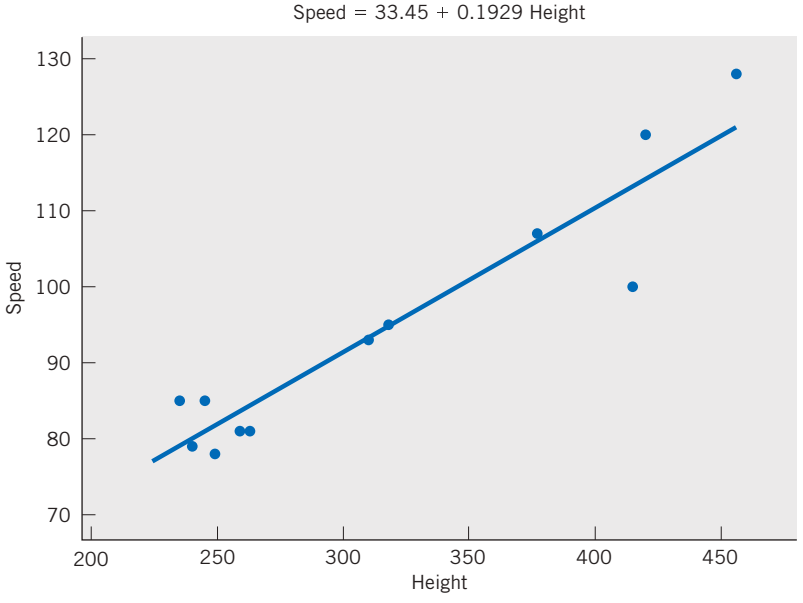
11

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 .19 miles per hour for each foot of height, or 19 miles per hour for each 100 feet in height.

This relation can be used to predict the top speed of the next 410 foot roller coaster.



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.

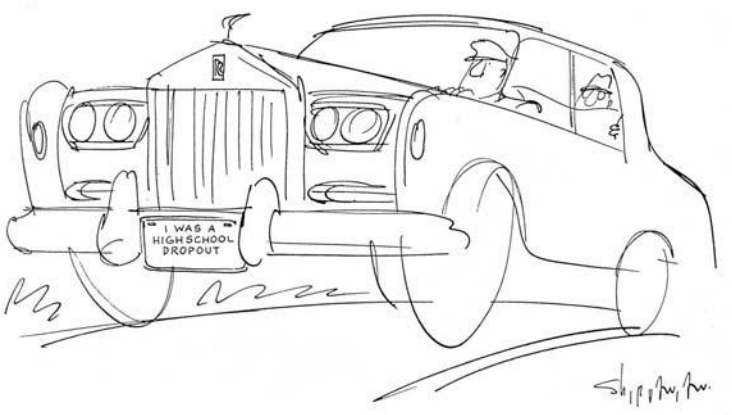
Regression analysis concerns the study of relationships between quantitative variables with the object of identifying, estimating, and validating the relationship. The estimated relationship can then be used to predict one variable from the value of the other variable(s). In this chapter, we introduce 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



Fine but you are an exception. Statistics[1], based on extensive data, confirm that earnings typically increase with each additional step in education Vahan Shirvanian, www.CartoonStock.com

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.

[1] Median weekly earnings in 2008, Bureau of Labor Statistics, Current Population Survey.

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, explanatory variable, causal variable,** or **input variable**

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

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

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 Hours of Relief y from Allergy for Ten Patients

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

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

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

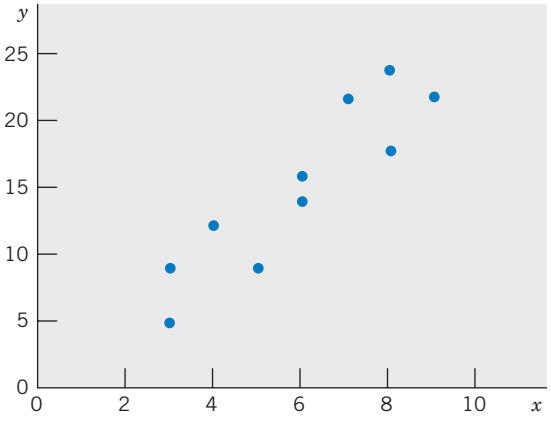


Figure 1 Scatter diagram of the data of Table 1.

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).

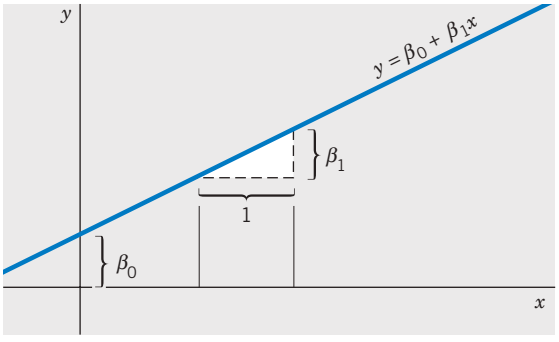


Figure 2 Graph of straight line $y = \beta_0 + \beta_1 x$.

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 the variables that influence the response, days of relief, are not even known, yet alone measured. The effects of all these variables are modeled as unobservable random variables. Given this viewpoint, we formulate the following linear regression model as a tentative representation of the mode of relationship between y and x.

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.

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

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

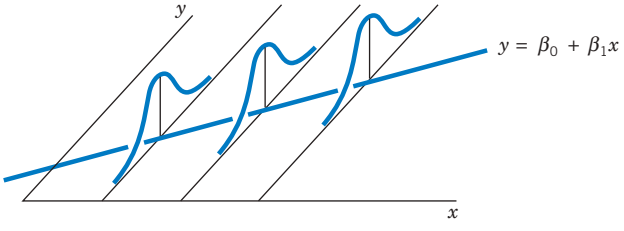


Figure 3 Normal distributions of Y with means on a straight line.

Exercises

11.1 Plot the line $y = 3 + 2x$ 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

$$y = 12x - 75$$

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.

(*continued*)

(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

$$Y = 4 + 3x + e$$

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

$$Y = 6 - 3x + e$$

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 $\beta_0 = 2$, $\beta_1 = -3$, 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 a 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

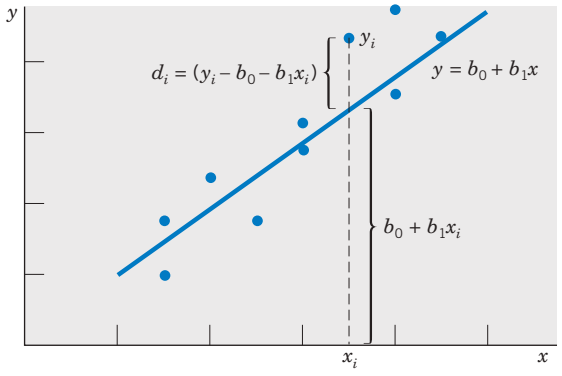


Figure 4 Deviations of the observations from a line $y = b_0 + b_1x$.

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

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

where the hat over y indicates that it is an estimated quantity.

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

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

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

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)

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

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

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

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

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

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

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

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

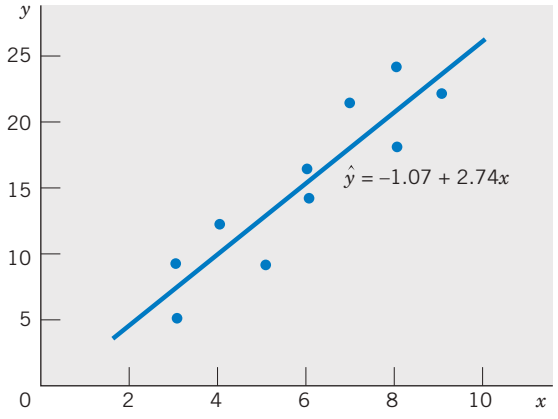


Figure 5 The least squares regression line for the data given in Table 1.

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 txt

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 A student collected data on the number of large pizzas consumed, y, while x students are watching a professional football game on TV. Suppose that the data from five games are:

x	2	5	6	3	4
y	1	6	10	3	5

(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 The office manager at a real estate firm makes a pot of coffee every morning. The time before it runs out, y, in hours depends on the number of persons working inside that day, x. Suppose that the pairs of (x, y) values from six days are:

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

(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 A help desk devoted to student software problems also receives phone calls. The number of persons that can be served in person, within one hour, is the response y. The predictor variable, x, is the number of phone calls answered.

x	0	1	2	3	4
y	7	8	5	4	1

(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.

(d) Use the fitted line to predict the number of persons served when 3 calls are answered.

11.17 A student hourly employee does small secretarial projects. The number of projects she completes in a day is the response variable y. The number of hours she works in a day is the predictor variable x.

x	1	2	4	6	7
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.

(d) Use the fitted line to predict the number of projects completed when she works 6.5 hours.

11.18 Crime is becoming more of a problem on many college campuses. The U.S. Department of Education reports data on alleged crimes at universities and colleges. Table 5 gives the data

TABLE 5 Crime at the Largest Universities

University	Enrollment (1000)	Forcible Rape	Robbery	Burglary	Arson
Ohio State	52.57	65	11	212	17
Minnesota	50.88	12	7	272	3
Central Florida	48.40	3	1	43	0
Michigan State	46.05	18	11	122	1
Pennsylvania State	43.25	8	3	112	7
Wisconsin-Madison	41.56	7	5	167	6
Florida State	40.56	10	8	69	3
Washington-Seattle	40.22	0	3	70	3
Florida International	38.18	3	5	90	1
Arizona	37.22	8	0	43	0
San Diego State	35.70	10	4	61	1
California-Berkeley	34.94	4	24	74	11
Rutgers-New Brunswick	34.80	4	4	97	5
Georgia	33.83	1	0	17	0
Boston	32.05	7	1	62	0
North Carolina State-Raleigh	31.80	6	8	57	1
San Francisco State	30.13	9	1	45	1
Purdue University-Indianapolis	29.85	1	0	93	0
California-Davis	29.80	7	9	42	2
Iowa	29.12	6	1	36	1

from the year 2007, for the twenty universities with the largest enrollments.

When y is the number of burglaries and the predictor variable x is enrollment, we have

$$n = 20 \quad \bar{x} = 38.046 \quad \bar{y} = 89.20$$
$$S_{xx} = 994.038 \quad S_{xy} = 6191.04 \quad S_{yy} = 76{,}293.2$$

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

(b) Calculate the residual sum of squares.

(c) Estimate σ^2.

