Regression

Linear (straight-line) relationships between two quantitative variables are easy to understand and quite common. In Chapter 4, we found linear relationships in settings as varied as counting carnivores, icicle growth, and heating a home. Correlation measures the direction and strength of these relationships. When a scatterplot shows a linear relationship, we would like to summarize the overall pattern by drawing a line on the scatterplot.

**Regression lines**

A regression line summarizes the relationship between two variables, but only in a specific setting: one of the variables helps explain or predict the other. That is, regression describes a relationship between an explanatory variable and a response variable.

**Regression Line**

A regression line is a straight line that describes how a response variable $y$ changes as an explanatory variable $x$ changes. We often use a regression line to predict the value of $y$ for a given value of $x$. 

In this chapter we cover...

Regression lines
The least-squares regression line
Using technology
Facts about least-squares regression
Residuals
Influential observations
Cautions about correlation and regression
Association does not imply causation
EXAMPLE 5.1  Does fidgeting keep you slim?

Obesity is a growing problem around the world. Here, following our four-step process (page 53), is an account of a study that sheds some light on gaining weight.

STATE: Some people don’t gain weight even when they overeat. Perhaps fidgeting and other “nonexercise activity” (NEA) explains why—some people may spontaneously increase nonexercise activity when fed more. Researchers deliberately overfed 16 healthy young adults for 8 weeks. They measured fat gain (in kilograms) and, as an explanatory variable, change in energy use (in calories) from activity other than deliberate exercise—fidgeting, daily living, and the like. Here are the data:

<table>
<thead>
<tr>
<th>NEA change (cal)</th>
<th>−94</th>
<th>−57</th>
<th>−29</th>
<th>135</th>
<th>143</th>
<th>151</th>
<th>245</th>
<th>355</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fat gain (kg)</td>
<td>4.2</td>
<td>3.0</td>
<td>3.7</td>
<td>2.7</td>
<td>3.2</td>
<td>3.6</td>
<td>2.4</td>
<td>1.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NEA change (cal)</th>
<th>392</th>
<th>473</th>
<th>486</th>
<th>535</th>
<th>571</th>
<th>580</th>
<th>620</th>
<th>690</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fat gain (kg)</td>
<td>3.8</td>
<td>1.7</td>
<td>1.6</td>
<td>2.2</td>
<td>1.0</td>
<td>0.4</td>
<td>2.3</td>
<td>1.1</td>
</tr>
</tbody>
</table>

Do people with larger increases in NEA tend to gain less fat?

FORMULATE: Make a scatterplot of the data and examine the pattern. If it is linear, use correlation to measure its strength and draw a regression line on the scatterplot to predict fat gain from change in NEA.

FIGURE 5.1  Weight gain after 8 weeks of overeating, plotted against increase in nonexercise activity over the same period.
**SOLVE:** Figure 5.1 is a scatterplot of these data. The plot shows a moderately strong negative linear association with no outliers. The correlation is \( r = -0.7786 \). The line on the plot is a regression line for predicting fat gain from change in NEA.

**CONCLUDE:** People with larger increases in nonexercise activity do indeed gain less fat. To add to this conclusion, we must study regression lines in more detail.

We can, however, already use the regression line to predict fat gain from NEA. Suppose that an individual’s NEA increases by 400 calories when she overeats. Go “up and over” on the graph in Figure 5.1. From 400 calories on the \( x \) axis, go up to the regression line and then over to the \( y \) axis. The graph shows that the predicted gain in fat is a bit more than 2 kilograms.

Many calculators and software programs will give you the equation of a regression line from keyed-in data. Understanding and using the line is more important than the details of where the equation comes from.

**REVIEW OF STRAIGHT LINES**

Suppose that \( y \) is a response variable (plotted on the vertical axis) and \( x \) is an explanatory variable (plotted on the horizontal axis). A straight line relating \( y \) to \( x \) has an equation of the form

\[
y = a + bx
\]

In this equation, \( b \) is the **slope**, the amount by which \( y \) changes when \( x \) increases by one unit. The number \( a \) is the **intercept**, the value of \( y \) when \( x = 0 \).

**EXAMPLE 5.2 Using a regression line**

Any straight line describing the NEA data has the form

\[
fat \text{ gain} = a + (b \times \text{NEA change})
\]

The line in Figure 5.1 is the regression line with the equation

\[
fat \text{ gain} = 3.505 - (0.00344 \times \text{NEA change})
\]

Be sure you understand the role of the two numbers in this equation:

- The slope \( b = -0.00344 \) tells us that fat gained goes down by 0.00344 kilogram for each added calorie of NEA. The slope of a regression line is the **rate of change** in the response as the explanatory variable changes.
- The intercept, \( a = 3.505 \) kilograms, is the estimated fat gain if NEA does not change when a person overeats.

The slope of a regression line is an important numerical description of the relationship between the two variables. Although we need the value of the intercept to draw the line, this value is statistically meaningful only when, as in this example, the explanatory variable can actually take values close to zero.
The equation of the regression line makes it easy to predict fat gain. If a person’s NEA increases by 400 calories when she overeats, substitute \( x = 400 \) in the equation. The predicted fat gain is

\[
\text{fat gain} = 3.505 - (0.00344 \times 400) = 2.13 \text{ kilograms}
\]

To plot the line on the scatterplot, use the equation to find the predicted \( y \) for two values of \( x \), one near each end of the range of \( x \) in the data. Plot each \( y \) above its \( x \) and draw the line through the two points.

