2.8 Linear Regression

Business, industry, and governments are interested in answers to such questions as, "How many cell phones will be needed in the next three years?" "What will be the population growth of Boulder, Colorado, over the next five years?" "Will the Midway School District need to build an additional elementary school within three years?" "How much will the number of 18-wheelers on I-35 increase over the next decade?"

Answers to these kinds of questions are important because it may take years to build the manufacturing plants, schools, or highways that will be needed. In some cases, information from past years can indicate a trend from which reasonable estimates of future growth can be obtained. For example, the problem of air quality has been addressed by cities in recent years. The cities of Los Angeles and Long Beach, California have made significant progress, as shown by the following table, which gives the number of days during the year that the Air Quality Index exceeded 100. As the index exceeds 100, the environment becomes unhealthy for sensitive groups.

Year	Days
1994	139
1995	113
1996	94
1997	60
1998	56
1999	27

When we plot this information using the year (1 for 1994, 2 for 1995, and so on) for the x-value and the number of days for the y-value, we get the graph of Figure 2–12.

We call a graph of points such as this a **scatter plot.** Notice that the points have a general, approximate linear downward trend, even though the points do not lie on a line. Even when the pattern of data points is not exactly linear, it may be useful to approximate their trend with a line so that we can estimate future behavior.

There may be cases when we believe the variables are related in a linear manner, but the data deviate from a line because (1) the data collected may not be accurate or (2) the assumption of a linear relationship is not valid. In either case, it may be useful to find the line that best approximates the trend and use it to obtain additional information and make predictions.

Figure 2–13 shows a line that approximates the trend of the points in Figure 2–12.

