CHAPTER 13

Curve Fitting and the Method of Least Squares

RELATIONSHIP BETWEEN VARIABLES

Very often in practice a relationship is found to exist between two (or more) variables. For example, weights of adult males depend to some degree on their heights, the circumferences of circles depend on their radii, and the pressure of a given mass of gas depends on its temperature and volume.

It is frequently desirable to express this relationship in mathematical form by determining an equation that connects the variables.

CURVE FITTING

To determine an equation that connects variables, a first step is to collect data that show corresponding values of the variables under consideration. For example, suppose that X and Y denote, respectively, the height and weight of adult males; then a sample of N individuals would reveal the heights $X_1, X_2, \ldots, X_N$ and the corresponding weights $Y_1, Y_2, \ldots, Y_N$.

A next step is to plot the points $(X_1, Y_1), (X_2, Y_2), \ldots, (X_N, Y_N)$ on a rectangular coordinate system. The resulting set of points is sometimes called a *scatter diagram*.

From the scatter diagram it is often possible to visualize a smooth curve that approximates the data. Such a curve is called an *approximating curve*. In Fig. 13-1, for example, the data appear to be approximated well by a straight line, and so we say that a *linear relationship* exists between the variables. In Fig. 13-2, however, although a relationship exists between the variables, it is not a linear relationship, and so we call it a *nonlinear relationship*.

The general problem of finding equations of approximating curves that fit given sets of data is called *curve fitting*.

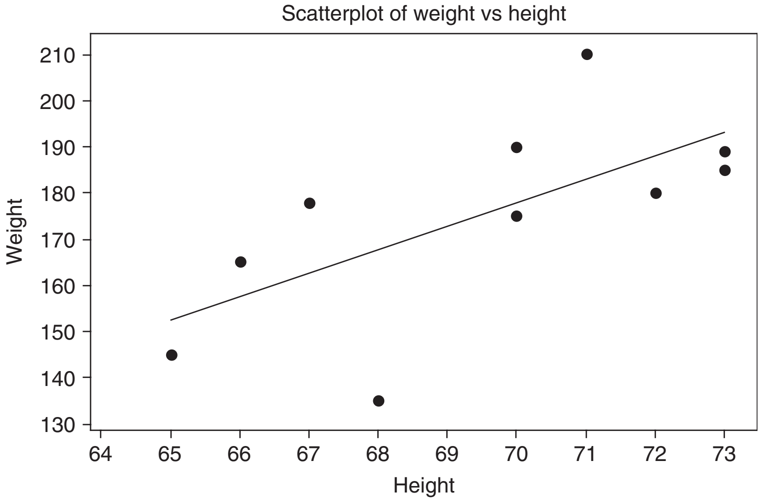


Fig. 13-1 Straight lines sometimes describe the relationship between two variables.

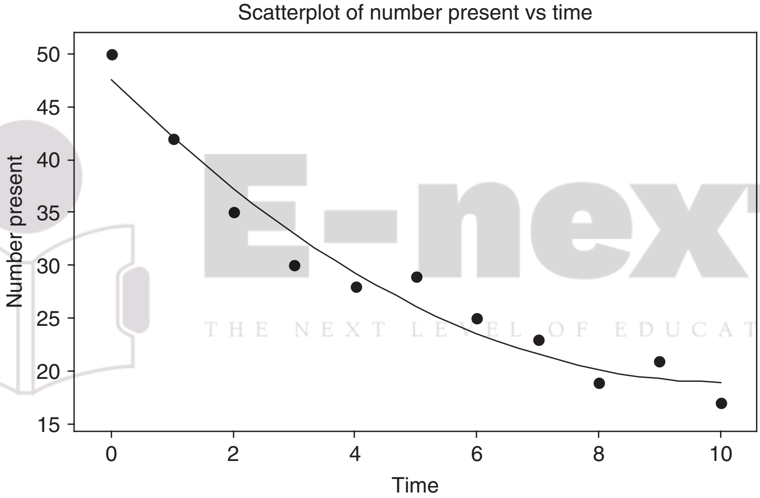


Fig. 13-2 Nonlinear relationships sometimes describe the relationship between two variables.

EQUATIONS OF APPROXIMATING CURVES

Several common types of approximating curves and their equations are listed below for reference purposes. All letters other than X and Y represent constants. The variables X and Y are often referred to as *independent* and *dependent variables*, respectively, although these roles can be interchanged.

Straight line	$Y = a_0 + a_1 X$	(*1*)
Parabola, or quadratic curve	$Y = a_0 + a_1 X + a_2 X^2$	(*2*)
Cubic curve	$Y = a_0 + a_1 X + a_2 X^2 + a_3 X^3$	(*3*)
Quartic curve	$Y = a_0 + a_1 X + a_2 X^2 + a_3 X^3 + a_4 X^4$	(*4*)
*n*th-Degree curve	$Y = a_0 + a_1 X + a_2 X^2 + \cdots + a_n X^n$	(*5*)

The right sides of the above equations are called *polynomials* of the first, second, third, fourth, and nth degrees, respectively. The functions defined by the first four equations are sometimes called *linear*, *quadratic*, *cubic*, and *quartic* functions, respectively.

The following are some of the many other equations frequently used in practice:

Hyperbola $$Y = \frac{1}{a_0 + a_1 X} \quad \text{or} \quad \frac{1}{Y} = a_0 + a_1 X \tag{6}$$

Exponential curve $$Y = ab^X \quad \text{or} \quad \log Y = \log a + (\log b)X = a_0 + a_1 X \tag{7}$$

Geometric curve $$Y = aX^b \quad \text{or} \quad \log Y = \log a + b(\log X) \tag{8}$$

Modified exponential curve $$Y = ab^X + g \tag{9}$$

Modified geometric curve $$Y = aX^b + g \tag{10}$$

Gompertz curve $$Y = pq^{b^X} \quad \text{or} \quad \log Y = \log p + b^X(\log q) = ab^X + g \tag{11}$$

Modified Gompertz curve $$Y = pq^{b^X} + h \tag{12}$$

Logistic curve $$Y = \frac{1}{ab^X + g} \quad \text{or} \quad \frac{1}{Y} = ab^X + g \tag{13}$$

$$Y = a_0 + a_1(\log X) + a_2(\log X)^2 \tag{14}$$

To decide which curve should be used, it is helpful to obtain scatter diagrams of transformed variables. For example, if a scatter diagram of log Y versus X shows a linear relationship, the equation has the form (*7*), while if log Y versus log X shows a linear relationship, the equation has the form (*8*). Special graph paper is often used in order to make it easy to decide which curve to use. Graph paper having one scale calibrated logarithmically is called *semilogarithmic* (or *semilog*) *graph paper*, and that having both scales calibrated logarithmically is called *log-log graph paper*.

FREEHAND METHOD OF CURVE FITTING

Individual judgment can often be used to draw an approximating curve to fit a set of data. This is called a *freehand method of curve fitting*. If the type of equation of this curve is known, it is possible to obtain the constants in the equation by choosing as many points on the curve as there are constants in the equation. For example, if the curve is a straight line, two points are necessary; if it is a parabola, three points are necessary. The method has the disadvantage that different observers will obtain different curves and equations.

THE STRAIGHT LINE

The simplest type of approximating curve is a straight line, whose equation can be written

$$Y = a_0 + a_1 X \tag{15}$$

Given any two points (X_1, Y_1) and (X_2, Y_2) on the line, the constants a_0 and a_1 can be determined. The resulting equation of the line can be written

$$Y - Y_1 = \left(\frac{Y_2 - Y_1}{X_2 - X_1}\right)(X - X_1) \qquad \text{or} \qquad Y - Y_1 = m(X - X_1) \tag{16}$$

where $$m = \frac{Y_2 - Y_1}{X_2 - X_1}$$

is called the *slope* of the line and represents the change in Y divided by the corresponding change in X.

When the equation is written in the form (*15*), the constant a_1 is the slope m. The constant a_0, which is the value of Y when $X = 0$, is called the *Y intercept*.

THE METHOD OF LEAST SQUARES

To avoid individual judgment in constructing lines, parabolas, or other approximating curves to fit sets of data, it is necessary to agree on a definition of a "best-fitting line," "best-fitting parabola," etc.

By way of forming a definition, consider Fig. 13-3, in which the data points are given by (X_1, Y_1), $(X_2, Y_2), \ldots, (X_N, Y_N)$. For a given value of X, say X_1, there will be a difference between the value Y_1 and the corresponding value as determined from the curve C. As shown in the figure, we denote this difference by D_1, which is sometimes referred to as a *deviation*, *error*, or *residual* and may be positive, negative, or zero. Similarly, corresponding to the values $X_2, \ldots, X_N$ we obtain the deviations $D_2, \ldots, D_N$.

A measure of the "goodness of fit" of the curve C to the given data is provided by the quantity $D_1^2 + D_2^2 + \cdots + D_N^2$. If this is small, the fit is good; if it is large, the fit is bad. We therefore make the following

> **Definition:** Of all curves approximating a given set of data points, the curve having the property that $D_1^2 + D_2^2 + \cdots + D_N^2$ is a minimum is called a *best-fitting curve*.

A curve having this property is said to fit the data in the *least-squares sense* and is called a *least-squares curve*. Thus a line having this property is called a *least-squares line*, a parabola with this property is called a *least-squares parabola*, etc.

It is customary to employ the above definition when X is the independent variable and Y is the dependent variable. If X is the dependent variable, the definition is modified by considering horizontal instead of vertical deviations, which amounts to an interchange of the X and Y axes. These two definitions generally lead to different least-squares curves. Unless otherwise specified, we shall consider Y the dependent variable and X the independent variable.

It is possible to define another least-squares curve by considering perpendicular distances from each of the data points to the curve instead of either vertical or horizontal distances. However, this is not used very often.