The slope \( b = -0.00344 \) in Example 5.2 is small. This does not mean that change in NEA has little effect on fat gain. The size of the slope depends on the units in which we measure the two variables. In this example, the slope is the change in fat gain in kilograms when NEA increases by one calorie. There are 1000 grams in a kilogram. If we measured fat gain in grams, the slope would be 1000 times larger, \( b = 3.44 \). You can’t say how important a relationship is by looking at the size of the slope of the regression line.

**APPLY YOUR KNOWLEDGE**

5.1 **IQ and reading scores.** Data on the IQ test scores and reading test scores for a group of fifth-grade children give the regression line

\[
\text{reading score} = -33.4 + (0.882 \times \text{IQ score})
\]

for predicting reading score from IQ score.

(a) Say in words what the slope of this line tells you.

(b) Explain why the value of the intercept is not statistically meaningful.

(c) Find the predicted reading scores for children with IQ scores 90 and 130.

(d) Draw a graph of the regression line for IQs between 90 and 130. (Be sure to show the scales for the \( x \) and \( y \) axes.)

5.2 **The equation of a line.** An eccentric professor believes that a child with IQ 100 should have reading score 50, and that reading score should increase by 1 point for every additional point of IQ. What is the equation of the professor’s regression line for predicting reading score from IQ?

The least-squares regression line

In most cases, no line will pass exactly through all the points in a scatterplot. Different people will draw different lines by eye. We need a way to draw a regression line that doesn’t depend on our guess as to where the line should go. Because we use the line to predict \( y \) from \( x \), the prediction errors we make are errors in \( y \), the vertical direction in the scatterplot. A good regression line makes the vertical distances of the points from the line as small as possible.

Figure 5.2 illustrates the idea. This plot shows three of the points from Figure 5.1, along with the line, on an expanded scale. The line passes above one of the points and below two of them. The three prediction errors appear as vertical line
The least-squares regression line

The least-squares regression line of $y$ on $x$ is the line that makes the sum of the squares of the vertical distances of the data points from the line as small as possible.

**FIGURE 5.2** The least-squares idea. For each observation, find the vertical distance of each point on the scatterplot from a regression line. The least-squares regression line makes the sum of the squares of these distances as small as possible.

segments. For example, one subject had $x = -57$, a decrease of 57 calories in NEA. The line predicts a fat gain of 3.7 kilograms, but the actual fat gain for this subject was 3.0 kilograms. The prediction error is

$$\text{error} = \text{observed response} - \text{predicted response} = 3.0 - 3.7 = -0.7 \text{ kilogram}$$

There are many ways to make the collection of vertical distances “as small as possible.” The most common is the least-squares method.
One reason for the popularity of the least-squares regression line is that the problem of finding the line has a simple answer. We can give the equation for the least-squares line in terms of the means and standard deviations of the two variables and the correlation between them.

**EQUATION OF THE LEAST-SQUARES REGRESSION LINE**

We have data on an explanatory variable $x$ and a response variable $y$ for $n$ individuals. From the data, calculate the means $\bar{x}$ and $\bar{y}$ and the standard deviations $s_x$ and $s_y$ of the two variables, and their correlation $r$. The least-squares regression line is the line

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

with slope

$$b = r \frac{s_y}{s_x}$$

and intercept

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

We write $\hat{y}$ (read “y hat”) in the equation of the regression line to emphasize that the line gives a predicted response $\hat{y}$ for any $x$. Because of the scatter of points about the line, the predicted response will usually not be exactly the same as the actually observed response $y$. In practice, you don’t need to calculate the means, standard deviations, and correlation first. Software or your calculator will give the slope $b$ and intercept $a$ of the least-squares line from the values of the variables $x$ and $y$. You can then concentrate on understanding and using the regression line.

**Using technology**

Least-squares regression is one of the most common statistical procedures. Any technology you use for statistical calculations will give you the least-squares line and related information. Figure 5.3 displays the regression output for the data of Examples 5.1 and 5.2 from a graphing calculator, two statistical programs, and a spreadsheet program. Each output records the slope and intercept of the least-squares line. The software also provides information that we do not yet need, although we will use much of it later. (In fact, we left out part of the Minitab and Excel outputs.) Be sure that you can locate the slope and intercept on all four outputs. Once you understand the statistical ideas, you can read and work with almost any software output.
The regression equation is
fat = 3.505122916 - 0.003441487

*FIGURE 5.3* Least-squares regression for the nonexercise activity data: output from a graphing calculator, two statistical programs, and a spreadsheet program (continued).
Microsoft Excel

![Microsoft Excel Table]

**FIGURE 5.3** (continued)

**APPLY YOUR KNOWLEDGE**

5.3 **Verify our claims.** Example 5.2 gives the equation of the regression line of fat gain y on change in NEA x as

\[ \hat{y} = 3.505 - 0.00344x \]

Enter the data from Example 5.1 into your software or calculator.

(a) Use the regression function to find the equation of the least-squares regression line.

(b) Also find the mean and standard deviation of both x and y and their correlation r. Calculate the slope b and intercept a of the regression line from these, using the facts in the box Equation of the Least-Squares Regression Line. Verify that in both part (a) and part (b) you get the equation in Example 5.2. (Results may differ slightly because of rounding off.)

