Chapter 7

Correlation and Simple Linear Regression

Introduction

Scientists and engineers often collect data in order to determine the nature of a relationship between two quantities. For example, a chemical engineer may run a chemical process several times in order to study the relationship between the concentration of a certain catalyst and the yield of the process. Each time the process is run, the concentration x and the yield y are recorded. The experiment thus generates **bivariate** data; a collection of ordered pairs $(x_1, y_1), \ldots, (x_n, y_n)$. In many cases, ordered pairs generated in a scientific experiment will fall approximately along a straight line when plotted. In these situations the data can be used to compute an equation for the line. This equation can be used for many purposes; for example, in the catalyst versus yield experiment just described, it could be used to predict the yield y that will be obtained the next time the process is run with a specific catalyst concentration x.

The methods of correlation and simple linear regression, which are the subject of this chapter, are used to analyze bivariate data in order to determine whether a straight-line fit is appropriate, to compute the equation of the line if appropriate, and to use that equation to draw inferences about the relationship between the two quantities.

7.1 Correlation

One of the earliest applications of statistics was to study the variation in physical characteristics in human populations. To this end, statisticians invented a quantity called the **correlation coef cient** as a way of describing how closely related two physical

characteristics were. The first published correlation coefficient was due to the English statistician Sir Francis Galton, who in 1888 measured the heights and forearm lengths of 348 adult men. (Actually, he measured the distance from the elbow to the tip of the middle finger, which is called a cubit.) If we denote the height of the ith man by x_i, and the length of his forearm by y_i, then Galton's data consist of 348 ordered pairs (x_i, y_i). Figure 7.1 presents a simulated re-creation of these data, based on a table constructed by Galton.

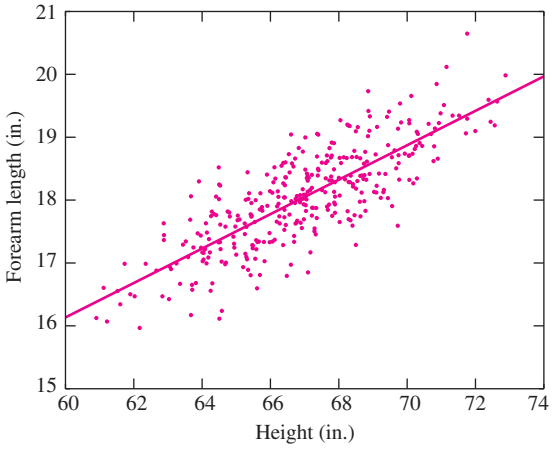

FIGURE 7.1 Heights and forearm lengths of 348 men.

The points tend to slope upward and to the right, indicating that taller men tend to have longer forearms. We say that there is a **positive association** between height and forearm length. The slope is approximately constant throughout the plot, indicating that the points are clustered around a straight line. The line superimposed on the plot is a special line known as the **least-squares line**. It is the line that fits the data best, in a sense to be described in Section 7.2. We will learn how to compute the least-squares line in Section 7.2.

Figure 7.2 presents the results of a study of the relationship between the mean daily temperature and the mean daily humidity at a site near Riverside, California, during a recent winter. Again the points are clustered around the least-squares line. The line has a negative slope, indicating that days with higher humidity tend to have lower temperatures.

The degree to which the points in a scatterplot tend to cluster around a line reflects the strength of the linear relationship between x and y. The visual impression of a scatterplot can be misleading in this regard, because changing the scale of the axes can make the clustering appear tighter or looser. For this reason, we define the **correlation coef cient** , which is a numerical measure of the strength of the linear relationship between two variables. The correlation coefficient is usually denoted by the letter r. There are several equivalent formulas for r. They are all a bit complicated, and it is not immediately obvious how they work. We will present the formulas, then show how they work.

