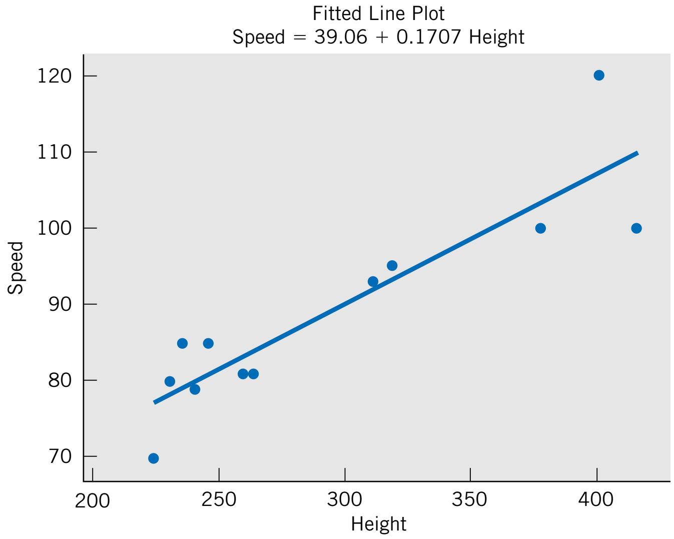
11

Regression Analysis I

Simple Linear Regression

The Highest Roller Coasters are Fastest

Some roller coasters are designed to twist riders and turn them upside down. Others are designed to provide fast rides over large drops. Among the 12 tallest roller coasters in the world, the maximum height (inches) is related to top speed (miles per hour). Each data point, consisting of the pair of values (height, speed), represents one roller coaster. The fitted line predicts an increase in top speed of .17 miles per hour for each foot of height, or 17 miles per hour for each 100 feet in height.



© Rafael Macia/Photo Researchers, Inc.

45688_11_p429-474 7/25/05 9:57 AM Page 431

1. INTRODUCTION

Except for the brief treatment in Sections 5 to 8 of Chapter 3, we have discussed statistical inferences based on the sample measurements of a single variable. In many investigations, two or more variables are observed for each experimental unit in order to determine:

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

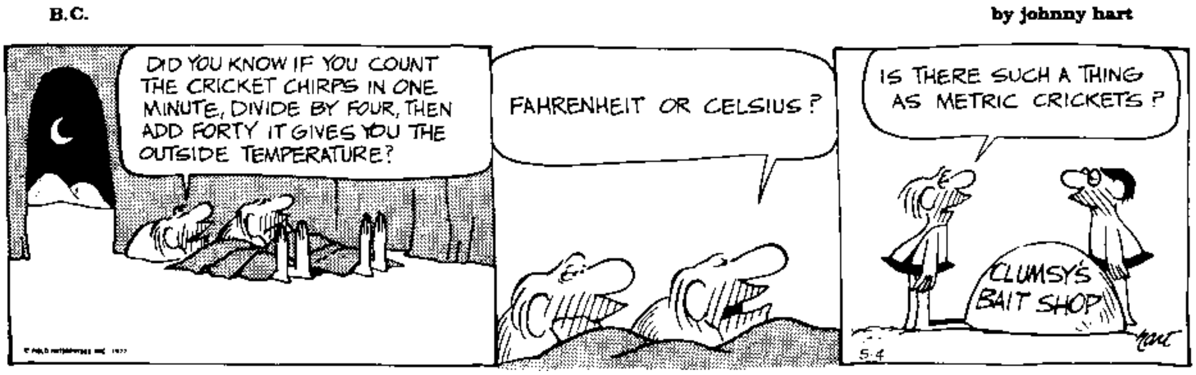
Regression analysis concerns the study of relationships between variables with the object of identifying, estimating, and validating the relationship. The estimated relationship can then be used to predict one variable from the value of the other variable(s). In this chapter, we introduce the subject with specific reference to the straight-line model. Chapter 3 treated the subject of fitting a line from a descriptive statistics viewpoint. Here, we take the additional step of including the omnipresent random variation as an error term in the model. Then, on the basis of the model, we can test whether one variable actually influences the other. Further, we produce confidence interval answers when using the estimated straight line for prediction. The correlation coefficient is shown to measure the strength of the linear relationship.

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

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

Example 1 A Straight Line Model Masked by Random Disturbances

A factory manufactures items in batches and the production manager wishes to relate the production cost y of a batch to the batch size x. Certain costs are practically constant, regardless of the batch size x. Building costs and


B.C.
by johnny hart
DID YOU KNOW IF YOU COUNT THE CRICKET CHIRPS IN ONE MINUTE, DIVIDE BY FOUR, THEN ADD FORTY IT GIVES YOU THE OUTSIDE TEMPERATURE?
FAHRENHEIT OR CELSIUS?
IS THERE SUCH A THING AS METRIC CRICKETS?
CLUMSY'S BAIT SHOP
5-4

Regression analysis allows us to predict one variable from the value of another variable.
(By permission of Johnny Hart and Field Enterprises, Inc.)

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

$$y = F + Cx$$

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

Example 2 Expect an Increasing Relation But Not Necessarily a Straight Line

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

Example 3 A Scatter Diagram May Reveal an Empirical Relation

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

These examples illustrate the simplest settings for regression analysis where one wishes to determine how one variable is related to one other variable. In more complex situations several variables may be interrelated, or one variable of major interest may depend on several influencing variables. Regression analysis extends to these multivariate problems. (See Section 3, Chapter 12). Even though randomness is omnipresent, regression analysis allows us to identify it and estimate relationships.

2. REGRESSION WITH A SINGLE PREDICTOR

A regression problem involving a single predictor (also called simple regression) arises when we wish to study the relation between two variables x and y and use it to predict y from x. The variable x acts as an independent variable whose values are controlled by the experimenter. The variable y depends on x and is also subjected to unaccountable variations or errors.

Notation

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

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

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

Example 4 Relief from Symptoms of Allergy Related to Dosage

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

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

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

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

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

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

TABLE 2 Data Structure for a Simple Regression

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

First Step in the Analysis

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

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

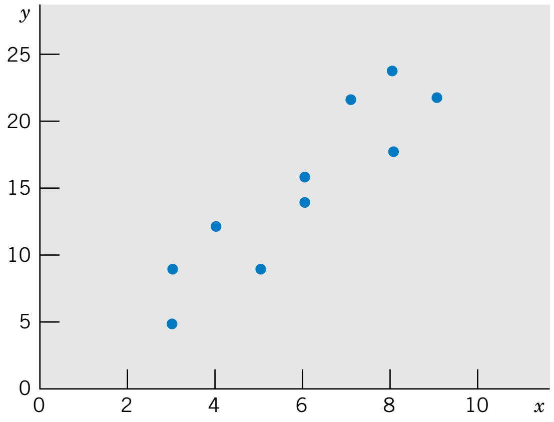


Figure 1 Scatter diagram of the data of Table 1.

3. A STRAIGHT LINE REGRESSION MODEL

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

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

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

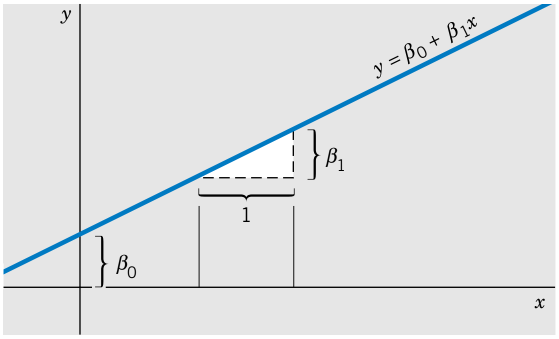


Figure 2 Graph of straight line $y = \beta_0 + \beta_1 x$.

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

Statistical Model for a Straight Line Regression

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

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

where:

1. Y_i denotes the response corresponding to the ith experimental run in which the input variable x is set at the value x_i.
2. $e_1, \ldots, e_n$ are the unknown error components that are superimposed on the true linear relation. These are **unobservable random**

variables, which we assume are independently and normally distributed with mean zero and an unknown standard deviation σ.

