

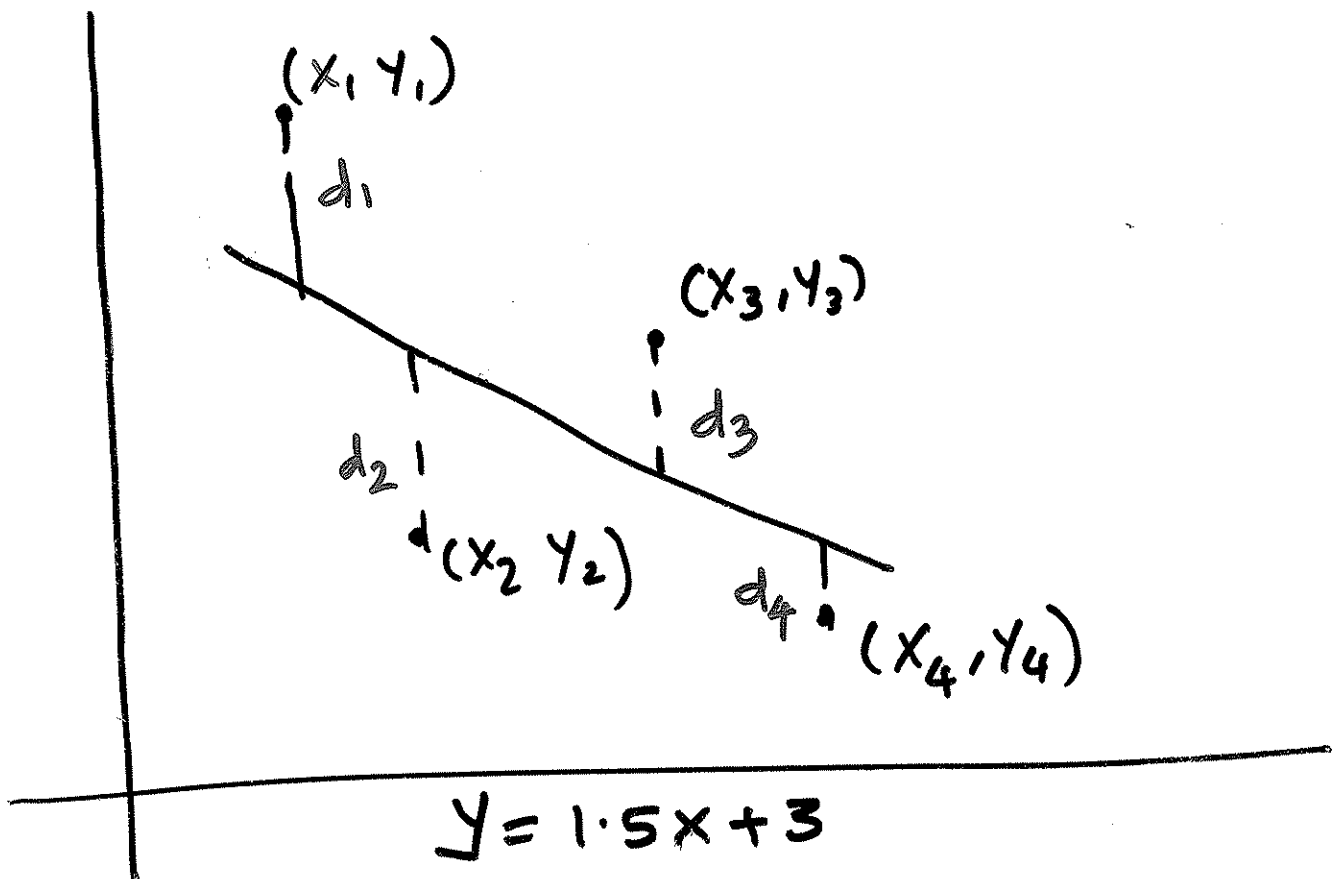
1.5 Method of Least Squares ①

Best fit line

Least Squares Error

method of Least Squares

} Key con
&
Appl'n



Data Point	Point on line	Vertical Dist
1, 6	1, 4.5	1.5
4, 5	4, 9	4
6, 14	6, 12	2

$$d = d_1^2 + d_2^2 + d_3^2$$

$$= 1.5^2 + 4^2 + 2^2 = 22.25$$

2

Finding Best Fit line using
Linear regression operation

$$m = \frac{N \cdot \sum xy - \sum x \cdot \sum y}{N \cdot \sum x^2 - (\sum x)^2}$$

$$b = \frac{\sum y - m \sum x}{N}$$

X_i	Y_i	X_i^2	$X_i Y_i$
1	10	1	10
2	9	4	18
3	8	9	24
5	7	25	35
9	5	81	45
Sum	20	39	120
			132