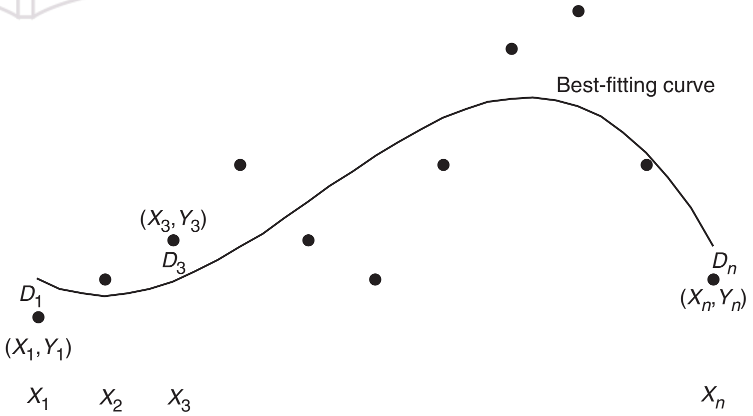


Fig. 13-3 D_1 is the distance from data point (X_1, Y_1) to the best-fitting curve, ..., D_n is the distance from data point (X_n, Y_n) to the best-fitting curve.

THE LEAST-SQUARES LINE

The least-squares line approximating the set of points $(X_1, Y_1), (X_2, Y_2), \ldots, (X_N, Y_N)$ has the equation

$$Y = a_0 + a_1 X \qquad (17)$$

where the constants a_0 and a_1 are determined by solving simultaneously the equations

$$\begin{aligned} \sum Y &= a_0 N \qquad + a_1 \sum X \\ \sum XY &= a_0 \sum X + a_1 \sum X^2 \end{aligned} \tag{18}$$

which are called the *normal equations for the least-squares line* (*17*). The constants a_0 and a_1 of equations (*18*) can, if desired, be found from the formulas

$$a_0 = \frac{(\sum Y)(\sum X^2) - (\sum X)(\sum XY)}{N \sum X^2 - (\sum X)^2} \qquad a_1 = \frac{N \sum XY - (\sum X)(\sum Y)}{N \sum X^2 - (\sum X)^2} \tag{19}$$

The normal equations (*18*) are easily remembered by observing that the first equation can be obtained formally by summing on both sides of (*17*) [i.e., $\sum Y = \sum (a_0 + a_1 X) = a_0 N + a_1 \sum X$], while the second equation is obtained formally by first multiplying both sides of (*17*) by X and then summing [i.e., $\sum XY = \sum X(a_0 + a_1 X) = a_0 \sum X + a_1 \sum X^2$]. Note that this is not a derivation of the normal equations, but simply a means for remembering them. Note also that in equations (*18*) and (*19*) we have used the short notation $\sum X$, $\sum XY$, etc., in place of $\sum_{j=1}^{N} X_j$, $\sum_{j=1}^{N} X_j Y_j$, etc.

The labor involved in finding a least-squares line can sometimes be shortened by transforming the data so that $x = X - \bar{X}$ and $y = Y - \bar{Y}$. The equation of the least-squares line can then be written (see Problem 13.15)

$$y = \left(\frac{\sum xy}{\sum x^2} \right) x \qquad \text{or} \qquad y = \left(\frac{\sum xY}{\sum x^2} \right) x \tag{20}$$

In particular, if X is such that $\sum X = 0$ (i.e., $\bar{X} = 0$), this becomes

$$Y = \bar{Y} + \left(\frac{\sum XY}{\sum X^2} \right) X \tag{21}$$

Equation (*20*) implies that $y = 0$ when $x = 0$; thus the least-squares line passes through the point $(\bar{X}, \bar{Y})$, called the *centroid*, or *center of gravity*, of the data.

If the variable X is taken to be the dependent instead of the independent variable, we write equation (*17*) as $X = b_0 + b_1 Y$. Then the above results hold if X and Y are interchanged and a_0 and a_1 are replaced by b_0 and b_1, respectively. The resulting least-squares line, however, is generally not the same as that obtained above [see Problems 13.11 and 13.15(*d*)].

NONLINEAR RELATIONSHIPS

Nonlinear relationships can sometimes be reduced to linear relationships by an appropriate transformation of the variables (see Problem 13.21).

THE LEAST-SQUARES PARABOLA

The least-squares parabola approximating the set of points (X_1, Y_1), $(X_2, Y_2), \ldots, (X_N, Y_N)$ has the equation

$$Y = a_0 + a_1 X + a_2 X^2 \tag{22}$$

where the constants a_0, a_1, and a_2 are determined by solving simultaneously the equations

$$\begin{aligned} \sum Y &= a_0 N \qquad + a_1 \sum X \ + a_2 \sum X^2 \\ \sum XY &= a_0 \sum X \ + a_1 \sum X^2 + a_2 \sum X^3 \\ \sum X^2 Y &= a_0 \sum X^2 + a_1 \sum X^3 + a_2 \sum X^4 \end{aligned} \tag{23}$$

called the *normal equations for the least-squares parabola* (*22*).

Equations (*23*) are easily remembered by observing that they can be obtained formally by multiplying equation (*22*) by 1, X, and X^2, respectively, and summing on both sides of the resulting equations. This technique can be extended to obtain normal equations for least-squares cubic curves, least-squares quartic curves, and in general any of the least-squares curves corresponding to equation (*5*).

As in the case of the least-squares line, simplifications of equations (*23*) occur if X is chosen so that $\sum X = 0$. Simplification also occurs by choosing the new variables $x = X - \bar{X}$ and $y = Y - \bar{Y}$.

REGRESSION

Often, on the basis of sample data, we wish to estimate the value of a variable Y corresponding to a given value of a variable X. This can be accomplished by estimating the value of Y from a least-squares curve that fits the sample data. The resulting curve is called a *regression curve of Y on X*, since Y is estimated from X.

If we wanted to estimate the value of X from a given value of Y, we would use a *regression curve of X on Y*, which amounts to interchanging the variables in the scatter diagram so that X is the dependent variable and Y is the independent variable. This is equivalent to replacing the vertical deviations in the definition of the least-squares curve on page 284 with horizontal deviations.

In general, the regression line or curve of Y on X is not the same as the regression line or curve of X on Y.

APPLICATIONS TO TIME SERIES

If the independent variable X is time, the data show the values of Y at various times. Data arranged according to time are called *time series*. The regression line or curve of Y on X in this case is often called a *trend line* or *trend curve* and is often used for purposes of *estimation*, *prediction*, or *forecasting*.

PROBLEMS INVOLVING MORE THAN TWO VARIABLES

Problems involving more than two variables can be treated in a manner analogous to that for two variables. For example, there may be a relationship between the three variables X, Y, and Z that can be described by the equation

$$Z = a_0 + a_1 X + a_2 Y \tag{24}$$

which is called a *linear equation in the variables X, Y, and Z*.

In a three-dimensional rectangular coordinate system this equation represents a plane, and the actual sample points (X_1, Y_1, Z_1), $(X_2, Y_2, Z_2), \ldots, (X_N, Y_N, Z_N)$ may "scatter" not too far from this plane, which we call an *approximating plane*.

By extension of the method of least squares, we can speak of a *least-squares plane* approximating the data. If we are estimating Z from given values of X and Y, this would be called a *regression plane of Z on X and Y*. The normal equations corresponding to the least-squares plane (*24*) are given by

$$\begin{aligned} \sum Z &= a_0 N + a_1 \sum X + a_2 \sum Y \\ \sum XZ &= a_0 \sum X + a_1 \sum X^2 + a_2 \sum XY \\ \sum YZ &= a_0 \sum Y + a_1 \sum XY + a_2 \sum Y^2 \end{aligned} \tag{25}$$

and can be remembered as being obtained from equation (*24*) by multiplying by 1, X, and Y successively and then summing.

More complicated equations than (*24*) can also be considered. These represent *regression surfaces*. If the number of variables exceeds three, geometric intuition is lost since we then require four-, five-, . . . n-dimensional spaces.

Problems involving the estimation of a variable from two or more variables are called problems of *multiple regression* and will be considered in more detail in Chapter 15.

Solved Problems

STRAIGHT LINES

13.1 Thirty high school students were surveyed in a study involving time spent on the Internet and their grade point average (GPA). The results are shown in Table 13.1. X is the amount of time spent on the Internet weekly and Y is the GPA of the student.

Table 13.1

Hours	GPA	Hours	GPA	Hours	GPA
11	2.84	9	2.85	25	1.85
5	3.20	5	3.35	6	3.14
22	2.18	14	2.60	9	2.96
23	2.12	18	2.35	20	2.30
20	2.55	6	3.14	14	2.66
20	2.24	9	3.05	19	2.36
10	2.90	24	2.06	21	2.24
19	2.36	25	2.00	7	3.08
15	2.60	12	2.78	11	2.84
18	2.42	6	2.90	20	2.45

Use MINITAB to do the following:

(*a*) Make a scatter plot of the data.

(*b*) Fit a straight line to the data and give the values of a_0 and a_1.

SOLUTION

(*a*) The data are entered into columns `C1` and `C2` of the MINITAB worksheet. `C1` is named `Internet-hours` and `C2` is named `GPA`. The pull-down menu **Stat** → **Regression** → **Regression** is given and the results shown in Fig. 13-4 are formed.

(*b*) The value for a_0 is 3.49 and the value for a_1 is -0.0594.