3. The parameters β_0 and β_1, which together locate the straight line, are unknown.

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

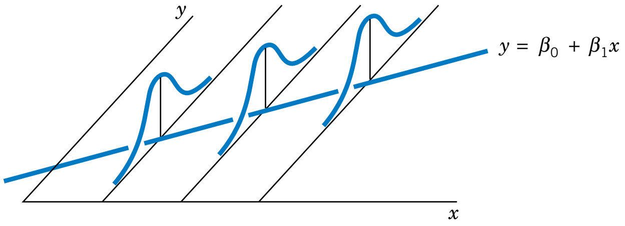


Figure 3 Normal distributions of Y with means on a straight line.

Exercises

11.1 Plot the line $y = 2 + 3x$ on graph paper by locating the points for $x = 1$ and $x = 4$. What is its intercept? What is its slope?

11.2 A store manager has determined that the monthly profit y realized from selling a particular brand of car battery is given by

$$y = 10x - 155$$

where x denotes the number of these batteries sold in a month.

(a) If 41 batteries were sold in a month, what was the profit?

(b) At least how many batteries must be sold in a month in order to make a profit?

11.3 Identify the predictor variable x and the response variable y in each of the following situations.

(a) A training director wishes to study the relationship between the duration of training for new recruits and their performance in a skilled job.

(b) The aim of a study is to relate the carbon monoxide level in blood samples from smokers with the average number of cigarettes they smoke per day.

(c) An agronomist wishes to investigate the growth rate of a fungus in relation to the level of humidity in the environment.

(d) A market analyst wishes to relate the expenditures incurred in promoting a product in test markets and the subsequent amount of product sales.

11.4 Identify the values of the parameters β_0, β_1, and σ in the statistical model

$$Y = 2 + 4x + e$$

where e is a normal random variable with mean 0 and standard deviation 5.

11.5 Identify the values of the parameters β_0, β_1, and σ in the statistical model

$$Y = 7 - 5x + e$$

where e is a normal random variable with mean 0 and standard deviation 3.

11.6 Under the linear regression model:

(a) Determine the mean and standard deviation of Y, for $x = 4$, when $\beta_0 = 1$, $\beta_1 = 3$, and $\sigma = 2$.

(b) Repeat part (a) with $x = 2$.

11.7 Under the linear regression model:

(a) Determine the mean and standard deviation of Y, for $x = 1$, when $\beta_0 = 3$, $\beta_1 = -4$, and $\sigma = 4$.

(b) Repeat part (a) with $x = 2$.

11.8 Graph the straight line for the means of the linear regression model

$$Y = \beta_0 + \beta_1 x + e$$

having $\beta_0 = -3$, $\beta_1 = 4$, and the normal random variable e has standard deviation 3.

11.9 Graph the straight line for the means of the linear regression model $Y = \beta_0 + \beta_1 x + e$ having $\beta_0 = 7$ and $\beta_1 = 2$.

11.10 Consider the linear regression model

$$Y = \beta_0 + \beta_1 x + e$$

where $\beta_0 = -2$, $\beta_1 = -1$, and the normal random variable e has standard deviation 3.

(a) What is the mean of the response Y when $x = 3$? When $x = 6$?

(b) Will the response at $x = 3$ always be larger than that at $x = 6$? Explain.

11.11 Consider the following linear regression model $Y = \beta_0 + \beta_1 x + e$, where $\beta_0 = 4$, $\beta_1 = 3$, and the normal random variable e has the standard deviation 4.

(a) What is the mean of the response Y when $x = 4$? When $x = 5$?

(b) Will the response at $x = 5$ always be larger than that at $x = 4$? Explain.

4. THE METHOD OF LEAST SQUARES

Let us tentatively assume that the preceding formulation of the model is correct. We can then proceed to estimate the regression line and solve a few related inference problems. The problem of estimating the regression parameters β_0 and β_1 can be viewed as fitting the best straight line of the y to x relationship in the scatter diagram. One can draw a line by eyeballing the scatter diagram, but such a judgment may be open to dispute. Moreover, statistical inferences cannot be based on a line that is estimated subjectively. On the other hand, the **method of least squares** is an objective and efficient method of determining the best fitting straight line. Moreover, this method is quite versatile because its application extends beyond the simple **straight line regression model.**

Suppose that an arbitrary line $y = b_0 + b_1 x$ is drawn on the scatter diagram as it is in Figure 4. At the value x_i of the independent variable, the y value predicted by this line is $b_0 + b_1 x_i$ whereas the observed value is y_i. The discrepancy

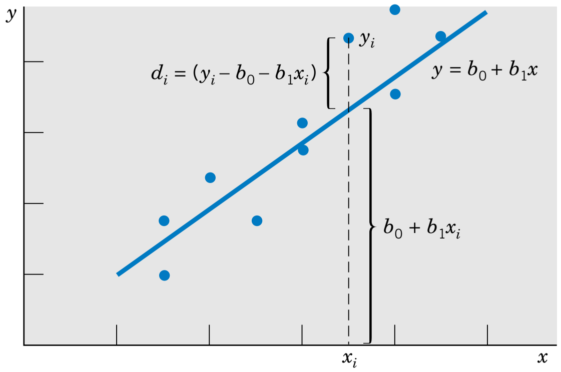


Figure 4 Deviations of the observations from a line $y = b_0 + b_1x$.

between the observed and predicted y values is then $y_i - b_0 - b_1x_i = d_i$, which is the **vertical** distance of the point from the line.

Considering such discrepancies at all the n points, we take

$$D = \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1x_i)^2$$

as an overall measure of the discrepancy of the observed points from the trial line $y = b_0 + b_1x$. The magnitude of D obviously depends on the line that is drawn. In other words, it depends on b_0 and b_1, the two quantities that determine the trial line. A good fit will make D as small as possible. We now state the principle of least squares in general terms to indicate its usefulness to fitting many other models.

The Principle of Least Squares

Determine the values for the parameters so that the overall discrepancy

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

is minimized.

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

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

$$D = \sum_{i=1}^{n} (y_i - b_0 - b_1x_i)^2$$

The quantities b_0 and b_1 thus determined are denoted by $\hat{\beta}_0$ and $\hat{\beta}_1$, respectively, and called the **least squares estimates** of the regression parameters β_0 and β_1. The **best fitting straight line** or **best fitting regression line** is then given by the equation

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

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

Basic Notation

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

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

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

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

The quantities $\bar{x}$ and $\bar{y}$ are the sample means of the x and y values; S_{xx} and S_{yy} are the sums of squared deviations from the means, and S_{xy} is the sum of the cross products of deviations. These five summary statistics are the key ingredients for calculating the least squares estimates and handling the inference problems associated with the linear regression model. The reader may review Sections 5 and 6 of Chapter 3 where calculations of these statistics were illustrated.

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

Least squares estimator of β_0

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

Least squares estimator of β_1

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

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

Fitted (or estimated) regression line

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

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

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

is the smallest.

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

Residuals

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

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

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

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

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

The second expression for SSE, which follows after some algebraic manipulations (see Exercise 11.24), is handy for directly calculating SSE. However, we stress the importance of determining the individual residuals for their role in model checking (see Section 4, Chapter 12).

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

Estimate of Variance

The estimator of the error variance σ^2 is

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

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

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

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

$$\bar{x} = 5.9, \qquad \bar{y} = 15.1 \qquad\qquad \hat{\beta}_1 = \frac{112.1}{40.9} = 2.74$$

$$S_{xx} = 389 - \frac{(59)^2}{10} = 40.9 \qquad\qquad \hat{\beta}_0 = 15.1 - 2.74 \times 5.9 = -1.07$$

$$S_{yy} = 2651 - \frac{(151)^2}{10} = 370.9 \qquad\qquad \text{SSE} = 370.9 - \frac{(112.1)^2}{40.9} = 63.6528$$

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

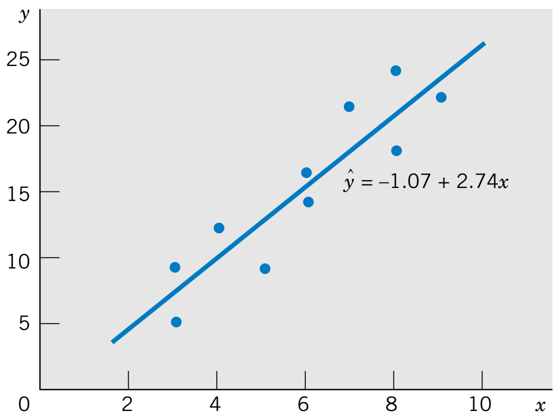


Figure 5 The least squares regression line for the data given in Table 1.