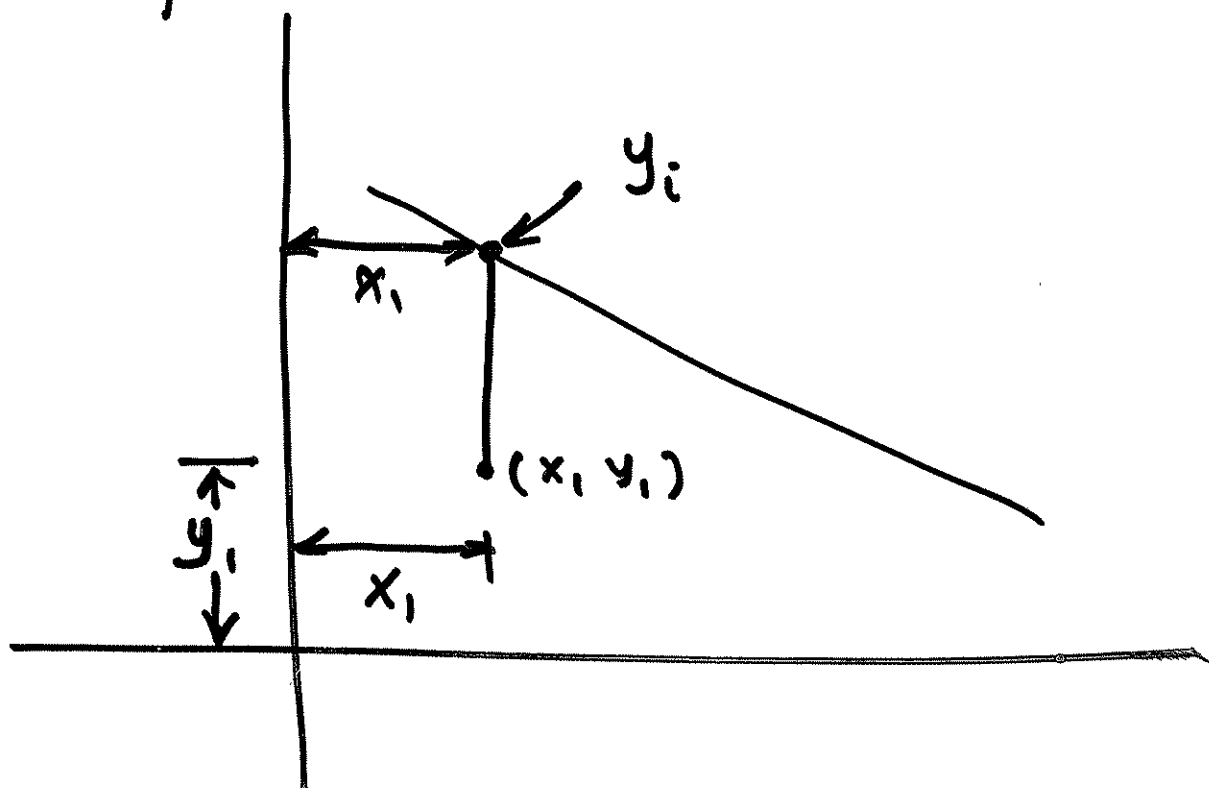
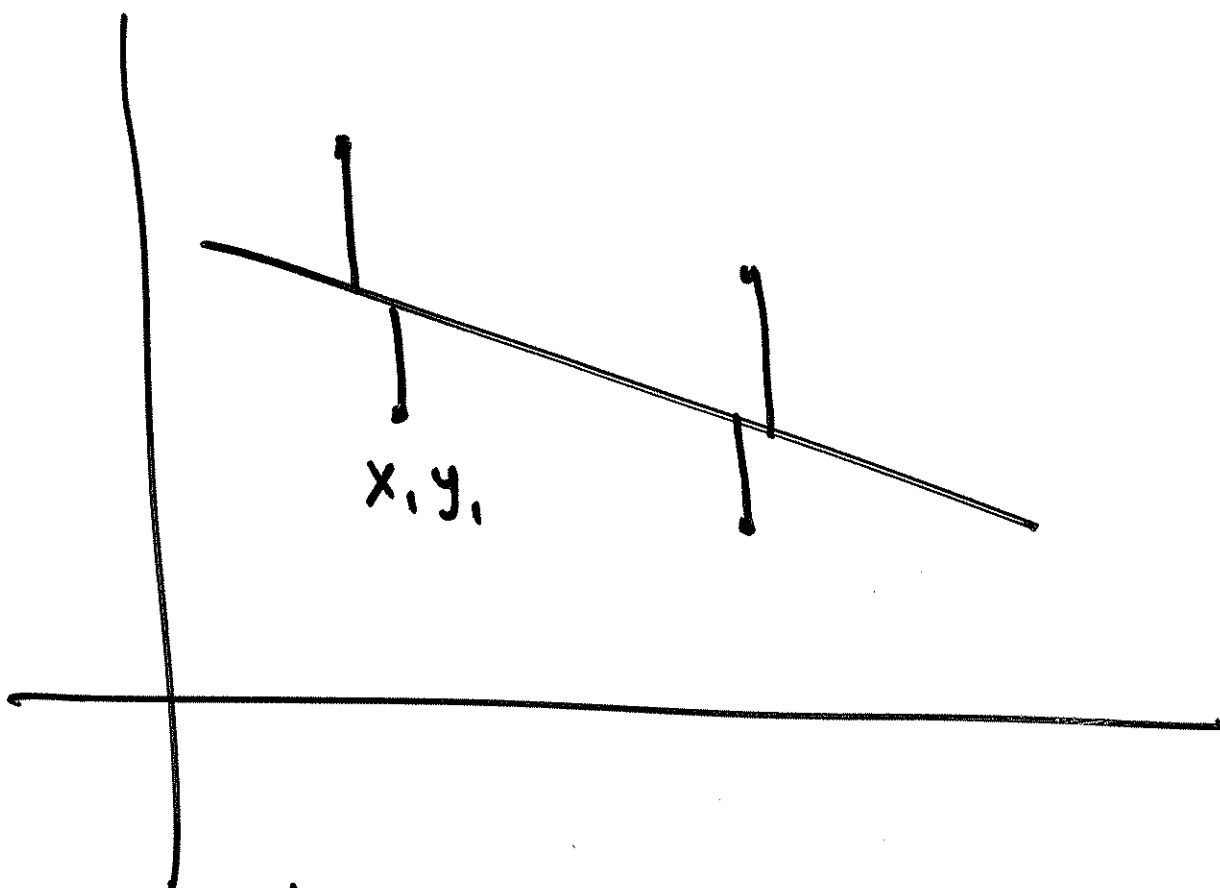
X	Y	X Y	X ²	
0	220	0	0	
1	205	205	1	
2	195	390	4	
3	180	540	9	
4	187	748	16	
10	987	1883	30	Σ

