the model can be improved by including either of these variables. What do you conclude?

i:	1	2	3	...	118	119	120
X_2:	122	132	119	...	140	111	110
X_3:	99	71	75	...	97	65	85

*3.4. Refer to **Copier maintenance** Problem 1.20.

a. Prepare a dot plot for the number of copiers serviced X_i. What information is provided by this plot? Are there any outlying cases with respect to this variable?

b. The cases are given in time order. Prepare a time plot for the number of copiers serviced. What does your plot show?

c. Prepare a stem-and-leaf plot of the residuals. Are there any noteworthy features in this plot?

d. Prepare residual plots of e_i versus $\hat{Y}_i$ and e_i versus X_i on separate graphs. Do these plots provide the same information? What departures from regression model (2.1) can be studied from these plots? State your findings.

e. Prepare a normal probability plot of the residuals. Also obtain the coefficient of correlation between the ordered residuals and their expected values under normality. Does the normality assumption appear to be tenable here? Use Table B.6 and $\alpha = .10$.

f. Prepare a time plot of the residuals to ascertain whether the error terms are correlated over time. What is your conclusion?

g. Assume that (3.10) is applicable and conduct the Breusch-Pagan test to determine whether or not the error variance varies with the level of X. Use $\alpha = .05$. State the alternatives, decision rule, and conclusion.

h. Information is given below on two variables not included in the regression model, namely, mean operational age of copiers serviced on the call (X_2, in months) and years of experience of the service person making the call (X_3). Plot the residuals against X_2 and X_3 on separate graphs to ascertain whether the model can be improved by including either or both of these variables. What do you conclude?

i:	1	2	3	...	43	44	45
X_2:	20	19	27	...	28	26	33
X_3:	4	5	4	...	3	3	6

*3.5. Refer to **Airfreight breakage** Problem 1.21.

a. Prepare a dot plot for the number of transfers X_i. Does the distribution of number of transfers appear to be asymmetrical?

b. The cases are given in time order. Prepare a time plot for the number of transfers. Is any systematic pattern evident in your plot? Discuss.

c. Obtain the residuals e_i and prepare a stem-and-leaf plot of the residuals. What information is provided by your plot?

d. Plot the residuals e_i against X_i to ascertain whether any departures from regression model (2.1) are evident. What is your conclusion?

e. Prepare a normal probability plot of the residuals. Also obtain the coefficient of correlation between the ordered residuals and their expected values under normality to ascertain whether the normality assumption is reasonable here. Use Table B.6 and $\alpha = .01$. What do you conclude?

f. Prepare a time plot of the residuals. What information is provided by your plot?

g. Assume that (3.10) is applicable and conduct the Breusch-Pagan test to determine whether or not the error variance varies with the level of X. Use $\alpha = .10$. State the alternatives, decision rule, and conclusion. Does your conclusion support your preliminary findings in part (d)?

3.6. Refer to **Plastic hardness** Problem 1.22.

a. Obtain the residuals e_i and prepare a box plot of the residuals. What information is provided by your plot?

b. Plot the residuals e_i against the fitted values $\hat{Y}_i$ to ascertain whether any departures from regression model (2.1) are evident. State your findings.

c. Prepare a normal probability plot of the residuals. Also obtain the coefficient of correlation between the ordered residuals and their expected values under normality. Does the normality assumption appear to be reasonable here? Use Table B.6 and $\alpha = .05$.

d. Compare the frequencies of the residuals against the expected frequencies under normality, using the 25th, 50th, and 75th percentiles of the relevant t distribution. Is the information provided by these comparisons consistent with the findings from the normal probability plot in part (c)?

e. Use the Brown-Forsythe test to determine whether or not the error variance varies with the level of X. Divide the data into the two groups, $X \leq 24$, $X > 24$, and use $\alpha = .05$. State the decision rule and conclusion. Does your conclusion support your preliminary findings in part (b)?