Let $(x_1, y_1), \ldots, (x_n, y_n)$ represent n points on a scatterplot. To compute the correlation, first compute the means and standard deviations of the xs and ys, that is,

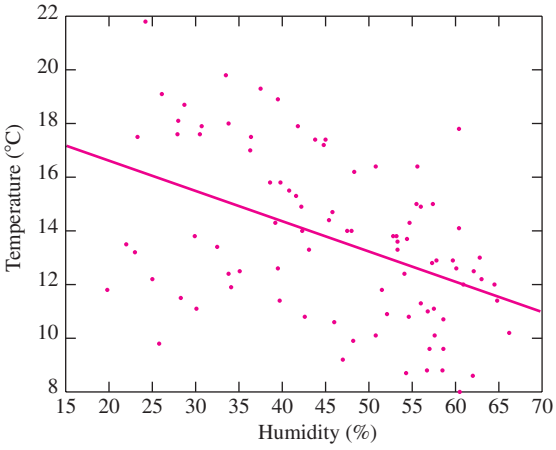

FIGURE 7.2 Humidity (in percent) and temperature (in °C) for days in a recent winter in Riverside, California.

$\overline{x}$, $\overline{y}$, s_x, and s_y. Then convert each x and y to standard units, or, in other words, compute the z-scores: $(x_i - \overline{x})/s_x$, $(y_i - \overline{y})/s_y$. The correlation coefficient is the average of the products of the z-scores, except that we divide by $n - 1$ instead of n:

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s_x} \right) \left(\frac{y_i - \overline{y}}{s_y} \right) \tag{7.1}$$

We can rewrite Equation (7.1) in a way that is sometimes useful. By substituting $\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2/(n-1)}$ for s_x and $\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2/(n-1)}$ for s_y, we obtain

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{7.2}$$

By performing some algebra on the numerator and denominator of Equation (7.2), we arrive at yet another equivalent formula for r:

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}\sqrt{\sum_{i=1}^{n} y_i^2 - n\overline{y}^2}} \tag{7.3}$$

Equation (7.3) is often the easiest to use when computing by hand.

In principle, the correlation coefficient can be calculated for any set of points. In many cases, the points constitute a random sample from a population of points. In these cases the correlation coefficient is often called the **sample correlation**, and it is an estimate of the population correlation. (Population correlation was discussed formally in Section 2.6; intuitively, you may imagine the population to consist of a large finite collection of points, and the population correlation to be the quantity computed using Equation (7.2) on the whole population, with sample means replaced by population means.) The sample correlation can be used to construct confidence intervals and perform hypothesis tests on the population correlation; these will be discussed later in this section. We point out that the correlation coefficient can also be used to measure the strength of

a linear relationship in many cases where the points are not a random sample from a population; see the discussion of the coefficient of determination in Section 7.2.

It is a mathematical fact that the correlation coefficient is always between -1 and 1. Positive values of the correlation coefficient indicate that the least-squares line has a positive slope, which means that greater values of one variable are associated with greater values of the other. Negative values of the correlation coefficient indicate that the least-squares line has a negative slope, which means that greater values of one variable are associated with lesser values of the other. Values of the correlation coefficient close to 1 or to -1 indicate a strong linear relationship; values close to 0 indicate a weak linear relationship. The correlation coefficient is equal to 1 (or to -1) only when the points in the scatterplot lie exactly on a straight line of positive (or negative) slope, in other words, when there is a perfect linear relationship. As a technical note, if the points lie exactly on a horizontal or a vertical line, the correlation coefficient is undefined, because one of the standard deviations is equal to zero. Finally, a bit of terminology: Whenever $r \neq 0$, x and y are said to be **correlated**. If $r = 0$, x and y are said to be **uncorrelated**.

The correlation between height and forearm length in Figure 7.1 is 0.80. The correlation between temperature and humidity in Figure 7.2 is -0.46. Figures 7.3 and 7.4 (pages 509 and 510) present some examples of scatterplots with various correlations. In each plot, both x and y have mean 0 and standard deviation 1. All plots are drawn to the same scale.