5.4 **Bird colonies.** One of nature’s patterns connects the percent of adult birds in a colony that return from the previous year and the number of new adults that join the colony. Here are data for 13 colonies of sparrowhawks:

<table>
<thead>
<tr>
<th>Percent return</th>
<th>74</th>
<th>66</th>
<th>81</th>
<th>52</th>
<th>73</th>
<th>62</th>
<th>52</th>
<th>45</th>
<th>62</th>
<th>46</th>
<th>60</th>
<th>46</th>
<th>38</th>
</tr>
</thead>
<tbody>
<tr>
<td>New adults</td>
<td>5</td>
<td>6</td>
<td>8</td>
<td>11</td>
<td>12</td>
<td>15</td>
<td>16</td>
<td>17</td>
<td>18</td>
<td>18</td>
<td>19</td>
<td>20</td>
<td>20</td>
</tr>
</tbody>
</table>

As you saw in Exercise 4.4 (page 93), there is a linear relationship between the percent x of adult sparrowhawks that return to a colony from the previous year and the number y of new adult birds that join the colony.
(a) Find the correlation $r$ for these data. The straight-line pattern is moderately strong.

(b) Find the least-squares regression line for predicting $y$ from $x$. Make a scatterplot and draw your line on the plot.

(c) Explain in words what the slope of the regression line tells us.

(d) An ecologist uses the line, based on 13 colonies, to predict how many birds will join another colony, to which 60% of the adults from the previous year return. What is the prediction?

---

Facts about least-squares regression

One reason for the popularity of least-squares regression lines is that they have many convenient special properties. Here are some facts about least-squares regression lines.

**Fact 1.** The distinction between explanatory and response variables is essential in regression. Least-squares regression makes the distances of the data points from the line small only in the $y$ direction. If we reverse the roles of the two variables, we get a different least-squares regression line.

**Example 5.3 Predicting fat, predicting NEA**

Figure 5.4 repeats the scatterplot of the nonexercise activity data in Figure 5.1, but with two least-squares regression lines. The solid line is the regression line for predicting fat gain from change in NEA. This is the line that appeared in Figure 5.1.

We might also use the data on these 16 subjects to predict the change in NEA for another subject from that subject’s fat gain when overfed for 8 weeks. Now the roles of the variables are reversed: fat gain is the explanatory variable and change in NEA is the response variable. The dashed line in Figure 5.4 is the least-squares line for predicting NEA change from fat gain. The two regression lines are not the same. In the regression setting, you must know clearly which variable is explanatory.

**Fact 2.** There is a close connection between correlation and the slope of the least-squares line. The slope is

$$b = r \frac{s_y}{s_x}$$

This equation says that along the regression line, a change of one standard deviation in $x$ corresponds to a change of $r$ standard deviations in $y$. When the variables are perfectly correlated ($r = 1$ or $r = -1$), the change in the predicted response $\hat{y}$ is the same (in standard deviation units) as the change in $x$. Otherwise, because $-1 \leq r \leq 1$, the change in $\hat{y}$ is less than the change in $x$. As the correlation grows less strong, the prediction $\hat{y}$ moves less in response to changes in $x$.

**Fact 3.** The least-squares regression line always passes through the point $(\bar{x}, \bar{y})$ on the graph of $y$ against $x$. 
CHAPTER 5 • Regression

Regression toward the mean

To “regress” means to go backward. Why are statistical methods for predicting a response from an explanatory variable called “regression”? Sir Francis Galton (1822–1911), who was the first to apply regression to biological and psychological data, looked at examples such as the heights of children versus the heights of their parents. He found that the taller-than-average parents tended to have children who were also taller than average but not as tall as their parents. Galton called this fact “regression toward the mean,” and the name came to be applied to the statistical method.

Fact 4. The correlation $r$ describes the strength of a straight-line relationship. In the regression setting, this description takes a specific form: the square of the correlation, $r^2$, is the fraction of the variation in the values of $y$ that is explained by the least-squares regression of $y$ on $x$.

The idea is that when there is a linear relationship, some of the variation in $y$ is accounted for by the fact that as $x$ changes it pulls $y$ along with it. Look again at Figure 5.4, the scatterplot of the NEA data. The variation in $y$ appears as the spread of fat gains from 0.4 kg to 4.2 kg. Some of this variation is explained by the fact that $x$ (change in NEA) varies from a loss of 94 calories to a gain of 690 calories. As $x$ moves from −94 to 690, it pulls $y$ along the solid regression line. You would predict a smaller fat gain for a subject whose NEA increased by 600 calories than for someone with 0 change in NEA. But the straight-line tie of $y$ to $x$ doesn’t explain all of the variation in $y$. The remaining variation appears as the scatter of points above and below the line.

Although we won’t do the algebra, it is possible to break the variation in the observed values of $y$ into two parts. One part measures the variation in $\hat{y}$ as $x$ moves and pulls $\hat{y}$ with it along the regression line. The other measures the vertical scatter of points.
Facts about least-squares regression

of the data points above and below the line. The squared correlation $r^2$ is the first of these as a fraction of the whole:

$$r^2 = \frac{\text{variation in } \hat{y} \text{ as } x \text{ pulls it along the line}}{\text{total variation in observed values of } y}$$

**Example 5.4 Using $r^2$**

For the NEA data, $r = -0.7786$ and $r^2 = 0.6062$. About 61% of the variation in fat gained is accounted for by the linear relationship with change in NEA. The other 39% is individual variation among subjects that is not explained by the linear relationship.