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

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

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

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

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

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

The estimate of the variance σ^2 is

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

The calculations involved in a regression analysis become increasingly tedious with larger data sets. Access to a computer proves to be a considerable advantage. Table 4 illustrates a part of the computer-based analysis of linear regression using the data of Example 4 and the MINITAB package. For a more complete regression analysis, see Table 5 in Section 6.4.

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

Data: C11T3 DAT

C1: 3 3 4 5 6 6 7 8 8 9
C2: 9 5 12 9 14 16 22 18 24 22

Dialog box:

Stat > **Regression** > **Regression**
Type *C2* in **Response**
Type *C1* in **Predictors.** Click **OK.**

Output:

Regression Analysis

The regression equation is
$y = -1.07 + 2.74x$

Exercises

11.12 Given the five pairs of (x, y) values

x	0	1	6	3	5
y	6	5	2	4	3

(a) Construct a scatter diagram.

(b) Calculate $\bar{x}$, $\bar{y}$, S_{xx}, S_{xy}, and S_{yy}.

(c) Calculate the least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$.

(d) Determine the fitted line and draw the line on the scatter diagram.

11.13 Given these six pairs of (x, y) values,

x	1	2	3	3	4	5
y	8	4	5	2	2	0

(a) Plot the scatter diagram.

(b) Calculate $\bar{x}$, $\bar{y}$, S_{xx}, S_{xy}, and S_{yy}.

(c) Calculate the least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$.

(d) Determine the fitted line and draw the line on the scatter diagram.

11.14 Refer to Exercise 11.12.

(a) Find the residuals and verify that they sum to zero.

(b) Calculate the residual sum of squares SSE by

(i) Adding the squares of the residuals.

(ii) Using the formula
$\text{SSE} = S_{yy} - S_{xy}^2 / S_{xx}$

(c) Obtain the estimate of σ^2.

11.15 Refer to Exercise 11.13.

(a) Find the residuals and verify that they sum to zero.

(b) Calculate the residual sums of squares SSE by

(i) Adding the squares of the residuals.

(ii) Using the formula
$\text{SSE} = S_{yy} - S_{xy}^2 / S_{xx}$

(c) Obtain the estimate of σ^2.

11.16 Given the five pairs of (x, y) values

x	0	1	2	3	4
y	3	2	5	6	9

(a) Calculate $\bar{x}, \bar{y}, S_{xx}, S_{xy}$, and S_{yy}.

(b) Calculate the least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$.

(c) Determine the fitted line.

11.17 Given the five pairs of (x, y) values

x	0	2	4	6	8
y	4	3	6	8	9

(a) Calculate $\bar{x}, \bar{y}, S_{xx}, S_{xy}$, and S_{yy}.

(b) Calculate the least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$.

(c) Determine the fitted line.

11.18 Computing from a data set of (x, y) values, we obtained the following summary statistics.

$$\begin{array}{lll} n = 14 & \bar{x} = 3.5 & \bar{y} = 2.84 \\ S_{xx} = 10.82 & S_{xy} = 2.677 & S_{yy} = 1.125 \end{array}$$

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

(b) Calculate the residual sum of squares.

(c) Estimate σ^2.

11.19 Computing from a data set of (x, y) values, we obtained the following summary statistics.

$$\begin{array}{lll} n = 20 & \bar{x} = 1.4 & \bar{y} = 5.2 \\ S_{xx} = 28.2 & S_{xy} = 3.25 & S_{yy} = 2.01 \end{array}$$

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

(b) Calculate the residual sum of squares.

(c) Estimate σ^2.

11.20 The data on female wolves in Table D.9 of the Data Bank concerning body weight (lb) and body length (cm) are

Weight	57	84	90	71	77	68	73
Body length	123	129	143	125	122	125	122

(a) Obtain the least squares fit of body weight to the predictor body length.

(b) Calculate the residual sum of squares.

(c) Estimate σ^2.

11.21 Refer to the data on female wolves in Exercise 11.20.

(a) Obtain the least squares fit of body length to the predictor body weight.

(b) Calculate the residual sum of squares.

(c) Estimate σ^2.

(d) Compare your answer in part (a) with your answer to part (a) of Exercise 11.20. Should the two answers be the same? Why or why not?

11.22 Using the formulas of $\hat{\beta}_1$ and SSE, show that SSE can also be expressed as

(a) $\text{SSE} = S_{yy} - \hat{\beta}_1 S_{xy}$

(b) $\text{SSE} = S_{yy} - \hat{\beta}_1^2 S_{xx}$

11.23 Referring to the formulas of $\hat{\beta}_0$ and $\hat{\beta}_1$, show that the point $(\bar{x}, \bar{y})$ lies on the fitted regression line.

11.24 To see why the residuals always sum to zero, refer to the formulas of $\hat{\beta}_0$ and $\hat{\beta}_1$ and verify that

(a) The predicted values are $\hat{y}_i = \bar{y} + \hat{\beta}_1(x_i - \bar{x})$.

(b) The residuals are

$$\hat{e}_i = y_i - \hat{y}_i = (y_i - \bar{y}) - \hat{\beta}_1(x_i - \bar{x})$$

Then show that $\sum_{i=1}^{n} \hat{e}_i = 0$.

(c) Verify that $\sum_{i=1}^{n} \hat{e}_i^2 = S_{yy} + \hat{\beta}_1^2 S_{xx} - 2\hat{\beta}_1 S_{xy} = S_{yy} - S_{xy}^2 / S_{xx}$.

5. THE SAMPLING VARIABILITY OF THE LEAST SQUARES ESTIMATORS—TOOLS FOR INFERENCE

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

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

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

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

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

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

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

To estimate the standard error, use

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

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

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

Inferences about the intercept β_0 are based on the t distribution

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

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

Estimated standard error

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

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

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

6. IMPORTANT INFERENCE PROBLEMS

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

6.1. INFERENCE CONCERNING THE SLOPE β_1

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

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

This does not change with a change in x if and only if $\beta_1 = 0$. We can therefore test the null hypothesis $H_0\colon \beta_1 = 0$ against a one- or a two-sided alternative, depending on the nature of the relation that is anticipated. If we refer to the boxed statement (2) of Section 5, the null hypothesis $H_0\colon \beta_1 = 0$ is to be tested using the test statistic

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

Example 5 A Test to Establish That Duration of Relief Increases with Dosage

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

SOLUTION For an increasing relation, we must have $\beta_1 > 0$. Therefore, we are to test the null hypothesis $H_0 : \beta_1 = 0$ versus the one-sided alternative $H_1 : \beta_1 > 0$. We select $\alpha = .05$. Since $t_{.05} = 1.860$, with d.f. $= 8$ we set the rejection region $R : T \geq 1.860$. Using the calculations that follow Table 2, we have

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

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

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

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

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

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

A warning is in order here concerning the interpretation of the test of $H_0 : \beta_1 = 0$. If H_0 is not rejected, we may be tempted to conclude that y does not depend on x. Such an unqualified statement may be erroneous. First, the absence of a linear relation has only been established over the range of the x values in the experiment. It may be that x was just not varied enough to influence y. Second, the interpretation of lack of dependence on x is valid only if our model formulation is correct. If the scatter diagram depicts a relation on a curve but we inadvertently formulate a linear model and test $H_0 : \beta_1 = 0$, the conclusion that H_0 is not rejected should be interpreted to mean “no linear relation,” rather than “no relation.” We elaborate on this point further in Section 7. Our present viewpoint is to assume that the model is correctly formulated and discuss the various inference problems associated with it.