How the Correlation Coefficient Works

Why does the formula (Equation 7.1) for the correlation coefficient r measure the strength of the linear association between two variables? Figure 7.5 (page 511) illustrates how the correlation coefficient works. In this scatterplot, the origin is placed at the point of averages $(\overline{x}, \overline{y})$. Therefore, in the first quadrant, the z-scores $(x_i - \overline{x})/s_x$ and $(y_i - \overline{y})/s_y$ are both positive, so their product is positive as well. Thus each point in the first quadrant contributes a positive amount to the sum in Equation (7.1). In the second quadrant, the z-scores for the x coordinates of the points are negative, while the z-scores for the y coordinates are positive. Therefore the products of the z-scores are negative, so each point in the second quadrant contributes a negative amount to the sum in Equation (7.1). Similarly, points in the third quadrant contribute positive amounts, and points in the fourth quadrant contribute negative amounts. Clearly, in Figure 7.5 there are more points in the first and third quadrants than in the second and fourth, so the correlation will be positive. If the plot had a negative slope, there would be more points in the second and fourth quadrants, and the correlation coefficient would be negative.

The Correlation Coefficient Is Unitless

In any sample $x_1, \ldots, x_n$, the mean $\overline{x}$ and the standard deviation s_x have the same units as $x_1, \ldots, x_n$. For this reason the z-scores $(x_i - \overline{x})/s_x$ are unitless. Since the correlation coefficient r is the average of products of z-scores, it too is unitless. This fact is crucial to the usefulness of r. For example, the units for the x and y coordinates in Figure 7.1 are both inches, while the corresponding units in Figure 7.2 are percent and degrees Celsius. If the correlation coefficients for the two plots had different units, it would