11.19 Refer to the crime data in Exercise 11.18, Table 5.

When y is the number of robberies and the predictor variable x is arson incidents, we have

$$n = 20 \quad \bar{x} = 3.15 \quad \bar{y} = 5.30$$
$$S_{xx} = 358.55 \quad S_{xy} = 290.10 \quad S_{yy} = 618.2$$

(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 hours for each unit dosage of the drug. Also, if we were to estimate the expected duration of relief for a specified dosage $x^* = 4.5$ milligrams, we would naturally

use the fitted regression line to calculate the estimate $-1.07 + 2.74 \times 4.5 = 11.26$ 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

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

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

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

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

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

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 $H_0 : \beta_1 = 0$ 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 $H_0 : \beta_1 = 0$ 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 3, 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 drug trial 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

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

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

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

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

Note that β_0 represents the mean response corresponding to the value 0 for the input variable x. In the drug evaluation problem 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

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

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 (a) The fitted regression line is

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

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

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

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

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

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

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

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

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

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

A 95% confidence interval is

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

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

The formula for the standard error shows that when x^* is close to $\bar{x}$, 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 concerning extrapolation: Extreme caution should be exercised in extending a fitted regression line to make long-range predictions far away from the range of x values covered in the experiment. Not only does the confidence interval become so wide that predictions based on it can be extremely unreliable, but an even greater danger exists. If the pattern of the relationship between the variables changes drastically at a distant value of x, the data provide no information with which to detect such a change. Figure 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.

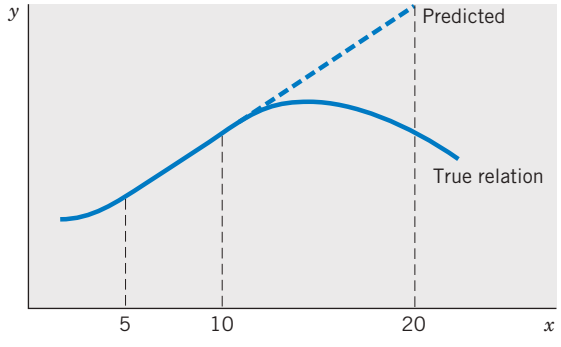


Figure 6 Danger in long-range prediction.

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

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

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

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

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

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

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

One measure of the development of a country is the Human Development Index (HDI) which combines life expectancy, literacy, educational attainment, and gross domestic product per capita into an index whose values lie between 0 and 1, inclusive.

We randomly selected fifteen countries, of the 152 countries, below the top twenty-five most developed countries on the list. HDI is the response variable y, and Internet usage per 100 persons, x, is the predictor variable The data, given in Exercise 11.31, have the summary statistics

$$n = 15 \qquad \bar{x} = 9.953 \qquad \bar{y} = .6670$$
$$S_{xx} = 1173.46 \qquad S_{yy} = 20.471 \qquad S_{xy} = .41772$$

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

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

(c) Estimate the mean value of HDI for 18 Internet users per 100 persons and construct a 95% confidence interval.

(d) Find the predicted y for $x = 43$ Internet users per 100 persons.

SOLUTION

(a) $$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{20.471}{1173.46} = .017445$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = .6670 - .017445 \times 9.953 = .4934$$

So, the equation of the fitted line is

$$\hat{y} = .493 - .0174x$$

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

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

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

$$\text{SSE} = S_{yy} - \frac{S_{xy}^2}{S_{xx}} = .41772 - \frac{(20.471)^2}{1173.46} = .06060$$

$$s = \sqrt{\frac{\text{SSE}}{n-2}} = \sqrt{\frac{.06060}{13}} = .0683$$

$$\text{Estimated S.E.}\,(\hat{\beta}_1) = \frac{s}{\sqrt{S_{xx}}} = \frac{.0683}{\sqrt{1173.46}} = .00199$$

The t statistic has the value

$$t = \frac{.01744}{.00199} = 8.77$$

Since the observed $t = 8.77$ is greater than 2.650, H_0 is rejected with $\alpha = .01$. The P–value is much less than .0001.

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

(c) The expected value of the HDI corresponding to x* = 18 Internet users per 100 is estimated as

$$\hat{\beta}_0 + \hat{\beta}_1 x^* = .4934 + .01745(18) = .8075$$

and its

$$\text{Estimated S.E.} = s\sqrt{\frac{1}{15} + \frac{(18 - 9.953)^2}{1173.46}} = .0238$$

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

$$.8075 \pm 2.160 \times .0238 = .8075 \pm .0514 \quad \text{or} \quad (.756, .859)$$

We are 95% confident that the expected value of HDI, for $x^* = 18$, is between .756 and .859.

(d) Since $x = 43$ is far above the largest value of 26.2 users per 100, it is not sensible to predict y at $x = 43$ using the fitted regression line. Here a formal calculation gives

$$\text{Predicted HDI} = .4934 + .01745(43) = 1.244$$

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

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 6.

TABLE 6 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 7. Notice the similarity of information in Tables 6 and 7. 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 7 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

University students taking a physical fitness class were asked to count the number of situps they could do at the start of the class and again at the end of the semester.

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

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

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

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

SOLUTION The scatter plot in Figure 7 suggests that a straight line may model the expected value of posttest situps given the number of pretest situps. Here x is the number of pretest situps and y is the number of posttest situps. We use MINITAB statistical software to obtain the output

Regression Analysis: Post Situps versus Pre Situps

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

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

S = 5.17893 R-Sq = 71.5%

Analysis of Variance

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

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

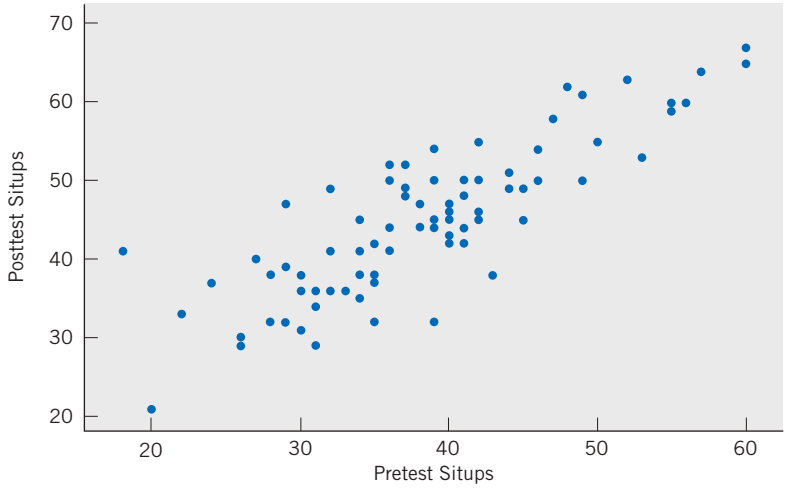


Figure 7 Scatter plot of number of situps.

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

Exercises

11.25 We all typically go to the shortest line in the grocery store. Data were collected on the number of carts ahead in line and the total time to check out (minutes), including time in line, on five occasions.

Number of Carts	Time to Check Out
1	5
2	11
3	9
4	14
5	16

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

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

(c) Estimate the expected y value corresponding to $x = 3$ carts and give a 90% confidence interval.

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

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

11.28 An engineer found that by adding small amounts of a compound to rechargeable batteries during manufacture, she could extend their lifetimes. She experimented with different amounts of the additive (g) and measured the hours they lasted in a laptop.

Amount of Additive	Life (hours)
0	1.9
1	2.0
2	2.5
3	2.6
4	3.0

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

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

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

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

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
283.5	288.0	310.2	311.0
290.0	291.2	294.6	299.0
270.5	276.2	320.0	318.0
300.8	307.0		

(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. Interpret.

11.30 Refer to the data in Exercise 11.29.

(a) Estimate the expected selling price of homes that were assessed at $290,000 and construct a 95% confidence interval. Interpret.

(b) For a single home that was assessed at $290,000, give a 95% prediction interval for the selling price. Interpret.

11.31 One measure of the development of a country is the Human Development Index (HDI). Life expectancy, literacy, educational attainment, and gross domestic product per capita are combined into an index between 0 and 1, inclusive with 1 being the highest development. The United Nations Development Program reports values for 177 countries. We randomly selected fifteen countries, below the top twenty-five. Both HDI and the predictor variable x = Internet usage per 100 persons are obtained from their reports.

TABLE 8 Human Development Index

Country	Internet/100	HDI
Bahrain	21.3	.866
Poland	26.2	.870
Uruguay	14.3	.852
Bulgaria	20.6	.824
Brazil	19.5	.800
Ukraine	9.7	.788
Dominican Republic	16.9	.799
Moldova	9.6	.708
India	5.5	.619
Madagascar	0.5	.533
Nepal	0.4	.534
Tanzania	0.9	.467
Uganda	1.7	.505
Zambia	2.0	.434
Ethiopia	0.2	.406

Source: Human Development 2007–2008 reports at UNDP web site http://hdr.undp.org

$$n = 15 \quad \bar{x} = 9.953 \quad \bar{y} = .6670$$
$$S_{xx} = 1173.46 \quad S_{xy} = 20.471 \quad S_{yy} = .41772$$

(a) Obtain a 95% confidence interval for the mean HDI when Internet usage per 100 persons is 22. Compare the width of the interval with that of the interval in Example 9.

(b) Obtain a 95% prediction interval for a single country with Internet usage 22 per one hundred persons. Interpret.

(c) Does your analysis show that Internet availability causes HDI to increase?

11.32 Refer to Exercise 11.31.

(a) Obtain the least squares estimates by fitting a straight line to the response Internet usage using the predictor variable HDI.

(b) Test, with $\alpha = .05$, $H_0 : \beta_1 = 0$ versus a two-sided alternative.

(c) Obtain a 95% confidence interval for the mean Internet usage per 100 persons, when the HDI is .650. Interpret.

(d) Obtain a 95% prediction interval for Internet usage in a single country with HDI .650. Interpret.

11.33 According to the computer output in Table 9:

(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 9:

(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 10:

(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 10:

(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 $x = 2$. Interpret.

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 9 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 = 299.4     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 10 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.

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

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

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

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

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

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

Decomposition of Variability

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

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

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

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

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

From Section 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

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

which is the square of the sample correlation coefficient r.

The value of r is always between –1 and 1, inclusive whereas r^2 is always between 0 and 1.