13.2 Solve Problem 13.1 using EXCEL.

SOLUTION

The data is entered into columns A and B of the EXCEL worksheet. The pull-down menu **Tools** → **Data Analysis** → **Regression** produces the dialog box in Fig. 13-5 which is filled in as shown. The part of the output which is currently of interest is

```
Intercept          3.488753
Internet-hours    −0.05935
```

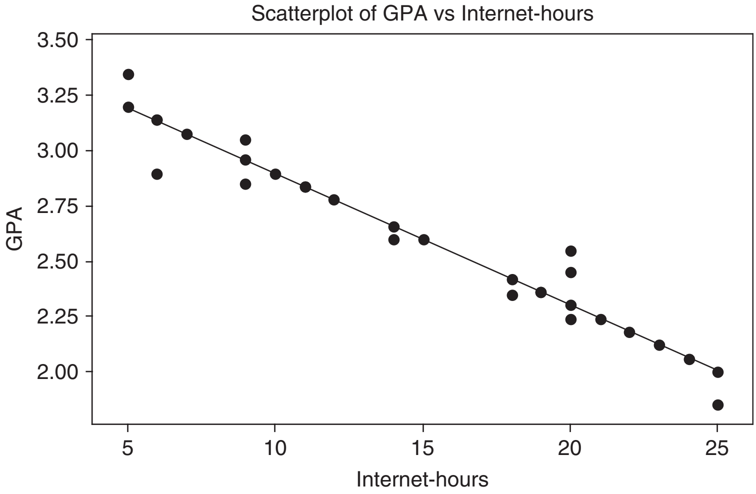
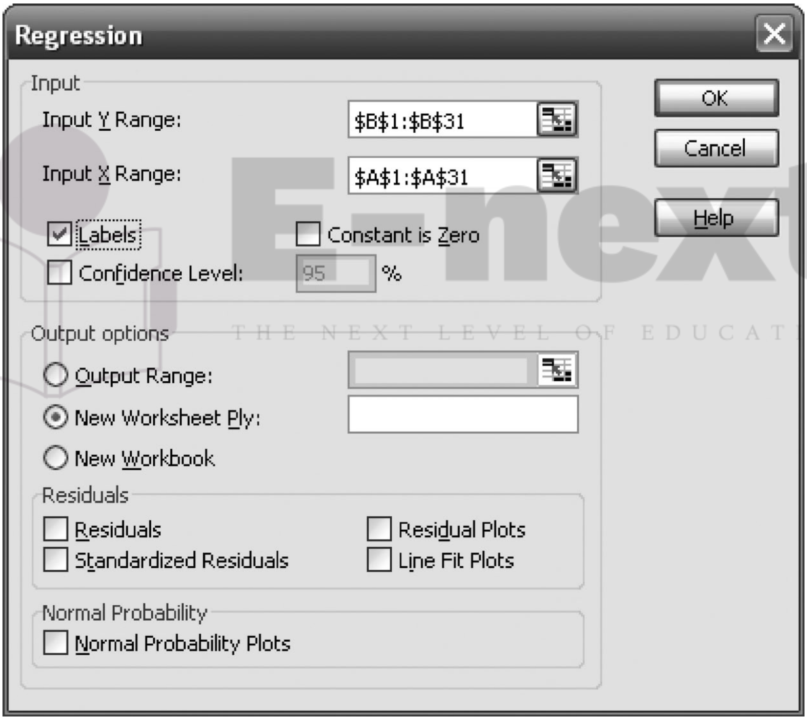


Fig. 13-4 The sum of the squares of the distances of the points from the best-fitting line is a minimum for the line GPA = 3.49 – 0.0594 Internet-hours.



Input

Input Y Range: $B$1:$B$31

Input X Range: $A$1:$A$31

☑ Labels □ Constant is Zero

□ Confidence Level: 95 %

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Residuals

□ Residuals □ Residual Plots

□ Standardized Residuals □ Line Fit Plots

Normal Probability

□ Normal Probability Plots

Fig. 13-5 EXCEL dialog box for Problem 13.2.

The constant a_0 is called the *intercept* and a_1 is the *slope*. The same values as given by MINITAB are obtained.

13.3 (*a*) Show that the equation of a straight line that passes through the points (X_1, Y_1) and (X_2, Y_2) is given by

$$Y - Y_1 = \frac{Y_2 - Y_1}{X_2 - X_1}(X - X_1)$$

(*b*) Find the equation of a straight line that passes through the points $(2, -3)$ and $(4, 5)$.

SOLUTION

(*a*) The equation of a straight line is

$$Y = a_0 + a_1 X \tag{29}$$

Since (X_1, Y_1) lies on the line,

$$Y_1 = a_0 + a_1 X_1 \tag{30}$$

Since (X_2, Y_2) lies on the line,

$$Y_2 = a_0 + a_1 X_2 \tag{31}$$

Subtracting equation (*30*) from (*29*),

$$Y - Y_1 = a_1(X - X_1) \tag{32}$$

Subtracting equation (*30*) from (*31*),

$$Y_2 - Y_1 = a_1(X_2 - X_1) \qquad \text{or} \qquad a_1 = \frac{Y_2 - Y_1}{X_2 - X_1}$$

Substituting this value of a_1 into equation (*32*), we obtain

$$Y - Y_1 = \frac{Y_2 - Y_1}{X_2 - X_1}(X - X_1)$$

as required. The quantity

$$\frac{Y_2 - Y_1}{X_2 - X_1}$$

often abbreviated m, represents the change in Y divided by the corresponding change in X and is the *slope* of the line. The required equation can be written $Y - Y_1 = m(X - X_1)$.

(*b*) **First method** [using the result of part (*a*)]

Corresponding to the first point $(2, -3)$, we have $X_1 = 2$ and $Y_1 = -3$; corresponding to the second point $(4, 5)$, we have $X_2 = 4$ and $Y_2 = 5$. Thus the slope is

$$m = \frac{Y_2 - Y_1}{X_2 - X_1} = \frac{5 - (-3)}{4 - 2} = \frac{8}{2} = 4$$

and the required equation is

$$Y - Y_1 = m(X - X_1) \qquad \text{or} \qquad Y - (-3) = 4(X - 2)$$

which can be written $Y + 3 = 4(X - 2)$, or $Y = 4X - 11$.

Second method

The equation of a straight line is $Y = a_0 + a_1 X$. Since the point $(2, -3)$ is on the line $-3 = a_0 + 2a_1$, and since the point $(4, 5)$ is on the line, $5 = a_0 + 4a_1$; solving these two equations simultaneously, we obtain $a_1 = 4$ and $a_0 = -11$. Thus the required equation is

$$Y = -11 + 4X \quad \text{or} \quad Y = 4X - 11$$

13.4 Wheat is grown on 9 equal-sized plots. The amount of fertilizer put on each plot is given in Table 13.2 along with the yield of wheat.

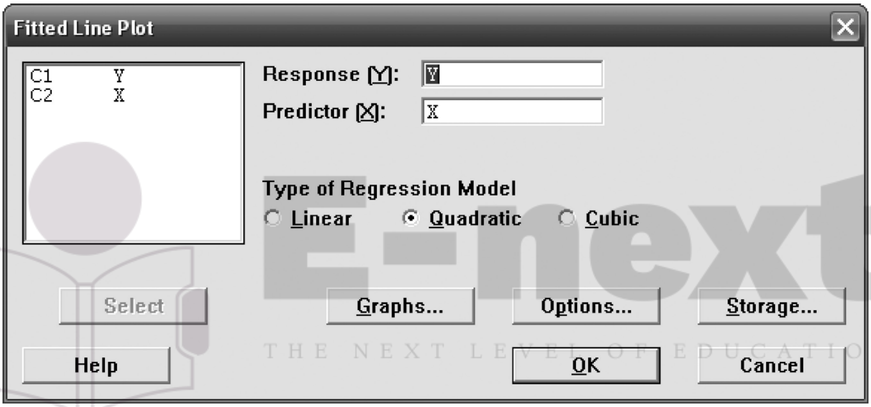
Use MINITAB to fit the parabolic curve $Y = a_0 + a_1 X + a_2 X^2$ to the data.

Table 13.2

Amount of Wheat (y)	Fertilizer (x)
2.4	1.2
3.4	2.3
4.4	3.3
5.1	4.1
5.5	4.8
5.2	5.0
4.9	5.5
4.4	6.1
3.9	6.9

SOLUTION

The wheat yield is entered into `C1` and the fertilizer is entered into `C2`. The pull-down **Stat** → **Regression** → **Fitted Line Plot** gives the dialog box in Fig. 13-6.



Response (Y): Y

Predictor (X): X

Type of Regression Model

○ Linear ● Quadratic ○ Cubic

Fig. 13-6 MINITAB dialog box for Problem 13.4.

This dialog box gives the output in Fig. 13-7.

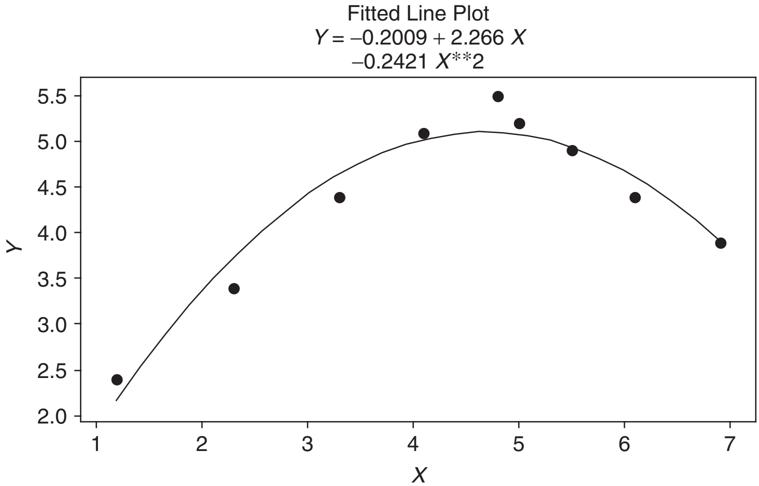


Fig. 13-7 Fitting the least-squares parabolic curve to a set of data using MINITAB.

13.5 Find (*a*) the slope, (*b*) the equation, (*c*) the Y intercept, and (*d*) the X intercept of the line that passes through the points $(1, 5)$ and $(4, -1)$.

SOLUTION

(*a*) $(X_1 = 1,\ Y_1 = 5)$ and $(X_2 = 4,\ Y_2 = -1)$. Thus

$$m = \text{slope} = \frac{Y_2 - Y_1}{X_2 - X_1} = \frac{-1 - 5}{4 - 1} = \frac{-6}{3} = -2$$

The negative sign of the slope indicates that as X increases, Y decreases, as shown in Fig. 13-8.

(*b*) The equation of the line is

$$Y - Y_1 = m(X - X_1) \qquad \text{or} \qquad Y - 5 = -2(X - 1)$$

That is,

$$Y - 5 = -2X + 2 \qquad \text{or} \qquad Y = 7 - 2X$$

This can also be obtained by the second method of Problem 13.3(*b*).

(*c*) The Y intercept, which is the value of Y when $X = 0$, is given by $Y = 7 - 2(0) = 7$. This can also be seen directly from Fig. 13-8.