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

The **test of the null hypothesis**

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

is based on

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

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

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

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

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

Example 6 A Confidence Interval for β_1

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

SOLUTION In Example 5, we found that $\hat{\beta}_1 = 2.74$ and $s/\sqrt{S_{xx}} = .441$. The required confidence interval is given by

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

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

6.2. INFERENCE ABOUT THE INTERCEPT β_0

Although somewhat less important in practice, inferences similar to those outlined in Section 6.1 can be provided for the parameter β_0. The procedures are again based on the t distribution with d.f. $= n - 2$, stated for $\hat{\beta}_0$ in Section 5. In particular,

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

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

To illustrate this formula, let us consider the data of Table 1. In Table 3, we have found $\hat{\beta}_0 = -1.07$, $\bar{x} = 5.9$, and $S_{xx} = 40.9$. Also, $s = 2.8207$. Therefore, a 95% confidence interval for β_0 is calculated as

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

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

Note that β_0 represents the mean response corresponding to the value 0 for the input variable x. In the drug evaluation problem of Example 4, the parameter β_0 is of little practical interest because the range of x values covered in the experiment was 3 to 9 and it would be unrealistic to extend the line to $x = 0$. In fact, the estimate $\hat{\beta}_0 = -1.07$ does not have an interpretation as a (time) duration of relief.

6.3. ESTIMATION OF THE MEAN RESPONSE FOR A SPECIFIED *x* VALUE

Often, the objective in a regression study is to employ the fitted regression in estimating the expected response corresponding to a specified level of the input variable. For example, we may want to estimate the expected duration of relief for a specified dosage x^* of the drug. According to the linear model described in Section 3, the expected response at a value x^* of the input variable x is given by $\beta_0 + \beta_1 x^*$. The expected response is estimated by $\hat{\beta}_0 + \hat{\beta}_1 x^*$ which is the ordinate of the fitted regression line at $x = x^*$. Referring to statement (3) of Section 5, we determine that the t distribution can be used to construct confidence intervals or test hypotheses.

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

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

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

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

Example 7 A Confidence Interval for the Expected Duration of Relief

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

SOLUTION The fitted regression line is

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

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

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

$$\begin{aligned}\text{Estimated standard error} &= s\sqrt{\frac{1}{10} + \frac{(6 - 5.9)^2}{40.9}} \\ &= 2.8207 \times .3166 = .893\end{aligned}$$

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

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

We are 95% confident that 6 milligrams of the drug produces an average duration of relief that is between about 13.3 and 17.4 days.

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

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

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

A 95% confidence interval is

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

The formula for the standard error shows that when x^* is close to $\bar{x}$, the standard error is smaller than it is when x^* is far removed from $\bar{x}$. This is confirmed by Example 7, where the standard error at $x^* = 9.5$ can be seen to be more than twice as large as the value at $x^* = 6$. Consequently, the confidence interval for the former is also wider. In general, estimation is more precise near the mean $\bar{x}$ than it is for values of the x variable that lie far from the mean.

Caution: Extreme caution should be exercised in extending a fitted regression line to make long-range predictions far away from the range of x values covered in the experiment. Not only does the confidence interval become so wide that predictions based on it can be extremely unreliable, but an even greater danger exists. If the pattern of the relationship between the variables changes drastically at a distant value of x, the data provide no information with which to detect such a change. Figure 6 illustrates this situation. We would observe a good linear relationship if we experimented with x values in the 5 to 10 range, but if the fitted line were extended to estimate the response at $x^* = 20$, then our estimate would drastically miss the mark.

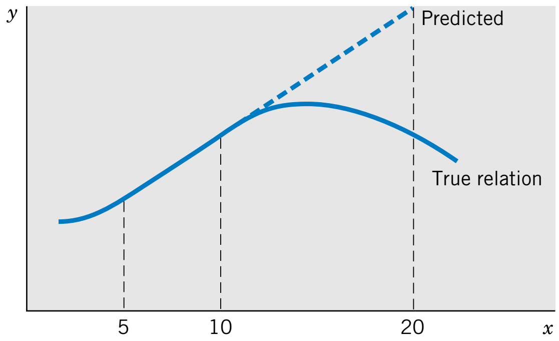


Figure 6 Danger in long-range prediction.

6.4. PREDICTION OF A SINGLE RESPONSE FOR A SPECIFIED *x* VALUE

Suppose that we give a specified dosage x^* of the drug to a **single** patient and we want to predict the duration of relief from the symptoms of allergy. This problem is different from the one considered in Section 6.3, where we were interested in estimating the mean duration of relief for the population of **all** patients given the dosage x^*. The prediction is still determined from the fitted line; that is, the predicted value of the response is $\hat{\beta}_0 + \hat{\beta}_1 x^*$ as it was in the preceding case. However, the standard error of the prediction here is larger, because a single observation is more uncertain than the mean of the population distribution. We now give the formula of the estimated standard error for this case.

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

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

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

Example 8 Calculating a Prediction Interval for a Future Trial

Once again, consider the drug trial data given in Table 1. A new trial is to be made on a single patient with the dosage $x^* = 6.5$ milligrams. Predict the duration of relief and give a 95% prediction interval for the duration of relief.

SOLUTION The predicted duration of relief is

$$\hat{\beta}_0 + \hat{\beta}_1 x^* = -1.07 + 2.74 \times 6.5 = 16.74 \text{ days}$$

Since $t_{.025} = 2.306$ with d.f. $= 8$, a 95% prediction interval for the new patient's duration of relief is

$$16.74 \pm 2.306 \times 2.8207 \sqrt{1 + \frac{1}{10} + \frac{(6.5 - 5.9)^2}{40.9}}$$

$$= 16.74 \pm 6.85 \qquad \text{or} \qquad (9.89, 23.59)$$

This means we are 95% confident that this particular patient will have relief from symptoms of allergy for about 9.9 to 23.6 days.

In the preceding discussion, we have used the data of Example 4 to illustrate the various inferences associated with a straight-line regression model. Example 9 gives applications to a different data set.

Example 9 Fitting a Straight Line Relation of Skill to the Amount of Training

In a study to determine how the skill in doing a complex assembly job is influenced by the amount of training, 15 new recruits were given varying amounts of training ranging between 3 and 12 hours. After the training, their times to perform the job were recorded. After denoting x = duration of training (in hours) and y = time to do the job (in minutes), the following summary statistics were calculated.

$$\bar{x} = 7.2 \qquad S_{xx} = 33.6 \qquad S_{xy} = -57.2$$
$$\bar{y} = 45.6 \qquad S_{yy} = 160.2$$

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

(b) Do the data substantiate the claim that the job time decreases with more hours of training?

(c) Estimate the mean job time for 9 hours of training and construct a 95% confidence interval.

(d) Find the predicted y for $x = 35$ hours and comment on the result.

SOLUTION Using the summary statistics we find:

(a) The least squares estimates are

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{-57.2}{33.6} = -1.702$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 45.6 - (-1.702) \times 7.2 = 57.85$$

So, the equation of the fitted line is

$$\hat{y} = 57.85 - 1.702x$$

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

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

We select $\alpha = .01$. Since $t_{.01} = 2.650$ with d.f. $= 13$, we set the left-sided rejection region $R : T \le -2.650$. We calculate

$$\text{SSE} = S_{yy} - \frac{S_{xy}^2}{S_{xx}} = 160.2 - \frac{(-57.2)^2}{33.6} = 62.824$$

$$s = \sqrt{\frac{\text{SSE}}{n - 2}} = \sqrt{\frac{62.824}{13}} = 2.198$$

$$\text{Estimated S.E. } (\hat{\beta}_1) = \frac{s}{\sqrt{S_{xx}}} = \frac{2.198}{\sqrt{33.6}} = .379$$

The t statistic has the value

$$t = \frac{-1.702}{.379} = -4.49$$

Since the observed $t = -4.49$ is less than -2.650, H_0 is rejected with $\alpha = .01$. The P–value is smaller than .010. (See Figure 7.)