$$\frac{5(1883) - 10(987)}{5(30) - (10)^2} = \underline{\underline{-9.1}}$$

X	Y	X ²	XY
2000	1704	4000,000	3408000
2002	1923	4008004	3849846
2003	2014	4012009	4034042
2004	2105	4016016	4218420
8009	7746	16036029	15510308

$$\frac{4(15510308) - (8009)(7746)}{4(16036029) - (8009)^2}$$

$$b = \frac{100.51}{\frac{7746 - 100.51(8009)}{4}}$$



1.5 Method of Least Squares

APPLICATION Technology and Productivity

In 1995 Cohen published a study correlating corporate spending on communications and computers (as a percent of all spending on equipment) with annual productivity growth. He collected data on 11 companies for the period from 1985-1989. This data is found in the following table.

x	0.06	0.11	0.16	0.20	0.22	0.25	0.33	0.33	0.47	0.62	0.78
y	-1.0	4.5	-0.6	4.2	0.4	0.1	0.4	1.4	1.1	3.4	5.5

x is the spending on communications and computers as a percent of all spending on equipment, and y is the annual productivity growth as a percent.

What is the equation of a line that best approximates this data? What can you conclude about the relationship between spending on communications and computers and annual productivity growth? See Example 3 for the answer.

Source: Cohen 1995

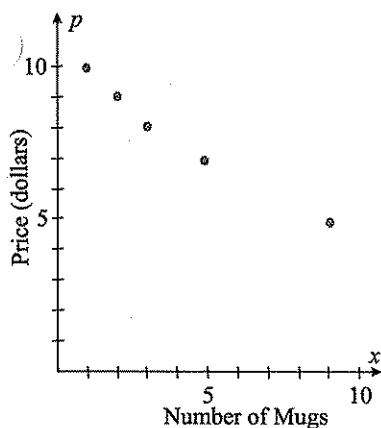


Figure 1.26

✧ The Method of Least Squares

Let x be the number of insulated mugs produced and sold, and let p be the price of each mug. Suppose x_1 mugs were sold at a price of p_1 and x_2 mugs were sold at a price of p_2 . If we then *assume* that the demand equation is linear, then of course there is only one straight line through these two points, and we can easily calculate the equation $y = ax + b$ of this line.

What if we have more than two data points? Suppose, as shown in Table 1.3, that we have five points available. Here, p_i are the prices in dollars for insulated mugs, and x_i is the corresponding demand for these insulated mugs in thousands of mugs sold per day.

x_i	1	2	3	5	9
p_i	10	9	8	7	5

Table 1.3

These are plotted in Figure 1.26, which is called a scatter diagram.

If we examine the scatter diagram, we see clearly that the points do not lie on any single straight line but seem to be scattered in a more or less linear fashion. Under such circumstances we might be justified in assuming that the demand equation was more or less a straight line. But what straight line? Any line that we draw will miss most of the points. We might then think to draw a line that is somehow closest to the data points. To actually follow such a procedure, we need to state exactly how we are to measure this closeness. We will measure this closeness in a manner that will lead us to the method of least squares.

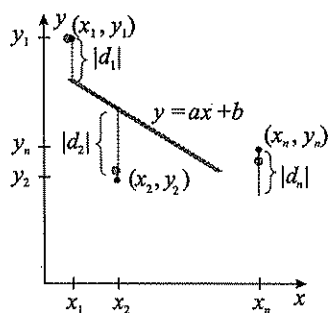


Figure 1.27

HISTORICAL NOTE**Carl Gauss (1777–1855)**

In 1781 Sir William Herschel discovered the planet Uranus, bringing the number of planets known at that time to seven. At that time astronomers were certain that an eighth planet existed between Mars and Jupiter, but the search for the missing planet had proved fruitless. On January 1, 1801, Giuseppe Piazzi, director of the Palermo Observatory in Italy, announced the discovery of a new planet Ceres in that location. (It was subsequently realized that Ceres was actually a large asteroid.) Unfortunately, he lost sight of it a few weeks later. Carl Gauss, one of the greatest mathematicians of all times, took up the challenge of determining its orbit from the few recorded observations. Gauss had in his possession one remarkable mathematical tool — the method of least squares — which he formulated in 1794 (at the age of 16!) but had not bothered to publish. Using this method, he predicted the orbit. Astronomers around the world were astonished to find Ceres exactly where Gauss said it would be. This incident catapulted him to fame.

First notice that to be given a non-vertical straight line is the same as to be given two numbers a and b with the equation of the straight line given as $y = ax + b$. Suppose now we are given n data points $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ and a line $y = ax + b$. We then define $d_1 = y_1 - (ax_1 + b)$, and note from the figure that $|d_1|$ is just the vertical distance from the first data point (x_1, y_1) to the line $y = ax + b$. Doing the same for all the data points, we then have

$$d_1 = y_1 - (ax_1 + b)$$

$$d_2 = y_2 - (ax_2 + b)$$

$$\vdots$$

$$d_n = y_n - (ax_n + b)$$

where $|d_2|$ is the vertical distance from the second data point (x_2, y_2) to the line $y = ax + b$, and so on. Refer to Figure 1.27.

Now if all the data points were on the line $y = ax + b$, then all the distances $|d_1|, |d_2|, \dots, |d_n|$ would be zero. Unfortunately, this will rarely be the case. We then use the sum of the squares of these distances

$$d = d_1^2 + d_2^2 + \dots + d_n^2$$

as a measure of how close the set of data points is to the line $y = ax + b$. Notice that this number d will be different for different straight lines: large if the straight line is far removed from the data points and small if the straight line passes close to all the data points. We then seek this line; that is, we need to find the two numbers a and b that will make this sum of squares the least. Thus the name least squares.

EXAMPLE 1 A Demand Function Find the best-fitting line through the data points in Table 1.3 and thus find a linear demand function. Then use the linear demand function to estimate the price if the demand is 6000 mugs. Finally, if the price is \$8.00 per mug, use the linear demand function to estimate the number of mugs that will be sold.