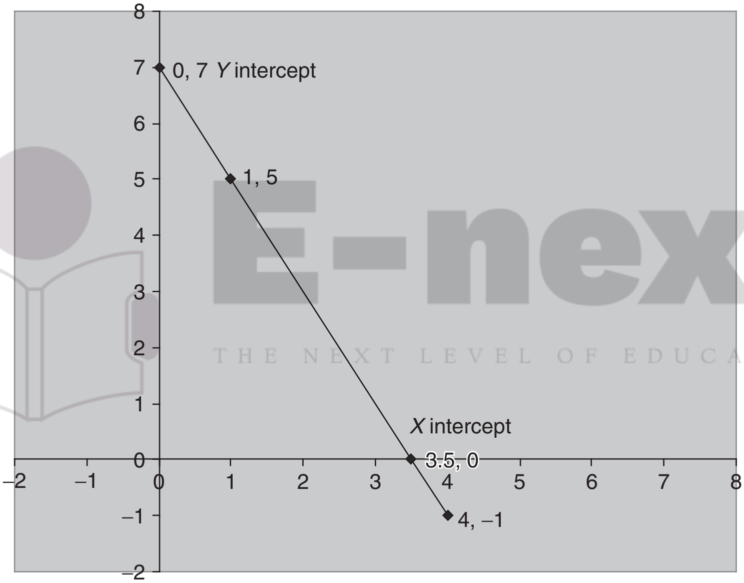


Fig. 13-8 Straight line showing X intercept and Y intercept.

(*d*) The X intercept is the value of X when $Y = 0$. Substituting $Y = 0$ in the equation $Y = 7 - 2X$, we have $0 = 7 - 2X$, or $2X = 7$ and $X = 3.5$. This can also be seen directly from Fig. 13-8.

13.6 Find the equation of a line passing through the point $(4, 2)$ that is parallel to the line $2X + 3Y = 6$.

SOLUTION

If two lines are parallel, their slopes are equal. From $2X + 3Y = 6$ we have $3Y = 6 - 2X$, or $Y = 2 - \frac{2}{3}X$, so that the slope of the line is $m = -\frac{2}{3}$. Thus the equation of the required line is

$$Y - Y_1 = m(X - X_1) \qquad \text{or} \qquad Y - 2 = -\tfrac{2}{3}(X - 4)$$

which can also be written $2X + 3Y = 14$.

Another method

Any line parallel to $2X + 3Y = 6$ has the equation $2X + 3Y = c$. To find c, let $X = 4$ and $Y = 2$. Then $2(4) + 3(2) = c$, or $c = 14$, and the required equation is $2X + 3Y = 14$.

13.7 Find the equation of a line whose slope is -4 and whose Y intercept is 16.

SOLUTION

In the equation $Y = a_0 + a_1X$, $a_0 = 16$ is the Y intercept and $a_1 = -4$ is the slope. Thus the required equation is $Y = 16 - 4X$.

13.8 (*a*) Construct a straight line that approximates the data of Table 13.3.

(*b*) Find an equation for this line.

Table 13.3

X	1	3	4	6	8	9	11	14
Y	1	2	4	4	5	7	8	9

SOLUTION

(*a*) Plot the points (1, 1), (3, 2), (4, 4), (6, 4), (8, 5), (9, 7), (11, 8), and (14, 9) on a rectangular coordinate system, as shown in Fig. 13-9. A straight line approximating the data is drawn *freehand* in the figure. For a method eliminating the need for individual judgment, see Problem 13.11, which uses the method of least squares.

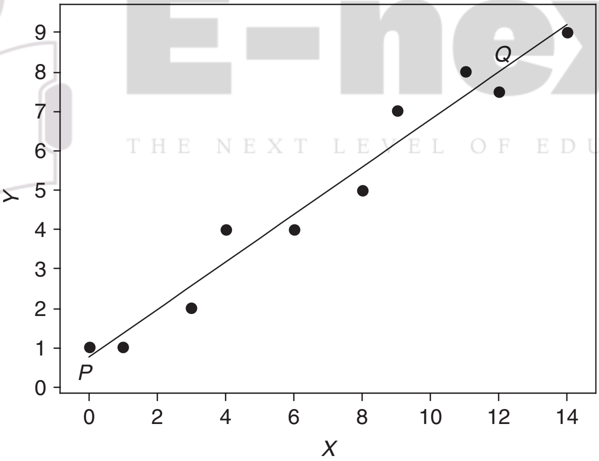


Fig. 13-9 Freehand method of curve fitting.

(*b*) To obtain the equation of the line constructed in part (*a*), choose any two points on the line, such as P and Q; the coordinates of points P and Q, as read from the graph, are approximately (0, 1) and (12, 7.5). The equation of the line is $Y = a_0 + a_1X$. Thus for point (0, 1) we have $1 = a_0 + a_1(0)$, and for point (12, 7.5) we have $7.5 = a_0 + 12a_1$; since the first of these equations gives us $a_0 = 1$, the second gives us $a_1 = 6.5/12 = 0.542$. Thus the required equation is $Y = 1 + 0.542X$.

Another method

$$Y - Y_1 = \frac{Y_2 - Y_1}{X_2 - X_1}(X - X_1) \qquad \text{and} \qquad Y - 1 = \frac{7.4 - 1}{12 - 0}(X - 0) = 0.542X$$

Thus $Y = 1 + 0.542X$.

13.9 (*a*) Compare the values of Y obtained from the approximating line with those given in Table 13.2.

(*b*) Estimate the value of Y when $X = 10$.

SOLUTION

(*a*) For $X = 1$, $Y = 1 + 0.542(1) = 1.542$, or 1.5. For $X = 3$, $Y = 1 + 0.542(3) = 2.626$ or 2.6. The values of Y corresponding to other values of X can be obtained similarly. The values of Y estimated from the equation $Y = 1 + 0.542X$ are denoted by Y_{est}. These estimated values, together with the actual data from Table 13.3, are shown in Table 13.4.

(*b*) The estimated value of Y when $X = 10$ is $Y = 1 + 0.542(10) = 6.42$, or 6.4.

Table 13.4

X	1	3	4	6	8	9	11	14
Y	1	2	4	4	5	7	8	9
Y_{est}	1.5	2.6	3.2	4.3	5.3	5.9	7.0	8.6

13.10 Table 13.5 shows the heights to the nearest inch (in) and the weights to the nearest pound (lb) of a sample of 12 male students drawn at random from the first-year students at State College.

Table 13.5

Height X (in)	70	63	72	60	66	70	74	65	62	67	65	68
Weight Y (lb)	155	150	180	135	156	168	178	160	132	145	139	152

(*a*) Obtain a scatter diagram of the data.

(*b*) Construct a line that approximates the data.

(*c*) Find the equation of the line constructed in part (*b*).

(*d*) Estimate the weight of a student whose height is known to be 63 in.

(*e*) Estimate the height of a student whose weight is known to be 168 lb.

SOLUTION

(*a*) The scatter diagram, shown in Fig. 13-10, is obtained by plotting the points (70, 155), (63, 150), ..., (68, 152).

(*b*) A straight line that approximates the data is shown dashed in Fig. 13-10. This is but one of the many possible lines that could have been constructed.

(*c*) Choose any two points on the line constructed in part (*b*), such as P and Q, for example. The coordinates of these points as read from the graph are approximately (60, 130) and (72, 170). Thus

$$Y - Y_1 = \frac{Y_2 - Y_1}{X_2 - X_1}(X - X_1) \qquad Y - 130 = \frac{170 - 130}{72 - 60}(X - 60) \qquad Y = \frac{10}{3}X - 70$$

(*d*) If $X = 63$, then $Y = \frac{10}{3}(63) - 70 = 140$ lb.

(*e*) If $Y = 168$, then $168 = \frac{10}{3}X - 70$, $\frac{10}{3}X = 238$, and $X = 71.4$, or 71 in.

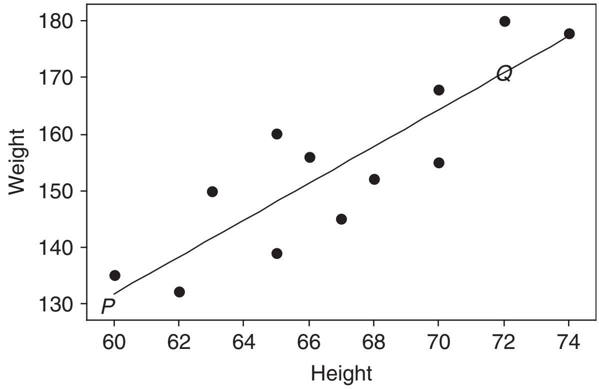


Fig. 13-10 Freehand method of curve fitting.

THE LEAST-SQUARES LINE

13.11 Fit a least-squares line to the data of Problem 13.8 by using (*a*) X as the independent variable and (*b*) X as the dependent variable.

SOLUTION

(*a*) The equation of the line is $Y = a_0 + a_1X$. The normal equations are

$$\begin{aligned} \sum Y &= a_0N + a_1 \sum X \\ \sum XY &= a_0 \sum X + a_1 \sum X^2 \end{aligned}$$

The work involved in computing the sums can be arranged as in Table 13.6. Although the right-hand column is not needed for this part of the problem, it has been added to the table for use in part (*b*).

Since there are eight pairs of values of X and Y, $N = 8$ and the normal equations become

$$\begin{aligned} 8a_0 + 56a_1 &= 40 \\ 56a_0 + 524a_1 &= 364 \end{aligned}$$

Solving simultaneously, $a_0 = \frac{6}{11}$, or 0.545; $a_1 = \frac{7}{11}$, or 0.636; and the required least-squares line is $Y = \frac{6}{11} + \frac{7}{11}X$, or $Y = 0.545 + 0.636X$.