Figure 4.2 (page 96) shows a stronger linear relationship in which the points are more tightly concentrated along a line. Here, $r = -0.9124$ and $r^2 = 0.8325$. More than 83% of the variation in carnivore abundance is explained by regression on body mass. Only 17% is variation among species with the same mass.

When you report a regression, give $r^2$ as a measure of how successful the regression was in explaining the response. Three of the outputs in Figure 5.3 include $r^2$, either in decimal form or as a percent. (CrunchIt! gives $r$ instead.) When you see a correlation, square it to get a better feel for the strength of the association. Perfect correlation ($r = -1$ or $r = 1$) means the points lie exactly on a line. Then $r^2 = 1$ and all of the variation in one variable is accounted for by the linear relationship with the other variable. If $r = -0.7$ or $r = 0.7$, $r^2 = 0.49$ and about half the variation is accounted for by the linear relationship. In the $r^2$ scale, correlation ±0.7 is about halfway between 0 and ±1.

Facts 2, 3, and 4 are special properties of least-squares regression. They are not true for other methods of fitting a line to data.

**Apply Your Knowledge**

5.5 Growing corn. Exercise 4.28 (page 110) gives data from an agricultural experiment. The purpose of the study was to see how the yield of corn changes as we change the planting rate (plants per acre).

(a) Make a scatterplot of the data. (Use a scale of yields from 100 to 200 bushels per acre.) Find the least-squares regression line for predicting yield from planting rate and add this line to your plot. Why should we not use the regression line for prediction in this setting?

(b) What is $r^2$? What does this value say about the success of the regression line in predicting yield?

(c) Even regression lines that make no practical sense obey Facts 2, 3, and 4. Use the equation of the regression line you found in (a) to show that when $x$ is the mean planting rate, the predicted yield $\hat{y}$ is the mean of the observed yields.

5.6 How useful is regression? Figure 4.7 (page 107) displays the returns on common stocks and Treasury bills over a period of more than 50 years. The correlation is $r = -0.113$. Exercise 4.27 (page 110) gives data on outside temperature and natural gas used by a home during the heating season. The correlation is $r = 0.995$. Explain in simple language why knowing only these correlations enables you to say that prediction of gas used from outside
temperature will be much more accurate than prediction of return on stocks from
return on T-bills.

**Residuals**

One of the first principles of data analysis is to look for an overall pattern and also
for striking deviations from the pattern. A regression line describes the overall
pattern of a linear relationship between an explanatory variable and a response
variable. We see deviations from this pattern by looking at the scatter of the data
points about the regression line. The vertical distances from the points to the least-
squares regression line are as small as possible, in the sense that they have the
smallest possible sum of squares. Because they represent “left-over” variation in
the response after fitting the regression line, these distances are called *residuals*.

**RESIDUALS**

A residual is the difference between an observed value of the response
variable and the value predicted by the regression line. That is, a residual is
the prediction error that remains after we have chosen the regression line:

\[
\text{residual} = \text{observed } y - \text{predicted } y = y - \hat{y}
\]

**EXAMPLE 5.5  I feel your pain**

“Empathy” means being able to understand what others feel. To see how the brain ex-
presses empathy, researchers recruited 16 couples in their midtwenties who were married
or had been dating for at least two years. They zapped the man’s hand with an electrode
while the woman watched, and measured the activity in several parts of the woman’s
brain that would respond to her own pain. Brain activity was recorded as a fraction of the
activity observed when the woman herself was zapped with the electrode. The women
also completed a psychological test that measures empathy. Will women who are higher
in empathy respond more strongly when their partner has a painful experience? Here
are data for one brain region:

<table>
<thead>
<tr>
<th>Subject</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Empathy score</td>
<td>38</td>
<td>53</td>
<td>41</td>
<td>55</td>
<td>56</td>
<td>61</td>
<td>62</td>
<td>48</td>
</tr>
<tr>
<td>Brain activity</td>
<td>-0.120</td>
<td>0.392</td>
<td>0.005</td>
<td>0.369</td>
<td>0.016</td>
<td>0.415</td>
<td>0.107</td>
<td>0.506</td>
</tr>
<tr>
<td>Subject</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>16</td>
</tr>
<tr>
<td>Empathy score</td>
<td>43</td>
<td>47</td>
<td>56</td>
<td>65</td>
<td>19</td>
<td>61</td>
<td>32</td>
<td>105</td>
</tr>
<tr>
<td>Brain activity</td>
<td>0.153</td>
<td>0.745</td>
<td>0.255</td>
<td>0.574</td>
<td>0.210</td>
<td>0.722</td>
<td>0.358</td>
<td>0.779</td>
</tr>
</tbody>
</table>

Figure 5.5 is a scatterplot, with empathy score as the explanatory variable \(x\) and
brain activity as the response variable \(y\). The plot shows a positive association. That is,
women who are higher in empathy do indeed react more strongly to their partner’s pain. The overall pattern is moderately linear, correlation $r = 0.515$.