Solution Find the best-fitting line using the linear regression operation on a spreadsheet or graphing calculator. The steps are detailed in Technology Notes 1 and 2. You will find that $a = -0.6$ and $b = 10.2$. Thus, the equation of the best-fitting straight line that we are seeking is

$$y = p(x) = -0.6x + 10.2$$

The graph is shown in Figure 1.28. If technology is unavailable, the best-fitting line can be found by hand.

In general, the line $y = ax + b$ closest to the data points $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ can be found by solving the following two linear equations for a and b :

$$\begin{aligned} (x_1^2 + \dots + x_n^2)a + (x_1 + \dots + x_n)b &= x_1y_1 + \dots + x_ny_n \\ (x_1 + \dots + x_n)a + nb &= y_1 + \dots + y_n \end{aligned}$$

For our example we indicate these calculations by the following table.

3

$N = \text{no of data points}$

$$N = 5$$

$$m = \frac{5 \cdot (132) - 20(39)}{5 \cdot (120) - (20)^2}$$

$$= -0.6$$

$$b = \frac{39 - (-0.6)(20)}{5} = 10.2$$

$$y = mx + b$$

$$y = -0.6x + 10.2$$

x - represents thousands of mugs

$$\therefore 6000 \text{ mugs } x = 6$$

$$y = -0.6(6) + 10.2$$
$$= 6.6$$

$$p = 8 = -0.6x + 10.2$$

$$x = 3.667$$

which means 3667 thousand mugs

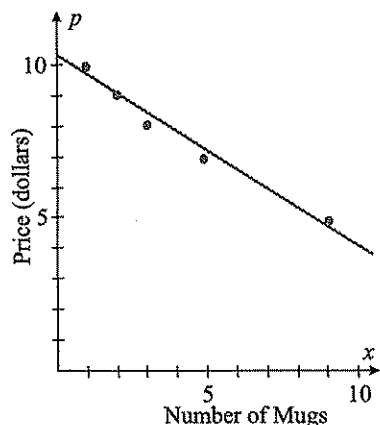


Figure 1.28

x_i	y_i	x_i^2	$x_i y_i$
1	10	1	10
2	9	4	18
3	8	9	24
5	7	25	35
9	5	81	45
Sum	20	39	120
		120	132

We then have the system of two equations in the two unknowns a and b .

$$120a + 20b = 132$$

$$20a + 5b = 39$$

These equations can be readily solved using the techniques we used in the last several sections.

To find the price for 6000 mugs, we will use $x = 6$ in the demand equation.

$$p(6) = (-0.6)(6) + 10.2 = 6.6$$

That is, if 6000 mugs are to be sold, then the price should be \$6.60 each.

For a price of \$8.00 we set $p = 8$ and solve for x ,

$$8 = -0.6x + 10.2$$

$$0.6x = 2.2$$

$$x = \frac{2.2}{0.6} = \frac{11}{3} \approx 3.6667$$

Next note that since x represents thousands of mugs, we will expect to sell 3667 mugs at a price of \$8.00 each. ♦

✧ Correlation

We have just seen how to determine a functional relationship between two variables. This is called regression analysis. We now wish to determine the strength or degree of association between two variables. This is referred to as correlation analysis. The strength of association is measured by the correlation coefficient which is defined as

$$r = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{[n \sum x_i^2 - (\sum x_i)^2][n \sum y_i^2 - (\sum y_i)^2]}}$$

where $\sum y_i$, for example, is $y_1 + y_2 + \cdots + y_n$. The value of this correlation coefficient ranges from +1 for two variables with perfect positive correlation to -1 for two variables with perfect negative correlation. See Figure 1.29 for examples.

Technology Option

Finding the correlation coefficient using a graphing calculator or Microsoft Excel is discussed in Technology Note 3 on page 77.

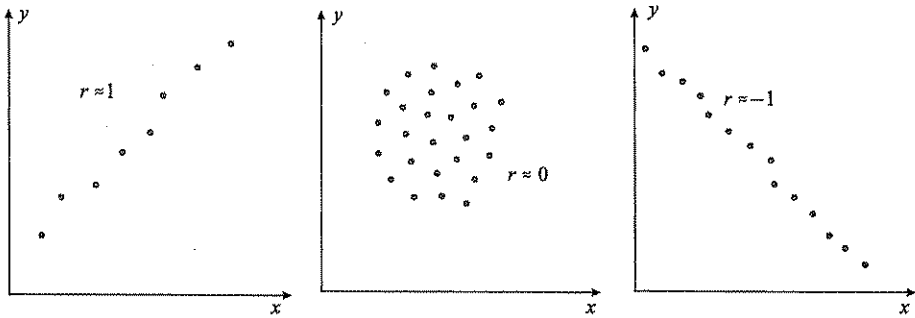


Figure 1.29a
Positive Correlation

Figure 1.29a
No Correlation

Figure 1.29a
Negative Correlation

EXAMPLE 2 Correlation Find the correlation coefficient for the data in Table 1.3.