Table 13.6

X	Y	X^2	XY	Y^2
1	1	1	1	1
3	2	9	6	4
4	4	16	16	16
6	4	36	24	16
8	5	64	40	25
9	7	81	63	49
11	8	121	88	64
14	9	196	126	81
$\sum X = 56$	$\sum Y = 40$	$\sum X^2 = 524$	$\sum XY = 364$	$\sum Y^2 = 256$

Another method

$$a_0 = \frac{(\sum Y)(\sum X^2) - (\sum X)(\sum XY)}{N\sum X^2 - (\sum X)^2} = \frac{(40)(524) - (56)(364)}{(8)(524) - (56)^2} = \frac{6}{11} \quad \text{or} \quad 0.545$$

$$a_1 = \frac{N\sum XY - (\sum X)(\sum Y)}{N\sum X^2 - (\sum X)^2} = \frac{(8)(364) - (56)(40)}{(8)(524) - (56)^2} = \frac{7}{11} \quad \text{or} \quad 0.636$$

Thus $Y = a_0 + a_1X$, or $Y = 0.545 + 0.636X$, as before.

(*b*) If X is considered the dependent variable, and Y the independent variable, the equation of the least-squares line is $X = b_0 + b_1Y$ and the normal equations are

$$\sum X = b_0N + b_1\sum Y$$
$$\sum XY = b_0\sum Y + b_1\sum Y^2$$

Then from Table 13.6 the normal equations become

$$8b_0 + 40b_1 = 56$$
$$40b_0 + 256b_1 = 364$$

from which $b_0 = -\frac{1}{2}$, or -0.50, and $b_1 = \frac{3}{2}$, or 1.50. These values can also be obtained from

$$b_0 = \frac{(\sum X)(\sum Y^2) - (\sum Y)(\sum XY)}{N\sum Y^2 - (\sum Y)^2} = \frac{(56)(256) - (40)(364)}{(8)(256) - (40)^2} = -0.50$$

$$b_1 = \frac{N\sum XY - (\sum X)(\sum Y)}{N\sum Y^2 - (\sum Y)^2} = \frac{(8)(364) - (56)(40)}{(8)(256) - (40)^2} = 1.50$$

Thus the required equation of the least-squares line is $X = b_0 + b_1Y$, or $X = -0.50 + 1.50Y$.

Note that by solving this equation for Y we obtain $Y = \frac{1}{3} + \frac{2}{3}X$, or $Y = 0.333 + 0.667X$, which is not the same as the line obtained in part (*a*).

13.12 For the height/weight data in Problem 13.10 use the statistical package SAS to plot the observed data points and the least-squares line on the same graph.

SOLUTION

In Fig. 13-11, the observed data values are shown as open circles, and the least-squares line is shown as a dashed line.

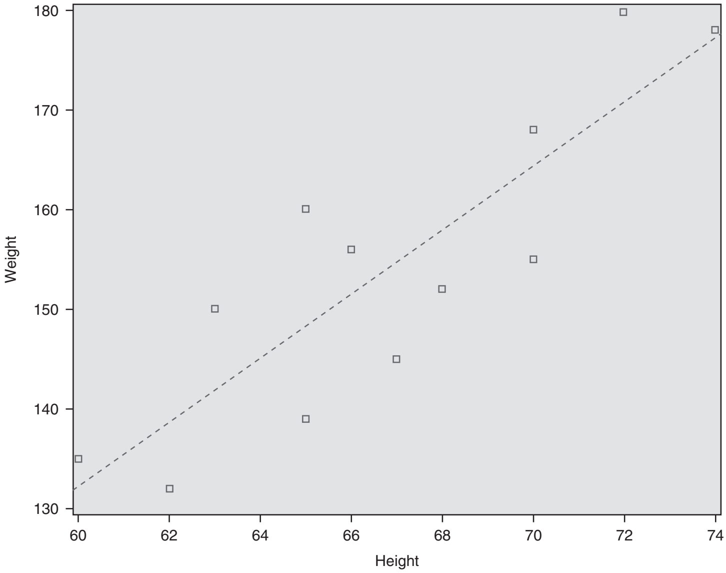


Fig. 13-11 SAS plot of the data points from Table 13.5 and the least-squares line.

13.13 (*a*) Show that the two least-squares lines obtained in Problem 13.11 intersect at point $(\bar{X}, \bar{Y})$.

(*b*) Estimate the value of Y when $X = 12$.

(*c*) Estimate the value of X when $Y = 3$.

SOLUTION

$$\bar{X} = \frac{\sum X}{N} = \frac{56}{8} = 7 \qquad \bar{Y} = \frac{\sum Y}{N} = \frac{40}{8} = 5$$

Thus point $(\bar{X}, \bar{Y})$, called the *centroid*, is (7, 5).

(*a*) Point (7, 5) lies on line $Y = 0.545 + 0.636X$; or, more exactly, $Y = \frac{6}{11} + \frac{7}{11}X$, since $5 = \frac{6}{11} + \frac{7}{11}(7)$. Point (7, 5) lies on line $X = -\frac{1}{2} + \frac{3}{2}Y$, since $7 = -\frac{1}{2} + \frac{3}{2}(5)$.

Another method

The equations of the two lines are $Y = \frac{6}{11} + \frac{7}{11}X$ and $X = -\frac{1}{2} + \frac{3}{2}Y$. Solving simultaneously, we find that $X = 7$ and $Y = 5$. Thus the lines intersect at point (7,5).

(*b*) Putting $X = 12$ into the regression line of Y (Problem 13.11), $Y = 0.545 + 0.636(12) = 8.2$.

(*c*) Putting $Y = 3$ into the regression line of X (Problem 13.11), $X = -0.50 + 1.50(3) = 4.0$.

13.14 Prove that a least-squares line always passes through the point $(\bar{X}, \bar{Y})$.

SOLUTION

Case 1 (X is the independent variable)

The equation of the least-squares line is

$$Y = a_0 + a_1 X \tag{34}$$

A normal equation for the least-squares line is

$$\sum Y = a_0 N + a_1 \sum X \tag{35}$$

Dividing both sides of equation (*35*) by N gives

$$\bar{Y} = a_0 + a_1 \bar{X} \tag{36}$$

Subtracting equation (*36*) from equation (*34*), the least-squares line can be written

$$Y - \bar{Y} = a_1(X - \bar{X}) \tag{37}$$

which shows that the line passes through the point $(\bar{X}, \bar{Y})$.

Case 2 (Y is the independent variable)

Proceeding as in Case 1, but interchanging X and Y and replacing the constants a_0 and a_1 with b_0 and b_1, respectively, we find that the least-squares line can be written

$$X - \bar{X} = b_1(Y - \bar{Y}) \tag{38}$$

which indicates that the line passes through the point $(\bar{X}, \bar{Y})$.

Note that lines (*37*) and (*38*) are not coincident, but intersect in $(\bar{X}, \bar{Y})$.

13.15 (*a*) Considering X to be the independent variable, show that the equation of the least-squares line can be written

$$y = \left(\frac{\sum xy}{\sum x^2}\right)x \qquad \text{or} \qquad y = \left(\frac{\sum xY}{\sum x^2}\right)x$$

where $x = X - \bar{X}$ and $y = Y - \bar{Y}$.

(*b*) If $\bar{X} = 0$, show that the least-squares line in part (*a*) can be written

$$Y = \bar{Y} + \left(\frac{\sum XY}{\sum X^2}\right)X$$

(*c*) Write the equation of the least-squares line corresponding to that in part (*a*) if Y is the independent variable.

(*d*) Verify that the lines in parts (*a*) and (*c*) are not necessarily the same.

SOLUTION

(*a*) Equation (*37*) can be written $y = a_1 x$, where $x = X - \bar{X}$ and $y = Y - \bar{Y}$. Also, from the simultaneous solution of the normal equations (*18*) we have

$$\begin{aligned} a_1 &= \frac{N\sum XY - (\sum X)(\sum Y)}{N\sum X^2 - (\sum X)^2} = \frac{N\sum(x+\bar{X})(y+\bar{Y}) - [\sum(x+\bar{X}][\sum(y+\bar{Y})]}{N\sum(x+\bar{X})^2 - [\sum(x+\bar{X})]^2} \\ &= \frac{N\sum(xy + x\bar{Y} + \bar{X}y + \bar{X}\bar{Y}) - (\sum x + N\bar{X})(\sum y + N\bar{Y})}{N\sum(x^2 + 2x\bar{X} + \bar{X}^2) - (\sum x + N\bar{X})^2} \\ &= \frac{N\sum xy + N\bar{Y}\sum x + N\bar{X}\sum y + N^2\bar{X}\bar{Y} - (\sum x + N\bar{X})(\sum y + N\bar{Y})}{N\sum x^2 + 2N\bar{X}\sum x + N^2\bar{X}^2 - (\sum x + N\bar{X})^2} \end{aligned}$$

But $\sum x = \sum(X - \bar{X}) = 0$ and $\sum y = \sum(Y - \bar{Y}) = 0$; hence the above simplifies to

$$a_1 = \frac{N\sum xy + N^2\bar{X}\bar{Y} - N^2\bar{X}\bar{Y}}{N\sum x^2 + N^2\bar{X}^2 - N^2\bar{X}^2} = \frac{\sum xy}{\sum x^2}$$

This can also be written

$$a_1 = \frac{\sum xy}{\sum x^2} = \frac{\sum x(Y - \bar{Y})}{\sum x^2} = \frac{\sum xY - \bar{Y}\sum x}{\sum x^2} = \frac{\sum xY}{\sum x^2}$$

Thus the least-squares line is $y = a_1 x$; that is,

$$y = \left(\frac{\sum xy}{\sum x^2}\right)x \qquad \text{or} \qquad y = \left(\frac{\sum xY}{\sum x^2}\right)x$$

(*b*) If $\bar{X} = 0$, $x = X - \bar{X} = X$. Then from

$$y = \left(\frac{\sum xY}{\sum x^2}\right)$$

we have

$$y = \left(\frac{\sum XY}{\sum X^2}\right)X \qquad \text{or} \qquad Y = \bar{Y} + \left(\frac{\sum XY}{\sum X^2}\right)X$$

Another method

The normal equations of the least-squares line $Y = a_0 + a_1 X$ are

$$\sum Y = a_0 N + a_1 \sum X \qquad \text{and} \qquad \sum XY = a_0 \sum X + a_1 \sum X^2$$

If $\bar{X} = (\sum X)/N = 0$, then $\sum X = 0$ and the normal equations become

$$\sum Y = a_0 N \qquad \text{and} \qquad \sum XY = a_1 \sum X^2$$

from which

$$a_0 = \frac{\sum Y}{N} = \bar{Y} \qquad \text{and} \qquad a_1 = \frac{\sum XY}{\sum X^2}$$

Thus the required equation of the least-squares line is

$$Y = a_0 + a_1 X \qquad \text{or} \qquad Y = \bar{Y} + \left(\frac{\sum XY}{\sum X^2}\right)X$$

(*c*) By interchanging X and Y or x and y, we can show as in part (*a*) that

$$x = \left(\frac{\sum xy}{\sum y^2} \right) y$$

(*d*) From part (*a*), the least-squares line is

$$y = \left(\frac{\sum xy}{\sum x^2} \right) x \tag{39}$$

From part (*c*), the least-squares line is

$$x = \left(\frac{\sum xy}{\sum y^2} \right) y$$

or

$$y = \left(\frac{\sum y^2}{\sum xy} \right) x \tag{40}$$

Since in general

$$\frac{\sum xy}{\sum x^2} \neq \frac{\sum y^2}{\sum xy}$$

the least-squares lines (*39*) and (*40*) are different in general. Note, however, that they intersect at $x = 0$ and $y = 0$ [i.e., at the point $(\bar{X}, \bar{Y})$].