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%
```

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 8*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}.

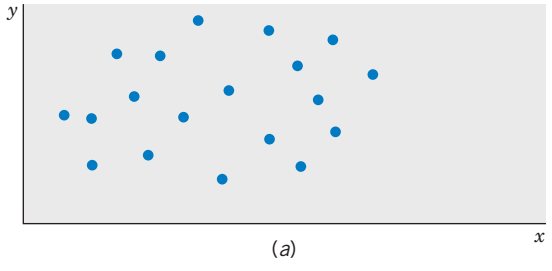


Figure 8 Scatter diagram patterns:
(*a*) No relation. (*b*) A nonlinear relation.

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 8*b* illustrates such a case, where the SSE can be reduced by fitting a suitable curve to the data.

Exercises

11.38 Refer to Exercise 11.18 and Table 5 for crimes during the year 2007, at the twenty universities with the largest enrollments. When predicting number of burglaries from the predictor enrollment, we have

$$n = 20 \quad \bar{x} = 38.046 \quad \bar{y} = 89.20$$
$$S_{xx} = 994.038 \quad S_{xy} = 6191.04 \quad S_{yy} = 76{,}293.2$$

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

11.39 Refer to Exercise 11.20 and Table 5 for crimes during the year 2007, at the twenty universities with the largest enrollments. When y is the number of robberies and the predictor variable x is arson incidents, we have

$$n = 20 \quad \bar{x} = 3.15 \quad \bar{y} = 5.30$$
$$S_{xx} = 358.55 \quad S_{xy} = 290.10 \quad S_{yy} = 618.2$$

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

11.40 Refer to Example 9 and Exercise 11.31, concerning the prediction of a human development index by Internet usage, where

$$n = 15 \quad \bar{x} = 9.953 \quad \bar{y} = .6670$$
$$S_{xx} = 1173.46 \quad S_{xy} = 20.471 \quad S_{yy} = .41772$$

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

11.41 Refer to Exercise 11.40 but consider the prediction of Internet usage when the human development index is the predictor variable.

(a) Determine the proportion of variation in Internet usage that is explained by linear regression.

(b) Compare your answer in Part (a) with that of Exercise 11.40. Comment.

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.

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 9, find the proportion of y variability explained by x. Also, calculate R^2 from the analysis of variance table.

11.45 Refer to Exercise 11.35. According to the computer output in Table 10, 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 $\text{SSE} = (1 - r^2)\, S_{yy}$.

*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 some 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 perform a test of hypotheses and look at

P–values to decide whether or not the parameters are 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.
5. Don't assume a large r^2, or correlation, implies a causal relationship.

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** $\hat{\beta}_0$ and **least squares estimate** $\hat{\beta}_1$ are obtained by the **method of least squares,** which minimizes the sum of squared deviations $\Sigma\,(y_i - b_0 - b_1 x_i)^2$. 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

$$\text{Estimator } \hat{\beta}_1$$
$$\text{Estimated S.E.} = \frac{S}{\sqrt{S_{xx}}}$$

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

$$\text{Estimator } \hat{\beta}_0$$
$$\text{Estimated S.E.} = S\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$$

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

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

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

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**

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

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

$$\text{Residual or unexplained variability} = \text{SSE}$$

$$\text{Total } y \text{ variability} = S_{yy}$$

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 Last week's total number of hours worked by a student, y, depends on the number of days, x, he reported to work last week. Suppose the data from nine students provided

x	1	1	1	2	3	3	3	4	5
y	8	6	7	10	15	12	13	19	18

(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 number of hours y corresponding to $x = 3$ days.

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 $x = 4$ days.

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	750	1000	850
925	775	1033	875
932	820	1050	915
940	820	1100	1040

(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	17.9	5	9.9
2	13.9	7	6.6
2	14.9	7	6.7
4	14.0	8	7.0
4	9.8		

(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 Refer to Table 5 for crimes on campus during the year 2007. When predicting number of arson incidents from the predictor number of robberies, we have

$$n = 20 \qquad \bar{x} = 5.30 \qquad \bar{y} = 3.15$$
$$\Sigma x^2 = 1180 \qquad \Sigma xy = 624 \qquad \Sigma y^2 = 557$$

(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.txt

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 11.

Compare Table 11 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 (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.

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

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

TABLE 11 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			

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

(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. 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 (iv) find a 95% confidence interval for the population mean when the freshwater growth is 100.

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

Height	Speed
456	128
420	120
415	100
377	107
318	95
310	93
263	81
259	81
249	78
245	85
240	79
235	85

(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 425 feet?

(c) What top speed is predicted for a new roller coaster of height 490 feet? What additional danger is there in this prediction?

12

Regression Analysis II

Multiple Linear Regression and Other Topics

Micronutrients and Kelp Cultures: Evidence for Cobalt and Manganese Deficiency in Southern California Deep Seawater

Abstract. *It has been suggested that naturally occurring copper and zinc concentrations in deep seawater are toxic to marine organisms when the free ion forms are overabundant. The effects of micronutrients on the growth of gametophytes of the ecologically and commercially significant giant kelp* Macrocystis pyrifera *were studied in defined media. The results indicate that toxic copper and zinc ion concentrations as well as cobalt and manganese deficiencies may be among the factors controlling the growth of marine organisms in nature.*

A least squares fit of gametophytic growth data in the defined medium generated the expression

$$\begin{aligned} Y = {} & 136 + 8x_{\mathrm{Mn}} - 5x_{\mathrm{Cu}} + 7x_{\mathrm{Co}} \\ & - 7x_{\mathrm{Zn}}x_{\mathrm{Cu}} - 15x_{\mathrm{Zn}}^2 - 27x_{\mathrm{Mn}}^2 - 12x_{\mathrm{Cu}}^2 \\ & - 18x_{\mathrm{Co}}^2 - 6x_{\mathrm{Cu}}x_{\mathrm{Zn}}^2 - 6x_{\mathrm{Cu}}x_{\mathrm{Mn}}^2 \end{aligned} \qquad (1)$$

where Y is mean gametophytic length in micrometers. The fit of the experimental data to Eq. (1) was considered excellent.

Here, several variables are important for predicting growth.

Source: J. S. Kuwabara, "Micronutrients and Kelp Cultures: Evidence for Cobalt and Manganese Deficiency in Southern California Deep Sea Waters," *Science,* **216** (June 11, 1982), pp. 1219–1221.

1. INTRODUCTION

The basic ideas of regression analysis have a much broader scope of application than the straight line model of Chapter 11. In this chapter, our goal is to extend the ideas of regression analysis in two important directions.

1. To handle nonlinear relations by means of appropriate transformations applied to one or both variables.
2. To accommodate several predictor variables into a regression model.

These extensions enable the reader to appreciate the breadth of regression techniques that are applicable to real-life problems. We then discuss some graphical procedures that are helpful in detecting any serious violation of the assumptions that underlie a regression analysis.

2. NONLINEAR RELATIONS AND LINEARIZING TRANSFORMATIONS

When studying the relation between two variables y and x, a scatter plot of the data often indicates that a relationship, although present, is far from linear. This can be established on a statistical basis by checking that the value of r^2 is small so a straight line fit is not adequate.

Statistical procedures for handling nonlinear relationships are more complicated than those for handling linear relationships, with the exception of a specific type of model called the **polynomial regression model,** which is discussed in Section 3. In some situations, however, it may be possible to transform the variables x and/or y in such a way that the new relationship is close to being linear. A linear regression model can then be formulated in terms of the transformed variables, and the appropriate analysis can be based on the transformed data.

Transformations are often motivated by the pattern of data. Sometimes, when the scatter diagram exhibits a relationship on a curve in which the y values increase too fast in comparison with the x values, a plot of $\sqrt{y}$ or some other fractional power of y can help to linearize the relation. This situation is illustrated in Example 1.

Example 1 Transforming the Response to Approximate a Linear Relation

To determine the maximum stopping ability of cars when their brakes are fully applied, 10 cars are driven each at a specified speed and the distance each requires to come to a complete stop is measured. The various initial speeds selected for each of the 10 cars and the stopping distances recorded are given in Table 1. Can the data be transformed to a nearly straight line relationship?

TABLE 1 Data on Speed and Stopping Distance

Initial speed x (mph)	20	20	30	30	30	40	40	50	50	60
Stopping distance y (ft)	16.3	26.7	39.2	63.5	51.3	98.4	65.7	104.1	155.6	217.2

SOLUTION The scatter diagram for the data appears in Figure 1. The relation deviates from a straight line most markedly in that y increases at a much faster rate at large x than at small x. This suggests that we can try to linearize the relation by plotting $\sqrt{y}$ or some other fractional power of y with x.

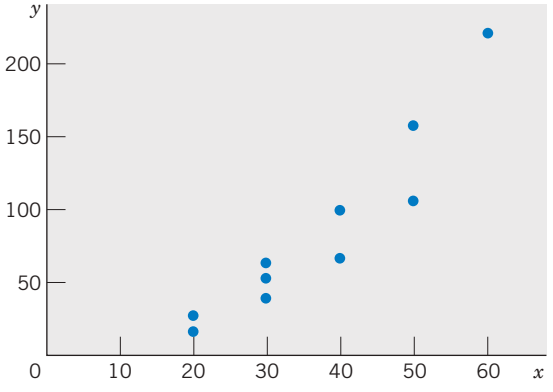


Figure 1 Scatter diagram of the data given in Table 1.

We try the transformed data $\sqrt{y}$ given in Table 2. The scatter diagram for these data, which exhibits an approximate linear relation, appears in Figure 2.

TABLE 2 Data on Speed and Square Root of Stopping Distance

x	20	20	30	30	30	40	40	50	50	60
$y' = \sqrt{y}$	4.037	5.167	6.261	7.969	7.162	9.920	8.106	10.203	12.474	14.738

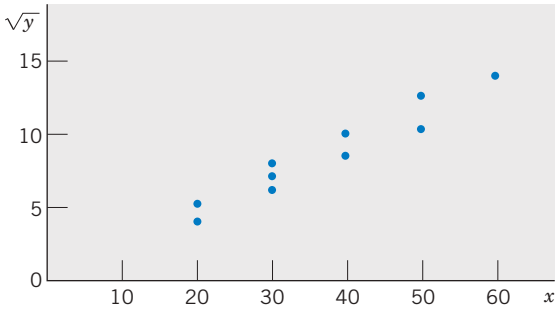


Figure 2 Scatter diagram of the transformed data given in Table 2.

With the aid of a standard computer program for regression analysis (see Exercise 12.27), the following results are obtained by transforming the original data.

$$\bar{x} = 37 \qquad \bar{y}' = 8.604$$
$$S_{xx} = 1610 \qquad S_{y'y'} = 97.773 \qquad S_{xy'} = 381.621$$
$$\hat{\beta}_0 = -.167 \qquad \hat{\beta}_1 = .237$$

Thus, the equation of the fitted line is

$$\hat{y}' = -.167 + .237x$$

The proportion of the y' variation that is explained by the straight line model is

$$r^2 = \frac{(381.621)^2}{(1610)(97.773)} = .925$$

A few common nonlinear models and their corresponding linearizing transformations are given in Table 3.

TABLE 3 Some Nonlinear Models and Their Linearizing Transformations

			Transformed Model
Nonlinear Model	Transformation		$y' = \beta_0 + \beta_1 x'$
(a) $y = ae^{bx}$	$y' = \log_e y$	$x' = x$	$\beta_0 = \log_e a$ $\beta_1 = b$
(b) $y = ax^b$	$y' = \log y$	$x' = \log x$	$\beta_0 = \log a$ $\beta_1 = b$
(c) $y = \frac{1}{a + bx}$	$y' = \frac{1}{y}$	$x' = x$	$\beta_0 = a$ $\beta_1 = b$
(d) $y = a + b\sqrt{x}$	$y' = y$	$x' = \sqrt{x}$	$\beta_0 = a$ $\beta_1 = b$

In some situations, a specific nonlinear relation is strongly suggested by either the data or a theoretical consideration. Even when initial information about the form is lacking, a study of the scatter diagram often indicates the appropriate linearizing transformation.

Once the data are entered on a computer, it is easy to obtain the transformed data $1/y$, $\log_e y$, $y^{1/2}$, and $y^{1/4}$. Note $y^{1/4}$ is obtained by taking the square root of $y^{1/2}$. A scatter plot of $\sqrt{y}$ versus $\log_e x$ or any number of others can then be constructed and examined for a linear relation. Under relation (a) in Table 3, the graph of $\log_e y$ versus x would be linear.