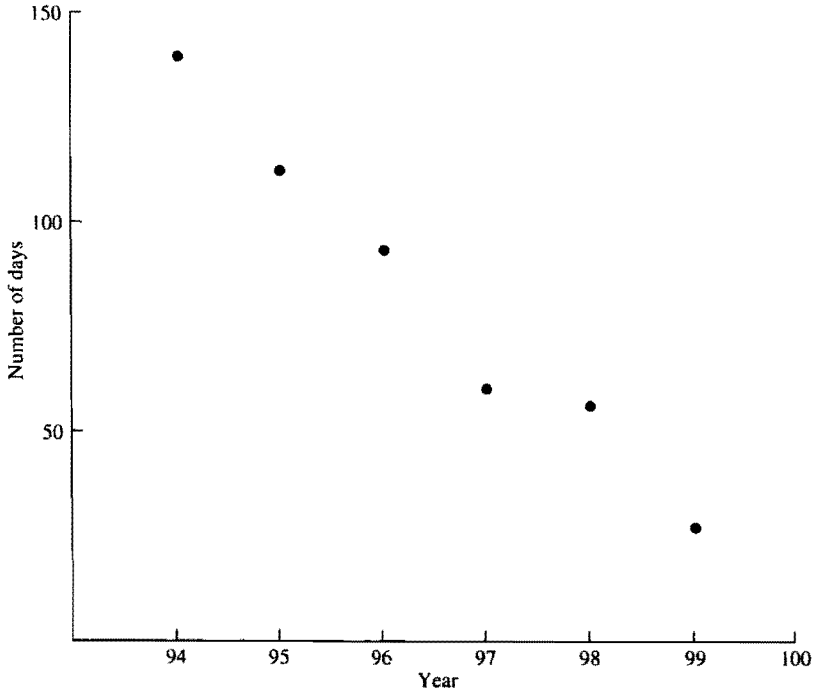


FIGURE 2–12

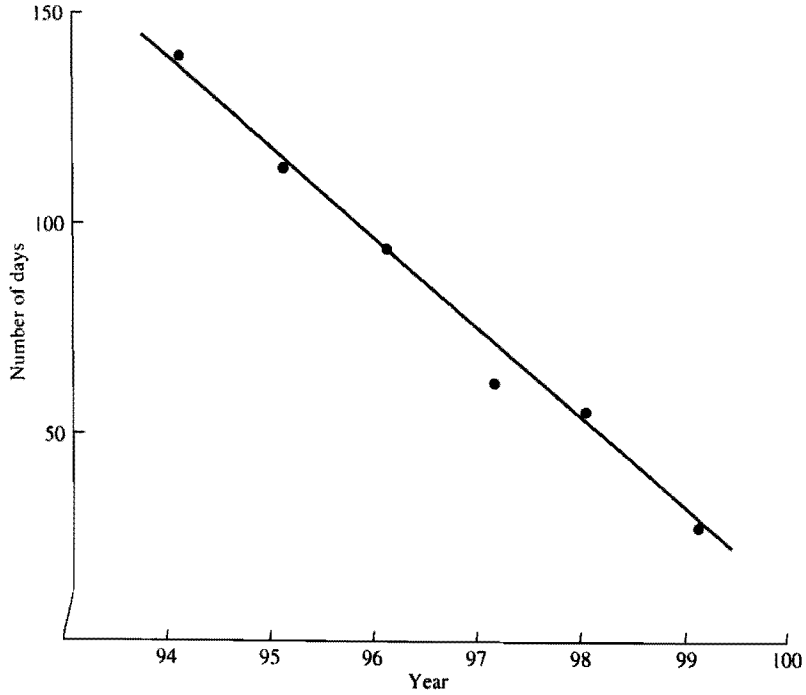


FIGURE 2–13

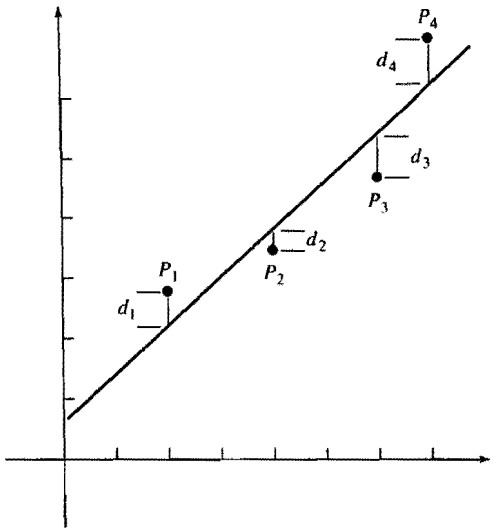


FIGURE 2–14

Although the line seems to be a reasonable representation of the general trend, we would like to know if this is the *best* approximation of the trend. Mathematicians use the **least squares line,** also called the **regression line,** for the line that best fits the data. In Figure 2–13, the line shown is the least squares, or regression, line $y = -21.9x + 158$. You have not been told how to find that equation. Before doing so, let's look at the idea behind the least squares procedure.

In Figure 2–14, we look at a simple case where we have drawn a line in the general direction of the trend of the scatter plot of the four points P_1, P_2, P_3, and P_4. For each of the points, we have indicated the vertical distance from each point to the line and labeled the distances d_1, d_2, d_3, and d_4. The distances to points above the line will be positive, and the distances to the points below the line will be negative. The basic idea of the least squares procedure is to find a line that somehow minimizes the entirety of these distances. To do so, find the line that gives the smallest possible sum of the *squares* of the d's — that is, the line $y = mx + b$ that makes

$$d_1^2 + d_2^2 + d_3^2 + d_4^2$$

the least value possible.

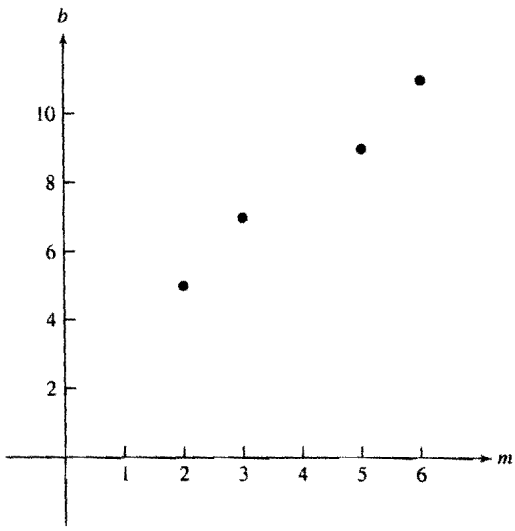
EXAMPLE 1

We find the values of m and b of the regression line $y = mx + b$ from a system of two equations in the variables m and b. Let's illustrate the procedure using the points (2, 5), (3, 7), (5, 9), and (6, 11). The scatter plot is shown in Figure 2–15.

