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

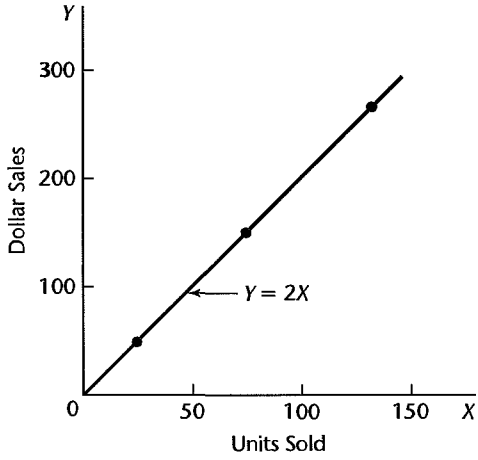


FIGURE 1.1
Example of Functional Relation.

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.

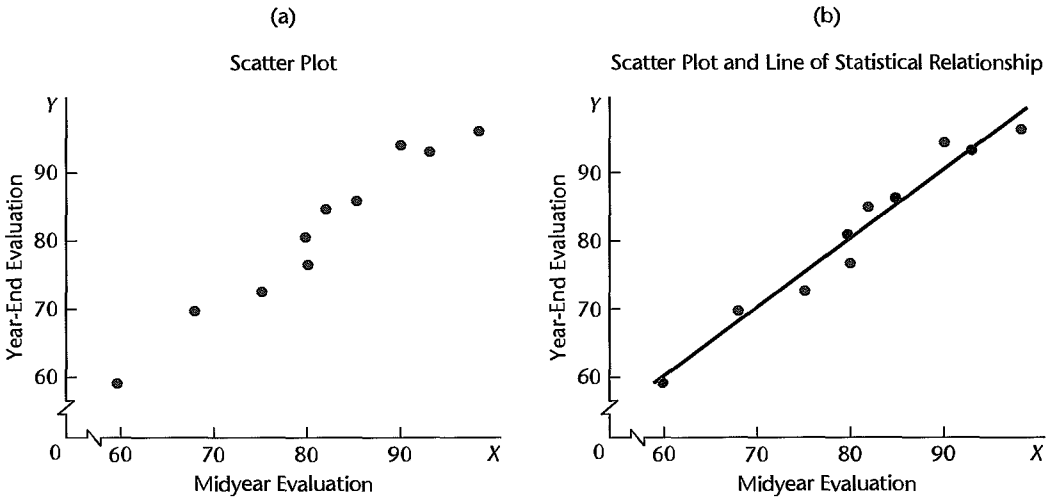
Statistical Relation between Two Variables

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 $X = 90$, $Y = 94$.

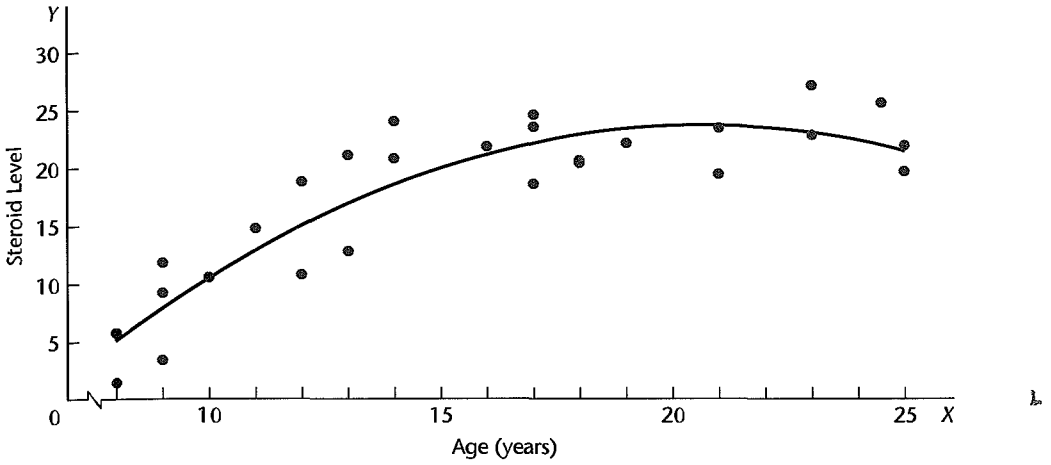
Figure 1.2a clearly suggests that there is a relation between midyear and year-end evaluations, in the sense that the higher the midyear evaluation, the higher tends to be the year-end evaluation. However, the relation is not a perfect one. There is a scattering of points, suggesting that some of the 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.