*3.7. Refer to **Muscle mass** Problem 1.27.

a. Prepare a stem-and-leaf plot for the ages X_i. Is this plot consistent with the random selection of women from each 10-year age group? Explain.

b. Obtain the residuals e_i and prepare a dot plot of the residuals. What does your plot show?

c. Plot the residuals e_i against $\hat{Y}_i$ and also against X_i on separate graphs to ascertain whether any departures from regression model (2.1) are evident. Do the two plots provide the same information? State your conclusions.

d. Prepare a normal probability plot of the residuals. Also obtain the coefficient of correlation between the ordered residuals and their expected values under normality to ascertain whether the normality assumption is tenable here. Use Table B.6 and $\alpha = .10$. What do you conclude?

e. Assume that (3.10) is applicable and conduct the Breusch-Pagan test to determine whether or not the error variance varies with the level of X. Use $\alpha = .01$. State the alternatives, decision rule, and conclusion. Is your conclusion consistent with your preliminary findings in part (c)?

3.8. Refer to **Crime rate** Problem 1.28.

a. Prepare a stem-and-leaf plot for the percentage of individuals in the county having at least a high school diploma X_i. What information does your plot provide?

b. Obtain the residuals e_i and prepare a box plot of the residuals. Does the distribution of the residuals appear to be symmetrical?

c. Make a residual plot of e_i versus $\hat{Y}_i$. What does the plot show?

d. Prepare a normal probability plot of the residuals. Also obtain the coefficient of correlation between the ordered residuals and their expected values under normality. Test the reasonableness of the normality assumption using Table B.6 and $\alpha = .05$. What do you conclude?

e. Conduct the Brown-Forsythe test to determine whether or not the error variance varies with the level of X. Divide the data into the two groups, $X \leq 69$, $X > 69$, and use $\alpha = .05$. State the decision rule and conclusion. Does your conclusion support your preliminary findings in part (c)?

3.9. **Electricity consumption.** An economist studying the relation between household electricity consumption (Y) and number of rooms in the home (X) employed linear regression model (2.1) and obtained the following residuals:

i:	1	2	3	4	5	6	7	8	9	10
X_i:	2	3	4	5	6	7	8	9	10	11
e_i:	3.2	2.9	−1.7	−2.0	−2.3	−1.2	−.9	.8	.7	.5

Plot the residuals e_i against X_i. What problem appears to be present here? Might a transformation alleviate this problem?

3.10. **Per capita earnings.** A sociologist employed linear regression model (2.1) to relate per capita earnings (Y) to average number of years of schooling (X) for 12 cities. The fitted values $\hat{Y}_i$ and the semistudentized residuals e_i^* follow.

i:	1	2	3	...	10	11	12
$\hat{Y}_i$:	9.9	9.3	10.2	...	15.6	11.2	13.1
e_i^*:	−1.12	.81	−.76	...	−3.78	.74	.32

a. Plot the semistudentized residuals against the fitted values. What does the plot suggest?

b. How many semistudentized residuals are outside ±1 standard deviation? Approximately how many would you expect to see if the normal error model is appropriate?

3.11. **Drug concentration.** A pharmacologist employed linear regression model (2.1) to study the relation between the concentration of a drug in plasma (Y) and the log-dose of the drug (X). The residuals and log-dose levels follow.

i:	1	2	3	4	5	6	7	8	9
X_i:	−1	0	1	−1	0	1	−1	0	1
e_i:	.5	2.1	−3.4	.3	−1.7	4.2	−.6	2.6	−4.0

a. Plot the residuals e_i against X_i. What conclusions do you draw from the plot?

b. Assume that (3.10) is applicable and conduct the Breusch-Pagan test to determine whether or not the error variance varies with log-dose of the drug (X). Use $\alpha = .05$. State the alternatives, decision rule, and conclusion. Does your conclusion support your preliminary findings in part (a)?

3.12. A student does not understand why the sum of squares defined in (3.16) is called a pure error sum of squares "since the formula looks like one for an ordinary sum of squares." Explain.