The system takes the form

$$Am + Bb = C$$
$$Dm + Eb = F$$

FIGURE 2–15



where

A = the sum of the squares of the x-coordinates; in this case, $2^2 + 3^2 + 5^2 + 6^2 = 74$.

B = the sum of the x-coordinates of the given points; in this case, $2 + 3 + 5 + 6 = 16$.

C = the sum of the products of the x- and y-coordinates of the given points; in this case, $(2 \times 5) + (3 \times 7) + (5 \times 9) + (6 \times 11) = 142$.

D = B, the sum of the x-coordinates, 16.

E = the number of given points, in this case, four points.

F = the sum of the y-coordinates; in this case, $5 + 7 + 9 + 11 = 32$.

Thus, the solution to the system

$$\begin{aligned} 74m + 16b &= 142 \\ 16m + 4b &= 32 \end{aligned}$$

gives the coefficients of the regression line that best fits the four given points. The solution is $m = 1.4$ and $b = 2.4$ (be sure you can find the solution), and the linear regression line is $y = 1.4x + 2.4$ (see Figure 2–16).

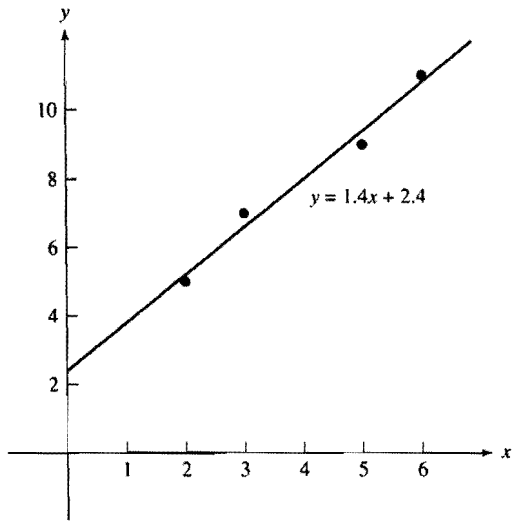
Least Squares Line

The linear function $y = mx + b$ is the **least squares line** for the points (x_1, y_1), $(x_2, y_2), \ldots, (x_n, y_n)$ when m and b are solutions of the system of equations.

$$\begin{aligned} (x_1^2 + x_2^2 + \cdots + x_n^2)m + (x_1 + x_2 + \cdots + x_n)b &= (x_1y_1 + x_2y_2 + \cdots + x_ny_n) \\ (x_1 + x_2 + \cdots + x_n)m + nb &= (y_1 + y_2 + \cdots + y_n) \end{aligned}$$

We now show you two ways to organize the data that makes it easier to keep track of the computations needed to find the coefficients of the system.

FIGURE 2-16



Method I. Form a table like the following.

	x	y	x^2	xy
	2	5	4	10
	3	7	9	21
	5	9	25	45
	6	11	36	66
Sum	16	32	74	142

This gives the system

$$\begin{aligned} 74m + 16b &= 142 \\ 16m + 4b &= 32 \end{aligned}$$

Method II. We find the augmented matrix A of the system of equations, with a matrix product. For this example, it is

$$A = \begin{bmatrix} 2 & 3 & 5 & 6 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 & 5 \\ 3 & 1 & 7 \\ 5 & 1 & 9 \\ 6 & 1 & 11 \end{bmatrix} = \begin{bmatrix} 74 & 16 & 142 \\ 16 & 4 & 32 \end{bmatrix}$$

We state the general case as follows: Given the points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, the augmented matrix M of the system

$$\begin{aligned} Am + Bb &= C \\ Dm + Eb &= F \end{aligned}$$

whose solution gives the least squares line of best fit for the given points is the product

$$M = \begin{bmatrix} x_1 & x_2 \cdots x_n \\ 1 & 1 \cdots 1 \end{bmatrix} \begin{bmatrix} x_1 & 1 & y_1 \\ x_2 & 1 & y_2 \\ \vdots & \vdots & \vdots \\ x_n & 1 & y_n \end{bmatrix}$$

This method is useful when using a calculator with matrix operations. ■

Let's use the data on the number of computers in use in the United States and work an example.