We must remember that all inferences about the transformed model are based on the assumptions of a linear relation and independent normal errors with constant variance. Before we can trust these inferences, this transformed model must be scrutinized to determine whether any serious violation of these assumptions may have occurred (see Section 4).

Exercises

12.1 Developers have built a small robotic vehicle that can travel over rough terrain. They recorded the time y, in minutes, that it takes to travel a fixed distance over various but similar terrains. For a fixed run, the robot's motor is set at a nominal speed x, in feet per second, but this varied from run to run.

x	.5	1	2	4	5	6	7
y	4.6	3.2	2.1	1.7	.9	.7	.8

(a) Plot the scatter diagram.

(b) Obtain the best fitting straight line and draw it on the scatter diagram.

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

12.2 Refer to the data of Exercise 12.1.

(a) Consider the reciprocal transformation $y' = 1/y$ and plot the scatter diagram of y' versus x.

(b) Fit a straight line regression to the transformed data.

(c) Calculate r^2 and comment on the adequacy of the fit.

12.3 Find a linearizing transformation in each case.

(a) $y = \dfrac{1}{(a + bx)^3}$

(b) $\dfrac{1}{y} = a + \dfrac{b}{1 + x}$

12.4 An experiment was conducted for the purpose of studying the effect of temperature on the life-length of an electrical insulation. Specimens of the insulation were tested under fixed temperatures, and their times to failure recorded.

Temperature x (°C)	Failure Time y (thousand hours)
180	7.3, 7.9, 8.5, 9.6, 10.3
210	1.7, 2.5, 2.6, 3.1
230	1.2, 1.4, 1.6, 1.9
250	.6, .7, 1.0, 1.1, 1.2

(a) Fit a straight line regression to the transformed data

$$x' = \frac{1}{x} \quad \text{and} \quad y' = \log y$$

(b) Is there strong evidence that an increase in temperature reduces the life of the insulation?

(c) Comment on the adequacy of the fitted line.

12.5 In an experiment (courtesy of W. Burkholder) involving stored-product beetles (*Trogoderma glabrum*) and their sex-attractant pheromone, the pheromone is placed in a pit-trap in the centers of identical square arenas. Marked beetles are then released along the diagonals of each square at various distances from the pheromone source. After 48 hours, the pit-traps are inspected. Control pit-traps containing no pheromone capture no beetles.

Release Distance (centimeters)	No. of Beetles Captured out of 8
6.25	5, 3, 4, 6
12.5	5, 2, 5, 4
24	4, 5, 3, 0
50	3, 4, 2, 2
100	1, 2, 2, 3

(a) Plot the original data with y = number of beetles captured. Also plot y with $x = \log_e$ (distance).

(b) Fit a straight line by least squares to the appropriate graph in part (a).

(c) Construct a 95% confidence interval for β_1.

(d) Establish a 95% confidence interval for the mean at a release distance of 18 cm.

3. MULTIPLE LINEAR REGRESSION

A response variable y may depend on a predictor variable x but, after a straight line fit, it may turn out that the unexplained variation is large, so r^2 is small and a poor fit is indicated. At the same time, an attempt to transform one or both of the variables may fail to dramatically improve the value of r^2. This difficulty may well be due to the fact that the response depends on not just x but other factors as well. When used alone, x fails to be a good predictor of y because of the effects of those other influencing variables. For instance, the yield of a crop depends on not only the amount of fertilizer but also on the rainfall and average temperature during the growing season. Cool weather and no rain could completely cancel the choice of a correct fertilizer.

To obtain a useful prediction model, one should record the observations of all variables that may significantly affect the response. These other variables may then be incorporated explicitly into the regression analysis. The name **multiple regression** refers to a model of relationship where the response depends on two or more predictor variables. Here, we discuss the main ideas of a multiple regression analysis in the setting of two predictor variables.

Suppose that the response variable y in an experiment is expected to be influenced by two input variables x_1 and x_2, and the data relevant to these input variables are recorded along with the measurements of y. With n runs of an experiment, we would have a data set of the form shown in Table 4.

TABLE 4 Data Structure for Multiple Regression with Two Input Variables

Experimental Run	Input Variables		Response
	x_1	x_2	y
1	x_{11}	x_{12}	y_1
2	x_{21}	x_{22}	y_2
.	.	.	.
.	.	.	.
.	.	.	.
i	x_{i1}	x_{i2}	y_i
.	.	.	.
.	.	.	.
.	.	.	.
n	x_{n1}	x_{n2}	y_n

By analogy with the simple linear regression model, we can then tentatively formulate:

A Multiple Regression Model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i \qquad i = 1, \ldots, n$$

where x_{i1} and x_{i2} are the values of the input variables for the ith experimental run and Y_i is the corresponding response.

The error components e_i are assumed to be independent normal variables with mean 0 and variance σ^2.

The regression parameters β_0, β_1, and β_2 are unknown and so is σ^2.

This model suggests that aside from the random error, the response varies linearly with each of the independent variables when the other remains fixed.

The principle of least squares is again useful in estimating the regression parameters. For this model, we are required to vary b_0, b_1, and b_2 simultaneously to minimize the sum of squared deviations

$$\sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i1} - b_2 x_{i2})^2$$

The least squares estimates $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ are the solutions to the following equations, which are extensions of the corresponding equations for fitting the straight line model (see Section 4 of Chapter 11.)

$$\begin{aligned}\hat{\beta}_1 S_{11} + \hat{\beta}_2 S_{12} &= S_{1y}\\ \hat{\beta}_1 S_{12} + \hat{\beta}_2 S_{22} &= S_{2y}\\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2 &\end{aligned}$$

where S_{11}, S_{12}, and so on, are the sums of squares and cross products of deviations of the variables indicated in the suffix. They are computed just as in a straight line regression model. Methods are available for interval estimation, hypothesis testing, and examining the adequacy of fit. In principle, these methods are similar to those used in the simple regression model, but the algebraic formulas are more complex and hand computations become more tedious. However, a multiple regression analysis is easily performed on a computer with the aid of the standard packages such as MINITAB, SAS, or SPSS. We illustrate the various aspects of a multiple regression analysis with the data of Example 2 and computer-based calculations.

Example 2 Interpreting the Regression of Blood Pressure on Weight and Age

We are interested in studying the systolic blood pressure y in relation to weight x_1 and age x_2 in a class of males of approximately the same height. From 13 subjects preselected according to weight and age, the data set listed in Table 5 was obtained.

TABLE 5 The Data of x_1 = Weight in Pounds, x_2 = Age, and y = Blood Pressure of 13 Males

x_1	x_2	y
152	50	120
183	20	141
171	20	124
165	30	126
158	30	117
161	50	129
149	60	123
158	50	125
170	40	132
153	55	123
164	40	132
190	40	155
185	20	147

Use a computer package to perform a regression analysis using the model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i$$

SOLUTION To use MINITAB, we first enter the data of x_1, x_2, and y in three different columns and then use the regression command,

Data: C12T5.txt
C1: 152 183 171 ⋯ 185
C2: 50 20 20 ⋯ 20
C3: 120 141 124 ⋯ 147
Dialog box:
Stat > Regression > Regression
Type *C3* in **Response.**
Type *C1* and *C2* in **Predictors.**
Click **OK.**

With the last command, the computer executes a multiple regression analysis. We focus our attention on the principal aspects of the output, as shown in Table 6.

TABLE 6 Regression Analysis of the Data in Table 5: Selected MINITAB Output

① THE REGRESSION EQUATION IS
Y = − 65.1 + 1.08 X1 + 0.425 X2

PREDICTOR		COEF	STDEV	T-RATIO	P
CONSTANT		− 65.10	14.94	− 4.36	0.001
X1	②	1.07710	0.07707	13.98	0.000
X2		0.42541	0.07315	④ 5.82	0.000

③ S = 2.509 ⑤ R-SQ = 95.8%

ANALYSIS OF VARIANCE

SOURCE	DF		SS	MS	F	P
REGRESSION	2		1423.84	711.92	113.13	0.000
ERROR	10	⑥	62.93	6.29		
TOTAL	12		1486.77			

We now proceed to interpret the results in Table 6 and use them to make further statistical inferences.

(i) The equation of the fitted linear regression is

$$① \quad \hat{y} = -65.1 + 1.08x_1 + .425x_2$$

This means that the mean blood pressure increases by 1.08 if weight x_1 increases by one pound and age x_2 remains fixed. Similarly, a 1-year increase in age with the weight held fixed will only increase the mean blood pressure by .425.

(ii) The estimated regression coefficient and the corresponding estimated standard errors are

$$② \quad \begin{aligned} \hat{\beta}_0 &= -65.10 \qquad \text{Estimated S.E.}\,(\hat{\beta}_0) = 14.94 \\ \hat{\beta}_1 &= 1.07710 \qquad \text{Estimated S.E.}\,(\hat{\beta}_1) = .07707 \\ \hat{\beta}_2 &= .42541 \qquad \text{Estimated S.E.}\,(\hat{\beta}_2) = .07315 \end{aligned}$$

③ Further, the error standard deviation σ is estimated by $s = 2.509$ with

$$\begin{aligned} \text{Degrees of freedom} &= n - (\text{No. of input variables}) - 1 \\ &= 13 - 2 - 1 \\ &= 10 \end{aligned}$$

These results are useful in interval estimation and hypothesis tests about the regression coefficients. In particular, a $100(1-\alpha)\%$ confidence interval for a coefficient β is given by

$$\text{Estimated coefficient} \pm t_{\alpha/2}\,(\text{Estimated S.E.})$$

where $t_{\alpha/2}$ is the upper $\alpha/2$ point of the distribution with d.f. $= 10$. For instance, a 95% confidence interval for β_1 is

$$\begin{aligned} &1.07710 \pm 2.228 \times .07707 \\ &= 1.07710 \pm .17171 \qquad \text{or} \qquad (.905, 1.249) \end{aligned}$$

To test the null hypothesis that a particular coefficient β is zero, we employ the test statistic

$$t = \frac{\text{Estimated coefficient} - 0}{\text{Estimated S.E.}} \qquad \text{d.f.} = 10$$

These t-ratios appear in Table 6. Suppose that we wish to examine whether the mean blood pressure significantly increases with age. In the language of hypothesis testing, this problem translates to one of testing $H_0: \beta_2 = 0$ versus $H_1: \beta_2 > 0$. The observed value of the ④ test statistic is $t = 5.82$ with d.f. $= 10$. Since this is larger than the tabulated value $t_{.01} = 2.764$, the null hypothesis is rejected in favor of H_1, with $\alpha = .01$. In fact, it is rejected even with $\alpha = .005$.

(iii) In Table 6, the result "R-SQ = 95.8%" or

$$⑤ \quad R^2 = .958$$

tells us that 95.8% of the variability of y is explained by the fitted multiple regression of y on x_1 and x_2. The "analysis of variance" shows the decomposition of the total variability $\Sigma(y_i - \bar{y})^2 = 1486.77$ into the two components.

⑥ 1486.77	=	1423.84	+	62.93
Total variability of y		Variability explained by the regression of y on x_1 and x_2		Residual or unexplained variability

Thus,

$$R^2 = \frac{1423.84}{1486.77} = .958$$

and σ^2 is estimated by $s^2 = 62.93/10 = 6.293$, so $s = 2.509$ [checks with s from (ii)].

TABLE 7 A Regression Analysis of the Data in Example 2 Using SAS

MODEL: MODEL 1
DEPENDENT VARIABLE: Y

ANALYSIS OF VARIANCE

SOURCE	DF		SUM OF SQUARES	MEAN SQUARE	F VALUE	PROB>F
MODEL	2		1423.83797	711.91898	113.126	0.0001
ERROR	10	⑥	62.93126	6.29313		
C TOTAL	12		1486.76923			

③ ROOT MSE 2.50861 ⑤ R-SQUARE 0.9577

PARAMETER ESTIMATES

VARIABLE	DF		PARAMETER ESTIMATE	STANDARD ERROR	T FOR H0: PARAMETER = 0	PROB > |T|
INTERCEP	1		−65.099678	14.94457547	−4.356	0.0014
X1	1	②	1.077101	0.07707220	13.975	0.0001
X2	1		0.425413	0.07315231	④ 5.815	0.0002

The **square of the multiple correlation coefficient** R^2 gives the proportion of variability in y explained by the fitted multiple regression.

The output from the SAS package is given in Table 7. The quantities needed in our analysis have been labeled with the same circled numbers as in the MINITAB output.

Example 3 Computer-Aided Regression Analysis—Two Predictors

The times for 81 students to complete a rowing test both before and after completing a one-semester conditioning course are given in Table D.5 in the Data Bank. It may be that not only the pretest rowing time but also gender would be useful for predicting the posttest rowing time. Perform a regression analysis.

SOLUTION We use MINITAB to obtain the output