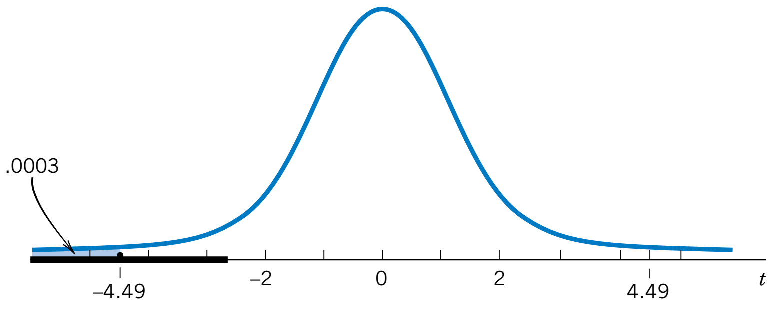


Figure 7 P–value $=$ .0003 for one-sided test.

A computer calculation gives

$$P[T \leq -4.49] = .0003$$

We conclude that increasing the duration of training significantly reduces the mean job time within the range covered in the experiment.

(c) The expected job time corresponding to $x^* = 9$ hours is estimated as

$$\begin{aligned}\hat{\beta}_0 + \hat{\beta}_1 x^* &= 57.85 + (-1.702) \times 9 \\ &= 42.53 \text{ minutes}\end{aligned}$$

and its

$$\text{Estimated S.E.} = s\sqrt{\frac{1}{15} + \frac{(9 - 7.2)^2}{33.6}} = .888$$

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

$$42.53 \pm 2.160 \times .888 = 42.53 \pm 1.92 \qquad \text{or} \qquad (40.6, 44.5) \text{ minutes}$$

(d) Since $x = 35$ hours is far beyond the experimental range of 3 to 12 hours, it is not sensible to predict y at $x = 35$ using the fitted regression line. Here a formal calculation gives

$$\begin{aligned}\text{Predicted job time} &= 57.85 - 1.702 \times 35 \\ &= -1.72 \text{ minutes}\end{aligned}$$

which is a nonsensical result.

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

TABLE 5 MINITAB Computer Output for the Data in Example 4

```
THE REGRESSION EQUATION IS
Y = −1.07 + 2.74X

PREDICTOR        COEF      STDEV     T-RATIO       P
CONSTANT       −1.071      2.751       −0.39   0.707
X              2.7408     0.4411        6.21   0.000

S = 2.821  R-SQ = 82.8%

ANALYSIS OF VARIANCE

SOURCE           DF          SS          MS        F          P
REGRESSION        1      307.25      307.25    38.62      0.000
ERROR             8       63.65        7.96
TOTAL             9      370.90
```

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

TABLE 6 SAS Computer Output for the Data in Example 4

```
MODEL: MODEL 1
DEPENDENT VARIABLE:  Y

                      ANALYSIS OF VARIANCE

                        SUM OF          MEAN
SOURCE       DF        SQUARES        SQUARE     F VALUE  PROB > F

MODEL         1      307.24719     307.24719      38.615    0.0003
ERROR         8       63.65281       7.95660
C TOTAL       9      370.90000

   ROOT MSE      2.82074     R-SQUARE      0.8284

                      PARAMETER ESTIMATES

                  PARAMETER     STANDARD    T FOR HO:
VARIABLE  DF       ESTIMATE        ERROR  PARAMETER = 0  PROB > |T|

INTERCEP   1      −1.070905   2.75091359         −0.389      0.7072
X1         1       2.740831   0.44106455          6.214      0.0003
```

Example 10 Predicting the Number of Situps after a Semester of Conditioning

Refer to the physical fitness data in Table D.5 of the Data Bank. Using the data on numbers of situps:

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

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

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

SOLUTION The scatter plot in Figure 8 suggests that a straight line may model the expected value of posttest situps given the number of pretest situps. Here x is the number of pretest situps and y is the number of posttest situps. We use MINITAB statistical software to obtain the output

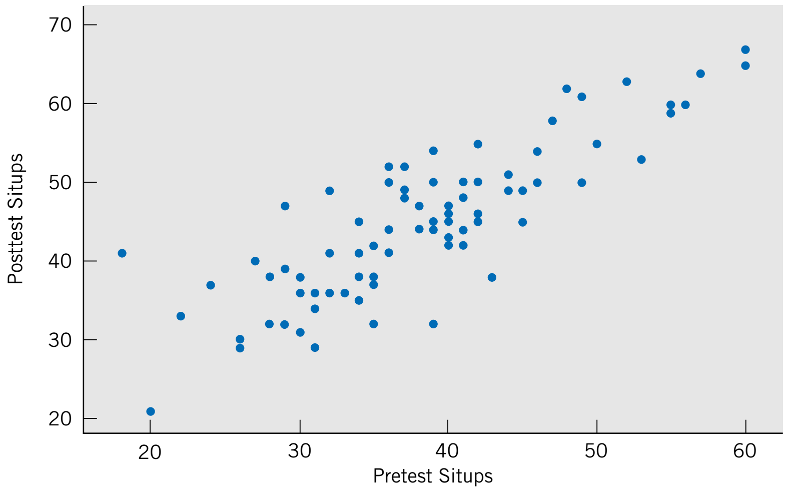


Figure 8 Scatter plot of number of situps.

```
Regression Analysis: Post Situps versus Pre Situps

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

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

S  =  5.17893     R-Sq  =  71.5%     R-Sq(adj)  =  71.1%

Analysis of Variance

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

Predicted Values for New Observations

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

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

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

Exercises

11.25 Given the five pairs of (x, y) values

x	0	1	6	3	5
y	5	4	1	3	2

(a) Calculate the least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$. Also estimate the error variance σ^2.

(b) Test $H_0\colon \beta_1 = 0$ versus $H_1\colon \beta_1 \neq 0$ with $\alpha = .05$.

(c) Estimate the expected y value corresponding to $x = 2.5$ and give a 90% confidence interval.

11.26 Refer to Exercise 11.25. Construct a 90% confidence interval for the intercept β_0.

11.27 Refer to Exercise 11.25. Obtain a 95% confidence interval for β_1.

11.28 Given these five pairs of (x, y) values,

x	2	3	4	5	6
y	.9	2.1	2.4	3.3	3.8

(a) Calculate the least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$. Also estimate the error variance σ^2.

(b) Test $H_0\colon \beta_1 = 1$ versus $H_1\colon \beta_1 \neq 1$ with $\alpha = .05$.

(c) Estimate the expected y value corresponding to $x = 3.5$ and give a 95% confidence interval.

(d) Construct a 90% confidence interval for the intercept β_0.

11.29 For a random sample of seven homes that are recently sold in a city suburb, the assessed values x and the selling prices y are

($1000)		($1000)	
x	y	x	y
183.5	188.0	210.2	211.0
190.0	191.2	194.6	199.0
170.5	176.2	220.0	218.0
200.8	207.0		

(a) Plot the scatter diagram.

(b) Determine the equation of the least squares regression line and draw this line on the scatter diagram.

(c) Construct a 95% confidence interval for the slope of the regression line.

11.30 Refer to the data in Exercise 11.29.

(a) Estimate the expected selling price of homes that were assessed at $190,000 and construct a 95% confidence interval.

(b) For a single home that was assessed at $190,000, give a 95% prediction interval for the selling price.

11.31 In an experiment designed to determine the relationship between the doses of a compost fertilizer x and the yield of a crop y, the following summary statistics are recorded:

$$n = 15 \qquad \bar{x} = 1.1 \qquad \bar{y} = 4.6$$
$$S_{xx} = 4.2 \qquad S_{yy} = 12.2 \qquad S_{xy} = 6.7$$

Assume a linear relationship.

(a) Find the equation of the least squares regression line.

(b) Compute the error sum of squares and estimate σ^2.

(c) Do the data establish the experimenter's conjecture that, over the range of x values covered in the study, the average increase in yield per unit increase in the compost dose is more than 1.3?

11.32 Refer to Exercise 11.31.

(a) Construct a 95% confidence interval for the expected yield corresponding to $x = 1.2$.

(b) Construct a 95% confidence interval for the expected yield corresponding to $x = 1.5$.

11.33 According to the computer output in Table 7:

(a) What model is fitted?

(b) Test, with $\alpha = .05$, if the x term is needed in the model.

11.34 According to the computer output in Table 7:

(a) Predict the mean response when $x = 5000$.

(b) Find a 90% confidence interval for the mean response when $x = 5000$. You will need the additional information $n = 30$, $\bar{x} = 8354$, and $\Sigma(x_i - \bar{x})^2 = 97{,}599{,}296$.

11.35 According to the computer output in Table 8:

(a) What model is fitted?

(b) Test, with $\alpha = .05$, if the x term is needed in the model.

11.36 According to the computer output in Table 8:

(a) Predict the mean response when $x = 3$.

(b) Find a 90% confidence interval for the mean response when $x = 3$. You will need the additional information $n = 25$, $\bar{x} = 1.793$, and $\Sigma(x_i - \bar{x})^2 = 1.848$.

(c) Find a 90% confidence interval for the mean response when $x = 2$.

11.37 Consider the data on male wolves in Table D.9 of the Data Bank concerning age (years) and canine length (mm).

(a) Obtain the least squares fit of canine length to the predictor age.

(b) Test $H_0: \beta_1 = 0$ versus $H_1: \beta_1 \neq 0$ with $\alpha = .05$.

(c) Obtain a 90% confidence interval for the canine length when age is $x = 4$.

(d) Obtain a 90% prediction interval for the canine length of an individual wolf when the age is $x = 4$.

TABLE 7 Computer Output for Exercises 11.33 and 11.34