13.16 If $X' = X + A$ and $Y' = Y + B$, where A and B are any constants, prove that

$$a_1 = \frac{N \sum XY - (\sum X)(\sum Y)}{N \sum X^2 - (\sum X)^2} = \frac{N \sum X'Y' - (\sum X')(\sum Y')}{N \sum X'^2 - (\sum X')^2} = a_1'$$

SOLUTION

$$x' = X' - \bar{X}' = (X + A) - (\bar{X} + A) = X - \bar{X} = x$$
$$y' = Y' - \bar{Y}' = (Y + B) - (\bar{Y} + B) = Y - \bar{Y} = y$$

Then

$$\frac{\sum xy}{\sum x^2} = \frac{\sum x'y'}{\sum x'^2}$$

and the result follows from Problem 13.15. A similar result holds for b_1.

This result is useful, since it enables us to simplify calculations in obtaining the regression line by subtracting suitable constants from the variables X and Y (see the second method of Problem 13.17).

Note: The result does not hold if $X' = c_1 X + A$ and $Y' = c_2 Y + B$ unless $c_1 = c_2$.

13.17 Fit a least-squares line to the data of Problem 13.10 by using (*a*) X as the independent variable and (*b*) X as the dependent variable.

SOLUTION

First method

(*a*) From Problem 13.15(*a*) the required line is

$$y = \left(\frac{\sum xy}{\sum x^2} \right) x$$

where $x = X - \bar{X}$ and $y = Y - \bar{Y}$. The work involved in computing the sums can be arranged as in Table 13.7. From the first two columns we find $\bar{X} = 802/12 = 66.8$ and $\bar{Y} = 1850/12 = 154.2$. The last column has been added for use in part (*b*).

Table 13.7

Height X	Weight Y	$x = X - \bar{X}$	$y = Y - \bar{Y}$	xy	x^2	y^2
70	155	3.2	0.8	2.56	10.24	0.64
63	150	−3.8	−4.2	15.96	14.44	17.64
72	180	5.2	25.8	134.16	27.04	665.64
60	135	−6.8	−19.2	130.56	46.24	368.64
66	156	−0.8	1.8	−1.44	0.64	3.24
70	168	3.2	13.8	44.16	10.24	190.44
74	178	7.2	23.8	171.36	51.84	566.44
65	160	−1.8	5.8	−10.44	3.24	33.64
62	132	−4.8	−22.2	106.56	23.04	492.84
67	145	0.2	−9.2	−1.84	0.04	84.64
65	139	−1.8	−15.2	27.36	3.24	231.04
68	152	1.2	−2.2	−2.64	1.44	4.84
$\sum X = 802$ $\bar{X} = 66.8$	$\sum Y = 1850$ $\bar{Y} = 154.2$			$\sum xy = 616.32$	$\sum x^2 = 191.68$	$\sum y^2 = 2659.68$

The required least-squares line is

$$y = \left(\frac{\sum xy}{\sum x^2}\right)x = \frac{616.32}{191.68}\,x = 3.22x$$

or $Y - 154.2 = 3.22(X - 66.8)$, which can be written $Y = 3.22X - 60.9$. This equation is called the *regression line of Y on X* and is used for estimating Y from given values of X.

(*b*) If X is the dependent variable, the required line is

$$x = \left(\frac{\sum xy}{\sum y^2}\right)y = \frac{616.32}{2659.68}\,y = 0.232y$$

which can be written $X - 66.8 = 0.232(Y - 154.2)$, or $X = 31.0 + 0.232Y$. This equation is called the *regression line of X on Y* and is used for estimating X from given values of Y.

Note that the method of Problem 13.11 can also be used if desired.

Second method

Using the result of Problem 13.16, we may subtract suitable constants from X and Y. We choose to subtract 65 from X and 150 from Y. Then the results can be arranged as in Table 13.7.

$$a_1 = \frac{N\sum X'Y' - (\sum X')(\sum Y')}{N\sum X'^2 - (\sum X')^2} = \frac{(12)(708) - (22)(50)}{(12)(232) - (22)^2} = 3.22$$

$$b_1 = \frac{N\sum X'Y' - (\sum Y')(\sum X')}{N\sum Y'^2 - (\sum Y')^2} = \frac{(12)(708) - (50)(22)}{(12)(2868) - (50)^2} = 0.232$$

Since $\bar{X} = 65 + 22/12 = 66.8$ and $\bar{Y} = 150 + 50/12 = 154.2$, the regression equations are $Y - 154.2 = 3.22(X - 66.8)$ and $X - 66.8 = 0.232(Y - 154.2)$; that is $Y = 3.22X - 60.9$ and $X = 0.232Y + 31.0$, in agreement with the first method.

13.18 Work Problem 13.17 using MINITAB. Plot the regression line of weight on height and the regression line of height on weight on the same set of axes. Show that the point $(\bar{X}, \bar{Y})$ satisfies both equations. The lines therefore intersect at $(\bar{X}, \bar{Y})$.

SOLUTION

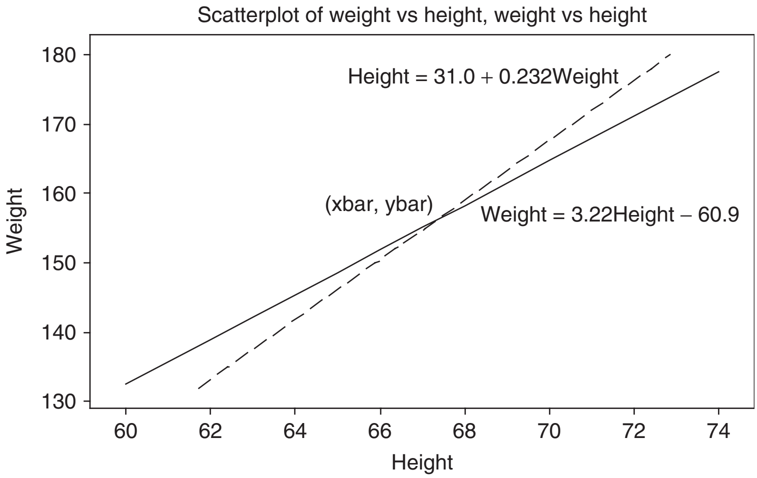


Fig. 13-12 The regression line of weight on height and the regression line of height on weight both pass through the point ($x$bar, $y$bar).

$(\bar{X}, \bar{Y})$ is the same as (`xbar, ybar`) and equals (66.83, 154.17). Note that weight = 3.22(**66.83**) − 60.9 = **154.17** and height = 31.0 + 0.232(**154.17**) = **66.83**. Therefore, both lines pass through (`xbar, ybar`).

Table 13.8

X'	Y'	X'^2	$X'Y'$	Y'^2
5	5	25	25	25
−2	0	4	0	0
7	30	49	210	900
−5	−15	25	75	225
1	6	1	6	36
5	18	25	90	324
9	28	81	252	784
0	10	0	0	100
−3	−18	9	54	324
2	−5	4	−10	25
0	−11	0	0	121
3	2	9	6	4
$\sum X' = 22$	$\sum Y' = 50$	$\sum X'^2 = 232$	$\sum X'Y' = 708$	$\sum Y'^2 = 2868$

APPLICATIONS TO TIME SERIES

13.19 The total agricultural exports in millions of dollars are given in Table 13.9. Use MINITAB to do the following.

Table 13.9

Year	2000	2001	2002	2003	2004	2005
Total value	51246	53659	53115	59364	61383	62958
Coded year	1	2	3	4	5	6

Source: The 2007 Statistical Abstract.

(*a*) Graph the data and show the least-squares regression line.

(*b*) Find and plot the trend line for the data.

(*c*) Give the *fitted values* and the *residuals* using the coded values for the years.

(*d*) Estimate the value of total agricultural exports in the year 2006.

SOLUTION

(*a*) The data and the regression line are shown in Fig. 13-13 (a). The pull-down **Stat** → **Regression** → **Fitted line plot** gives the graph shown in Fig. 13-13 (a).

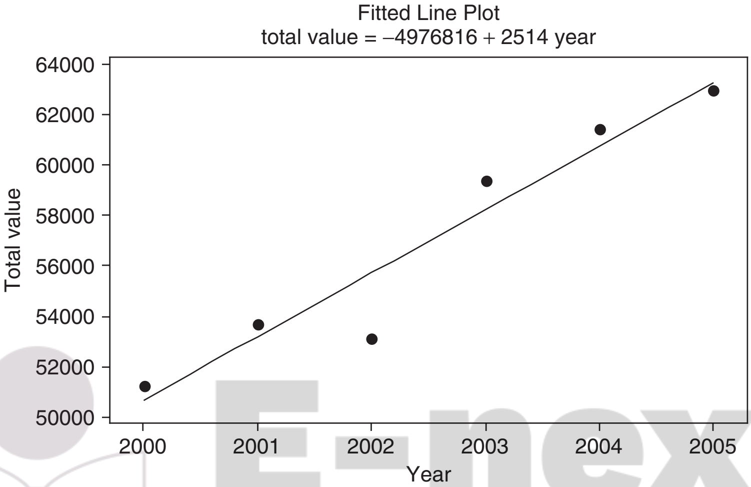


Fig. 13-13 (a) The regression line for total agricultural exports in millions of dollars.