```
Regression Analysis: Post row versus Pre row, Gender

The regression equation is
Post row = 97.3 + 0.726 Pre row + 32.1 Gender

Predictor       Coef   SE Coef       T       P
Constant       97.33     31.68    3.07   0.003
Pre row      0.72573   0.05487   13.23   0.000
Gender        32.083     9.756    3.29   0.002

S   =   31.8137 R-Sq   =   85.7%      R-Sq(adj)   =   85.3%

Analysis of Variance

Source           DF       SS       MS        F       P
Regression        2   471547   235774   232.95   0.000
Residual Error   78    78945     1012
Total            80   550492
```

Which variables should be used to predict posttest rowing time? Reading from the column of P–values for the individual coefficients, the largest is only .003. The constant term and the coefficients of pretest rowing time and gender are significantly different from 0. All three terms are needed in the model.

The plot of residuals versus fit in Figure 3 on page 498 reveals a constant width band so there is no evidence against the assumption of constant variance. The one large negative residual is case 17 and the two large positive residuals are cases 29 and 70.

POLYNOMIAL REGRESSION

A scatter diagram may exhibit a relationship on a curve for which a suitable linearizing transformation cannot be constructed. Another method of handling such a nonlinear relation is to include terms with higher powers of x in the model $Y = \beta_0 + \beta_1 x + e$. In this instance, by including the second power of x, we obtain the model

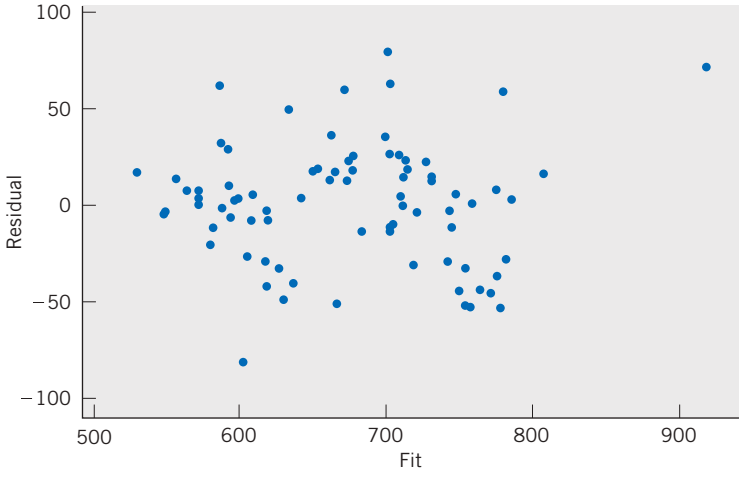


Figure 3 Residuals of posttest row times versus fits.

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + e_i \qquad i = 1, \ldots, n$$

which states that aside from the error components e_i, the response y is a **quadratic function** (or a **second-degree polynomial**) of the independent variable x. Such a model is called a **polynomial regression model** of y with x, and the highest power of x that occurs in the model is called the **degree** or the **order** of the polynomial regression. It is interesting to note that the analysis of a polynomial regression model does not require any special techniques other than those used in multiple regression analysis. By identifying x and x^2 as the two variables x_1 and x_2, respectively, this second-degree polynomial model reduces to the form of a multiple regression model:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i \qquad i = 1, \ldots, n$$

where $x_{i1} = x_i$ and $x_{i2} = x_i^2$. In fact, both these types of models and many more types are special cases of a class called **general linear models** [1, 2].

Before we talk about the general linear model, let's look at an example which analyzes a second-degree polynomial model.

Example 4 Fitting a Quadratic Relation of a Human Development Index to Internet Usage

Refer to Chapter 11, Example 9, concerning the development of a country measured by the Human Development Index (HDI) and the predictor variable Internet usage per 100 persons. Although we randomly selected only fifteen countries, of the 152 countries, below the twenty-five most developed there is still an indication that the relation is increasing less rapidly for high Internet usage. Fit a quadratic and test whether or not squared term is required.

SOLUTION We fit the quadratic model $Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + e_i$ using MINITAB and obtain the output.

```
The regression equation is
HDI = 0.452 + 0.0356 Internet − 0.000789 Internet Sq

Predictor          Coef   SE Coef      T      P
Constant        0.45213   0.02315  19.53  0.000
Internet       0.035648  0.005557   6.41  0.000
Internet Sq −0.0007893 0.0002322  -3.40  0.005