```
THE REGRESSION EQUATION IS
Y  =  994  +  0.104X

PREDICTOR            COEF          STDEV      T-RATIO          P
CONSTANT            994.0          254.7         3.90      0.001
X                 0.10373        0.02978         3.48      0.002

S   =  2.99.4    R-SQ  =  30.2%

ANALYSIS OF VARIANCE

SOURCE            DF           SS            MS         F          P
REGRESSION         1      1087765       1087765     12.14      0.002
ERROR             28      2509820         89636
TOTAL             29      3597585
```

TABLE 8 Computer Output for Exercises 11.35 and 11.36

```
THE REGRESSION EQUATION IS
Y  =  0.338  +  0.831X

PREDICTOR            COEF          STDEV        T-RATIO          P
CONSTANT           0.3381         0.1579           2.14      0.043
X                 0.83099        0.08702           9.55      0.000

S  =  0.1208    R-SQ  =  79.9%

ANALYSIS OF VARIANCE

SOURCE           DF         SS         MS          F          P
REGRESSION        1     1.3318     1.3318      91.20      0.000
ERROR            23     0.3359     0.0146
TOTAL            24     1.6676
```

7. THE STRENGTH OF A LINEAR RELATION

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

$$\underbrace{y_i}_{\substack{\text{Observed}\\ y \text{ value}}} = \underbrace{(\hat{\beta}_0 + \hat{\beta}_1 x_i)}_{\substack{\text{Explained by}\\ \text{linear relation}}} + \underbrace{(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)}_{\substack{\text{Residual or}\\ \text{deviation from}\\ \text{linear relation}}}$$

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

We can consider the sum of squares of the residuals

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

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

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

of which SSE forms a part. The difference

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

forms the other part. Motivated by the decomposition of the observation y_i, just given, we can now consider a decomposition of the variability of the y values.

Decomposition of Variability

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

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

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

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

From Section 6 of Chapter 3, recall that the quantity

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

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

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

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

which is the square of the sample correlation coefficient r.

Example 11 The Proportion of Variability in Duration Explained by Dosage

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

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

Fitted regression line

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

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

SOLUTION To answer this question, we calculate

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

This means that 83% of the variability in y is explained by linear regression, and the linear model seems satisfactory in this respect.

Example 12 Proportion of Variation Explained in Number of Situps

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

SOLUTION Repeating the relevant part of the computer output from Example 10,

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

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

S  =  5.17893   R-Sq  =  71.5%   R-Sq(adj)  =  71.1%

Analysis of Variance

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

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

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

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

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

1. There is little relation between the variables in the sense that the scatter diagram fails to exhibit any pattern, as illustrated in Figure 9*a*. In this case, the use of a different regression model is not likely to reduce the SSE or explain a substantial part of S_{yy}.

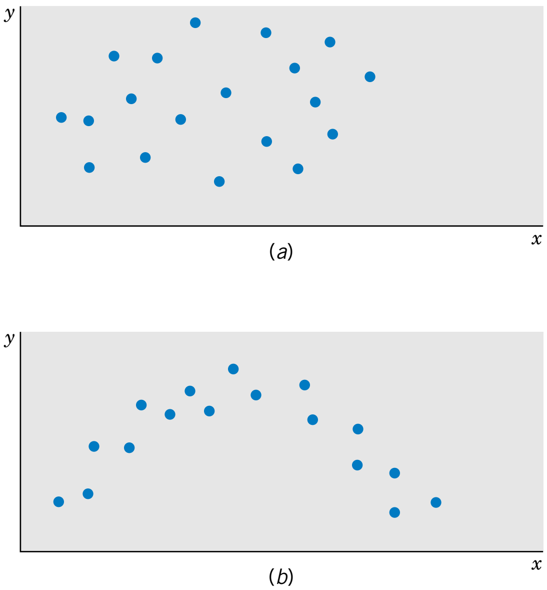


Figure 9 Scatter diagram patterns: (*a*) No relation. (*b*) A nonlinear relation.

2. There is a prominent relation but it is nonlinear in nature; that is, the scatter is banded around a curve rather than a line. The part of S_{yy} that is explained by straight line regression is small because the model is inappropriate. Some other relationship may improve the fit substantially. Figure 9*b* illustrates such a case, where the SSE can be reduced by fitting a suitable curve to the data.

Exercises

11.38 Computing from a data set of (x, y) values, the following summary statistics are obtained.

$$\begin{array}{lll} n = 12 & \bar{x} = 1.2 & \bar{y} = 5.1 \\ S_{xx} = 15.10 & S_{xy} = 2.31 & S_{yy} = 2.35 \end{array}$$

Determine the proportion of variation in y that is explained by linear regression.

11.39 Computing from a data set of (x, y) values, the following summary statistics are obtained:

$$\begin{array}{lll} n = 16 & \bar{x} = 7.3 & \bar{y} = 2.1 \\ S_{xx} = 43.2 & S_{xy} = 9.4 & S_{yy} = 6.7 \end{array}$$

Determine the proportion of variation in y that is explained by linear regression.

11.40 Given $S_{xx} = 14.2$, $S_{yy} = 18.3$, and $S_{xy} = 10.3$, determine the proportion of variation in y that is explained by linear regression.

11.41 A calculation shows that $S_{xx} = 9.2$, $S_{yy} = 49$, and $S_{xy} = 16$. Determine the proportion of variation in y that is explained by linear regression.

11.42 Refer to Exercise 11.25.

(a) What proportion of the y variability is explained by the linear regression on x?

(b) Find the sample correlation coefficient.

(c) Calculate the residual sum of squares.

(d) Estimate σ^2.

11.43 Refer to Exercise 11.28.

(a) What proportion of y variability is explained by the linear regression on x?

(b) Find the sample correlation coefficient.

11.44 Refer to Exercise 11.33. According to the computer output in Table 7, find the proportion of y variability explained by x.

11.45 Refer to Exercise 11.35. According to the computer output in Table 8, find the proportion of y variability explained by x.

11.46 Consider the data on wolves in Table D.9 of the Data Bank concerning body length (cm) and weight (lb). Calculate the correlation coefficient r and r^2 for

(a) all wolves.

(b) male wolves.

(c) female wolves.

(d) Comment on the differences in your answers. Make a multiple scatter diagram (see Chapter 3) to clarify the situation.

*11.47 (a) Show that the sample correlation coefficient r and the slope $\hat{\beta}_1$ of the fitted regression line are related as

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

(b) Show that SSE $= (1 - r^2)\, S_{yy}$.

*11.48 Show that the SS due to regression, S_{xy}^2 / S_{xx}, can also be expressed as $\hat{\beta}_1^2 S_{xx}$.

8. REMARKS ABOUT THE STRAIGHT LINE MODEL ASSUMPTIONS

A regression study is not completed by performing a few routine hypothesis tests and constructing confidence intervals for parameters on the basis of the formulas given in Section 5. Such conclusions can be seriously misleading if the assumptions made in the model formulations are grossly incompatible with the data. It is therefore essential to check the data carefully for indications of any violation of the assumptions. To review, the assumptions involved in the formulation of our straight line model are briefly stated again.

1. The underlying relation is linear.
2. Independence of errors.

3. Constant variance.
4. Normal distribution.

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

USING STATISTICS WISELY

1. As a first step, plot the response variable versus the predictor variable. Examine the plot to see if a linear or other relationship exists.
2. Apply the principal of least squares to obtain estimates of the coefficients when fitting a straight line model.
3. Determine the $100(1 - \alpha)\%$ confidence intervals for the slope and intercept parameters. You can also look at P–values to decide whether or not they are non-zero. If not, you can use the fitted line for prediction.
4. Don't use the fitted line to make predictions beyond the range of the data. The model may be different over that range.

KEY IDEAS AND FORMULAS

In its simplest form, **regression analysis** deals with studying the manner in which the **response variable** y depends on a **predictor variable** x. Sometimes, the response variable is called the **dependent variable** and predictor variable is called the **independent** or **input variable.**

The first important step in studying the relation between the variables y and x is to plot the **scatter diagram** of the data (x_i, y_i), $i = 1, \ldots, n$. If this plot indicates an approximate linear relation, a **straight line regression model** is formulated:

$$\begin{array}{ccccc} \text{Response} & = & \text{A straight line in } x & + & \text{Random error} \\ Y_i & = & \beta_0 + \beta_1 x_i & + & e_i \end{array}$$

The random errors are assumed to be independent, normally distributed, and have mean 0 and equal standard deviations σ.

The **least squares estimate of** $\hat{\beta}_0$ and **least squares estimate of** $\hat{\beta}_1$ are obtained by the **method of least squares,** which minimizes the sum of squared deviations $\Sigma\,(y_i - \beta_0 - \beta_1 x_i)^2$. The least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ determine the **best fitting regression line** $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, which serves to predict y from x.

The differences $y_i - \hat{y}_i$ = Observed response − Predicted response are called the **residuals.**

The adequacy of a straight line fit is measured by r^2, which represents the proportion of y variability that is explained by the linear relation between y and x. A low value of r^2 only indicates that a linear relation is not appropriate—there may still be a relation on a curve.

Least squares estimators

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Best fitting straight line

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

Residuals

$$\hat{e}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

Residual sum of squares

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

Estimate of variance σ^2

