

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.

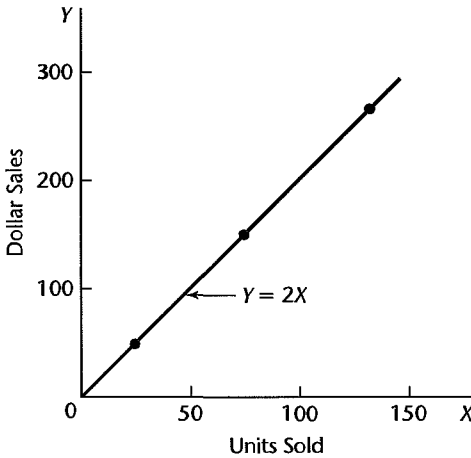
1.1 Relations between Variables

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

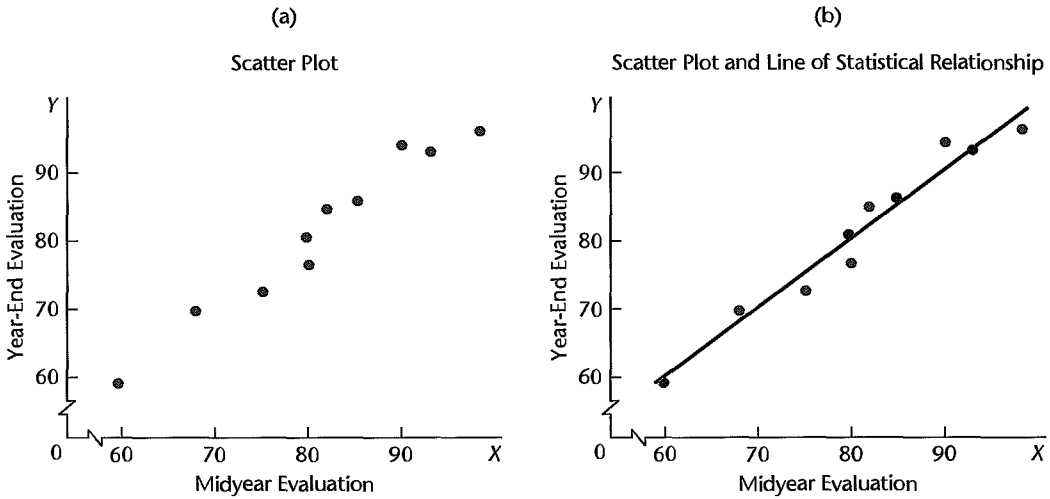
These observations are plotted also in Figure 1.1. Note that all fall directly on the line of functional relationship. This is characteristic of all functional relations.

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.

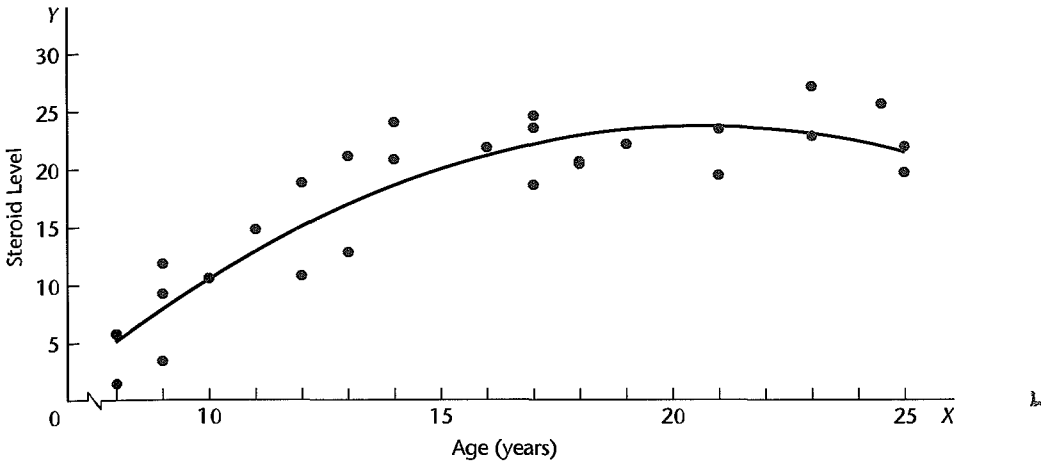
variable X . The plotting is done as before. For instance, the midyear and year-end performance evaluations for the first employee are plotted at $X = 90$, $Y = 94$.

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.