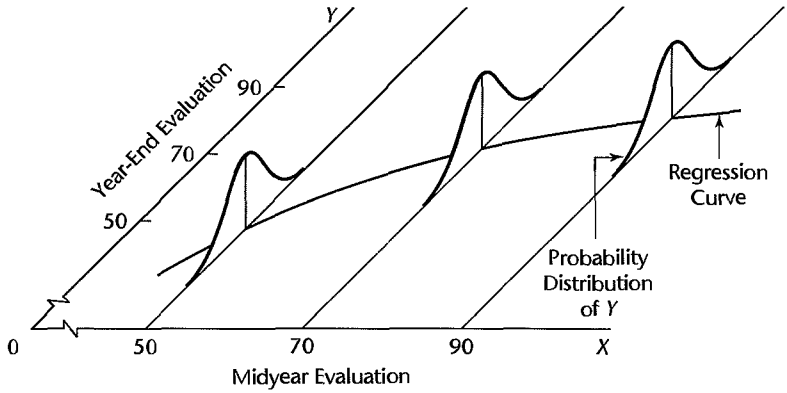


FIGURE 1.4 Pictorial Representation of Regression Model.

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.

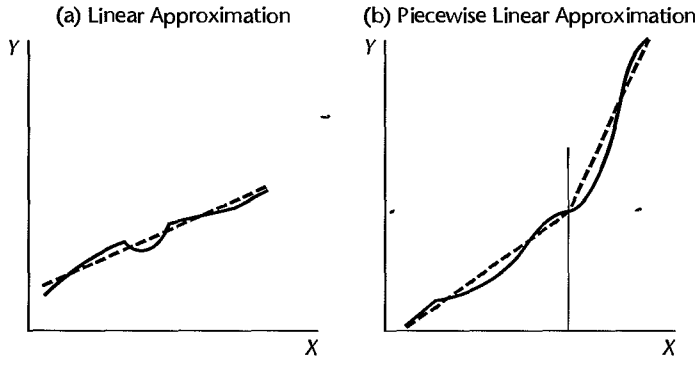
Construction of Regression Models

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 $E\{Y_i\}$.

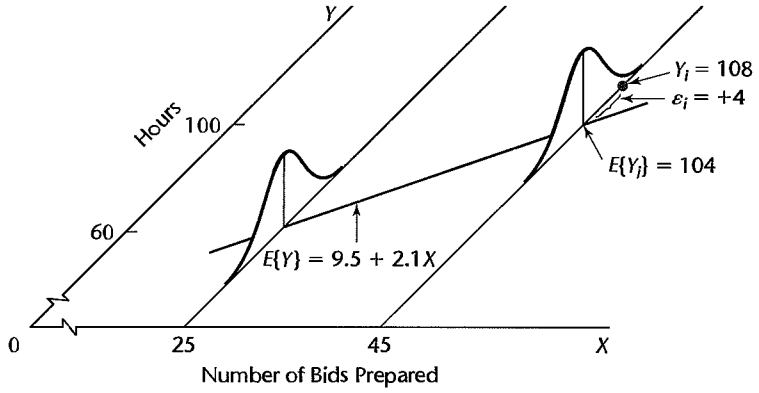


FIGURE 1.6
Illustration of Simple Linear Regression Model (1.1).

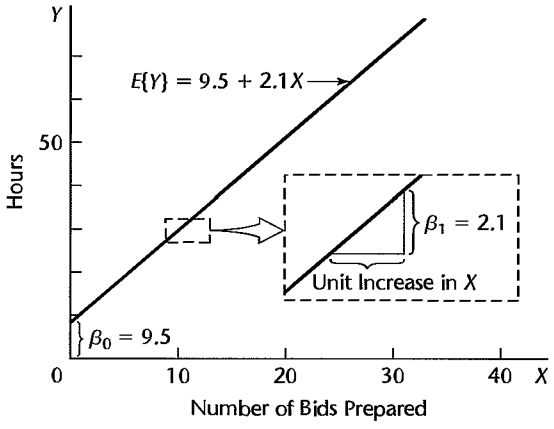


FIGURE 1.7
Meaning of Parameters of Simple Linear Regression Model (1.1).