EXAMPLE 2 ➤ The computer industry estimated that the number of computers in use in the United States for the years 1991 through 1995 was the following:

Year	Computers in Use (millions)
1991	62.0
1992	68.2
1993	76.5
1994	85.8
1995	96.2

Find the scatter plot and least squares line $y = mx + b$ where x is the year ($x = 1$ for 1991, $x = 2$ for 1992, etc.) and y is the number of computers in use.

SOLUTION

Figure 2–17 shows the scatter plot.

We will find the coefficients to the system of equations, using both methods illustrated.

Method I.

	x	y	x^2	xy
	1	62.0	1	62.0
	2	68.2	4	136.4
	3	76.5	9	229.5
	4	85.8	16	343.2
	5	96.2	25	481.0
Sum	15	388.7	55	1252.1

FIGURE 2–17

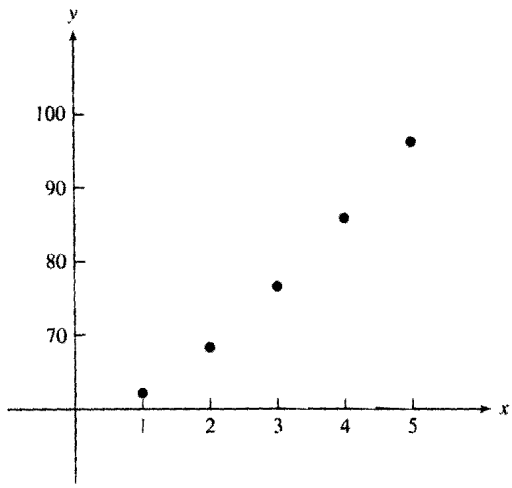
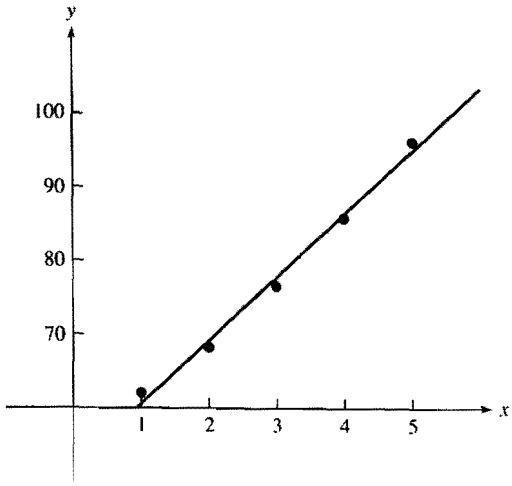


FIGURE 2–18



The system of equations is

$$\begin{aligned} 55m + 15b &= 1252.1 \\ 15m + 5b &= 388.7 \end{aligned}$$

Multiplying the second equation by 3, we have the system

$$\begin{aligned} 55m + 15b &= 1252.1 \\ 45m + 15b &= 1166.1 \quad \text{(Subtract)} \\ \hline 10m \qquad &= 86 \\ m &= 8.6 \\ b &= \frac{388.7 - 15(8.6)}{5} = 51.94 \end{aligned}$$

The least squares line is $y = 8.6x + 51.94$; its graph on the scatter plot is above.

Method II. Using matrices, the augmented matrix of the least squares system is

$$A = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 62.0 \\ 2 & 1 & 68.2 \\ 3 & 1 & 76.5 \\ 4 & 1 & 85.8 \\ 5 & 1 & 96.2 \end{bmatrix} = \begin{bmatrix} 55 & 15 & 1252.1 \\ 15 & 5 & 388.7 \end{bmatrix}$$

Using this augmented matrix, we obtain the same line as in method I, $y = 8.6x + 51.94$. ■

2.8 EXERCISES

Find the coefficients of the regression line to two decimals unless otherwise noted.