The line on the plot is the least-squares regression line of brain activity on empathy score. Its equation is

$$\hat{y} = -0.0578 + 0.00761x$$

For Subject 1, with empathy score 38, we predict

$$\hat{y} = -0.0578 + (0.00761)(38) = 0.231$$

This subject’s actual brain activity level was $-0.120$. The residual is

$$\text{residual} = \text{observed } y - \text{predicted } y = -0.120 - 0.231 = -0.351$$

The residual is negative because the data point lies below the regression line. The dashed line segment in Figure 5.5 shows the size of the residual.

There is a residual for each data point. Finding the residuals is a bit unpleasant because you must first find the predicted response for every $x$. Software or a graphing calculator gives you the residuals all at once. Following are the 16 residuals for the empathy study data, from software:
residuals:
-0.3515 -0.2494 -0.3526 -0.3072 -0.1166 -0.1136 0.1231 0.1721
 0.0463 0.0080 0.0084 0.1983 0.4449 0.1369 0.3154 0.0374

Because the residuals show how far the data fall from our regression line, examining the residuals helps assess how well the line describes the data. Although residuals can be calculated from any curve fitted to the data, the residuals from the least-squares line have a special property: the mean of the least-squares residuals is always zero.

Compare the scatterplot in Figure 5.5 with the residual plot for the same data in Figure 5.6. The horizontal line at zero in Figure 5.6 helps orient us. This “residual = 0” line corresponds to the regression line in Figure 5.5.

![Residual plot](image)

**FIGURE 5.6** Residual plot for the data shown in Figure 5.5. The horizontal line at zero residual corresponds to the regression line in Figure 5.5.

**RESIDUAL PLOTS**

A residual plot is a scatterplot of the regression residuals against the explanatory variable. Residual plots help us assess how well a regression line fits the data.
Influential observations

A residual plot in effect turns the regression line horizontal. It magnifies the deviations of the points from the line and makes it easier to see unusual observations and patterns.

A residual plot in effect turns the regression line horizontal. It magnifies the deviations of the points from the line and makes it easier to see unusual observations and patterns.

**APPLY YOUR KNOWLEDGE**

5.7 Does fast driving waste fuel? Exercise 4.6 (page 96) gives data on the fuel consumption \( y \) of a car at various speeds \( x \). Fuel consumption is measured in liters of gasoline per 100 kilometers driven, and speed is measured in kilometers per hour. Software tells us that the equation of the least-squares regression line is

\[
\hat{y} = 11.058 - 0.01466x
\]

Using this line we can add the residuals to the original data:

<table>
<thead>
<tr>
<th>Speed</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>60</th>
<th>70</th>
<th>80</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fuel</td>
<td>21.00</td>
<td>13.00</td>
<td>10.00</td>
<td>8.00</td>
<td>7.00</td>
<td>5.90</td>
<td>6.30</td>
<td>6.95</td>
</tr>
<tr>
<td>Residual</td>
<td>10.09</td>
<td>2.24</td>
<td>-0.62</td>
<td>-2.47</td>
<td>-3.33</td>
<td>-4.28</td>
<td>-3.73</td>
<td>-2.94</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Speed</th>
<th>90</th>
<th>100</th>
<th>110</th>
<th>120</th>
<th>130</th>
<th>140</th>
<th>150</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fuel</td>
<td>7.57</td>
<td>8.27</td>
<td>9.03</td>
<td>9.87</td>
<td>10.79</td>
<td>11.77</td>
<td>12.83</td>
</tr>
<tr>
<td>Residual</td>
<td>-2.17</td>
<td>-1.32</td>
<td>-0.42</td>
<td>0.57</td>
<td>1.64</td>
<td>2.76</td>
<td>3.97</td>
</tr>
</tbody>
</table>

(a) Make a scatterplot of the observations and draw the regression line on your plot.

(b) Would you use the regression line to predict \( y \) from \( x \)? Explain your answer.

(c) Verify the value of the first residual, for \( x = 10 \). Verify that the residuals have sum zero (up to roundoff error).

(d) Make a plot of the residuals against the values of \( x \). Draw a horizontal line at height zero on your plot. How does the pattern of the residuals about this line compare with the pattern of the data points about the regression line in the scatterplot in (a)?

**Influential observations**

Figures 5.5 and 5.6 show one unusual observation. Subject 16 is an outlier in the \( x \) direction, with empathy score 40 points higher than any other subject. Because of its extreme position on the empathy scale, this point has a strong influence on the correlation. Dropping Subject 16 reduces the correlation from \( r = 0.515 \) to \( r = 0.331 \). You can see that this point extends the linear pattern in Figure 5.5 and so increases the correlation. We say that Subject 16 is *influential* for calculating the correlation.
INFLUENTIAL OBSERVATIONS

An observation is influential for a statistical calculation if removing it would markedly change the result of the calculation.

Points that are outliers in either the x or y direction of a scatterplot are often influential for the correlation. Points that are outliers in the x direction are often influential for the least-squares regression line.

EXAMPLE 5.6 An influential observation?

Subject 16 is influential for the correlation because removing it greatly reduces the correlation coefficient. Is this observation also influential for the least-squares line? Figure 5.7 shows that it is not. The regression line calculated without Subject 16 (dashed) differs little from the line that uses all of the observations (solid). The reason that the outlier has little influence on the regression line is that it lies close to the dashed regression line calculated from the other observations.

