Chapter	Descriptive Statistics: Graphing and Summarizing Data
Section	Linear Regression and Correlation

Often, we want to know if the relationship between two things can be described in terms of a straight line. The process of determining the equation of that line is called **linear regression**.

The extent to which the equation works is called **correlation**.

If one increases as the other increases, they are **positively** correlated.

Example: The farther away something is, the longer it will take to get to it.

If one increases as the other decreases, they are **negatively** correlated.

Example: The farther away something is, the smaller it appears.

We have two kinds of variables:

1. Independent (also called predictor or explanatory) denoted as x.
   The distance in our two examples above.

2. Dependent (also called response) denoted as y.
   The time and size in our two examples above.

If we plot the points on an x-y coordinate axis, we get what is called a **scatter plot**.

We want a straight line through the "cloud" of points. The distance from each point to the line is called the **error**. We want the error to be as small as possible.

The line that minimizes the error is called the **line of best fit**, or (more commonly) the **regression equation**.

We use a method called the **least-squares criterion** to find the regression equation.

The general form of the regression equation is:

$$\hat{y} = b_0 + b_1 x$$

where:

$$b_1 = \frac{S_{xy}}{S_{xx}} \text{ and } b_0 = \bar{y} - b_1 \bar{x}$$

The linear correlation coefficient is given by:

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

In the above formulas:

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

$\bar{y}$ is the mean (average) of the y values
$\bar{x}$ is the mean (average) of the x values

The easiest way to get these values is to fill out a table:

n	x	y	xy	x^2	y^2
1					
2					
3					
…					
N					
TOTALS	$\sum x$	$\sum y$	$\sum (xy)$	$\sum (x^2)$	$\sum (y^2)$

Populate the "x" and "y" columns with the data pairs.
Calculate the values for each of the cells in the last three columns.
The "TOTALS" line is obtained by adding each column of numbers.