1. Display the scatter plot and find the regression line for the points (3, 5), (4, 6), (5, 8), (6, 10), and (7, 9). Display the graph of the regression line on the scatter plot.

2. Display the scatter plot and find the regression line for the points (2, 8), (4, 6), (6, 5), and (8, 3). Display the graph of the regression line on the scatter plot.

3. Display the scatter plot and find the regression line for the points (4, 8), (5, 7.5), (7, 6), and (9, 3.5). Display the graph of the regression line on the scatter plot.

4. Display the scatter plot and find the regression line for the points (10, 12), (11, 15), (12, 14), (13, 16), and (14, 18).

(a) Display the graph of the regression line on the scatter plot.

(b) Set the window so the x-range includes 20 and find the point on the regression line for $x = 20$.

5. The percentage of the United States population that uses the Internet is given for selected years from 1997 through 2001.

Year	% of U.S. Population
1997	22.2
1998	32.7
2000	44.4
2001	53.9

(a) Use $x = 0$ for 1997, $x = 1$ for 1998, and so on, and y = percentage of population to find the regression line for the data.

(b) Use the regression line to estimate the percentage of users in 2005.

(c) Use the regression line to estimate when the percentage will reach 100%.

6. Display the scatter plot and find the regression line for the points (12.7, 18.0), (9.5, 17.1), (15.3, 14.5), and (19.1, 12.2). Display the graph of the regression line on the scatter plot.

7. The following table gives the life expectancy at birth, for selected years, of females born in the United States:

Year	To Age
1940	65.2
1960	73.1
1980	77.4
1996	79.1
2000	79.1

(a) Using $x = 0$ for 1940, $x = 20$ for 1960, and so on, find the least squares line that fits these data.

(b) Using the least squares line, find the estimated life expectancy of a female born in 2010.

(c) Use the least squares line to estimate when a newborn's life expectancy will reach 100.

8. The following table gives the suicide rate (number of suicides per 100,000 population), for selected years, of the 15- to 24-year-old age group in the United States:

Year	Rate
1970	8.8
1980	12.3
1990	13.2
1996	12.0
1999	10.3

Use $x = 0$ for 1970, $x = 10$ for 1980, and so on.

(a) Find the least squares line for these data.

(b) Use the least squares line to estimate the suicide rate for the year 2015.

9. The average monthly temperature (in Fahrenheit) in Springfield, Illinois, for January, April, and July is January, 24.6; April, 53.3; and July, 76.5.

(a) Find the scatter plot for the data.

(b) Find the regression line for the data.

(c) Graph the regression line on the scatter plot. Does the line seem to fit the data well?

(d) Use the regression line to estimate the average monthly temperature for October.

(e) Is the October estimate realistic? Why or why not?

EXPLORATIONS

10. The number of miles that passenger cars travel annually in the United States is given for selected years.

Year	Miles (billions)
1960	0.587
1965	0.723
1970	0.917
1975	1.034
1980	1.112
1985	1.247
1990	1.408
1996	1.468
1999	1.569

(a) Display the scatter plot for the data.
(b) Find the regression line for the data with the coefficients to three decimal places.
(c) Based on the regression line, when will the annual mileage reach 2 billion miles?

11. The following table gives the birth rates, per 1000, for Israel and the United States for five different years.

	Birth Rate	
Year	Israel	U.S.
1975	28.2	14.0
1980	24.1	16.2
1985	23.5	15.7
1990	22.2	16.7
1998	20.0	14.4
2002	18.9	14.1

(a) Using $x = 0$ for 1975, $x = 5$ for 1980, and so on, find the birth-rate regression lines for each country.
(b) Using the regression lines, estimate when, if ever, the birth rate of the two countries will be the same.