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

Inferences

1. Inferences concerning the **slope** β_1 are based on the

   Estimator $\hat{\beta}_1$

   $$\text{Estimated S.E.} = \frac{S}{\sqrt{S_{xx}}}$$

   and the sampling distribution

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

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

$$\hat{\beta}_1 \pm t_{\alpha/2} \frac{S}{\sqrt{S_{xx}}}$$

To test $H_0 : \beta_1 = \beta_{10}$, the test statistic is

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

2. Inferences concerning the **intercept** β_0 are based on the

Estimator $\hat{\beta}_0$

$$\text{Estimated S.E.} = S\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$$

and the sampling distribution

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

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

$$\hat{\beta}_0 \pm t_{\alpha/2}\, S\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$$

3. At a specified $x = x^*$, the expected response is $\beta_0 + \beta_1 x^*$. Inferences about the **expected response** are based on the

Estimator $\hat{\beta}_0 + \hat{\beta}_1 x^*$

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

A $100(1 - \alpha)\%$ confidence interval for the expected response at x^* is given by

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

4. A **single response** at a specified $x = x^*$ is predicted by $\hat{\beta}_0 + \hat{\beta}_1 x^*$ with

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

A $100(1-\alpha)\%$ prediction interval for a single response is

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

Decomposition of Variability

The total sum of squares S_{yy} is the sum of two components, the sum of squares due to regression S_{xy}^2/S_{xx} and the sum of squares due to error

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

Variability explained by the linear relation $= \dfrac{S_{xy}^2}{S_{xx}} = \hat{\beta}_1^2 S_{xx}$

Residual or unexplained variability $=$ SSE

Total y variability $= S_{yy}$

The strength of a linear relation, **or** proportion of y variability explained by linear regression

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

Sample correlation coefficient

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

TECHNOLOGY

Fitting a straight line and calculating the correlation coefficient

MINITAB

Fitting a straight line—regression analysis

Begin with the values for the predictor variable x in *C1* and the response variable y in C2.

> **Stat > Regression < Regression.**
> Type C2 in **Response.** Type *C1* in **Predictors.**
> Click **OK.**

To calculate the correlation coefficient, start as above with data in *C1* and C2.

> **Stat > Basic Statistics > Correlation.**
> Type *C1 C2* in **Variables.** Click **OK.**

EXCEL

Fitting a straight line—regression analysis

Begin with the values of the predictor variable in column A and the values of the response variable in column B. To plot,

> Highlight the data and go to **Insert** and then **Chart.**
> Select **XY(Scatter)** and click **Finish.**
> Go to **Chart** and then **Add Trendline.**
> Click on the **Options** tab and check **Display equation on chart.**
> Click **OK.**

To obtain a more complete statistical analysis and diagnostic plots, instead use the following steps:

> Select **Tools** and then **Data Analysis.**
> Select **Regression.** Click **OK.**
> With the cursor in the **Y Range,** highlight the data in column B.
> With the cursor in the **X Range,** highlight the data in column A.
> Check boxes for **Residuals, Residual Plots,** and **Line Fit Plot.** Click **OK.**

To calculate the correlation coefficient, begin with the first variable in column A and the second in column B.

> Click on a blank cell. Select **Insert** and then **Function**
> (or click on the f_x icon).
> Select **Statistical** and then **CORREL.**
> **Highlight the data** in column A for **Array1** and highlight the data in column B for **Array2.** Click **OK.**

TI-84/-83 PLUS

Fitting a straight line—regression analysis

Enter the values of the predictor variable in **L1** and those of the response variable in **L2**.

> Select **STAT,** then **CALC,** and then **4:LinReg (ax + b).**
> With **LinReg** on the Home screen Press **Enter.**

The calculator will return the intercept *a*, slope *b*, and correlation coefficient *r*. If *r* is not shown, go to the **2nd 0:CATALOG** and select **Diagnostic.** Press **ENTER** twice. Then go back to **LinReg.**

9. REVIEW EXERCISES

11.49 Concerns that were raised for the environment near a government facility led to a study of plants. Since leaf area is difficult to measure, the leaf area (cm^2) was fit to

$$x = \text{Leaf length} \times \text{Leaf width}$$

using a least squares approach. For data collected one year, the fitted regression line is

$$\hat{y} = .2 + 0.5x$$

and $s^2 = (0.3)^2$. Comment on the size of the slope. Should it be positive or negative, less than one, equal to one, or greater than one?

11.50 Given these nine pairs of (x, y) values:

x	1	1	1	2	3	3	4	5	5
y	9	7	8	10	15	12	19	24	21

(a) Plot the scatter diagram.

(b) Calculate $\bar{x}, \bar{y}, S_{xx}, S_{yy}$, and S_{xy}.

(c) Determine the equation of the least squares fitted line and draw the line on the scatter diagram.

(d) Find the predicted y corresponding to $x = 3$.

11.51 Refer to Exercise 11.50.

(a) Find the residuals.

(b) Calculate the SSE by (i) summing the squares of the residuals and also (ii) using the formula SSE $= S_{yy} - S_{xy}^2 / S_{xx}$.

(c) Estimate the error variance.

11.52 Refer to Exercise 11.50.

(a) Construct a 95% confidence interval for the slope of the regression line.

(b) Obtain a 90% confidence interval for the expected y value corresponding to $x = 4$.

11.53 An experiment is conducted to determine how the strength y of plastic fiber depends on the size x of the droplets of a mixing polymer in suspension. Data of (x, y) values, obtained from 15 runs of the experiment, have yielded the following summary statistics.

$$\bar{x} = 8.3 \qquad \bar{y} = 54.8$$
$$S_{xx} = 5.6 \qquad S_{xy} = -12.4 \qquad S_{yy} = 38.7$$

(a) Obtain the equation of the least squares regression line.

(b) Test the null hypothesis $H_0: \beta_1 = -2$ against the alternative $H_1: \beta_1 < -2$, with $\alpha = .05$.

(c) Estimate the expected fiber strength for droplet size $x = 10$ and set a 95% confidence interval.

11.54 Refer to Exercise 11.53.

(a) Obtain the decomposition of the total y variability into two parts: one explained by linear relation and one not explained.

(b) What proportion of the y variability is explained by the straight line regression?

(c) Calculate the sample correlation coefficient between x and y.

11.55 A recent graduate moving to a new job collected a sample of monthly rent (dollars) and size (square feet) of 2-bedroom apartments in one area of a midwest city.

Size	Rent	Size	Rent
900	550	1000	650
925	575	1033	675
932	620	1050	715
940	620	1100	840

(a) Plot the scatter diagram and find the least squares fit of a straight line.

(b) Do these data substantiate the claim that the monthly rent increases with the size of the apartment? (Test with $\alpha = .05$).

(c) Give a 95% confidence interval for the expected increase in rent for one additional square foot.

(d) Give a 95% prediction interval for the monthly rent of a specific apartment having 1025 square feet.

11.56 Refer to Exercise 11.55.

(a) Calculate the sample correlation coefficient.

(b) What proportion of the y variability is explained by the fitted regression line?

11.57 A Sunday newspaper lists the following used-car prices for a foreign compact, with age x measured in years and selling price y measured in thousands of dollars.

x	y	x	y
1	16.9	5	8.9
2	12.9	7	5.6
2	13.9	7	5.7
4	13.0	8	6.0
4	8.8		

(a) Plot the scatter diagram.

(b) Determine the equation of the least squares regression line and draw this line on the scatter diagram.

(c) Construct a 95% confidence interval for the slope of the regression line.

11.58 Refer to Exercise 11.57.

(a) From the fitted regression line, determine the predicted value for the average selling price of a 5-year-old car and construct a 95% confidence interval.

(b) Determine the predicted value for a 5-year-old car to be listed in next week's paper. Construct a 90% prediction interval.

(c) Is it justifiable to predict the selling price of a 15-year-old car from the fitted regression line? Give reasons for your answer.

11.59 Again referring to Exercise 11.57, find the sample correlation coefficient between age and selling price. What proportion of the y variability is explained by the fitted straight line? Comment on the adequacy of the straight line fit.

11.60 Given

$$n = 20 \qquad \sum x = 17 \qquad \sum y = 31$$
$$\sum x^2 = 19 \qquad \sum xy = 21 \qquad \sum y^2 = 73$$

(a) Find the equation of the least squares regression line.

(b) Calculate the sample correlation coefficient between x and y.

(c) Comment on the adequacy of the straight line fit.

The Following Exercises Require a Computer

11.61 *Using the computer.* The calculations involved in a regression analysis become increasingly tedious with larger data sets. Access to a computer proves to be of considerable advantage. We repeat here a computer-based analysis of linear regression using the data of Example 4 and the MINITAB package.

The sequence of steps in MINITAB:

Data: C11T3.DAT

C1: 3 3 4 5 6 6 7 8 8 9
C2: 9 5 12 9 14 16 22 18 24 22
Dialog box:

Stat > Regression > Regression
Type *C2* in **Response**
Type *C1* in **Predictors.** Click **OK.**

produces all the results that are basic to a linear regression analysis. The important pieces in the output are shown in Table 9.

Compare Table 9 with the calculations illustrated in Sections 4 to 7. In particular, identify:

(a) The least squares estimates.

(b) The SSE.

(c) The estimated standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$.

(d) The t statistics for testing $H_0 : \beta_0 = 0$ and $H_0 : \beta_1 = 0$.

(e) r^2.

(f) The decomposition of the total sum of squares into the sum of squares explained by the linear regression and the residual sum of squares.

11.62 Consider the data on all of the wolves in Table D.9 of the Data Bank concerning body length

TABLE 9 MINITAB Regression Analysis of the Data in Example 4