To see why points that are outliers in the x direction are often influential, let's try an experiment. Pull Subject 16’s point in the scatterplot straight down and watch the

**Figure 5.7** Subject 16 is an outlier in the x direction. The outlier is not influential for least-squares regression, because removing it moves the regression line only a little.
Influential observations

FIGURE 5.8 An outlier in the $x$ direction pulls the least-squares line to itself because there are no other observations with similar values of $x$ to hold the line in place. When the outlier moves down, the original regression line (solid) chases it down to the dashed line.

We did not need the distinction between outliers and influential observations in Chapter 2. A single high salary that pulls up the mean salary $\bar{x}$ for a group of workers is an outlier because it lies far above the other salaries. It is also influential, because the mean changes when it is removed. In the regression setting, however, not all outliers are influential.

APPLY YOUR KNOWLEDGE

5.8 Bird colonies. Return to the data of Exercise 5.4 (page 122) on sparrowhawk colonies. We will use these data to illustrate influence.
(a) Make a scatterplot of the data suitable for predicting new adults from percent of returning adults. Then add two new points. Point A: 10% return, 15 new adults. Point B: 60% return, 28 new adults. In which direction is each new point an outlier?

(b) Add three least-squares regression lines to your plot: for the original 13 colonies, for the original colonies plus Point A, and for the original colonies plus Point B. Which new point is more influential for the regression line? Explain in simple language why each new point moves the line in the way your graph shows.

Cautions about correlation and regression

Correlation and regression are powerful tools for describing the relationship between two variables. When you use these tools, you must be aware of their limitations. You already know that

- Correlation and regression lines describe only linear relationships. You can do the calculations for any relationship between two quantitative variables, but the results are useful only if the scatterplot shows a linear pattern.

- Correlation and least-squares regression lines are not resistant. Always plot your data and look for observations that may be influential.

Here are more things to keep in mind when you use correlation and regression.

Beware extrapolation. Suppose that you have data on a child's growth between 3 and 8 years of age. You find a strong linear relationship between age x and height y. If you fit a regression line to these data and use it to predict height at age 25 years, you will predict that the child will be 8 feet tall. Growth slows down and then stops at maturity, so extending the straight line to adult ages is foolish. Few relationships are linear for all values of x. Don't make predictions far outside the range of x that actually appears in your data.

Extrapolation is the use of a regression line for prediction far outside the range of values of the explanatory variable x that you used to obtain the line. Such predictions are often not accurate.

Beware the lurking variable. Another caution is even more important: the relationship between two variables can often be understood only by taking other variables into account. Lurking variables can make a correlation or regression misleading.
**LURKING VARIABLE**

A *lurking variable* is a variable that is not among the explanatory or response variables in a study and yet may influence the interpretation of relationships among those variables.

You should always think about possible lurking variables before you draw conclusions based on correlation or regression.

---

**EXAMPLE 5.7  Magic Mozart?**

The Kalamazoo (Michigan) Symphony once advertised a “Mozart for Minors” program with this statement: “Question: Which students scored 51 points higher in verbal skills and 39 points higher in math? Answer: Students who had experience in music.”

We could as well answer “Students who played soccer.” Why? Children with prosperous and well-educated parents are more likely than poorer children to have experience with music and also to play soccer. They are also likely to attend good schools, get good health care, and be encouraged to study hard. These advantages lead to high test scores. Family background is a lurking variable that explains why test scores are related to experience with music.

---

**APPLY YOUR KNOWLEDGE**

**5.9  The declining farm population.** The number of people living on American farms has declined steadily during the last century. Here are data on the farm population (millions of persons) from 1935 to 1980:

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Population</td>
<td>32.1</td>
<td>30.5</td>
<td>24.4</td>
<td>23.0</td>
<td>19.1</td>
<td>15.6</td>
<td>12.4</td>
<td>9.7</td>
<td>8.9</td>
<td>7.2</td>
</tr>
</tbody>
</table>

(a) Make a scatterplot of these data and find the least-squares regression line of farm population on year.

(b) According to the regression line, how much did the farm population decline each year on the average during this period? What percent of the observed variation in farm population is accounted for by linear change over time?

(c) Use the regression equation to predict the number of people living on farms in 2000. Is this result reasonable? Why?

**5.10  Is math the key to success in college?** A College Board study of 15,941 high school graduates found a strong correlation between how much math minority students took in high school and their later success in college. News articles quoted the head of the College Board as saying that “math is the gatekeeper for success in college.” Maybe so, but we should also think about lurking variables. What might lead minority students to take more or fewer high school math courses? Would these same factors influence success in college?
Association does not imply causation

Thinking about lurking variables leads to the most important caution about correlation and regression. When we study the relationship between two variables, we often hope to show that changes in the explanatory variable cause changes in the response variable. A strong association between two variables is not enough to draw conclusions about cause and effect. Sometimes an observed association really does reflect cause and effect. A household that heats with natural gas uses more gas in colder months because cold weather requires burning more gas to stay warm. In other cases, an association is explained by lurking variables, and the conclusion that \( x \) causes \( y \) is either wrong or not proved.

**EXAMPLE 5.8** Does having more cars make you live longer?

A serious study once found that people with two cars live longer than people who own only one car.\(^6\) Owning three cars is even better, and so on. There is a substantial positive correlation between number of cars \( x \) and length of life \( y \).