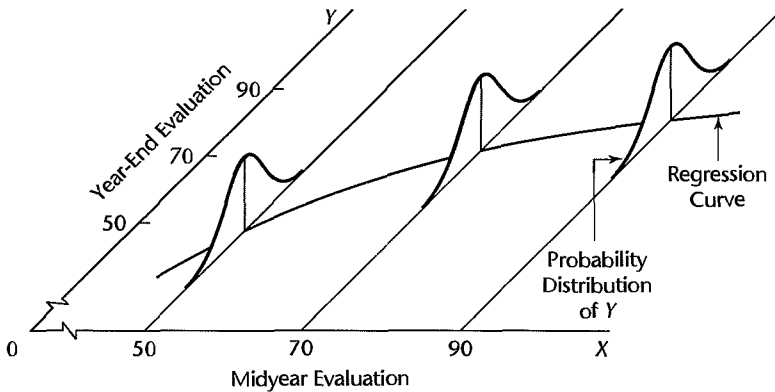
These two characteristics are embodied in a regression model by postulating that:

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.

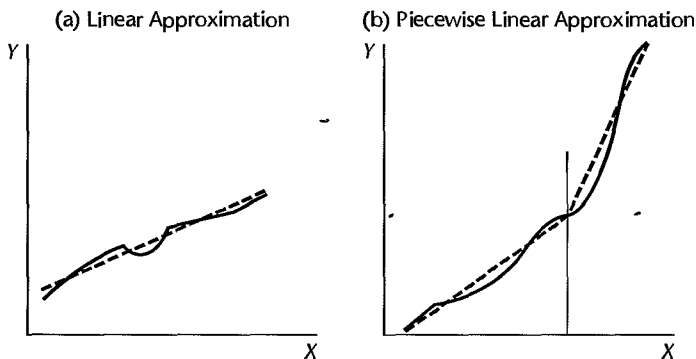
Construction of Regression Models

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:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad (1.1)$$

where:

Y_i is the value of the response variable in the i th trial

β_0 and β_1 are parameters

X_i is a known constant, namely, the value of the predictor variable in the i th 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 $i, j; i \neq j$)

$i = 1, \dots, n$

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 i th 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 i th trial is X_i , comes from a probability distribution whose mean is:

$$E\{Y_i\} = \beta_0 + \beta_1 X_i \quad (1.2)$$

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

$$E\{Y\} = \beta_0 + \beta_1 X \quad (1.3)$$

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 i th 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:

$$\sigma^2\{Y_i\} = \sigma^2 \quad (1.4)$$

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 i th 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).

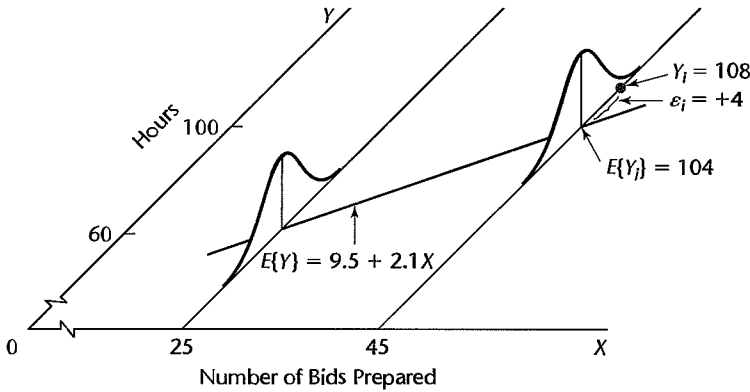


FIGURE 1.7
Meaning of Parameters of Simple Linear Regression Model (1.1).

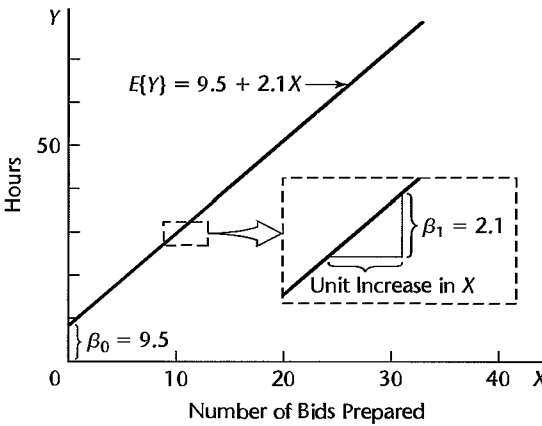


Figure 1.6 also shows the probability distribution of Y when $X = 25$. Note that this distribution exhibits the same variability as the probability distribution when $X = 45$, in conformance with the requirements of regression model (1.1).

