5.11 It may seem odd, but one of the ways biologists can tell how old a lobster is involves measuring the concentration of a pigment called neurolipofuscin in the eyestalk of a lobster. (We are not making this up!) The authors of the paper "Neurolipofuscin is a Measure of Age in *Panulirus argus,* the Caribbean Spiny Lobster, in Florida" (*Biological Bulletin* [2007]: 55–66) wondered if it was sufficient to measure the pigment in just one eye stalk, which would be the case if there is a strong relationship between the concentration in the right and left eyestalks. Pigment concentration (as a percentage of tissue sample) was measured in both eyestalks for 39 lobsters, resulting is the following summary quantities (based on data read from a graph that appeared in the paper):

$$\begin{array}{rrr} n = 39 & \sum x = 88.8 & \sum y = 86.1 \\ \sum xy = 281.1 & \sum x^2 = 288.0 & \sum y^2 = 286.6 \end{array}$$

An alternative formula for computing the correlation coefficient that is based on raw data and is algebraically equivalent to the one given in the text is

$$r = \frac{\sum xy - \dfrac{(\sum x)(\sum y)}{n}}{\sqrt{\sum x^2 - \dfrac{(\sum x)^2}{n}}\sqrt{\sum y^2 - \dfrac{(\sum y)^2}{n}}}$$

Use this formula to compute the value of the correlation coefficient, and interpret this value.

5.12 An auction house released a list of 25 recently sold paintings. Eight artists were represented in these sales. The sale price of each painting also appears on the list. Would the correlation coefficient be an appropriate way to summarize the relationship between artist (x) and sale price (y)? Why or why not?

5.13 A sample of automobiles traversing a certain stretch of highway is selected. Each one travels at roughly a constant rate of speed, although speed does vary from auto to auto. Let x = speed and y = time needed to traverse this segment of highway. Would the sample correlation coefficient be closest to .9, .3, −.3, or −.9? Explain.

Bold exercises answered in back ● Data set available online ✦ Video Solution available

5.2 Linear Regression: Fitting a Line to Bivariate Data

The objective of *regression analysis* is to use information about one variable, x, to draw some sort of conclusion concerning a second variable, y. For example, we might want to predict y = product sales during a given period when the amount spent on advertising is x = \$10,000. The two variables in a regression analysis play different roles: y is called the **dependent** or **response variable**, and x is referred to as the **independent**, **predictor**, or **explanatory variable**.

Scatterplots frequently exhibit a linear pattern. When this is the case, it makes sense to summarize the relationship between the variables by finding a line that is as close as possible to the points in the plot. Before seeing how this is done, let's review some elementary facts about lines and linear relationships.

The equation of a line is $y = a + bx$. A particular line is specified by choosing values of a and b. For example, one line is $y = 10 + 2x$; another is $y = 100 - 5x$. If we choose some x values and compute $y = a + bx$ for each value, the points in the plot of the resulting (x, y) pairs will fall exactly on a straight line.

DEFINITION

The equation of a line is

$$y = \overset{\text{Intercept}}{a} + \underset{\text{Slope}}{b}x$$

The value of b, called the **slope** of the line, is the amount by which y increases when x increases by 1 unit. The value of a, called the **intercept** (or sometimes the **y-intercept** or **vertical intercept**) of the line, is the height of the line above the value $x = 0$.

The line $y = 10 + 2x$ has slope $b = 2$, so each 1-unit increase in x is paired with an increase of 2 in y. When $x = 0$, $y = 10$, so the height at which the line crosses the vertical axis (where $x = 0$) is 10. This is illustrated in Figure 5.8(a). The slope of the line $y = 100 - 5x$ is -5, so y increases by -5 (or equivalently, decreases by 5) when x increases by 1. The height of the line above $x = 0$ is $a = 100$. The resulting line is pictured in Figure 5.8(b).