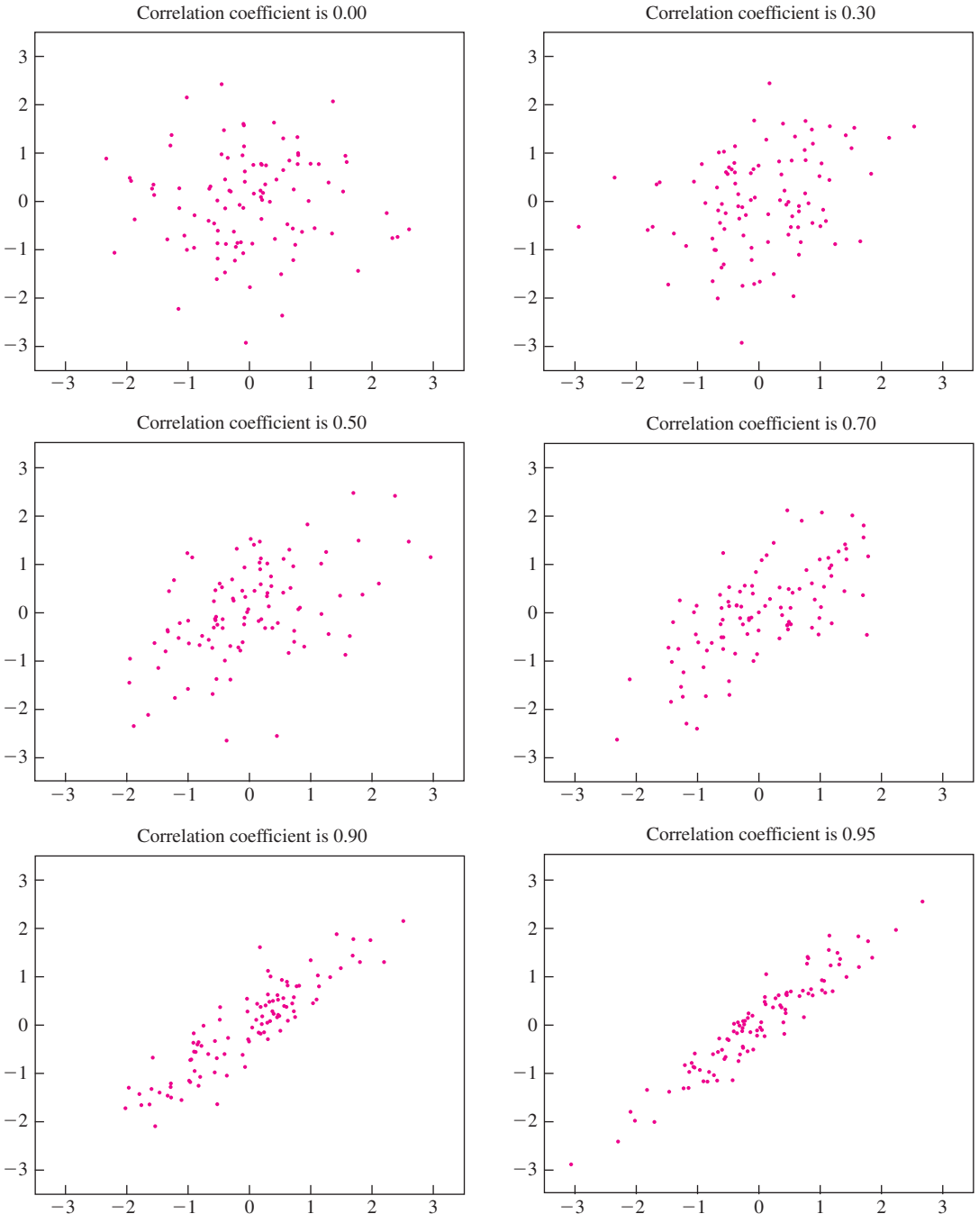

FIGURE 7.3 Examples of various levels of positive correlation.