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

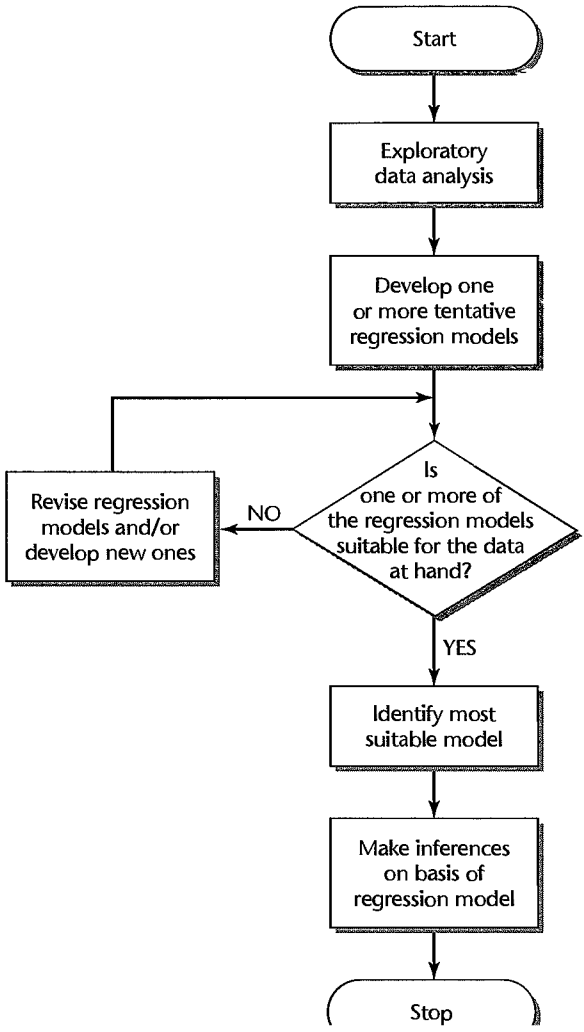


FIGURE 1.8 Typical Strategy for Regression Analysis.

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:

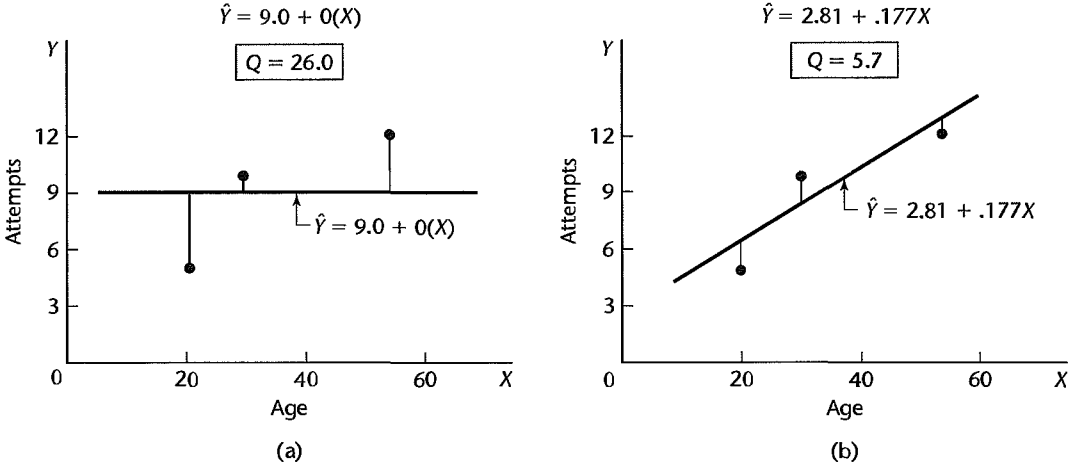
$$Y_i - (\beta_0 + \beta_1 X_i) \tag{1.7}$$