S  =  0.0507320      R-Sq  =  92.6%

Analysis of Variance

Source           DF       SS       MS       F       P
Regression        2  0.38684  0.19342   75.15   0.000
Residual Error   12  0.03088  0.00257
Total            14  0.41772
```

According to the output, the estimated coefficient of x^2 is $\hat{\beta}_2 = -.0007893$ and the t test for testing $H_0 : \beta_2 = 0$ versus $H_1 : \beta_2 \neq 0$ has $P-$ value

$$P[T < -3.4] + P[T > 3.4] = .005$$

as indicated under P in the MINITAB output. This gives strong evidence that a quadratic term is needed. Note that the proportion of variation in HDI explained by Internet usage has increased to .926 as indicated in the output R-Sq = 92.6%.

The data and fitted curve, shown in Figure 4, illustrate the bend in the straight line relation at the few higher values of Internet usage. The quadratic fit is still not ideal because it starts to turn down over the range of the experiment, whereas the underlying relation is likely to always increase. This reminds us that there is no "true model," but proposed models are only approximations.

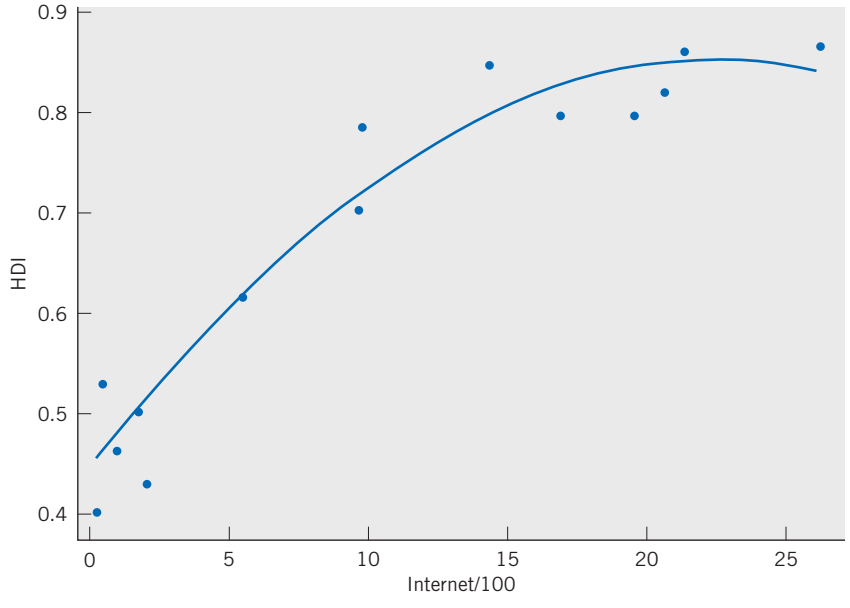


Figure 4 Quadratic Fit to HDI by Internet Usage

GENERAL LINEAR MODEL

By virtue of its wide applicability, the multiple linear regression model plays a prominent role in the portfolio of a statistician. Although a complete analysis cannot be given here, the general structure of a multiple regression model merits further attention. We have already mentioned that most least squares analyses of multiple linear regression models are carried out with the aid of a computer. All the programs for implementing the analysis require the investigator to provide the values of the response y_i and the p input variables $x_{i1}, \ldots, x_{ip}$ for each run $i = 1, 2, \ldots, n$. In writing $1 \cdot \beta_0$, where 1 is the known value of an extra "dummy" input variable corresponding to β_0, the model is

Observation — Input variables — Error

$$Y_i = 1 \cdot \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{ip}\beta_p + e_i$$

This is called a **linear model** because it is linear in the β_i 's. That is, there are no terms such as $\beta_1 \beta_2$ or β_1^2.

The basic quantities can be arranged in the form of these arrays, which are denoted by boldface letters.

Observation — Input variables

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_i \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ 1 & x_{i1} & \cdots & x_{ip} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}$$

Only the arrays $\mathbf{y}$ and $\mathbf{X}$ are required to obtain the least squares estimates of β_0, $\beta_1, \ldots, \beta_p$ that minimize

$$\sum_{i=1}^{n} (y_i - b_0 - x_{i1} b_1 - \cdots - x_{ip} b_p)^2$$

The input array $\mathbf{X}$ is called the **design matrix.**

In the same vein, setting

$$\mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \cdot \\ \cdot \\ \cdot \\ e_n \end{bmatrix} \quad \text{and} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \cdot \\ \cdot \\ \cdot \\ \beta_p \end{bmatrix}$$

we can write the model in the suggested form

$$\underset{\mathbf{y}}{\text{Observation}} = \underset{\mathbf{X}}{\text{Design matrix}} \underset{\boldsymbol{\beta}}{\text{Parameter}} + \underset{\mathbf{e}}{\text{Error}}$$

which forms the basis for a thorough but more advanced treatment of regression.

Exercises

12.6 A student fit the regression model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$ to data from the fifty states and Washington, D.C., so $n = 51$. The response y = median income in thousands of dollars and the two predictor variables are x_1 = median monthly housing costs for home owners and x_2 = percentage of persons below the poverty level in the last twelve months. The least squares estimates are

$$\hat{\beta}_0 = 38.413 \quad \hat{\beta}_1 = .0166 \quad \hat{\beta}_2 = 1.008$$

Predict the response for

(a) $x_1 = 1200, \quad x_2 = 13$

(b) $x_1 = 1200, \quad x_2 = 15$

(c) $x_1 = 1400, \quad x_2 = 15$

12.7 Consider the multiple linear regression model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$$

where $\beta_0 = -2$, $\beta_1 = -1$, $\beta_2 = 3$, and the normal random variable e has standard deviation 3. What is the mean of the response Y when $x_1 = 3$ and $x_2 = -2$?

12.8 In Exercise 12.6, suppose that the residual sum of squares (SSE) is 167.7 and the SS due to regression is 2538.7.

(a) Estimate the error standard deviation σ. State the degrees of freedom.

(b) Find R^2 and interpret the result.

12.9 Refer again to Exercise 12.6 and assume that the assumptions about the model prevail. The estimated standard errors of $\hat{\beta}_1$, and $\hat{\beta}_2$ are .00107, and .0977, respectively.

(a) Determine a 95% confidence interval for β_2.

(b) Test $H_0 : \beta_1 = .0140$ versus $H_1 : \beta_1 > .0140$, with $\alpha = .05$.

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

(a) Obtain the least squares fit of the straight line regression model $Y = \beta_0 + \beta_1 x + e$ to predict canine length from age.

(b) Obtain the least squares fit of the multiple regression model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$ to predict canine length using age x_1 and body length x_2.

(c) What is the predicted canine length for a wolf of age 2.5 and body length 127?

(d) What proportion of the y variability is explained by the fitted model in Part (b)?

(e) Obtain 95% confidence intervals for β_0, β_1, and β_2.

12.11 Consider the response variable miles per gallon on highways and the two predictor variables x_1 = engine volume (l) and x_2 = size of battery (v). Using the government *2009 Fuel Economy Guide*, and the data on hybrid-electric cars and SUVs, we obtain the regression analysis given in Table 8.

(a) Identify the least squares estimates $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$.

(b) What model is suggested from this analysis?

(c) What is the proportion of y variability explained by the regression on x_1 and x_2?

(d) Estimate σ^2.

12.12 Laptop computers are advertised every week by several stores. From one Sunday paper in March 2009, the response variable hard disk size(GB) and the two predictor variables x_1 = read-only memory(GB) and x_2 = screen size (in) were recorded. The output from a regression analysis is given in Table 9, page 502.

TABLE 8 Computer Output of a Regression Analysis to Be Used for Exercise 12.11

```
Regression Analysis: y versus x1, x2

The Regression equation is
y = 45.3 - 3.22 x1 - 0.0207 X2

Predictor      Coef   SE Coef        T      P
Constant     45.299     2.345    19.32  0.000
x1         -3.2243    0.4562    -7.07  0.000
x2       -0.020661  0.008574    -2.41  0.026

S = 3.40356 R-SQ = 79.5%

Analysis of Variance

Source      DF       SS      MS     F      P
Regression   2   897.53  448.77 38.74 0.000
Error       20   231.68   11.58
Total       22  1129.22
```

(a) How many laptops were included in the analysis?

(b) Identify the least squares estimates $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$.

(c) What model is suggested from this analysis?

(d) What is the proportion of y variability explained by the regression on x_1 and x_2?

12.13 With reference to Exercise 12.11:

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

(b) Test $H_0 : \beta_2 = 0$ versus $H_1 : \beta_2 \neq 0$ with $\alpha = .05$.

(c) Estimate the expected y value corresponding to $x_1 = 3.2$ and $x_2 = 200$.

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

TABLE 9 Computer Output of a Regression Analysis to Be Used for Exercise 12.12

```
Regression Analysis: y versus x1, x2

The regression equation is
Y = - 258 + 61.7 X1 + 20.7 X2

Predictor      Coef   SE Coef        T      P
Constant    -257.8     117.9    -2.19  0.048
x1           61.71     13.03     4.74  0.000
x2          20.716     7.204     2.88  0.013