12. The percentage of the U.S. adult population who smoked is given for selected years.

Year	Overall Population (%)	Males (%)	Females (%)
1974	37.1	43.1	32.1
1980	33.2	37.6	29.3
1985	30.1	32.6	27.9
1990	25.5	28.4	22.8
1994	25.5	28.2	23.1
1999	23.3	25.2	21.6

(a) Find the regression line for the percentage of males who smoke. Use $x = 0$ for 1974, $x = 6$ for 1980, and so on.
(b) Using the equation of the line, estimate the percentage of males who will smoke in 2005.
(c) Based on the regression line, will the percentage of males who smoke ever reach zero? If so, when? Why do you think this is or is not realistic?
(d) Find the regression line for the percentage of females who smoke.
(e) Based on the regression lines, will the percentage of females who smoke equal the percentage of males who smoke? If so, when?
(f) Find the regression line for the percentage of the overall population who smoke.
(g) For which of the three groups is the decline the greatest?

13. Throughout a person's working career, a portion of salary is withheld for Social Security. Upon retirement, a worker becomes eligible for Social Security benefits. The average monthly Social Security benefits for selected years are:

Year	1975	1980	1985	1990	1995	2000
Benefits	$146	$321	$432	$550	$672	$845

(a) Using these data, find the least squares regression line giving monthly benefits as a function of years since 1975 ($x = 0$ for 1975).
(b) Use the linear function to estimate the average monthly benefit for 2010. For 2030.

14. The United Stated national debt (in billions of dollars) for selected years from 1980 is shown in the following table:

Year	1980	1985	1990	1995	2000
Debt	930.2	1945.9	3233.3	4974.0	5674.2

(a) Find the least squares regression linear function giving national debt as a function of years since 1980 ($x = 0$ for 1980).
(b) Use the regression line to estimate the national debt in 2005. In 2015.
(c) Assume 300 million (0.3 billion) people in the United States in 2015. What is the estimated per capita debt?

15. The value of the American dollar declined during the second half of the 20th century. The table below shows the number of dollars required to equal the 1975 dollar value. For example, it cost $3.20 in 2000 to purchase what cost $1.00 in 1975.

	Amount required to equal $1 in 1975						
Year	**1975**	**1980**	**1985**	**1990**	**1995**	**2000**	**2002**
Dollar	1.00	1.53	2.00	2.43	2.83	3.20	3.29

(a) Find the least squares regression linear function giving the dollars required as a function of years since 1975 ($x = 0$ for 1975).
(b) Use the linear function to estimate the number of dollars required in 2010.
(c) Use the linear function to estimate when it will require $10.00 to equal the 1975 dollar.

16. The poverty level for a family of four, two parents with two children under 18, for selected years from 1990 is:

	Poverty Threshold for a Family of Four			
Year	**1990**	**1995**	**2000**	**2002**
Threshold	$13,300	$15,500	$17,500	$18,200

(a) Find the least squares linear regression function for threshold level as a function of years since 1990.
(b) Use the linear function to estimate when the poverty threshold level will reach $25,000.

17. The United States per capita income for selected years is:

	Per Capita Income				
Year	**1980**	**1985**	**1990**	**1995**	**2001**
Income	$7,787	$11,013	$14,387	$17,227	$22,851

(a) Find the least squares linear regression function for per capita income as a function of years since 1980.
(b) Use the linear function to estimate the per capita income level for 2008.
(c) Use the linear function to estimate when the per capita income level will reach $25,000.

18. Credit card companies aggressively compete for credit card holders. The percentage of Americans having credit cards has increased in recent years, as shown in the following table:

	Percentage of Americans Having a Credit Card			
Year	**1989**	**1992**	**1995**	**1998**
Percentage	56.0	62.4	66.4	67.5

(a) Use these data to determine the linear least squares regression function with percentage having a credit card as a function of years since 1989 ($x = 0$ for 1989).
(b) Use the linear function to estimate the percentage having a credit card in 1980.
(c) Use the linear function to estimate the percentage having a credit card in 2005.
(d) According to the function, when will 100% of Americans have a credit card? Is this reasonable?