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					

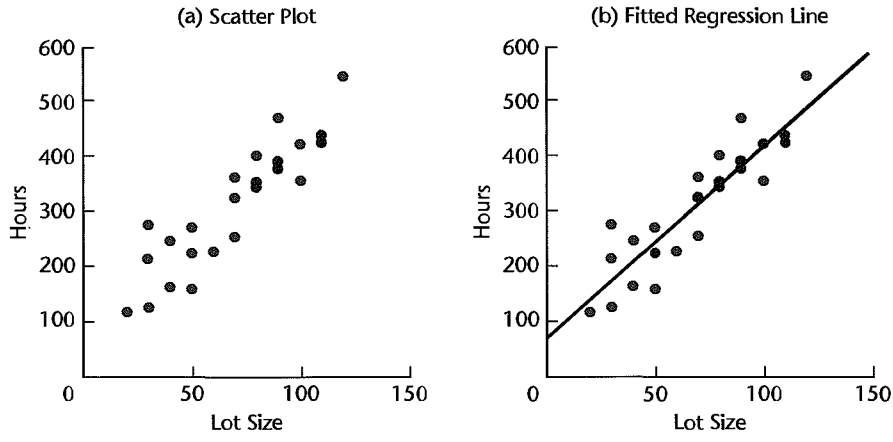


FIGURE 1.10 SYSTAT Scatter Plot and Fitted Regression Line—Toluca Company Example.

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:

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

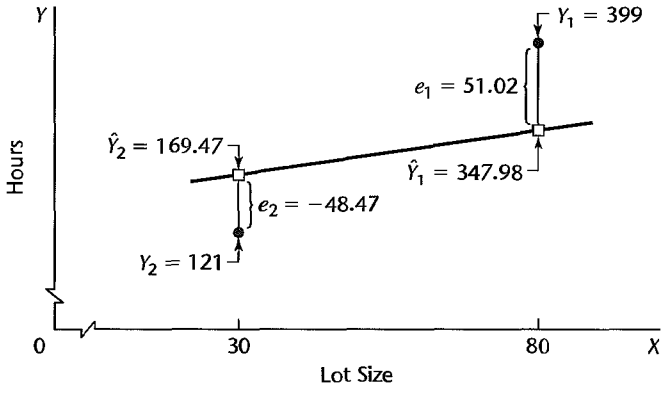


FIGURE 1.12 Illustration of Residuals—Toluca Company Example (not drawn to scale).

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

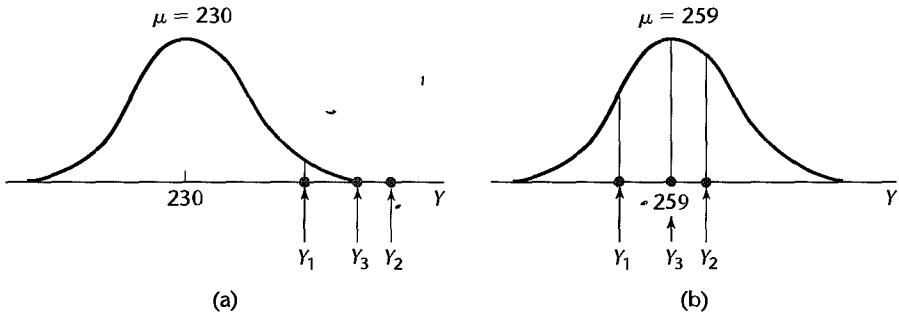


FIGURE 1.13 **Densities for Sample Observations for Two Possible Values of μ: $Y_1 = 250$, $Y_2 = 265$, $Y_3 = 259$.**

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

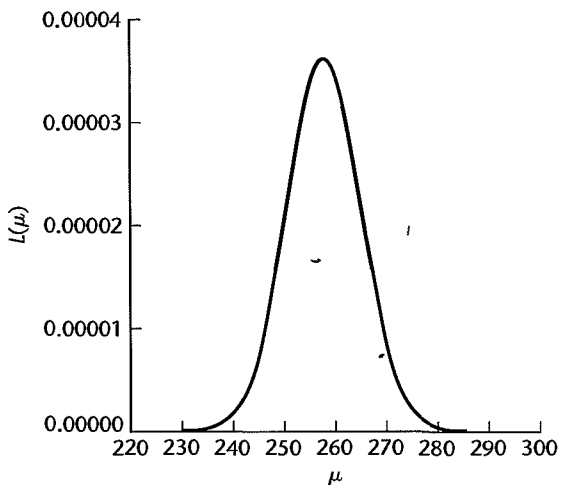


FIGURE 1.14
Likelihood Function for Estimation of Mean of Normal Population: $Y_1 = 250$, $Y_2 = 265$, $Y_3 = 259$.