S = 35.3371 R-Sq = 72.3%

Analysis of Variance

Source           DF       SS       MS       F       P
Regression        2    42361    21180   16.96   0.000
Residual Error   13    16233     1249
Total            15    58594
```

12.14 With reference to Exercise 12.12:

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

(b) Test $H_0 : \beta_2 = 0$ versus $H_1 : \beta_2 \neq 0$ with $\alpha = .05$.

(c) Estimate the expected y value corresponding to $x_1 = 2.0$ and $x_2 = 16.5$.

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

4. RESIDUAL PLOTS TO CHECK THE ADEQUACY OF A STATISTICAL MODEL

General Attitude Toward a Statistical Model

A regression analysis is not completed by fitting a model by least squares, providing confidence intervals, and testing various hypotheses. These steps tell only half the story: the statistical inferences that can be made when the postulated model is adequate. In most studies, we cannot be sure that a particular model is correct. Therefore, we should adopt the following strategy.

1. **Tentatively entertain a model.**
2. **Obtain least squares estimates and compute the residuals.**
3. **Review the model by examining the residuals.**

Step 3 often suggests methods of appropriately modifying the model. We then return to step 1, where the modified model is entertained, and this **iteration** is continued until a model is obtained for which the data do not seem to contradict the assumptions made about the model.

Once a model is fitted by least squares, all the information on variation that cannot be explained by the model is contained in the residuals

$$\hat{e}_i = y_i - \hat{y}_i \qquad i = 1, 2, \ldots, n$$

where y_i is the observed value and $\hat{y}_i$ denotes the corresponding value predicted by the fitted model. For example, in the case of a simple linear regression model, $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

Recall from our discussion of the straight line model in Chapter 11 that we have made the assumptions of independence, constant variance, and a normal distribution for the error components e_i. The inference procedures are based on these assumptions. When the model is correct, the residuals can be considered as estimates of the errors e_i that are distributed as $N(0, \sigma)$.

To determine the merits of the tentatively entertained model, we can examine the residuals by plotting them in various ways. Then if we recognize any systematic pattern formed by the plotted residuals, we would suspect that some assumptions regarding the model are invalid. There are many ways to plot the residuals, depending on what aspect is to be examined. We mention a few of these here to illustrate the techniques. A more comprehensive discussion can be found in Chapter 3 of Draper and Smith [2].

HISTOGRAM OR DOT DIAGRAM OF RESIDUALS

To picture the overall behavior of the residuals, we can plot a **histogram** for a large number of observations or a **dot diagram** for fewer observations. For example, in a dot diagram like the one in Figure 5*a*, the residuals seem to behave like a sample

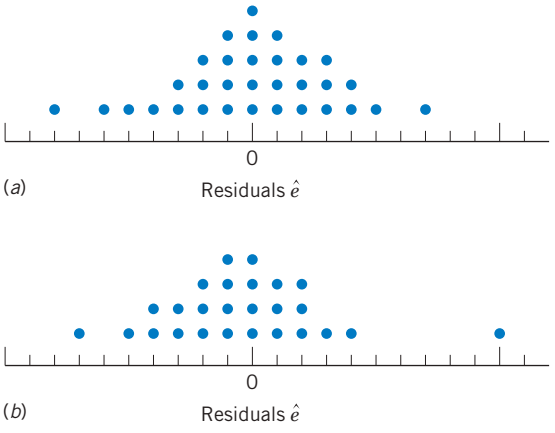


Figure 5 Dot diagram of residuals.
(*a*) Normal pattern. (*b*) One large residual.

from a normal population and there do not appear to be any "wild" observations. In contrast, Figure 5*b* illustrates a situation in which the distribution appears to be quite normal except for a single residual that lies far to the right of the others. The circumstances that produced the associated observation demand a close scrutiny.

PLOT OF RESIDUAL VERSUS PREDICTED VALUE

A plot of the residuals $\hat{e}_i$ versus the **predicted value** $\hat{y}_i$ often helps to detect the inadequacies of an assumed relation or a violation of the assumption of constant error variance. Figure 6 illustrates some typical phenomena. If the points form a horizontal band around zero, as in Figure 6*a*, then no abnormality is indicated. In Figure 6*b*, the width of the band increases noticeably with increasing values of $\hat{y}$. This indicates that the error variance σ^2 tends to increase with an increasing level of response. We would then suspect the validity of the assumption of constant variance in the model. Figure 6*c* shows residuals that form a systematic pattern. Instead of being randomly distributed around the $\hat{y}$ axis, they tend first to increase steadily and then decrease. This would lead us to suspect that the model is inadequate and a squared term or some other nonlinear x term should be considered.

PLOT OF RESIDUAL VERSUS TIME ORDER

The most crucial assumption in a regression analysis is that the errors e_i are independent. Lack of independence frequently occurs in business and economic applications, where the observations are collected in a time sequence with the intention of using regression techniques to predict future trends. In many other experiments, trials are conducted successively in time. In any event, a plot of the

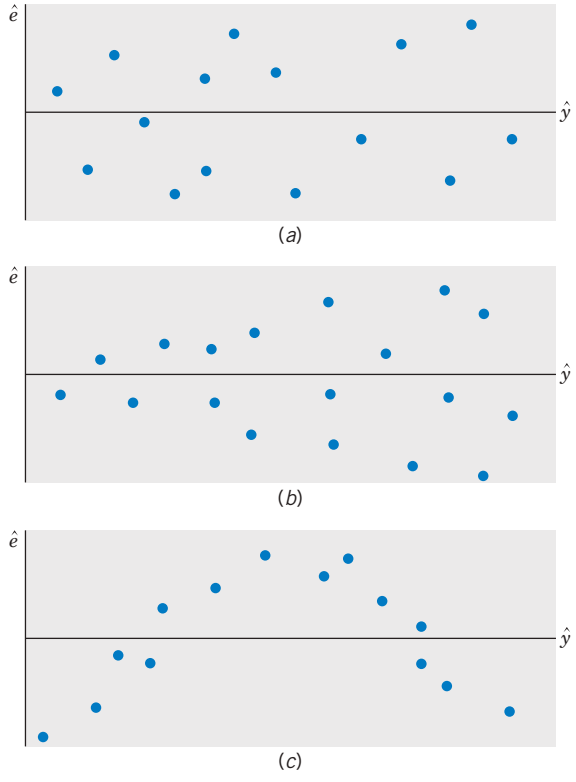


Figure 6 Plot of residual versus predicted value. (*a*) Constant spread. (*b*) Increasing spread. (*c*) Systematic curved pattern.

residuals versus **time order** often detects a violation of the assumption of independence. For example, the plot in Figure 7 exhibits a systematic pattern in that a string of high values is followed by a string of low values. This indicates that consecutive residuals are (positively) correlated, and we would suspect a violation of the independence assumption. Independence can also be checked by plotting the successive pairs $(\hat{e}_i, \hat{e}_{i-1})$, where $\hat{e}_1$ indicates the residual from the first y value observed, $\hat{e}_2$ indicates the second, and so on. Independence is suggested if the scatter diagram is a patternless cluster, whereas points clustered along a line suggest a lack of independence between adjacent observations.

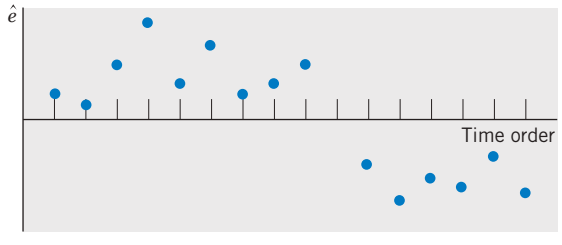


Figure 7 Plot of residual versus time order.

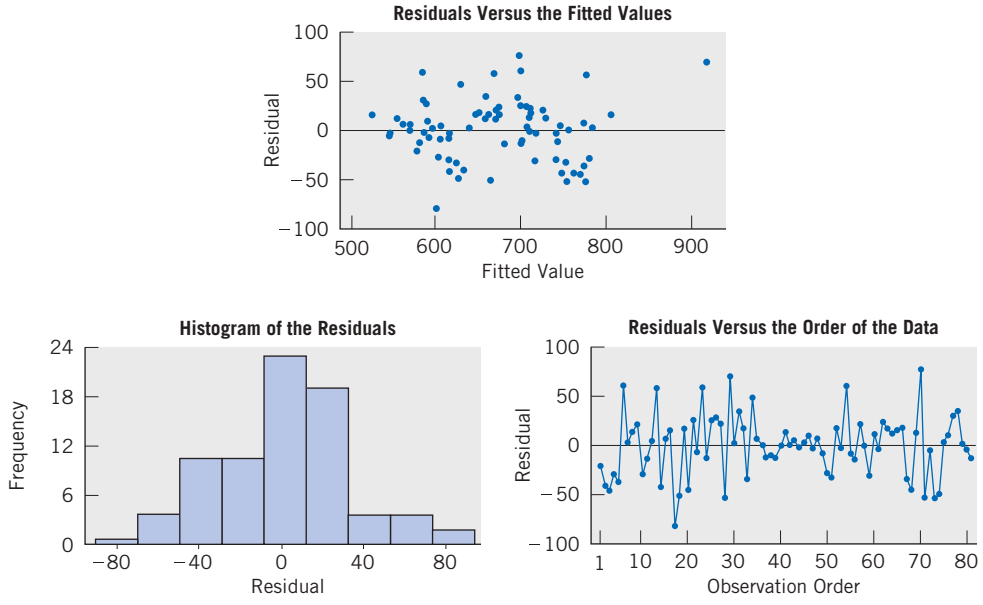


Figure 8 Three residual plots for posttest rowing time using MINITAB.

The MINITAB regression *four-in-one* graphics option created four residual plots, including the three in Figure 8 using the data and fit from Example 3.

It is important to remember that our confidence in statistical inference procedures is related to the validity of the assumptions about them. A mechanically made inference may be misleading if some model assumption is grossly violated. An **examination of the residuals** is an important part of regression analysis, because it helps to detect any inconsistency between the data and the postulated model.

If no serious violation of the assumption is exposed in the process of examining residuals, we consider the model adequate and proceed with the relevant inferences. Otherwise, we must search for a more appropriate model.

References

1. S. Chatterjee and A. Hadi, *Regression Analysis by Example,* 4th ed., John Wiley & Sons: New York, 2006.
2. N. R. Draper and H. Smith, *Applied Regression Analysis,* 3rd ed., John Wiley & Sons: New York, 1998.
3. D. Montgomery, E. Peck and G. Vinning, *Introduction to Linear Regression Analysis,* 3rd. ed., John Wiley & Sons: New York, 2001.

USING STATISTICS WISELY

1. Always, as a first step, plot the response variable versus the predictor variable. If there is more than one predictor variable, make separate plots for each. Examine the plot to see if a linear or other relationship exists.
2. Do not routinely accept the regression analysis presented in computer output. Instead, criticize the model by inspecting the residuals for outliers and moderate to severe lack of normality. A normal-scores plot is useful if there are more than 20 or so residuals. That plot may suggest a transformation.
3. Plot the residuals versus predicted value to check the assumption of constant variance. Plot the residuals in time order if that is appropriate. A trend over time would cast doubt on the assumption of independent errors.

KEY IDEAS AND FORMULAS

When a scatter diagram shows relationship on a curve, it may be possible to choose a **transformation** of one or both variables such that the transformed data exhibit a linear relation. A simple linear regression analysis can then be performed on the transformed data.

Multiple regression analysis is a versatile technique of building a prediction model with several input variables. In addition to obtaining the least squares fit, we can construct confidence intervals and test hypotheses about the influence of each input variable.

A **polynomial regression model** is a special case of multiple regression where the powers x, x^2, x^3, and so on, of a single predictor x play the role of the individual predictors.

The highest power of x that occurs in the model is called the **degree** or **order** of the regression model. A **quadratic function,** or **second-degree polynomial,** is commonly fit as an alternative to a straight line.

Both the polynomial regression model and the multiple regression with several predictors are special cases of **general linear models.**

The measure R^2, called the **square of the multiple correlation coefficient,** represents the proportion of y variability that is explained by the fitted multiple regression model.

$$R^2 = \frac{\text{Regression SS}}{\text{Total SS}}$$

To safeguard against a misuse of regression analysis, we must scrutinize the data for agreement with the model assumptions. An **examination of the residuals,** especially by graphical plots, including a **dot diagram** or **histogram,** a plot versus **predicted value,** and a plot versus **time order,** is essential for detecting possible violations of the assumptions and also identifying the appropriate modifications of an initial model.

TECHNOLOGY

Regression with two or more predictors and quadratic regression

MINITAB

Regression with two or more predictors

Begin with the values for the two predictor variables in C2 and C3 and the response variable *y* in C1.

Stat > Regression > Regression.
Type C1 in **Response.** Type C2 and C3 in **Predictors.**
Select **Graphs.** Click **Four in one.** Click **OK.**
Click **OK.**

The graphics step produces four residual plots: including histogram, residual versus fit, and residual versus order.

Transforming data

We illustrate with the predictor variable *x* in C2 and transforming to $\log(x)$, where the logarithm is base 10 in C3.

Calc > Calculator.
Type C3 in **Store results in variable** and *LOGT(C2)* in **Expression.**