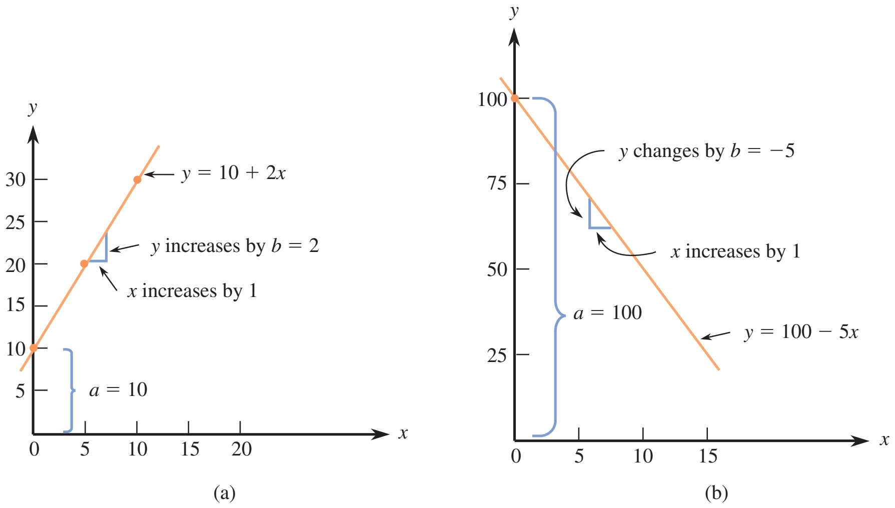


FIGURE 5.8
Graphs of two lines: (a) slope $b = 2$, intercept $a = 10$; (b) slope $b = -5$, intercept $a = 100$.

It is easy to draw the line corresponding to any particular linear equation. Choose any two x values and substitute them into the equation to obtain the corresponding y values. Then plot the resulting two (x, y) pairs as two points. The desired line is the one passing through these points. For the equation $y = 10 + 2x$, substituting $x = 5$ yields $y = 20$, whereas using $x = 10$ gives $y = 30$. The resulting two points are then (5, 20) and (10, 30). The line in Figure 5.8(a) passes through these points.

Fitting a Straight Line: The Principle of Least Squares

Figure 5.9 shows a scatterplot with two lines superimposed on the plot. Line II is a better fit to the data than Line I is. In order to measure the extent to which a particular line provides a good fit to data, we focus on the vertical deviations from the line. For example, Line II in Figure 5.9 has equation $y = 10 + 2x$, and the third and fourth points from the left in the scatterplot are (15, 44) and (20, 45). For these two points, the vertical deviations from this line are

$$\begin{aligned}\text{3rd deviation} &= y_3 - \text{height of the line above } x_3\\ &= 44 - [10 + 2(15)]\\ &= 4\end{aligned}$$

and

$$\text{4th deviation} = 45 - [10 + 2(20)] = -5$$

A positive vertical deviation results from a point that lies above the chosen line, and a negative deviation results from a point that lies below this line. A particular line is said to be a good fit to the data if the deviations from the line are small in magnitude. Line I in Figure 5.9 fits poorly, because all deviations from that line are larger in magnitude (some are much larger) than the corresponding deviations from Line II.

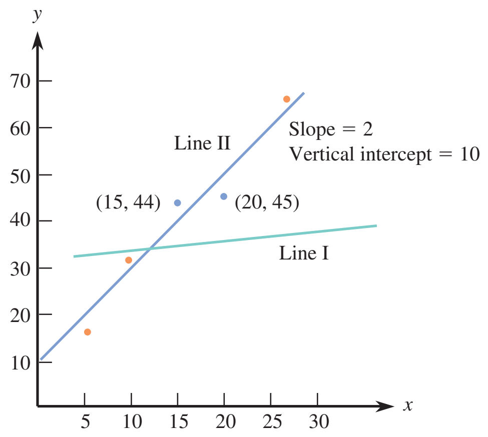


FIGURE 5.9
Line I gives a poor fit and Line II gives a good fit to the data.