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.**

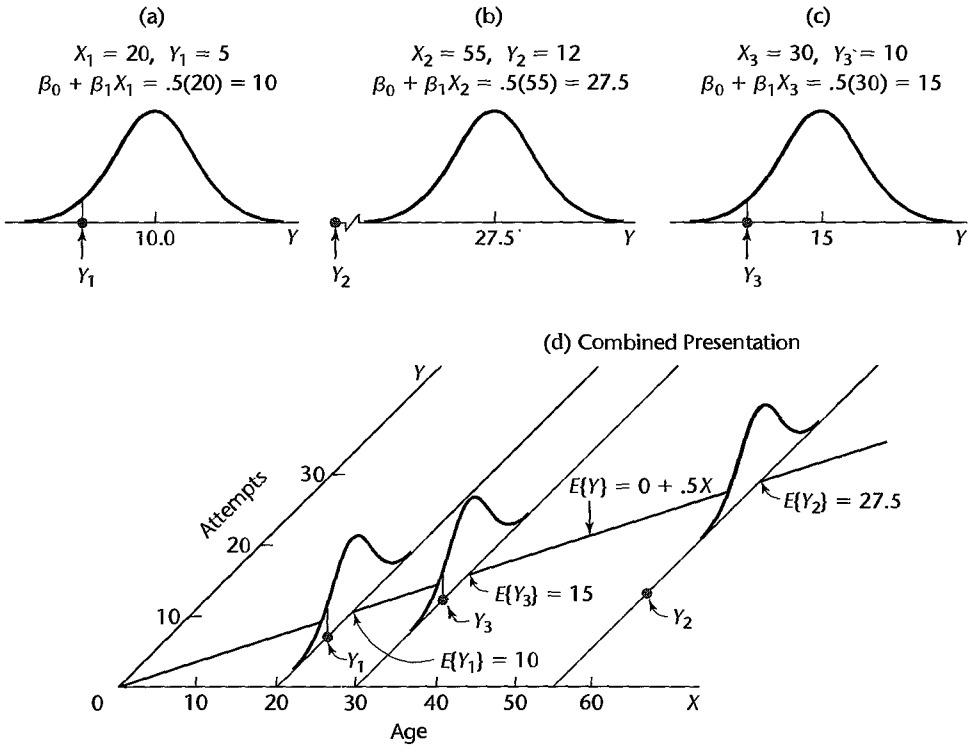


Figure 1.15d combines all of this information, showing the regression function $E\{Y\} = 0 + .5X$, the three sample cases, and the three normal distributions. Note how poorly the regression line fits the three sample cases, as was also indicated by the three small density values. Thus, it appears that $\beta_0 = 0$ 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:

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

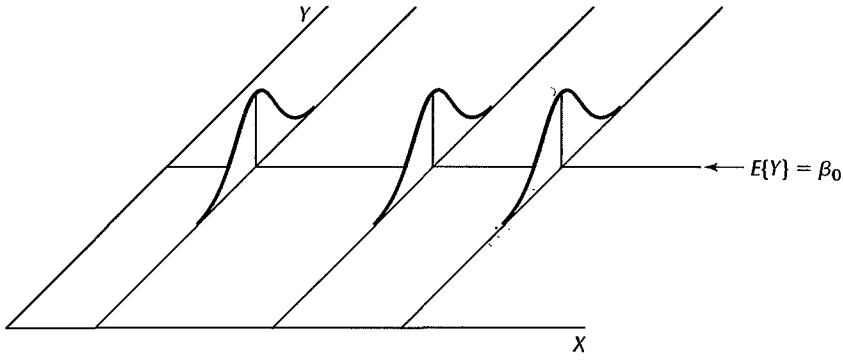


FIGURE 2.1
Regression Model (2.1) when $\beta_1 = 0$.