19. The median income for a four-person family for selected years is:

	Median Income					
Year	**1988**	**1990**	**1992**	**1994**	**1996**	**1998**
Income	$39,050	$41,150	$44,250	$47,010	$51,500	$56,050

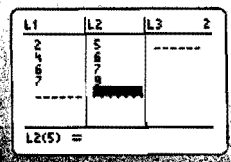
(a) Use these data to find the least squares linear function with income as a function of years since 1988.
(b) Use the linear function to estimate the median income for 1985. (The actual was $32,800)
(c) Use the linear function to estimate the median income for 2000. (The actual was $65,500)
(d) Based on the linear function, when will the median income reach $75,000?

USING YOUR TI-83

REGRESSION LINES

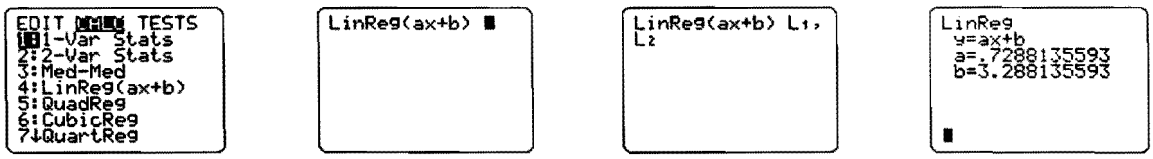
A TI graphing calculator can be used to find the equation of the regression line. We illustrate with the points (2, 5), (4, 6), (6, 7), and (7, 9).

Enter the points in the lists L1 and L2 with the x-coordinates in L1 and the y-coordinates in L2.



Find the regression coefficients.

This sequence of commands will give the following screens:



The last screen indicates that the least squares line is $y = 0.7288x + 3.288$.

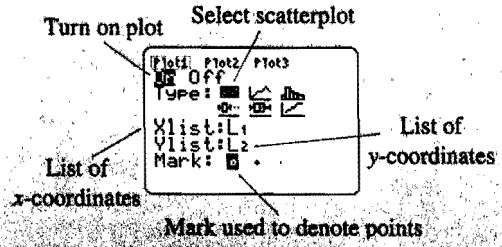
EXERCISES

1. Find the least squares line for the points (15, 22), (17, 25), and (18, 27).
2. Find the least squares line for the points (21, 44), (24, 40), (26, 38), and (30, 32).
3. Find the least squares line for the points (3.2, 5.7), (4.1, 6.3), (5.3, 6.7), and (6.0, 7.2).

SCATTERPLOT

The scatterplot of a set of points may be obtained by the following steps.

1. **Enter Data.** Enter the points with the x-coordinates in the list L1 and the corresponding y-coordinates in L2.
2. **Set the Horizontal and Vertical Scales.**
   Set Xmin, Xmax, Xscl, Ymin, Ymax, and Yscl using [WINDOW] in the same way it is used to set the screen for graphing functions.
3. **Define the Scatterplot.**
   Press [STAT PLOT] <1:PLOT1> [ENTER]
   You will see a screen similar to the following.



On that screen select, as shown in the figure, <ON>, scatterplot for Type, L1 for the list of x-coordinates, L2 for the list of y-coordinates, and the kind of mark that will show the location of the points. Press [ENTER] after each selection.

4. **Display the Scatterplot.**
   Press [GRAPH]

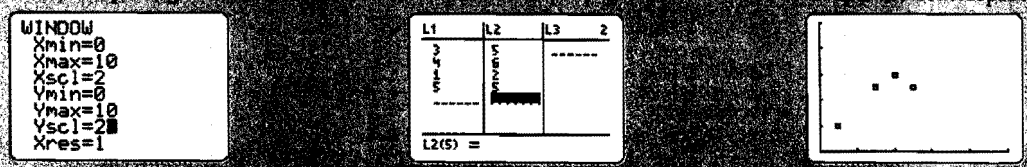
EXAMPLE

Show the scatterplot of the points (3, 5), (4, 6), (1, 2), (5, 7).

Set the Window to (0, 10) for the x and y-ranges.

Enter the points in the lists.

Graph the scatterplot.