The basic meaning of causation is that by changing \( x \) we can bring about a change in \( y \). Could we lengthen our lives by buying more cars? No. The study used number of cars as a quick indicator of affluence. Well-off people tend to have more cars. They also tend to live longer, probably because they are better educated, take better care of themselves, and get better medical care. The cars have nothing to do with it. There is no cause-and-effect tie between number of cars and length of life.

Correlations such as that in Example 5.8 are sometimes called “nonsense correlations.” The correlation is real. What is nonsense is the conclusion that changing one of the variables causes changes in the other. A lurking variable—such as personal affluence in Example 5.8—that influences both \( x \) and \( y \) can create a high correlation even though there is no direct connection between \( x \) and \( y \).

**ASSOCIATION DOES NOT IMPLY CAUSATION**

An association between an explanatory variable \( x \) and a response variable \( y \), even if it is very strong, is not by itself good evidence that changes in \( x \) actually cause changes in \( y \).

**EXAMPLE 5.9** Obesity in mothers and daughters

Obese parents tend to have obese children. The results of a study of Mexican American girls aged 9 to 12 years are typical. The investigators measured body mass index (BMI), a measure of weight relative to height, for both the girls and their mothers. People with high BMI are overweight or obese. The correlation between the BMI of daughters and the BMI of their mothers was \( r = 0.506 \).\(^7\)

Body type is in part determined by heredity. Daughters inherit half their genes from their mothers. There is therefore a direct causal link between the BMI of mothers and
daughters. But perhaps mothers who are overweight also set an example of little exercise, poor eating habits, and lots of television. Their daughters may pick up these habits, so the influence of heredity is mixed up with influences from the girls' environment. Both contribute to the mother-daughter correlation.

The lesson of Example 5.9 is more subtle than just “association does not imply causation.” Even when direct causation is present, it may not be the whole explanation for a correlation. You must still worry about lurking variables. Careful statistical studies try to anticipate lurking variables and measure them. The mother-daughter study did measure TV viewing, exercise, and diet. Elaborate statistical analysis can remove the effects of these variables to come closer to the direct effect of mother's BMI on daughter's BMI. This remains a second-best approach to causation. The best way to get good evidence that $x$ causes $y$ is to do an experiment in which we change $x$ and keep lurking variables under control. We will discuss experiments in Chapter 8.

When experiments cannot be done, explaining an observed association can be difficult and controversial. Many of the sharpest disputes in which statistics plays a role involve questions of causation that cannot be settled by experiment. Do gun control laws reduce violent crime? Does using cell phones cause brain tumors? Has increased free trade widened the gap between the incomes of more educated and less educated American workers? All of these questions have become public issues. All concern associations among variables. And all have this in common: they try to pinpoint cause and effect in a setting involving complex relations among many interacting variables.

**Example 5.10** Does smoking cause lung cancer?

Despite the difficulties, it is sometimes possible to build a strong case for causation in the absence of experiments. The evidence that smoking causes lung cancer is about as strong as nonexperimental evidence can be.

Doctors had long observed that most lung cancer patients were smokers. Comparison of smokers and "similar" nonsmokers showed a very strong association between smoking and death from lung cancer. Could the association be explained by lurking variables? Might there be, for example, a genetic factor that predisposes people both to nicotine addiction and to lung cancer? Smoking and lung cancer would then be positively associated even if smoking had no direct effect on the lungs. How were these objections overcome?

Let’s answer this question in general terms: what are the criteria for establishing causation when we cannot do an experiment?

- **The association is strong.** The association between smoking and lung cancer is very strong.
- **The association is consistent.** Many studies of different kinds of people in many countries link smoking to lung cancer. That reduces the chance that a lurking variable specific to one group or one study explains the association.
Higher doses are associated with stronger responses. People who smoke more cigarettes per day or who smoke over a longer period get lung cancer more often. People who stop smoking reduce their risk.

The alleged cause precedes the effect in time. Lung cancer develops after years of smoking. The number of men dying of lung cancer rose as smoking became more common, with a lag of about 30 years. Lung cancer kills more men than any other form of cancer. Lung cancer was rare among women until women began to smoke. Lung cancer in women rose along with smoking, again with a lag of about 30 years, and has now passed breast cancer as the leading cause of cancer death among women.

The alleged cause is plausible. Experiments with animals show that tars from cigarette smoke do cause cancer.

Medical authorities do not hesitate to say that smoking causes lung cancer. The U.S. surgeon general has long stated that cigarette smoking is “the largest avoidable cause of death and disability in the United States.” The evidence for causation is overwhelming—but it is not as strong as the evidence provided by well-designed experiments.

**APPLY YOUR KNOWLEDGE**

5.11 Education and income. There is a strong positive association between workers’ education and their income. For example, the Census Bureau reports that the median income of young adults (ages 25 to 34) who work full-time increases from $18,508 for those with less than a ninth-grade education, to $27,201 for high school graduates, to $41,628 for holders of a bachelor’s degree, and on up for yet more education. In part, this association reflects causation—education helps people qualify for better jobs. Suggest several lurking variables that also contribute. (Ask yourself what kinds of people tend to get more education.)

5.12 To earn more, get married? Data show that men who are married, and also divorced or widowed men, earn quite a bit more than men the same age who have never been married. This does not mean that a man can raise his income by getting married, because men who have never been married are different from married men in many ways other than marital status. Suggest several lurking variables that might help explain the association between marital status and income.