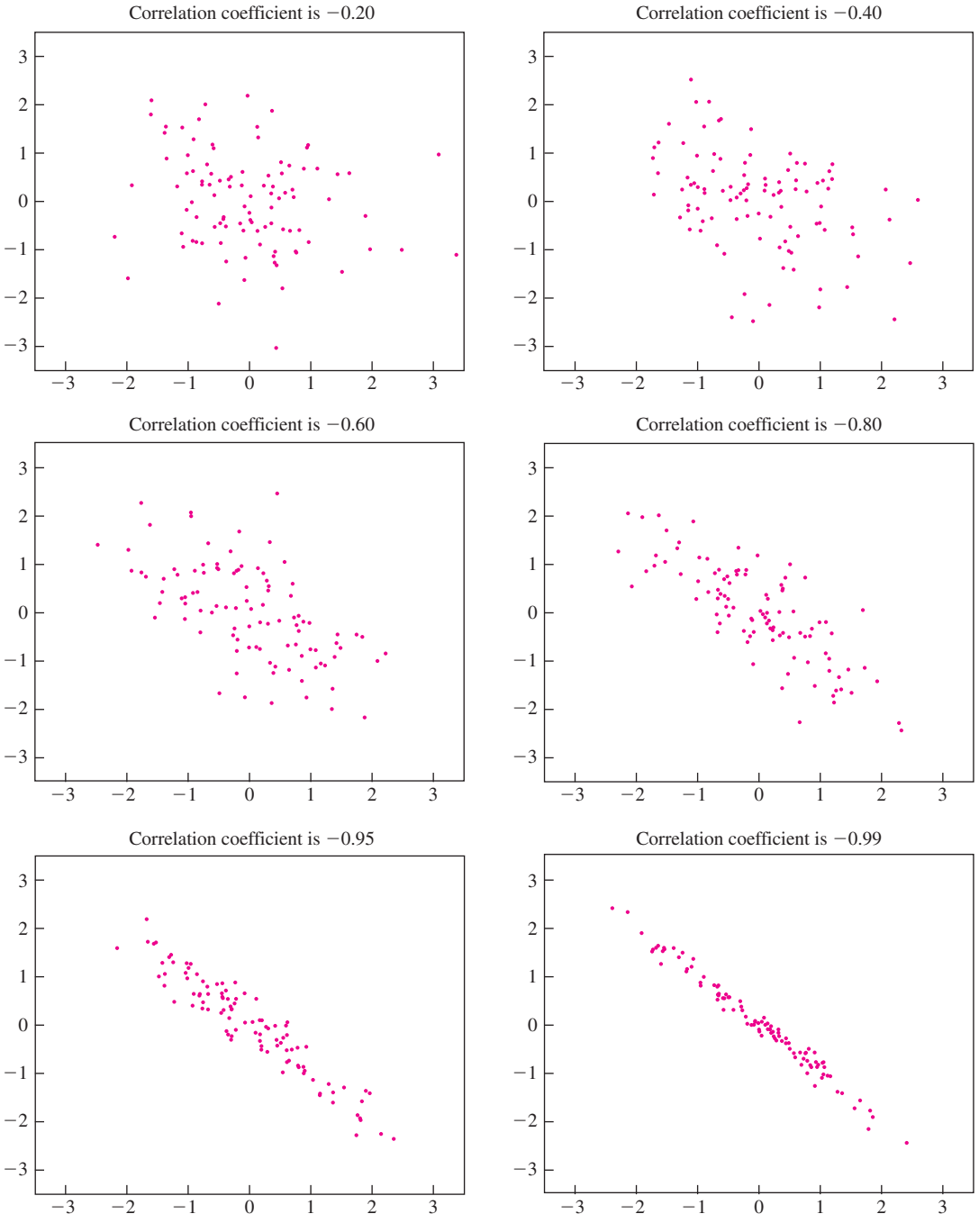

FIGURE 7.4 Examples of various levels of negative correlation.

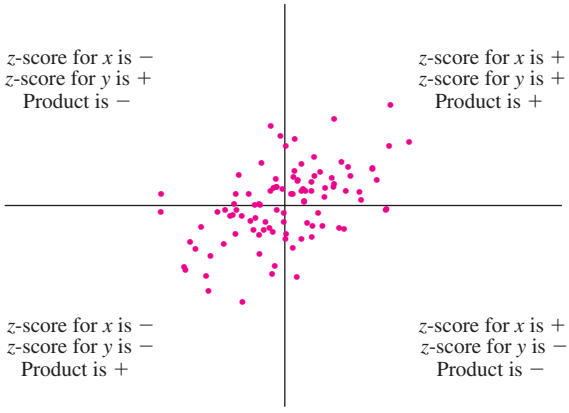

FIGURE 7.5 How the correlation coefficient works.

be impossible to compare their values to determine which plot exhibited the stronger linear relationship. But since the correlation coefficients have no units, they are directly comparable, and we can conclude that the relationship between heights of men and their forearm lengths in Figure 7.1 is more strongly linear than the relationship between temperature and humidity in Figure 7.2.

Another crucial property of the correlation coefficient is that it is unaffected by the units in which the measurements are made. For example, imagine that in Figure 7.1 the heights of the men were measured in centimeters rather than inches. Then each x_i would be multiplied by 2.54. But this would cause $\overline{x}$ and s_x to be multiplied by 2.54 as well, so the z-scores $(x_i - \overline{x})/s_x$ would be unchanged, and r would be unchanged as well. In a more fanciful example, imagine that each man stood on a platform 2 inches high while being measured. This would increase each x_i by 2, but the value of $\overline{x}$ would be increased by 2 as well. Thus the z-scores would be unchanged, so the correlation coefficient would be unchanged as well. Finally, imagine that we interchanged the values of x and y, using x to represent the forearm lengths, and y to represent the heights. Since the correlation coefficient is determined by the product of the z-scores, it does not matter which variable is represented by x and which by y.