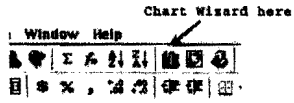


USING EXCEL

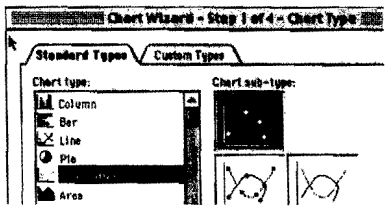
TO DRAW AND FIND A LINEAR REGRESSION LINE

We can use features of EXCEL to obtain a scatter plot of points that are given, find the equation of the regression line, and graph the line. We illustrate with the points (1, 3), (2, 5), (4, 9), (5, 8).

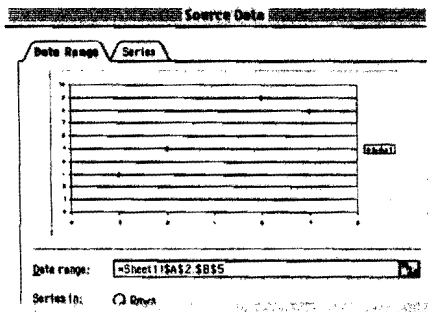
- For the given points, enter x in cells A2:A5 and y in cells B2:B5.
- Select the **Chart Wizard** icon in the Ruler.



- Under **Chart type,** select **XY Scatter**, then click on the first graph in the first column in **Chart Subtype.**



- Click **Next.** Be sure the cursor is in the **Data Range** box. Select the cells containing the data points, A2:B5 in this case.

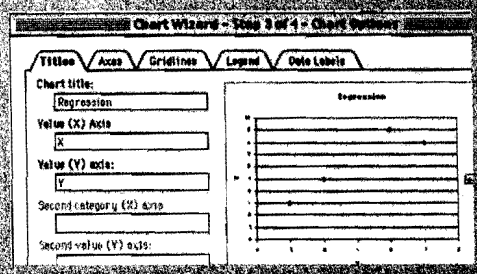


- Click **Next** and you will see a plot of the data points.

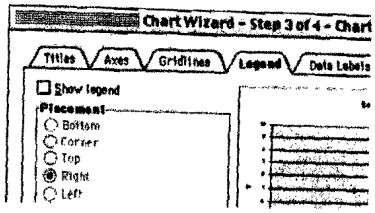
You can now enter some names and labels.

- Click on **Titles** at the top of the dialog box.
- In **Chart title,** enter the name of the line, such as *Regression.*

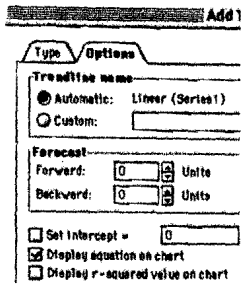
For **Value(X) Axis,** you may want to indicate what x represents such as "Number of ...." For **Value(Y) Axis,** you may identify what y represents.



- *Remove the legend by clicking on* **Legend** *at the top of the dialog box and removing the check mark by* **Show Legend.**

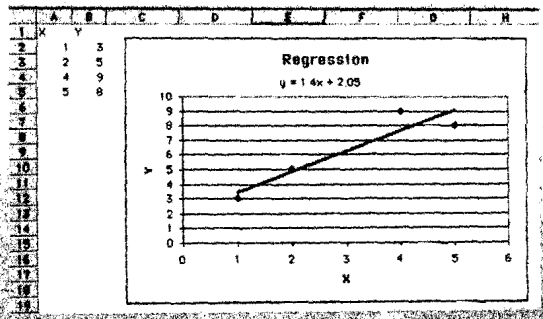


- Click **Finish.**
- Under the **Chart** menu, select **Add Trendline.**
- Under **Type** in the dialog box that appears, select **Linear.**
- Under **Options** in the dialog box click **Display equation on chart.**



- Click **OK.**

The line and its equation, $y = 1.4x + 2.05$, are shown on the chart.



EXERCISES

Find the least squares regression line for the points given.

1. (2, 4), (4, 5), (5, 6), (8, 11)
2. (10, 15), (13, 11), (14, 10), (17, 6), (20, 3)
3. (31, 6), (33, 2), (37, 10), (40, 15)