Solution We can use the above formula or let our computers or graphing calculators do the work. In doing Example 1 you will find that $r \approx -0.99$. This indicates a high negative correlation, which can easily be seen by observing the original data in Figure 1.26. We conclude that there is a strong negative correlation between price and demand in this instance. Thus, we expect increases in prices to lead to decreases in demand. ♦

♦ **Additional Examples**

There has been a debate recently as to whether corporate investment in computers has a positive effect on productivity. Some suggest that worker difficulties in adjusting to using computers might actually hinder productivity. The paper mentioned in the following example explores this question.

EXAMPLE 3 Technology and Productivity In 1995 Cohen published a study correlating corporate spending on communications and computers (as a percent of all spending on equipment) with annual productivity growth. He collected data on 11 companies for the period from 1985-1989. This data is found in the following table.

x	0.06	0.11	0.16	0.20	0.22	0.25	0.33	0.33	0.47	0.62	0.78
y	-1.0	4.5	-0.6	4.2	0.4	0.1	0.4	1.4	1.1	3.4	5.5

x is the spending on communications and computers as a percent of all spending on equipment, and y is the annual productivity growth as a percent.

Determine the best-fitting line using least squares and find the correlation coefficient. Discuss the results.

Source: Cohen 1995

Solution Using a spreadsheet or a graphing calculator, we find that $a \approx 5.246$, $b \approx 0.080$, and $r \approx 0.518$. Thus, the best-fitting straight line is

$$y = ax + b = 5.246x + 0.080$$

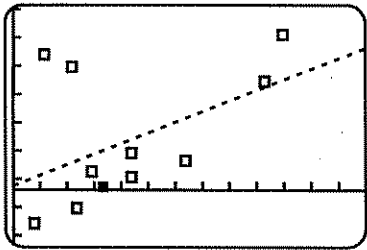


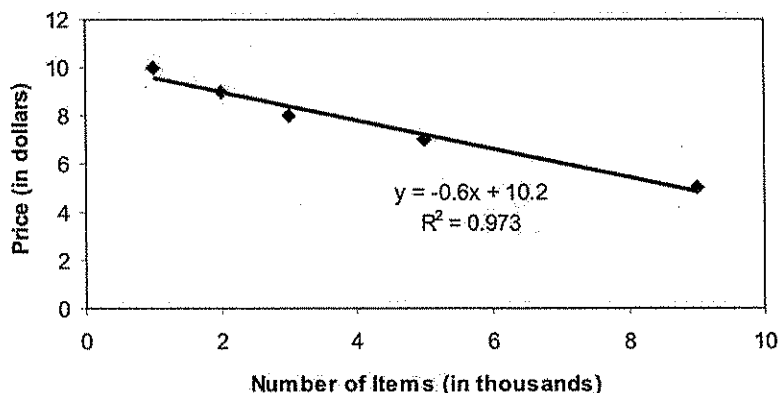
Figure 1.30

Technology Note 2 Example 1 with Microsoft Excel

Begin by opening a new worksheet. Label Column A with an x and Column B with a p. Enter the data from Table 1.3 as shown in Worksheet 1.1. Highlight the data and choose Chart Wizard from the tool bar or Insert and then Chart, if the Chart Wizard is not on the toolbar. Then pick the XY (Scatter) graph and choose the first chart type, Scatter. Compares pairs of values. Continue clicking Next until the chart is complete, adding titles and labels as needed. When the scatter plot is complete, click on the graph and then on Chart in the tool bar. Choose Add Trendline... Pick the linear trend line and under the options menu, choose Display equation on chart and Display R-squared value on chart. The completed graph is shown below:

	A	B
1	x	p
2	1	10
3	2	9
4	3	8
5	5	7
6	9	5

Worksheet 1



Technology Note 3 Correlation Coefficient

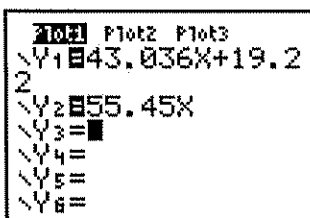


If your calculator does not display the correlation coefficient, r , it can be enabled by going to the CATALOG menu (above the $\square$) and scroll down to DiagnosticOn (see Screen 9) and press **ENTER**. Press **ENTER** again on the homescreen to enable the diagnostic. Next time a regression is done, the correlation coefficient, r , and the square of the correlation coefficient, r^2 , will automatically be displayed. The correlation coefficient will be displayed on the Excel chart when this option is enabled. See Worksheet 1.