Summary

The correlation coefficient remains unchanged under each of the following operations:

- Multiplying each value of a variable by a positive constant.
- Adding a constant to each value of a variable.
- Interchanging the values of x and y.

Figure 7.6 (page 512) presents plots of mean temperatures for the months of April and October for several U.S. cities. Whether the temperatures are measured in °C or °F, the correlation is the same. This is because converting from °C to °F involves multiplying by 1.8 and adding 32.

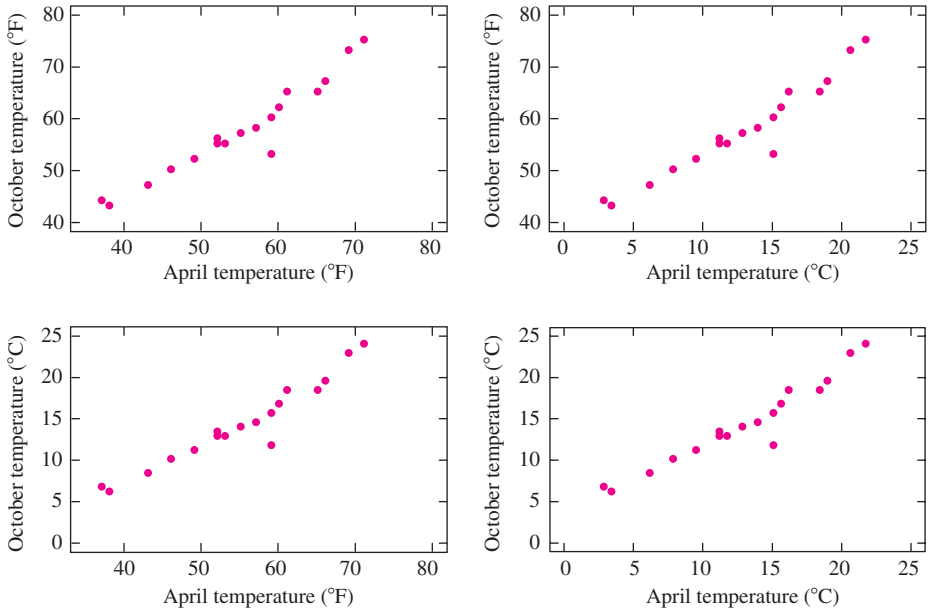

FIGURE 7.6 Mean April and October temperatures for several U.S. cities. The correlation coefficient is 0.96 for each plot; the choice of units does not matter.

The Correlation Coefficient Measures Only *Linear* Association

An object is fired upward from the ground with an initial velocity of 64 ft/s. At each of several times $x_1, \ldots, x_n$, the heights $y_1, \ldots, y_n$ of the object above the surface of the earth are measured. In the absence of friction, and assuming that there is no measurement error, the scatterplot of the points $(x_1, y_1), \ldots, (x_n, y_n)$ will look like Figure 7.7.

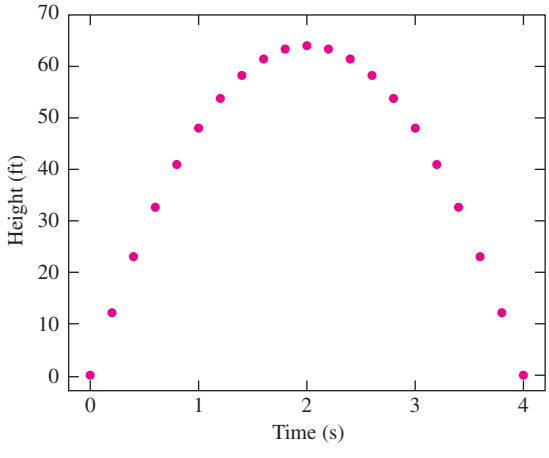

FIGURE 7.7 The relationship between the height of a free-falling object with a positive initial velocity and the time in free fall is quadratic. The correlation is equal to 0.