To assess the overall fit of a line, we need a way to combine the n deviations into a single measure of fit. The standard approach is to square the deviations (to obtain nonnegative numbers) and then to sum these squared deviations.

DEFINITION

The most widely used measure of the goodness of fit of a line $y = a + bx$ to bivariate data $(x_1, y_1), \ldots, (x_n, y_n)$ is the **sum of the squared deviations** about the line

$$\Sigma[y - (a + bx)]^2 = [y_1 - (a + bx_1)]^2 + [y_2 - (a + bx_2)]^2 + \cdots + [y_n - (a + bx_n)]^2$$

The **least-squares line**, also called the **sample regression line**, is the line that minimizes this sum of squared deviations.

Fortunately, the equation of the least-squares line can be obtained without having to calculate deviations from any particular line. The accompanying box gives relatively simple formulas for the slope and intercept of the least-squares line.

The slope of the least-squares line is

$$b = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2}$$

and the y intercept is

$$a = \bar{y} - b\bar{x}$$

We write the equation of the least-squares line as

$$\hat{y} = a + bx$$

where the ˆ above y indicates that $\hat{y}$ (read as y-hat) is the prediction of y that results from substituting a particular x value into the equation.

Statistical software packages and many calculators can compute the slope and intercept of the least-squares line. If the slope and intercept are to be computed by hand, the following computational formula can be used to reduce the amount of time required to perform the calculations.

Calculating Formula for the Slope of the Least-Squares Line

$$b = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

EXAMPLE 5.5 Pomegranate Juice and Tumor Growth

● Pomegranate, a fruit native to Persia, has been used in the folk medicines of many cultures to treat various ailments. Researchers are now studying pomegranate's antioxidant properties to see if it might have any beneficial effects in the treatment of cancer. One such study, described in the paper "Pomegranate Fruit Juice for Chemoprevention and Chemotherapy of Prostate Cancer" (*Proceedings of the National Academy of Sciences* [October 11, 2005]: 14813–14818), investigated whether pomegranate fruit extract (PFE) was effective in slowing the growth of prostate cancer tumors. In this study, 24 mice were injected with cancer cells. The mice were then randomly assigned to one of three treatment groups. One group of eight mice received normal drinking water, the second group of eight mice received drinking water supplemented with .1% PFE, and the third group received drinking water supplemented with .2% PFE. The average tumor volume for the mice in each group was recorded at several points in time. The accompanying data on y = average tumor volume (in mm^3) and x = number of days after injection of cancer cells for the mice that received plain drinking water was approximated from a graph that appeared in the paper:

x	11	15	19	23	27
y	150	270	450	580	740

A scatterplot of these data (Figure 5.10) shows that the relationship between number of days after injection of cancer cells and average tumor volume could reasonably be summarized by a straight line.

● Data set available online

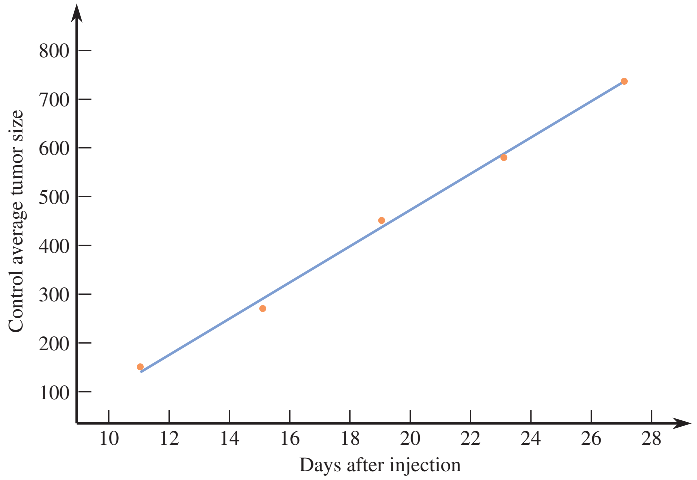


FIGURE 5.10
Minitab scatterplot for the data of Example 5.5.