*3.13. Refer to **Copier maintenance** Problem 1.20.

a. What are the alternative conclusions when testing for lack of fit of a linear regression function?

b. Perform the test indicated in part (a). Control the risk of Type I error at .05. State the decision rule and conclusion.

c. Does the test in part (b) detect other departures from regression model (2.1), such as lack of constant variance or lack of normality in the error terms? Could the results of the test of lack of fit be affected by such departures? Discuss.

3.14. Refer to **Plastic hardness** Problem 1.22.

a. Perform the F test to determine whether or not there is lack of fit of a linear regression function; use $\alpha = .01$. State the alternatives, decision rule, and conclusion.

b. Is there any advantage of having an equal number of replications at each of the X levels? Is there any disadvantage?

c. Does the test in part (a) indicate what regression function is appropriate when it leads to the conclusion that the regression function is not linear? How would you proceed?

3.15. **Solution concentration.** A chemist studied the concentration of a solution (Y) over time (X). Fifteen identical solutions were prepared. The 15 solutions were randomly divided into five sets of three, and the five sets were measured, respectively, after 1, 3, 5, 7, and 9 hours. The results follow.

i:	1	2	3	...	13	14	15
X_i:	9	9	9	...	1	1	1
Y_i:	.07	.09	.08	...	2.84	2.57	3.10

a. Fit a linear regression function.

b. Perform the F test to determine whether or not there is lack of fit of a linear regression function; use $\alpha = .025$. State the alternatives, decision rule, and conclusion.

c. Does the test in part (b) indicate what regression function is appropriate when it leads to the conclusion that lack of fit of a linear regression function exists? Explain.

3.16. Refer to **Solution concentration** Problem 3.15.

a. Prepare a scatter plot of the data. What transformation of Y might you try, using the prototype patterns in Figure 3.15 to achieve constant variance and linearity?

b. Use the Box-Cox procedure and standardization (3.36) to find an appropriate power transformation. Evaluate SSE for $\lambda = -.2, -.1, 0, .1, .2$. What transformation of Y is suggested?

c. Use the transformation $Y' = \log_{10} Y$ and obtain the estimated linear regression function for the transformed data.

d. Plot the estimated regression line and the transformed data. Does the regression line appear to be a good fit to the transformed data?

e. Obtain the residuals and plot them against the fitted values. Also prepare a normal probability plot. What do your plots show?

f. Express the estimated regression function in the original units.

*3.17. **Sales growth.** A marketing researcher studied annual sales of a product that had been introduced 10 years ago. The data are as follows, where X is the year (coded) and Y is sales in thousands

of units:

i:	1	2	3	4	5	6	7	8	9	10
X_i:	0	1	2	3	4	5	6	7	8	9
Y_i:	98	135	162	178	221	232	283	300	374	395

a. Prepare a scatter plot of the data. Does a linear relation appear adequate here?

b. Use the Box-Cox procedure and standardization (3.36) to find an appropriate power transformation of Y. Evaluate SSE for $\lambda = .3, .4, .5, .6, .7$. What transformation of Y is suggested?

c. Use the transformation $Y' = \sqrt{Y}$ and obtain the estimated linear regression function for the transformed data.

d. Plot the estimated regression line and the transformed data. Does the regression line appear to be a good fit to the transformed data?

e. Obtain the residuals and plot them against the fitted values. Also prepare a normal probability plot. What do your plots show?

f. Express the estimated regression function in the original units.

3.18. **Production time.** In a manufacturing study, the production times for 111 recent production runs were obtained. The table below lists for each run the production time in hours (Y) and the production lot size (X).

i:	1	2	3	...	109	110	111
X_i:	15	9	7	...	12	9	15
Y_i:	14.28	8.80	12.49	...	16.37	11.45	15.78

a. Prepare a scatter plot of the data. Does a linear relation appear adequate here? Would a transformation on X or Y be more appropriate here? Why?

b. Use the transformation $X' = \sqrt{X}$ and obtain the estimated linear regression function for the transformed data.

c. Plot the estimated regression line and the transformed data. Does the regression line appear to be a good fit to the transformed data?

d. Obtain the residuals and plot them against the fitted values. Also prepare a normal probability plot. What do your plots show?

e. Express the estimated regression function in the original units.