Click **OK.**

Fitting a quadratic regression model

With the values of *x* in C1 and the *y* values in C2, you must select:

Stat > Regression > Fitted Line Plot.
Enter C2 in **Response (Y)** and enter C1 in **Predictor (X).**
Under **Type of Regression Model** choose **Quadratic.** Click **OK.**

TI-84/-83 PLUS

Fitting a quadratic regression model

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

Select **STAT,** then **CALC,** and then **5: QuadReg (ax + b)**.
With **LinReg** on the Home screen press **Enter.**

The calculator will give *a*, *b*, and *c* in the equation

$$y = ax^2 + bx + c$$

5. REVIEW EXERCISES

12.15 An environmental scientist identified a point source for E. Coli at the edge of a stream. She then measured y = E. Coli, in colony forming units per 100 ml water, at different distances, in feet, downstream from the point source. Suppose she obtains the following pairs of (x, y).

x	100	150	250	250	400	650	1000	1600
y	21	20	24	17	18	10	11	9

(a) Transform the x values to $x' = \log_{10} x$ and plot the scatter diagram of y versus x'.

(b) Fit a straight line regression to the transformed data.

(c) Obtain a 90% confidence interval for the slope of the regression line.

(d) Estimate the expected y value corresponding to $x = 300$ and give a 95% confidence interval.

*12.16 Obtain a linearizing transformation in each case.

(a) $y = \dfrac{1}{(1 + ae^{bx})^2}$

(b) $y = e^{ax^b}$

12.17 A genetic experiment is undertaken to study the competition between two types of female *Drosophila melanogaster* in cages with one male genotype acting as a substrate. The independent variable x is the time spent in cages, and the dependent variable y is the ratio of the numbers of type 1 to type 2 females. The following data (courtesy of C. Denniston) are recorded.

Time x (days)	No. Type 1	No. Type 2	$y = \dfrac{\text{No. Type 1}}{\text{No. Type 2}}$
17	137	586	.23
31	278	479	.58
45	331	167	1.98
59	769	227	3.39
73	976	75	13.01

(a) Plot the scatter diagram of y versus x and determine if a linear model of relation is appropriate.

(b) Determine if a linear relation is plausible for the transformed data $y' = \log_{10} y$.

(c) Fit a straight line regression to the transformed data.

12.18 Refer to the 2007 campus crime data in Chapter 11, Table 5. Obtaining the least squares fit to the response y = number of arson incidents, using the two predictor variables x_1 = robbery and x_2 = forceable rape, gives the results

$\hat{\beta}_0 = -.3780 \quad \hat{\beta}_1 = .3401 \quad \hat{\beta}_2 = .1826$

SS due to regression = 245.40

SSE = 113.15

(a) Predict the response for

(i) $x_1 = 7$ and $x_2 = 5$

(ii) $x_1 = 15$ and $x_2 = 11$

(b) Estimate the error standard deviation σ and state the degrees of freedom.

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

12.19 Refer to Exercise 12.18. The estimated standard errors of $\hat{\beta}_1$ and $\hat{\beta}_2$ were .1085 and .0451, respectively.

(a) Obtain a 90% confidence interval for β_1.

(b) Test $H_0 : \beta_2 = .10$ versus $H_1 : \beta_2 \neq .10$ with $\alpha = .05$.

12.20 A second-degree polynomial $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2$ is fitted to a response y, and the following predicted values and residuals are obtained.

$\hat{y}$	Residuals
4.01	.28
5.53	−.33
6.21	−.21
6.85	.24
8.17	−.97
8.34	.46
8.81	.79
9.62	−1.02
10.05	1.35
10.55	−1.55
10.77	.63
10.77	1.73
10.94	−2.14
10.98	1.92
10.98	−1.18

Do the assumptions appear to be violated?

12.21 The following predicted values and residuals are obtained in an experiment conducted to determine the degree to which the yield of an important chemical in the manufacture of penicillin is dependent on sugar concentration (the time order of the experiments is given in parentheses).

Predicted	Residual
2.2(9)	−1
3.1(6)	−2
2.5(13)	3
3.3(1)	−3
2.3(7)	−1
3.6(14)	5
2.6(8)	0
2.5(3)	0
3.0(12)	3
3.2(4)	−2
2.9(11)	2
3.3(2)	−5
2.7(10)	0
3.2(5)	1

(a) Plot the residuals against the predicted values and also against the time order.

(b) Do the basic assumptions appear to be violated?

12.22 An experimenter obtains the following residuals after fitting a quadratic expression in x.

$x = 1$	$x = 2$	$x = 3$	$x = 4$	$x = 5$
−.1	1.3	−.1	0	−.2
0	−.2	−.3	.2	0
−.2	−.1	.1	−.1	−.2
.6	−.3	.4	0	−.2
−.1	.1	−.1	−.2	−.3
		.1		−.1

Do the basic assumptions appear to be violated?

12.23 An interested student used the method of least squares to fit the straight line $\hat{y} = 264.3 + 18.77x$ to gross national product, y, in real dollars. The results for 26 recent years, $x = 1, 2, \ldots, 26$, appear below. Which assumption(s) for a linear regression model appear to be seriously violated by the data? (*Note:* Regression methods are usually not appropriate for this type of data.)

Year	y	$\hat{y}$	Residual
1	309.9	283.1	26.8
2	323.7	301.9	21.8
3	324.1	320.6	3.5
4	355.3	339.4	15.9
5	383.4	358.2	25.2
6	395.1	376.9	18.2
7	412.8	395.7	17.1
8	407	414.5	−7.5
9	438	433.2	4.8
10	446.1	452.0	−5.9
11	452.5	470.8	−18.3
12	447.3	489.5	−42.2
13	475.9	508.3	−32.4
14	487.7	527.1	−39.4
15	497.2	545.8	−48.6
16	529.8	564.6	−34.8
17	551	583.4	−32.4
18	581.1	602.1	−21.0
19	617.8	620.9	−3.1
20	658.1	639.7	18.4
21	675.2	658.4	16.8
22	706.6	677.2	29.4
23	725.6	696.0	29.6
24	722.5	714.7	7.8
25	745.4	733.5	11.9
26	790.7	752.3	38.4

The Following Exercises Require a Computer

12.24 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) Obtain the least squares fit of canine length to a quadratic function of the predictor age. The MINITAB commands are

Data: DBT9.txt
C2: 4 2 4 ⋯ 0
C6: 28.7 27.0 27.2 ⋯ 24.5
Dialog box:

Stat > Regression > Fitted line plot
Type C6 in **Response.**
Type C2 in **Predictor.**
Click **Quadratic.** Click **OK.**

(c) What proportion of the y variability is explained by the quadratic regression model?

(d) Compare the estimated standard deviations, s, of the random error term in parts (a) and (b).

12.25 The resident population of the United States has grown over the last 100 years from 1910 to 2010 but the growth has not been linear. The response variable is y = population in millions and, to simplify the calculations, the predictor variable is x = year − 1900.

x	10	20	30	40	50	60
y	92.2	106.0	123.2	132.2	151.3	179.3

x	70	80	90	100	110
y	203.3	226.5	248.7	281.4	310.2

(a) Fit a quadratic regression model $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + e$ to these data.

(b) What proportion of the y variability is explained by the quadratic regression model?

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

12.26 Listed below are the price quotations for a midsize foreign used car along with their age and odometer mileage.

Age (years) x_1	Mileage x_2 (thousand miles)	Price y (thousand miles)
1	14	17.9
2	44	13.9
2	20	14.9
4	36	14.0
4	66	9.8
5	59	9.9
7	100	6.6
7	95	6.7
8	38	7.0

Perform a multiple regression analysis of these data. In particular

(a) Determine the equation for predicting the price from age and mileage. Interpret the meaning of the coefficients $\hat{\beta}_1$ and $\hat{\beta}_2$.

(b) Give 95% confidence intervals for β_1 and β_2.

(c) Obtain R^2 and interpret the result.

12.27 Refer to the data of speed x and stopping distance y given in Table 1. The MINITAB commands for fitting a straight line regression to $y' = \sqrt{y}$ and x are

Data: C12T1.txt

C1:	20	20	30	30	30	40	40	50	50	60
C2:	16.3	26.7	39.2	63.5	51.3	98.4	65.7	104.1	155.6	217.2

Dialog box:

Calc > Calculator
Type *SQRT(C2)* in the **Expression** box.
Type *C3* in **Store** box. Click **OK.**

Stat > Regression > Regression
Type *C3* in **Response.**
Type *C1* in **Predictors.** Click **OK.**

(a) Obtain the computer output and identify the equation of the fitted line and the value of r^2 (see Example 1).

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

(c) Obtain a 95% confidence interval for the expected y' value at $x = 45$.

12.28 A forester seeking information on basic tree dimensions obtains the following measurements of the diameters 4.5 feet above the ground and the heights of 12 sugar maple trees (courtesy of A. Ek). The forester wishes to determine if the diameter measurements can be used to predict the tree height.

(a) Plot the scatter diagram and determine if a straight line relation is appropriate.

(b) Determine an appropriate linearizing transformation. In particular, try $x' = \log x$, $y' = \log y$.

(c) Fit a straight-line regression to the transformed data.

Diameter x (inches)	Height y (feet)
.9	18
1.2	26
2.9	32
3.1	36
3.3	44.5
3.9	35.6
4.3	40.5
6.2	57.5
9.6	67.3
12.6	84
16.1	67
25.8	87.5

(d) What proportion of variability is explained by the fitted model?

12.29 Recorded here are the scores x_1 and x_2 in two midterm examinations, the GPA x_3, and the final examination score y for 20 students in a statistics class.

x_1	x_2	x_3	y	x_1	x_2	x_3	y
87	25	2.9	60	93	60	3.2	44
100	84	3.3	80	92	69	3.1	53
91	52	3.5	73	100	86	3.6	86
85	60	3.7	83	80	67	3.5	59
56	76	2.8	33	100	96	3.8	81
81	28	3.1	65	69	51	2.8	20
85	67	3.1	53	80	75	3.6	64
96	83	3.0	68	74	70	3.1	38
79	60	3.7	88	79	66	2.9	77
96	69	3.7	89	95	83	3.3	47

(a) Ignoring the data of GPA and the first midterm score, fit a simple linear regression of y on x_2. Compute r^2.

(b) Fit a multiple linear regression to predict the final examination score from the GPA and the scores in the midterms. Compute R^2.

(c) Interpret the values of r^2 and R^2 obtained in parts (a) and (b).

12.30 Refer to Exercise 11.64.

(a) Fit a quadratic model $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + e$ to the data for CLEP scores y and CQT scores x.

(b) Use the fitted regression to predict the expected CLEP score when $x = 160$.

(c) Compute r^2 for fitting a line and R^2 for fitting a quadratic expression. Interpret these values and comment on the improvement of fit.

*12.31 Write the design matrix $\mathbf{X}$ for fitting a multiple regression model to the data of Exercise 12.26.

*12.32 Write the design matrix $\mathbf{X}$ for fitting a quadratic regression model using the data of Exercise 12.25.

12.33 Refer to the physical conditioning data given in Table D.5 of the Data Bank. Use MINITAB or some other package to fit a regression of the final number of situps on the initial number of situps and the gender of the student.

12.34 Refer to the physical fitness data in Table D.5 of the Data Bank. Use both the data on the pretest run time and gender for predicting the posttest run time. Obtain the least squares fit and plot the residuals versus fitted value.