The summary quantities necessary to compute the equation of the least-squares line are

$$\sum x = 95 \qquad \sum x^2 = 1965 \qquad \sum xy = 47{,}570$$
$$\sum y = 2190 \qquad \sum y^2 = 1{,}181{,}900$$

From these quantities, we compute

$$\bar{x} = 19 \qquad \bar{y} = 438$$

$$b = \frac{\sum xy - \dfrac{(\sum x)(\sum y)}{n}}{\sum x^2 - \dfrac{(\sum x)^2}{n}} = \frac{47{,}570 - \dfrac{(95)(2190)}{5}}{1965 - \dfrac{(95)^2}{5}} = \frac{5960}{160} = 37.25$$

and

$$a = \bar{y} - b\bar{x} = 438 - (37.25)(19) = -269.75$$

The least-squares line is then

$$\hat{y} = -269.75 + 37.25x$$

This line is also shown on the scatterplot of Figure 5.10.

If we wanted to predict average tumor volume 20 days after injection of cancer cells, we could use the y value of the point on the least-squares line above $x = 20$:

$$\hat{y} = -269.75 + 37.25(20) = 475.25$$

Predicted average tumor volume for other numbers of days after injection of cancer cells could be computed in a similar way.

But, be careful in making predictions—the least-squares line should not be used to predict average tumor volume for times much outside the range 11 to 27 days (the range of x values in the data set) because we do not know whether the linear pattern observed in the scatterplot continues outside this range. This is sometimes referred to as the **danger of extrapolation.**

In this example, we can see that using the least-squares line to predict average tumor volume for fewer than 10 days after injection of cancer cells can lead to nonsensical predictions. For example, if the number of days after injection is five the predicted average tumor volume is negative:

$$\hat{y} = -269.75 + 37.25(5) = -83.5$$

Because it is impossible for average tumor volume to be negative, this is a clear indication that the pattern observed for x values in the 11 to 27 range does not continue outside this range. Nonetheless, the least-squares line can be a useful tool for making predictions for x values within the 11- to 27-day range.

Figure 5.11 shows a scatterplot for average tumor volume versus number of days after injection of cancer cells for both the group of mice that drank only water and the group that drank water supplemented by .2% PFE. Notice that the tumor growth seems to be much slower for the mice that drank water supplemented with PFE. For the .2% PFE group, the relationship between average tumor volume and number of days after injection of cancer cells appears to be curved rather than linear. We will see in Section 5.4 how a curve (rather than a straight line) can be used to summarize this relationship.

Calculations involving the least-squares line can obviously be tedious. This is when the computer or a graphing calculator comes to our rescue. All the standard statistical packages can fit a straight line to bivariate data.

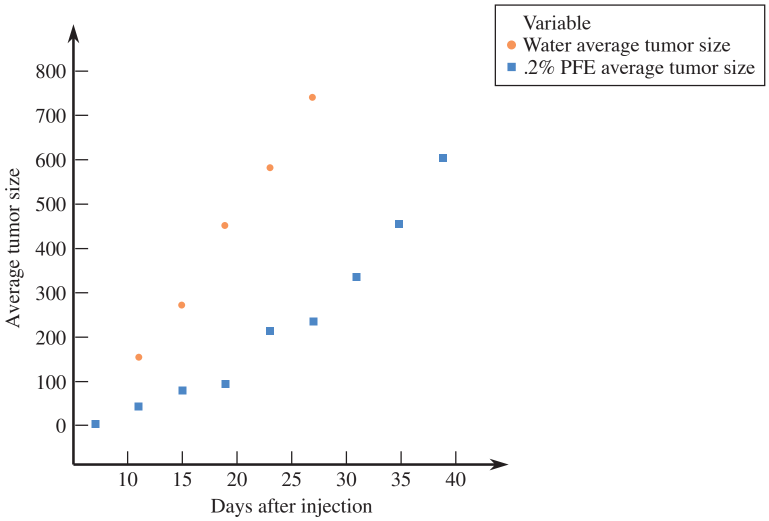


FIGURE 5.11
Scatterplot of average tumor volume versus number of days after injection of cancer cells for the water group and the .2% PFE group.