Exercises

3.19. A student fitted a linear regression function for a class assignment. The student plotted the residuals e_i against Y_i and found a positive relation. When the residuals were plotted against the fitted values $\hat{Y}_i$, the student found no relation. How could this difference arise? Which is the more meaningful plot?

3.20. If the error terms in a regression model are independent $N(0, \sigma^2)$, what can be said about the error terms after transformation $X' = 1/X$ is used? Is the situation the same after transformation $Y' = 1/Y$ is used?

3.21. Derive the result in (3.29).

3.22. Using (A.70), (A.41), and (A.42), show that $E\{MSPE\} = \sigma^2$ for normal error regression model (2.1).

3.23. A linear regression model with intercept $\beta_0 = 0$ is under consideration. Data have been obtained that contain replications. State the full and reduced models for testing the appropriateness of the regression function under consideration. What are the degrees of freedom associated with the full and reduced models if $n = 20$ and $c = 10$?

Projects

3.24. **Blood pressure.** The following data were obtained in a study of the relation between diastolic blood pressure (Y) and age (X) for boys 5 to 13 years old.

i:	1	2	3	4	5	6	7	8
X_i:	5	8	11	7	13	12	12	6
Y_i:	63	67	74	64	75	69	90	60

a. Assuming normal error regression model (2.1) is appropriate, obtain the estimated regression function and plot the residuals e_i against X_i. What does your residual plot show?

b. Omit case 7 from the data and obtain the estimated regression function based on the remaining seven cases. Compare this estimated regression function to that obtained in part (a). What can you conclude about the effect of case 7?

c. Using your fitted regression function in part (b), obtain a 99 percent prediction interval for a new Y observation at $X = 12$. Does observation Y_7 fall outside this prediction interval? What is the significance of this?

3.25. Refer to the **CDI** data set in Appendix C.2 and Project 1.43. For each of the three fitted regression models, obtain the residuals and prepare a residual plot against X and a normal probability plot. Summarize your conclusions. Is linear regression model (2.1) more appropriate in one case than in the others?

3.26. Refer to the **CDI** data set in Appendix C.2 and Project 1.44. For each geographic region, obtain the residuals and prepare a residual plot against X and a normal probability plot. Do the four regions appear to have similar error variances? What other conclusions do you draw from your plots?

3.27. Refer to the **SENIC** data set in Appendix C.1 and Project 1.45.

a. For each of the three fitted regression models, obtain the residuals and prepare a residual plot against X and a normal probability plot. Summarize your conclusions. Is linear regression model (2.1) more apt in one case than in the others?

b. Obtain the fitted regression function for the relation between length of stay and infection risk after deleting cases 47 ($X_{47} = 6.5$, $Y_{47} = 19.56$) and 112 ($X_{112} = 5.9$, $Y_{112} = 17.94$). From this fitted regression function obtain separate 95 percent prediction intervals for new Y observations at $X = 6.5$ and $X = 5.9$, respectively. Do observations Y_{47} and Y_{112} fall outside these prediction intervals? Discuss the significance of this.

3.28. Refer to the **SENIC** data set in Appendix C.1 and Project 1.46. For each geographic region, obtain the residuals and prepare a residual plot against X and a normal probability plot. Do the four regions appear to have similar error variances? What other conclusions do you draw from your plots?

3.29. Refer to **Copier maintenance** Problem 1.20.

a. Divide the data into four bands according to the number of copiers serviced (X). Band 1 ranges from $X = .5$ to $X = 2.5$; band 2 ranges from $X = 2.5$ to $X = 4.5$; and so forth. Determine the median value of X and the median value of Y in each of the bands and develop