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:

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

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)**

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

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

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.

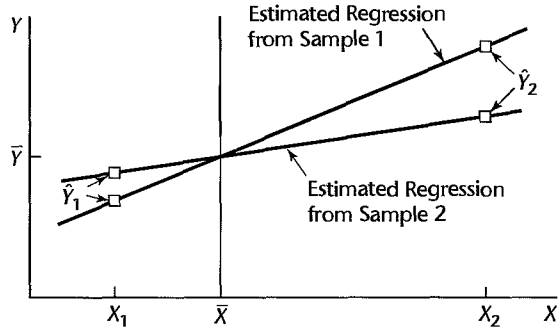


FIGURE 2.3
Effect on $\hat{Y}_h$ of Variation in b_1 from Sample to Sample in Two Samples with Same Means $\bar{Y}$ and $\bar{X}$.

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:

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

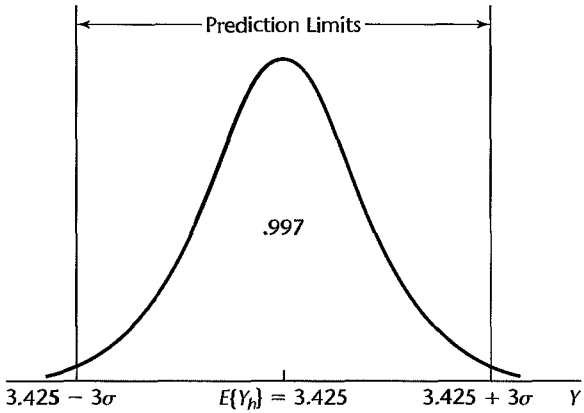


FIGURE 2.4 Prediction of $Y_{h(new)}$ when Parameters Known.

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

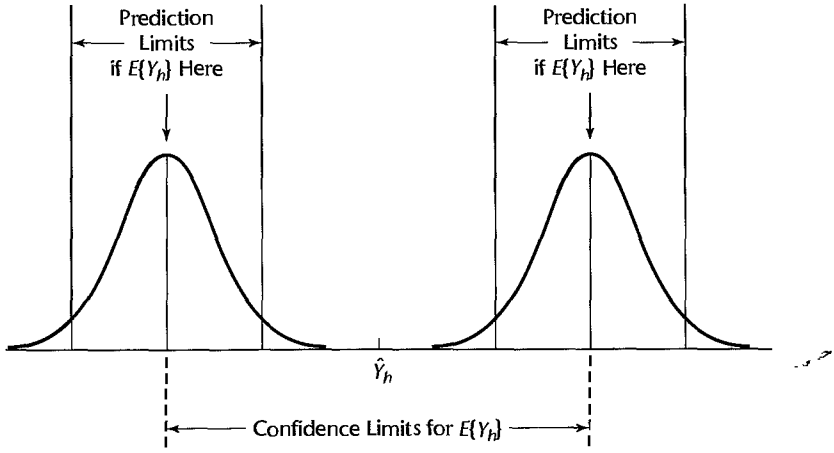


FIGURE 2.5
Prediction of $Y_{h(\text{new})}$ when Parameters Unknown.

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.

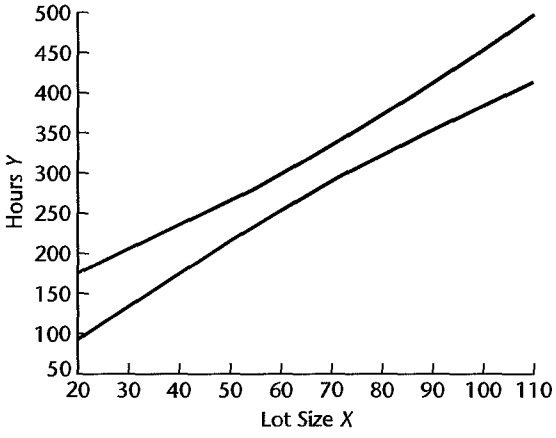


FIGURE 2.6 **Confidence Band for Regression Line—Toluca Company Example.**

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 $\bar{Y}$:

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