USE CAUTION—The Danger of Extrapolation

The least-squares line should not be used to make predictions outside the range of the *x* values in the data set because we have no evidence that the linear relationship continues outside this range.

EXAMPLE 5.6 Revisiting the Tannin Concentration Data

Data on x = tannin concentration and y = perceived astringency for n = 32 red wines was given in Example 5.2. In that example, we saw that the correlation coefficient was 0.916, indicating a strong positive linear relationship. This linear relationship can be summarized using the least-squares line, as shown in Figure 5.12.

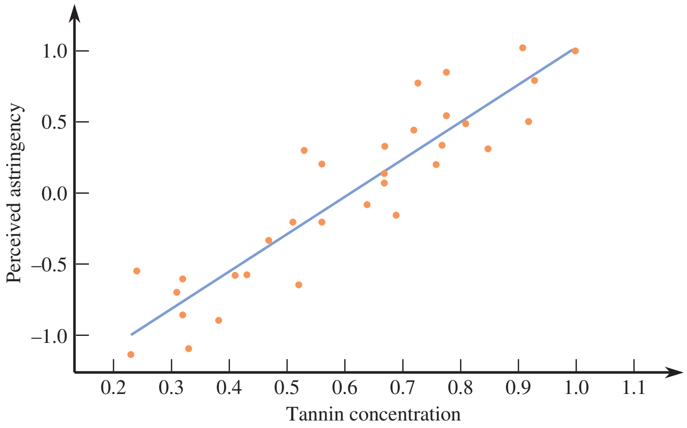
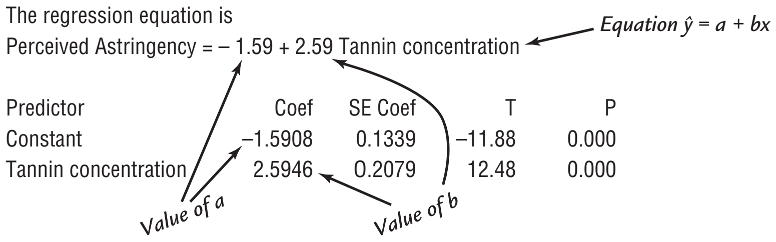


FIGURE 5.12
Scatterplot and least-squares line for the data of Example 5.6.

Minitab was used to fit the least-squares line, and Figure 5.13 shows part of the resulting output. Instead of x and y, the variable labels "Perceived Astringency" and "Tannin Concentration" are used. The equation at the top is that of the least-squares line. In the rectangular table just below the equation, the first row gives information about the intercept, a, and the second row gives information concerning the slope, b. In particular, the coefficient column labeled "Coef" contains the values of a and b using more digits than in the rounded values that appear in the equation.



```
The regression equation is
Perceived Astringency = – 1.59 + 2.59 Tannin concentration

Predictor                  Coef   SE Coef       T       P
Constant                –1.5908    0.1339  –11.88   0.000
Tannin concentration     2.5946    0.2079   12.48   0.000
```

Equation $\hat{y} = a + bx$

Value of a

Value of b

FIGURE 5.13
Partial Minitab output for Example 5.6.

The least-squares line should not be used to predict the perceived astringency for wines with tannin concentrations such as $x = 0.10$ or $x = 0.15$. These x values are well outside the range of the data, and we do not know if the linear relationship continues outside the observed range.

Regression

The least-squares line is often called the **sample regression line.** This terminology comes from the relationship between the least-squares line and Pearson's correlation coefficient. To understand this relationship, we first need alternative expressions for the slope b and the equation of the line itself. With s_x and s_y denoting the sample standard deviations of the x's and y's, respectively, a bit of algebraic manipulation gives

$$b = r\left(\frac{s_y}{s_x}\right)$$

$$\hat{y} = \bar{y} + r\left(\frac{s_y}{s_x}\right)(x - \bar{x})$$

You do not need to use these formulas in any computations, but several of their implications are important for appreciating what the least-squares line does.