```
THE REGRESSION EQUATION IS
Y = −1.07 + 2.74x

PREDICTOR        COEF     STDEV    T-RATIO        P
CONSTANT       −1.071     2.751     −0.39     0.707
X              2.7408    0.4411      6.21     0.000

S = 2.821      R-SQ = 82.8%

ANALYSIS OF VARIANCE

SOURCE         DF       SS        MS        F         P
REGRESSION      1   307.25    307.25    38.62     0.000
ERROR           8    63.65      7.96
TOTAL           9   370.90
```

(cm) and weight (lb). Using MINITAB or some other software program:

(a) Plot weight versus body length.

(b) Obtain the least squares fit of weight to the predictor variable body length.

(c) Test $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 > 0$ with $\alpha = .05$.

11.63 Refer to Exercise 11.62 and a least squares fit using the data on all of the wolves in Table D.9 of the Data Bank concerning body length (cm) and weight (lb). There is one obvious outlier, row 18 with body length 123 and weight 106, indicated in the MINITAB output. Drop this observation.

(a) Obtain the least squares fit of weight to the predictor variable body length.

(b) Test $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 > 0$ with $\alpha = .05$.

(c) Comment on any important differences between your answers to parts (a) and (b) and the answer to Exercise 11.62.

11.64 Many college students obtain college degree credits by demonstrating their proficiency on exams developed as part of the College Level Examination Program (CLEP). Based on their scores on the College Qualification Test (CQT), it would be helpful if students could predict their scores on a corresponding portion of the CLEP exam. The following data (courtesy of R. W. Johnson) are for x = Total CQT score and y = Mathematical CLEP score.

x	y	x	y
170	698	174	645
147	518	128	578
166	725	152	625
125	485	157	558
182	745	174	698
133	538	185	745
146	485	171	611
125	625	102	458
136	471	150	538
179	798	192	778

(a) Find the least squares fit of a straight line.

(b) Construct a 95% confidence interval for the slope.

(c) Construct a 95% prediction interval for the CLEP score of a student who obtains a CQT score of 150.

(d) Repeat part (c) with $x = 175$ and $x = 195$.

11.65 Crickets make a chirping sound with their wing covers. Scientists have recognized that there is a relationship between the frequency of chirps and the temperature. (There is some truth to the cartoon on p. 432.) Use the 15 measurements for the striped ground cricket to:

(a) Fit a least squares line.

(b) Obtain a 95% confidence interval for the slope.

(c) Predict the temperature when $x = 15$ chirps per second.

Chirps (per second) (x)	Temperature (°F) (y)
20.0	88.6
16.0	71.6
19.8	93.3
18.4	84.3
17.1	80.6
15.5	75.2
14.7	69.7
17.1	82.0
15.4	69.4
16.3	83.3
15.0	79.6
17.2	82.6
16.0	80.6
17.0	83.5
14.4	76.3

Source: G. Pierce, *The Songs of Insects*, Cambridge, MA: Harvard University Press, 1949, pp. 12–21.

11.66 Use MINITAB or some other software to obtain the scatter diagram, correlation coefficient, and the regression line of the final time to run 1.5 miles on the initial times given in Table D.5 of the Data Bank.

11.67 Use MINITAB or some other software program to regress the marine growth on freshwater growth for the fish growth data in Table D.7 of the Data Bank. Do separate regression analyses for:

(a) All fish.

(b) Males.

(c) Females.

Your analysis should include (i) a scatter diagram, (ii) a fitted line, (iii) a determination if β_1 differs from zero. Also find a 95% confidence interval for the population mean when the freshwater growth is 100.

11.68 The data on the maximum height and top speed of the 12 highest roller coasters, displayed in the chapter opener, are

Height	Speed
400	120
415	100
377	100
318	95
310	93
263	81
259	81
245	85
240	79
235	85
230	80
224	70

(a) Use MINITAB or some other software program to determine the proportion of variation in speed due to regression on height.

(b) What top speed is predicted for a new roller coaster of height 325 feet?

(c) What top speed is predicted for a new roller coaster of height 480 feet? What additional danger is there in this prediction?