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:

$$Y_i = \beta_0 X_0 + \beta_1 X_i + \varepsilon_i \quad \text{where } X_0 \equiv 1 \quad (1.5)$$

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:

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

where:

$$\beta_0^* = \beta_0 + \beta_1\bar{X} \quad (1.6a)$$

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.

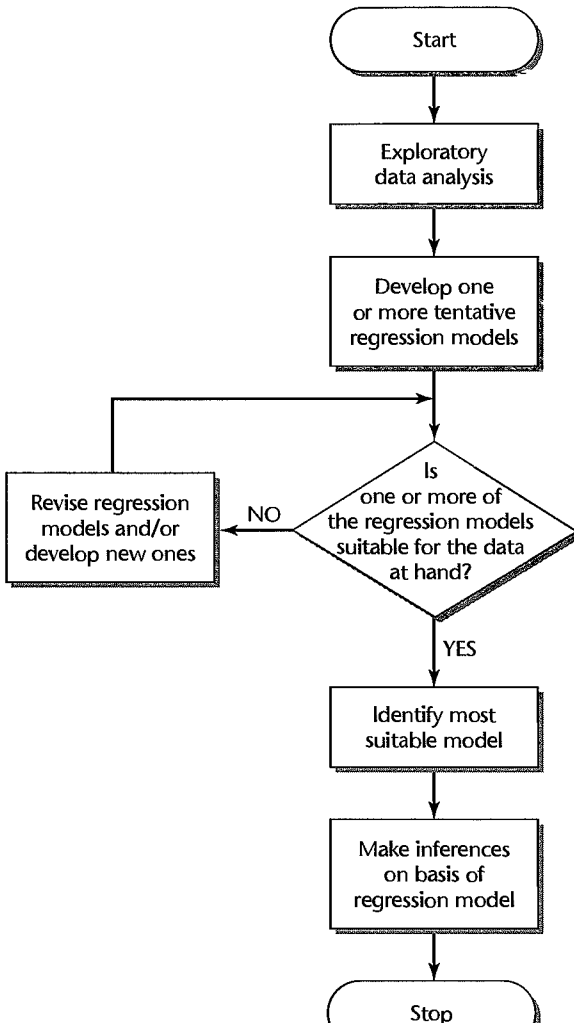
1.5 Overview of Steps in Regression Analysis

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 i th trial as (X_i, Y_i) , where $i = 1, \dots, n$.

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:

$$Y_i - (\beta_0 + \beta_1 X_i) \quad (1.7)$$

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

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

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 $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$.

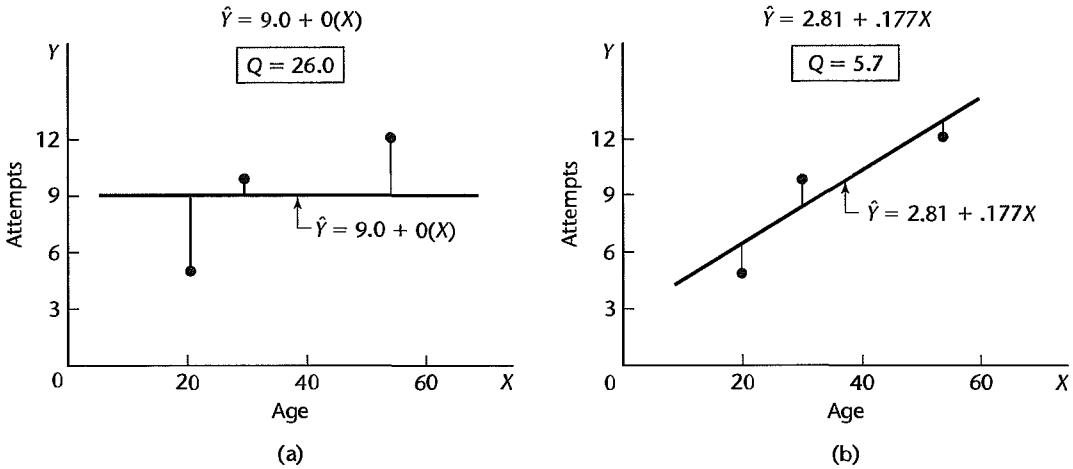
FIGURE 1.9 Illustration of Least Squares Criterion Q for Fit of a Regression Line—Persistence Study Example.**Example**

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:

$$\sum Y_i = nb_0 + b_1 \sum X_i \quad (1.9a)$$

$$\sum X_i Y_i = b_0 \sum X_i + b_1 \sum X_i^2 \quad (1.9b)$$

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 :

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

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

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:

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

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, \dots, 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	Lot Size	Work Hours					
i	X_i	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.