(*b*) The pull-down **Stat** → **Time series** → **Trend Analysis** gives the plot shown in Fig. 13-13 (b). It is a different way of looking at the same data. It may be a little easier to work with index numbers rather than years.

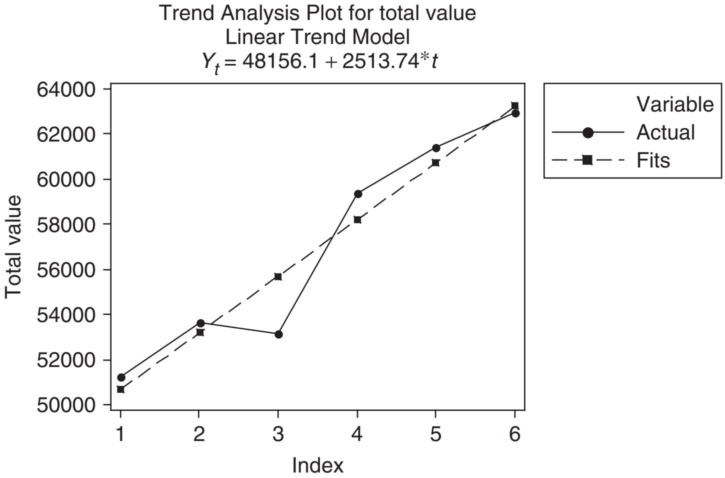


Fig. 13-13 (b) The trend line for total agricultural exports in millions of dollars.

(*c*) Table 13.10 gives the fitted values and the residuals for the data in Table 13.9 using coded values for the years.

Table 13.10

Year coded	Total value	Fitted value	Residual
1	51246	50669.8	576.19
2	53659	53183.6	475.45
3	53115	55697.3	−2582.30
4	59364	58211.0	1152.96
5	61383	60724.8	658.22
6	62958	63238.5	−280.52

(*d*) Using the coded value, the estimated value is $Y_t = 48156.1 + 2513.74(7) = 65752.3$.

13.20 Table 13.11 gives the purchasing power of the dollar as measured by consumer prices according to the U.S. Bureau of Labor Statistics, Survey of Current Business.

Table 13.11

Year	2000	2001	2002	2003	2004	2005
Consumer prices	0.581	0.565	0.556	0.544	0.530	0.512

Source: U.S. Bureau of Labor Statistics, Survey of Current Business.

(*a*) Graph the data and obtain the trend line using MINITAB.

(*b*) Find the equation of the trend line by hand.

(*c*) Estimate the consumer price in 2008 assuming the trend continues for 3 more years.

SOLUTION

(*a*) The solid line in Fig. 13-14 shows a plot of the data in Table 13.11 and the dashed line shows the graph of the least-squares line.

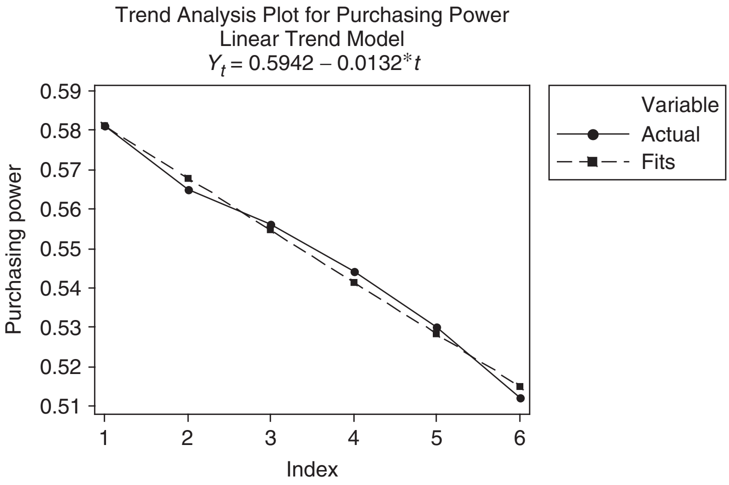


Fig. 13-14 Trend line for purchasing power.

(*b*) The computations for finding the trend line by hand are shown in Table 13.12. The equation is

$$y = \frac{\sum xy}{\sum x^2} x$$

where $x = X - \bar{X}$ and $y = Y - \bar{Y}$, which can be written as $Y - 0.548 = -0.0132(X - 3.5)$ or $Y = -0.0132X + 0.5942$. The computational work saved by statistical software is tremendous as illustrated by this problem.

Table 13.12

Year	X	Y	$x = X - \bar{X}$	$y = Y - \bar{Y}$	x^2	xy
2000	1	0.581	−2.5	0.033	6.25	−0.0825
2001	2	0.565	−1.5	0.017	2.25	−0.0255
2002	3	0.556	−0.5	0.008	0.25	−0.004
2003	4	0.544	0.5	−0.004	0.25	−0.002
2004	5	0.530	1.5	−0.018	2.25	−0.027
2005	6	0.512	2.5	−0.036	6.25	−0.09
	$\Sigma X = 21$	$\Sigma Y = 3.288$			Σx^2	Σxy
	$\bar{X} = 3.5$	$\bar{Y} = 0.548$			17.5	−0.231

(*c*) The estimation for 2008 is obtained by substituting $t = 9$ in the trend line equation. Estimated consumer price is $0.5942 - 0.0132(9) = 0.475$.

NONLINEAR EQUATIONS REDUCIBLE TO LINEAR FORM

13.21 Table 13.13 gives experimental values of the pressure P of a given mass of gas corresponding to various values of the volume V. According to thermodynamic principles, a relationship having the form $PV^{\gamma} = C$, where γ and C are constants, should exist between the variables.

(*a*) Find the values of γ and C.

(*b*) Write the equation connecting P and V.

Table 13.13

Volume V in cubic inches (in^3)	54.3	61.8	72.4	88.7	118.6	194.0
Pressure P in pounds per square inch (lb/in^2)	61.2	49.2	37.6	28.4	19.2	10.1

(*c*) Estimate P when $V = 100.0\,\text{in}^3$.

SOLUTION

Since $PV^{\gamma} = C$, we have

$$\log P + \gamma \log V = \log C \qquad \text{or} \qquad \log P = \log C - \gamma \log V$$

Calling $\log V = X$ and $\log P = Y$, the last equation can be written

$$Y = a_0 + a_1 X \tag{41}$$

where $a_0 = \log C$ and $a_1 = -\gamma$.

Table 13.14 gives $X = \log V$ and $Y = \log P$, corresponding to the values of V and P in Table 13.13, and also indicates the calculations involved in computing the least-squares line (*41*). The normal equations corresponding to the least-squares line (*41*) are

$$\sum Y = a_0 N + a_1 \sum X \qquad \text{and} \qquad \sum XY = a_0 \sum X + a_1 \sum X^2$$

from which

$$a_0 = \frac{(\sum Y)(\sum X^2) - (\sum X)(\sum XY)}{N \sum X^2 - (\sum X)^2} = 4.20 \qquad a_1 = \frac{N \sum XY - (\sum X)(\sum Y)}{N \sum X^2 - (\sum X)^2} = -1.40$$

Thus $Y = 4.20 - 1.40X$.

(*a*) Since $a_0 = 4.20 = \log C$ and $a_1 = -1.40 = -\gamma$, $C = 1.60 \times 10^4$ and $\gamma = 1.40$.

(*b*) The required equation in terms of P and V can be written $PV^{1.40} = 16{,}000$.

(*c*) When $V = 100$, $X = \log V = 2$ and $Y = \log P = 4.20 - 1.40(2) = 1.40$. Then $P = \text{antilog}\, 1.40 = 25.1\ \text{lb/in}^2$.

Table 13.14

$X = \log V$	$Y = \log P$	X^2	XY
1.7348	1.7868	3.0095	3.0997
1.7910	1.6946	3.2077	3.0350
1.8597	1.5752	3.4585	2.9294
1.9479	1.4533	3.7943	2.8309
2.0741	1.2833	4.3019	2.6617
2.2878	1.0043	5.2340	2.2976
$\sum X = 11.6953$	$\sum Y = 8.7975$	$\sum X^2 = 23.0059$	$\sum XY = 16.8543$

13.22 Use MINITAB to assist in the solution of Problem 13.21.

SOLUTION

The transformations $X = \log_t(V)$ and $Y = \log_t(P)$ converts the problem to a linear fit problem. The calculator in MINITAB is used to take common logarithms of volume and pressure. The following are in columns `C1` through `C4` of the MINITAB worksheet.

V	*P*	Log10*V*	Log10*P*
54.3	61.2	1.73480	1.78675
61.8	49.2	1.79099	1.69197
72.4	37.6	1.85974	1.57519
88.7	28.4	1.94792	1.45332
118.6	19.2	2.07408	1.28330
194.0	10.1	2.28780	1.00432

The least-squares fit gives the following: $\log_{10}(P) = 4.199 - 1.402 \log_{10}(V)$. See Fig. 13-15. $a_0 = \log C$ and $a_1 = -\gamma$. Taking antilogs gives $C = 10^{a_0}$ and $\gamma = -a_1$ or $C = 15812$ and $\gamma = 1.402$. The nonlinear equation is $PV^{1.402} = 15812$.

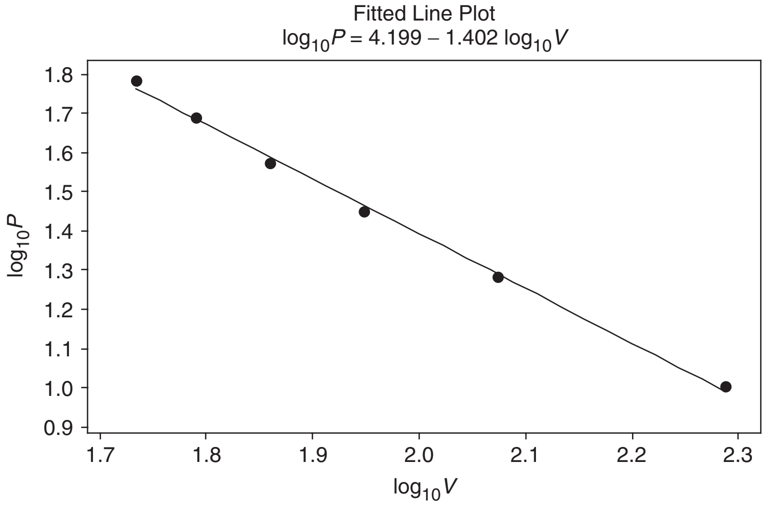


Fig. 13-15 Reducing nonlinear equations to linear form.