1. When $x = \bar{x}$ is substituted in the equation of the line, $\hat{y} = \bar{y}$ results. That is, the least-squares line passes through the *point of averages* $(\bar{x}, \bar{y})$.
2. Suppose for the moment that $r = 1$, so that all points lie exactly on the line whose equation is

   $$\hat{y} = \bar{y} + \frac{s_y}{s_x}(x - \bar{x})$$

   Now substitute $x = \bar{x} + s_x$, which is 1 standard deviation above $\bar{x}$:

   $$\hat{y} = \bar{y} + \frac{s_y}{s_x}(\bar{x} + s_x - \bar{x}) = \bar{y} + s_y$$

That is, with $r = 1$, when x is 1 standard deviation above its mean, we predict that the associated y value will be 1 standard deviation above its mean. Similarly, if $x = \bar{x} - 2s_x$ (2 standard deviations below its mean), then

$$\hat{y} = \bar{y} + \frac{s_y}{s_x}(\bar{x} - 2s_x - \bar{x}) = \bar{y} - 2s_y$$

which is also 2 standard deviations below the mean. If $r = -1$, then $x = \bar{x} + s_x$ results in $\hat{y} = \bar{y} - s_y$, so the predicted y is also 1 standard deviation from its mean but on the opposite side of $\bar{y}$ from where x is relative to $\bar{x}$. In general, if x and y are perfectly correlated, the predicted y value associated with a given x value will be the same number of standard deviations (of y) from its mean $\bar{y}$ as x is from its mean $\bar{x}$.

3. Now suppose that x and y are not perfectly correlated. For example, suppose $r = .5$, so the least-squares line has the equation

$$\hat{y} = \bar{y} + .5\left(\frac{s_y}{s_x}\right)(x - \bar{x})$$

Then substituting $x = \bar{x} + s_x$ gives

$$\hat{y} = \bar{y} + .5\left(\frac{s_y}{s_x}\right)(\bar{x} + s_x - \bar{x}) = \bar{y} + .5s_y$$

That is, for $r = .5$, when x lies 1 standard deviation above its mean, we predict that y will be only 0.5 standard deviation above its mean. Similarly, we can predict y when r is negative. If $r = -.5$, then the predicted y value will be only half the number of standard deviations from $\bar{y}$ that x is from $\bar{x}$ but x and the predicted y will now be on opposite sides of their respective means.

Consider using the least-squares line to predict the value of y associated with an x value some specified number of standard deviations away from $\bar{x}$. Then the predicted y value will be only r times this number of standard deviations from $\bar{y}$. In terms of standard deviations, except when $r = 1$ or -1, the predicted y will always be closer to $\bar{y}$ than x is to $\bar{x}$.

Using the least-squares line for prediction results in a predicted y that is pulled back in, or regressed, toward the mean of y compared to where x is relative to the mean of x. This regression effect was first noticed by Sir Francis Galton (1822–1911), a famous biologist, when he was studying the relationship between the heights of fathers and their sons. He found that predicted heights of sons whose fathers were above average in height were also above average (because r is positive here) but not by as much as the father's height; he found a similar relationship for fathers whose heights were below average. This regression effect has led to the term **regression analysis** for the collection of methods involving the fitting of lines, curves, and more complicated functions to bivariate and multivariate data.

The alternative form of the regression (least-squares) line emphasizes that predicting y from knowledge of x is not the same problem as predicting x from knowledge of y. The slope of the least-squares line for predicting x is $r(s_x/s_y)$ rather than $r(s_y/s_x)$ and the intercepts of the lines are almost always different. For purposes of prediction, it makes a difference whether y is regressed on x, as we have done, or x is regressed on y. *The regression line of y on x should not be used to predict x, because it is not the line that minimizes the sum of squared deviations in the x direction.*