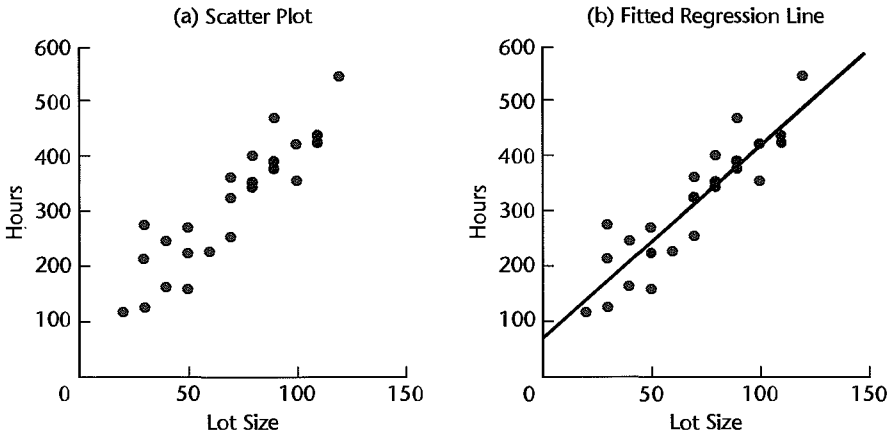


FIGURE 1.11
Portion of
MINITAB
Regression
Output—
Toluca
Company
Example.

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

$s = 48.82$ $R\text{-sq} = 82.2\%$ $R\text{-sq(adj)} = 81.4\%$

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

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

Using (1.10) we obtain:

$$\begin{aligned}b_1 &= \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \frac{70,690}{19,800} = 3.5702 \\ b_0 &= \bar{Y} - b_1 \bar{X} = 312.28 - 3.5702(70.0) = 62.37\end{aligned}$$

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:

$$\hat{Y} = b_0 + b_1 X \quad (1.12)$$

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 :

$$\hat{Y}_i = b_0 + b_1 X_i \quad i = 1, \dots, n \quad (1.13)$$

the *fitted value* for the i th 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:

$$b_0 = 62.37 \quad b_1 = 3.5702$$

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.

	(1)	(2)	(3)	(4)	(5)
Run	Lot	Work	Estimated		Squared
i	Size	Hours	Mean	Residual	Residual
	X_i	Y_i	Response	$Y_i - \hat{Y}_i = e_i$	$(Y_i - \hat{Y}_i)^2 = e_i^2$
			\hat{Y}_i		
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:

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

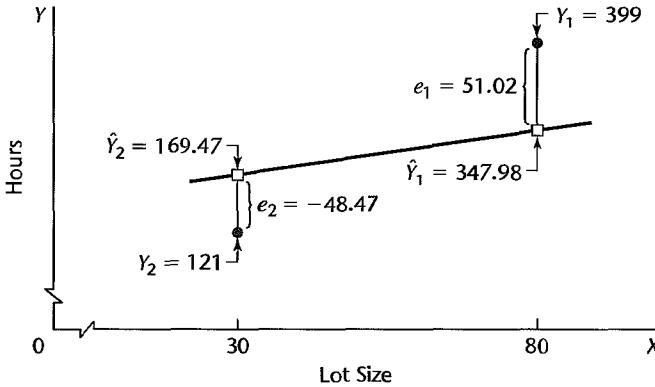
which, of course, is identical to our earlier result.

Residuals

The i th 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:

$$e_i = Y_i - (b_0 + b_1 X_i) = Y_i - b_0 - b_1 X_i \quad (1.16a)$$

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:

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

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 :

$$\sum_{i=1}^n Y_i = \sum_{i=1}^n \hat{Y}_i \quad (1.18)$$

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 i th trial is weighted by the level of the predictor variable in the i th trial:

$$\sum_{i=1}^n X_i e_i = 0 \quad (1.19)$$

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 i th trial is weighted by the fitted value of the response variable for the i th trial:

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

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:

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

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:

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

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:

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

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:

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

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):

$$E\{MSE\} = \sigma^2 \quad (1.23)$$

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 Toluca 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:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad (1.24)$$

where:

Y_i is the observed response in the i th trial

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

β_0 and β_1 are parameters

ε_i are independent $N(0, \sigma^2)$

$i = 1, \dots, n$

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 : $Y_1 = 250$,
 $Y_2 = 265$,
 $Y_3 = 259$.