13.23 Table 13.15 gives the population of the United States at 5-year intervals from 1960 to 2005 in millions. Fit a straight line as well as a parabola to the data and comment on the two fits. Use both models to predict the United States population in 2010.

Table 13.15

Year	1960	1965	1970	1975	1980	1985	1990	1995	2000	2005
Population	181	194	205	216	228	238	250	267	282	297

Source: U.S. Bureau of Census.

SOLUTION

A partial printout of the MINITAB solution for the least-squares line and least-squares parabola is given below.

```
Year   Population    x     xsquare

1960      181        1         1
1965      194        2         4
1970      205        3         9
1975      216        4        16
1980      228        5        25
1985      238        6        36
1990      250        7        49
1995      267        8        64
2000      282        9        81
2005      297       10       100
```

The straight line model is as follows:
The regression equation is

Population = 166 + 12.6 *x*

The quadratic model is as follows:
The regression equation is

Population = 174 + 9.3 *x* - 0.326 x^2

Table 13.16 gives the fitted values and residuals for the straight line fit to the data.

Table 13.16

Year	Population	Fitted value	Residual
1960	181	179.018	1.98182
1965	194	191.636	2.36364
1970	205	204.255	0.74545
1975	216	216.873	−0.87273
1980	228	229.491	−1.49091
1985	238	242.109	−4.10909
1990	250	254.727	−4.72727
1995	267	267.345	−0.34545
2000	282	279.964	2.03636
2005	297	292.582	4.41818

Table 13.17 gives the fitted values and residuals for the parabolic fit to the data. The sum of squares of the residuals for the straight line is 76.073 and the sum of squares of the residuals for the parabola is 20.042. It appears that, overall, the parabola fits the data better than the straight line.

Table 13.17

Year	Population	Fitted value	Residual
1960	181	182.927	−1.92727
1965	194	192.939	1.06061
1970	205	203.603	1.39697
1975	216	214.918	1.08182
1980	228	226.885	1.11515
1985	238	239.503	−1.50303
1990	250	252.773	−2.77273
1995	267	266.694	0.30606
2000	282	281.267	0.73333
2005	297	296.491	0.50909

To predict the population in the year 2010, note that the coded value for 2010 is 11. The straight line predicted value is population $= 166 + 12.6x = 166 + 138.6 = 304.6$ million and the parabola model predicts population $= 174 + 9.03x + 0.326x^2 = 174 + 99.33 + 39.446 = 312.776$.

Supplementary Problems

STRAIGHT LINES

13.24 If $3X + 2Y = 18$, find (*a*) X when $Y = 3$, (*b*) Y when $X = 2$, (*c*) X when $Y = -5$, (*d*) Y when $X = -1$, (*e*) the X intercept, and (*f*) the Y intercept.

13.25 Construct a graph of the equations (*a*) $Y = 3X - 5$ and (*b*) $X + 2Y = 4$ on the same set of axes. In what point do the graphs intersect?

13.26 (*a*) Find an equation for the straight line passing through the points $(3, -2)$ and $(-1, 6)$.

(*b*) Determine the X and Y intercepts of the line in part (*a*).

(*c*) Find the value of Y corresponding to $X = 3$ and to $X = 5$.

(*d*) Verify your answers to parts (*a*), (*b*), and (*c*) directly from a graph.

13.27 Find an equation for the straight line whose slope is $\frac{2}{3}$ and whose Y intercept is -3.

13.28 (*a*) Find the slope and Y intercept of the line whose equation is $3X - 5Y = 20$.

(*b*) What is the equation of a line which is parallel to the line in part (*a*) and which passes through the point $(2, -1)$?

13.29 Find (*a*) the slope, (*b*) the Y intercept, and (*c*) the equation of the line passing through the points $(5, 4)$ and $(2, 8)$.

13.30 Find the equation of a straight line whose X and Y intercepts are 3 and -5, respectively.

13.31 A temperature of 100 degrees Celsius (°C) corresponds to 212 degrees Fahrenheit (°F), while a temperature of 0°C corresponds to 32 °F. Assuming that a linear relationship exists between Celsius and Fahrenheit temperatures, find (*a*) the equation connecting Celsius and Fahrenheit temperatures, (*b*) the Fahrenheit temperature corresponding to 80 °C, and (*c*) the Celsius temperature corresponding to 68 °F.

THE LEAST-SQUARES LINE

13.32 Fit a least-squares line to the data in Table 13.18, using (*a*) X as the independent variable and (*b*) X as the dependent variable. Graph the data and the least-squares lines, using the same set of coordinate axes.

Table 13.18

X	3	5	6	8	9	11
Y	2	3	4	6	5	8

13.33 For the data of Problem 13.32, find (*a*) the values of Y when $X = 5$ and $X = 12$ and (*b*) the value of X when $Y = 7$.

13.34 (*a*) Use the freehand method to obtain an equation for a line fitting the data of Problem 13.32.
(*b*) Using the result of part (*a*), answer Problem 13.33.

13.35 Table 13.19 shows the final grades in algebra and physics obtained by 10 students selected at random from a large group of students.

(*a*) Graph the data.
(*b*) Find a least-squares line fitting the data, using X as the independent variable.

(*c*) Find a least-squares line fitting the data, using Y as the independent variable.

(*d*) If a student receives a grade of 75 in algebra, what is her expected grade in physics?

(*e*) If a student receives a grade of 95 in physics, what is her expected grade in algebra?

Table 13.19

Algebra (X)	75	80	93	65	87	71	98	68	84	77
Physics (Y)	82	78	86	72	91	80	95	72	89	74

13.36 Table 13.20 shows the birth rate per 1000 population during the years 1998 through 2004.

(*a*) Graph the data.

(*b*) Find the least-squares line fitting the data. Code the years 1998 through 2004 as the whole numbers 1 through 7.

(*c*) Compute the trend values (fitted values) and the residuals.

(*d*) Predict the birth rate in 2010, assuming the present trend continues.

Table 13.20

Year	1998	1999	2000	2001	2002	2003	2004
Birth rate per 1000	14.3	14.2	14.4	14.1	13.9	14.1	14.0

Source: U.S. National Center for Health Statistics, Vital Statistics of the United States, annual; National Vital Statistics Reports and unpublished data.

13.37 Table 13.21 shows the number in thousands of the United States population 85 years and over for the years 1999 through 2005.

(*a*) Graph the data.

(*b*) Find the least-squares line fitting the data. Code the years 1999 through 2005 as the whole numbers 1 through 7.

(*c*) Compute the trend values (fitted values) and the residuals.

(*d*) Predict the number of individuals 85 years and older in 2010, assuming the present trend continues.

Table 13.21

Year	1999	2000	2001	2002	2003	2004	2005
85 and over	4154	4240	4418	4547	4716	4867	5096

Source: U.S. Bureau of Census.

LEAST-SQUARES CURVES

13.38 Fit a least-squares parabola, $Y = a_0 + a_1X + a_2X^2$, to the data in Table 13.22.

Table 13.22

X	0	1	2	3	4	5	6
Y	2.4	2.1	3.2	5.6	9.3	14.6	21.9

13.39 The total time required to bring an automobile to a stop after one perceives danger is the reaction time (the time between recognizing danger and applying the brakes) plus the braking time (the time for stopping after applying the brakes). Table 13.23 gives the stopping distance D (in feet, of ft) of an automobile traveling at speeds V (in miles per hour, or mi/h) from the instant that danger is perceived.

(*a*) Graph D against V.

(*b*) Fit a least-squares parabola of the form $D = a_0 + a_1 V + a_2 V^2$ to the data.

(*c*) Estimate D when $V = 45$ mi/h and 80 mi/h.

Table 13.23

Speed V (mi/h)	20	30	40	50	60	70
Stopping distance D (ft)	54	90	138	206	292	396

13.40 Table 13.24 shows the male and female populations of the United States during the years 1940 through 2005 in millions. It also shows the years coded and the differences which equals male minus female.

(*a*) Graph the data points and the linear least-squares best fit.

(*b*) Graph the data points and the quadratic least-squares best fit.

(*c*) Graph the data points and the cubic least-squares best fit.

(*d*) Give the fitted values and the residuals using each of the three models. Give the sum of squares for residuals for all three models.

(*e*) Use each of the three models to predict the difference in 2010.

Table 13.24

Year	1940	1950	1960	1970	1980	1990	2000	2005
Coded	0	1	2	3	4	5	6	6.5
Male	66.1	75.2	88.3	98.9	110.1	121.2	138.1	146.0
Female	65.6	76.1	91.0	104.3	116.5	127.5	143.4	150.4
Difference	0.5	−0.9	−2.7	−5.4	−6.4	−6.3	−5.3	−4.4

Source: U.S. Bureau of Census.

13.41 Work Problem 13.40 using the ratio of females to males instead of differences.

13.42 Work Problem 13.40 by fitting a least squares parabola to the differences.

13.43 The number Y of bacteria per unit volume present in a culture after X hours is given in Table 13.25.

Table 13.25

Number of hours (X)	0	1	2	3	4	5	6
Number of bacteria per unit volume (Y)	32	47	65	92	132	190	275

(*a*) Graph the data on semilog graph paper, using the logarithmic scale for Y and the arithmetic scale for X.

(*b*) Fit a least-squares curve of the form $Y = ab^x$ to the data and explain why this particular equation should yield good results.

(*c*) Compare the values of Y obtained from this equation with the actual values.

(*d*) Estimate the value of Y when $X = 7$.

13.44 In Problem 13.43, show how a graph on semilog graph paper can be used to obtain the required equation without employing the method of least squares.