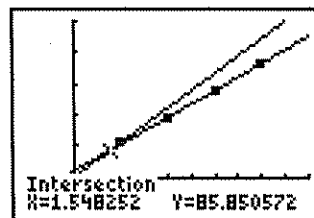
Screen 9

Technology Note 4 Example 4 on a Graphing Calculator

Enter the data into lists L1 and L2. Then, as in Technology Note 2, find the regression equation and enter it into Y_1 for our cost equation. In Y_2 enter the revenue equation, $55.45X$, as shown in Screen 10. Screen 11 has the graph.



Screen 10



Screen 11

$[-2, 10] \times [-100, 500]$

Self-Help Exercises 1.5

- Using the method of least squares, find the best-fitting line through the three points $(0, 0)$, $(2, 2)$, and $(3, 2)$. Find the correlation coefficient.
- The table below shows x , the number of boxes of cereal in thousands that will be supplied at a price y , in dollars. Use the method of least

squares and the supply information in the table to determine how many boxes of cereal will be supplied at a price of \$3.75.

x	9	7	6	4	3
y	4.49	3.05	2.49	2.10	1.92

1.5 Exercises

In Exercises 1 through 8, find the best-fitting straight line to the given set of data, using the method of least squares. Graph this straight line on a scatter diagram. Find the correlation coefficient.

- $(0, 0)$, $(1, 2)$, $(2, 1)$
- $(0, 1)$, $(1, 2)$, $(2, 2)$
- $(0, 0)$, $(1, 1)$, $(2, 3)$, $(3, 3)$
- $(0, 0)$, $(1, 2)$, $(2, 2)$, $(3, 0)$
- $(1, 4)$, $(2, 2)$, $(3, 2)$, $(4, 1)$
- $(0, 0)$, $(1, 1)$, $(2, 2)$, $(3, 4)$
- $(0, 4)$, $(1, 2)$, $(2, 2)$, $(3, 1)$, $(4, 1)$

Applications

- Selling Strawberries** The table below shows x , the number cartons of strawberries, that a fruit stand can sell at different prices y in dollars. Find the demand equation for strawberries using linear regression. Using the demand equation, find the price the stand should charge if they wish to sell 35 cartons of strawberries.

x	12	15	20	27	44	60
y	5.00	4.00	3.50	3.00	2.50	2.00

- Selling Puzzles** The table below shows x , the number of puzzles (in thousands), that the A-Mart company can sell at different prices, y in dollars. Find the demand equation for these puzzles using linear regression. Using that demand equation, find the price that A-Mart should charge if they wish to sell 10,000 puzzles.

- Supply of Mugs** The table below shows x , the number of mugs in thousands supplied by the Big Mug company for different prices in euros (y). Using linear regression find the supply equation and use it to determine the number of mugs that the Big Mug company will supply when the price is 3 euros.

x	8	5	12	3	15
y	5.00	6.00	3.50	8.00	3.00

- Purchase Price of a Car** The value of a 1998 car is given in the table below (value is in thousands of dollars). Find the depreciation equation using linear regression and use it to estimate to the nearest dollar the purchase price of this car.

Year	2000	2001	2002	2003	2004
Value	19	15	12	9	7

- Purchase Price of a Machine** The data in the table below is the value of a milling machine in thousands of dollars and the number of years since the item was purchased. Use linear regression to estimate the purchase price of this milling machine to the nearest dollar.

years since purchase	1	2	4	6	7
value in dollars	3.9	3.2	2.5	2	1.8

Referenced Applications

13. Deer Population and Deer-Vehicle Accidents

Rondeau and Conrad studied the relationship of deer population in urban areas to accidents with vehicles on roads. Below is the data where x is the deer population in Irondequoit, NY, and y is the number of collisions with deer in the same city.

x	340	350	480	510	515	600	650
y	150	60	90	130	120	110	140

x	700	760	765	820	850
y	120	210	170	180	230

- Determine the best-fitting line using least squares and the correlation coefficient.
- Explain what the linear model is saying about the deer population and collisions.

Source: Rondeau and Conrad 2003

14. Economies of Scale in Plant Size *Strategic Management* relates a study in economies of scale in the machine-tool industry. The data is found in the following table,

x	70	115	130	190	195	400	450
y	1.1	1.0	0.85	0.75	0.85	0.67	0.50

where x is the plant capacity in thousands of units, and y is the employee-hours per unit.

- Determine the best-fitting line using least squares and the correlation coefficient.
- Is there an advantage in having a large plant? Explain.
- What does this model predict the employee-hours per unit will be when the plant capacity is 300,000 units?
- What does this model predict the plant capacity will be if the employee-hours per unit is 0.90?

Source: Rowe, Mason, Dickel, Mann, and Mockler 1994

15. Cost Curve Dean made a statistical estimation of the cost-output relationship in a hosiery mill. The data is given in the following table,

x	16	31	48	57	63	103	110
y	30	60	100	130	135	230	230

x	114	116	117	118	123	126
y	235	245	250	235	250	260

where x is the output in thousands of dozens and y is the total cost in thousands of dollars.

- Determine the best-fitting line using least squares and find the correlation coefficient. Graph.
- What does this model predict the total cost will be when the output is 100,000 dozen?
- What does this model predict the output will be if the total cost is \$125,000?

Source: Dean 1976

16. Cost Curve Johnston estimated the cost-output relationship for 40 firms. The data for the fifth firm is given in the following table where x is the output in millions of units and y is the total cost in thousands of pounds sterling

x	180	210	215	230	260	290	340	400
y	130	180	205	190	215	220	250	300

x	405	430	430	450	470	490	510
y	285	305	325	330	330	340	375

- Determine the best-fitting line using least squares and find the correlation coefficient. Graph.
- What does this model predict the total cost will be when the output is 300 million units?
- What does this model predict the output will be if the total cost is 200,000 of pounds sterling?

Source: Johnston 1960

17. Productivity Bernstein studied the correlation between productivity growth and gross national product (GNP) growth of six countries. The countries were France (F), Germany (G), Italy (I), Japan (J), the United Kingdom (UK), and the United States (US). Productivity is given as output per employee-hour in manufacturing. The data they collected for the years 1950–1977 is given in the following table where x is the productivity growth (%) and y is the GNP growth (%).

	US	UK	F	I	G	J
x	2.5	2.7	5.2	5.6	5.7	9.0
y	3.5	2.3	4.9	4.9	5.7	8.5

Example 8

4

X	Y	X_i^2	$X_i Y_i$
8	5.58	64	44.64
11	4.63	121	50.93
12	4.0	144	48.0
13	3.58	169	46.54
16	2.83	256	45.28
20	2.39	400	47.8
Σ 80	23.01	1154	283.19

$$m = \frac{6(283.19) - (80)(23.01)}{6(1154) - (80)^2}$$

$$= -0.2703$$

$$b = \frac{23.01 - (-0.2703) 80}{6}$$

$$= 7.4390$$

$$-0.2703X + 7.4390$$

$$P(15) = -0.2703(15) + 7.4390 = 3.38$$

$$\sum_{i=1}^3 X_i$$

$$X_1 + X_2 + X_3$$

X	Y	X ²	XY	
12	5	144	60	
15	4	225	60	
20	3.5	400	70.	
27	3.0	729	81	
44	2.5	1936	110	
60	2.0	3600	120	
178	20	7034	501	Sum

$$N = 6$$

$$\sum x = 178$$

$$\sum y = 20$$

$$\sum xy = 501$$

$$\sum x^2 = 7034$$

$$m = \frac{6(501) - 178(20)}{6(7034) - (178)^2} = -0.05$$

$$b = 4.82 / 4.89$$

$$y = mx + b$$

$$D(x) = -0.05x + 4.82$$

H.W. 1.5

1) Graphing

2) Questions 28, 30 From the book

x	1	2	3	4
y	20	10	15	40

x	y	xy	x ²
---	---	----	----------------

$$m =$$

$$b =$$

$$y = mx + b$$

$$y = 3x + 2$$

$y = 3x + 2$	x
2	0
5	1
-1	-1
8	2

$$R = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{[n \sum x_i^2 - (\sum x_i)^2][n \sum y_i^2 - (\sum y_i)^2]}}$$

n = no of data points