5.13 Are big hospitals bad for you? A study shows that there is a positive correlation between the size of a hospital (measured by its number of beds) and the median number of days that patients remain in the hospital. Does this mean that you can shorten a hospital stay by choosing a small hospital? Why?

**CHAPTER 5 SUMMARY**

A regression line is a straight line that describes how a response variable $y$ changes as an explanatory variable $x$ changes. You can use a regression line to predict the value of $y$ for any value of $x$ by substituting this $x$ into the equation of the line.
The slope \( b \) of a regression line \( \hat{y} = a + bx \) is the rate at which the predicted response \( \hat{y} \) changes along the line as the explanatory variable \( x \) changes. Specifically, \( b \) is the change in \( \hat{y} \) when \( x \) increases by 1.

The intercept \( a \) of a regression line \( \hat{y} = a + bx \) is the predicted response \( \hat{y} \) when the explanatory variable \( x = 0 \). This prediction is of no statistical interest unless \( x \) can actually take values near 0.

The most common method of fitting a line to a scatterplot is least squares. The least-squares regression line is the straight line \( \hat{y} = a + bx \) that minimizes the sum of the squares of the vertical distances of the observed points from the line.

The least-squares regression line of \( y \) on \( x \) is the line with slope \( \frac{rs_y}{sx} \) and intercept \( a = \overline{y} - b\overline{x} \). This line always passes through the point \((\overline{x}, \overline{y})\).

Correlation and regression are closely connected. The correlation \( r \) is the slope of the least-squares regression line when we measure both \( x \) and \( y \) in standardized units. The square of the correlation \( r^2 \) is the fraction of the variation in one variable that is explained by least-squares regression on the other variable.

Correlation and regression must be interpreted with caution. Plot the data to be sure the relationship is roughly linear and to detect outliers and influential observations. A plot of the residuals makes these effects easier to see.

Look for influential observations, individual points that substantially change the correlation or the regression line. Outliers in the \( x \) direction are often influential for the regression line.

Avoid extrapolation, the use of a regression line for prediction for values of the explanatory variable far outside the range of the data from which the line was calculated.

Lurking variables may explain the relationship between the explanatory and response variables. Correlation and regression can be misleading if you ignore important lurking variables.

Most of all, be careful not to conclude that there is a cause-and-effect relationship between two variables just because they are strongly associated. High correlation does not imply causation. The best evidence that an association is due to causation comes from an experiment in which the explanatory variable is directly changed and other influences on the response are controlled.

**CHECK YOUR SKILLS**

5.14 Figure 5.9 is a scatterplot of reading test scores against IQ test scores for 14 fifth-grade children. The line is the least-squares regression line for predicting reading score from IQ score. If another child in this class has IQ score 110, you predict the reading score to be close to

(a) 50.  (b) 60.  (c) 70.

5.15 The slope of the line in Figure 5.9 is closest to

(a) \(-1\).  (b) 0.  (c) 1.
5.16 The points on a scatterplot lie close to the line whose equation is \( y = 4 - 3x \).
The slope of this line is
(a) 4.   (b) 3.   (c) \(-3\).

5.17 Fred keeps his savings in his mattress. He began with \$500 from his mother and adds \$100 each year. His total savings \( y \) after \( x \) years are given by the equation
(a) \( y = 500 + 100x \).   (b) \( y = 100 + 500x \).   (c) \( y = 500 + x \).

5.18 Starting with a fresh bar of soap, you weigh the bar each day after you take a shower. Then you find the regression line for predicting weight from number of days elapsed. The slope of this line will be
(a) positive.   (b) negative.   (c) Can’t tell without seeing the data.

5.19 For a biology project, you measure the weight in grams and the tail length in millimeters (mm) of a group of mice. The equation of the least-squares line for predicting tail length from weight is
\[
\text{predicted tail length} = 20 + 3 \times \text{weight}
\]
How much (on the average) does tail length increase for each additional gram of weight?
(a) 3 mm   (b) 20 mm   (c) 23 mm
5.20 According to the regression line in Exercise 5.19, the predicted tail length for a mouse weighing 18 grams is
(a) 74 mm.  (b) 54 mm.  (c) 34 mm.

5.21 By looking at the equation of the least-squares regression line in Exercise 5.19, you can see that the correlation between weight and tail length is
(a) greater than zero.
(b) less than zero.
(c) Can’t tell without seeing the data.

5.22 If you had measured the tail length in Exercise 5.19 in centimeters instead of millimeters, what would be the slope of the regression line? (There are 10 millimeters in a centimeter.)
(a) $\frac{3}{10} = 0.3$  (b) 3  (c) $3(10) = 30$

5.23 Because elderly people may have difficulty standing to have their heights measured, a study looked at predicting overall height from height to the knee. Here are data (in centimeters) for five elderly men:

<table>
<thead>
<tr>
<th>Knee height (x)</th>
<th>57.7</th>
<th>47.4</th>
<th>43.5</th>
<th>44.8</th>
<th>55.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height (y)</td>
<td>192.1</td>
<td>153.3</td>
<td>146.4</td>
<td>162.7</td>
<td>169.1</td>
</tr>
</tbody>
</table>

Use your calculator or software: what is the equation of the least-squares regression line for predicting height from knee height?
(a) $\hat{y} = 2.4 + 44.1x$  (b) $\hat{y} = 44.1 + 2.4x$  (c) $\hat{y} = -2.5 + 0.32x$

**Chapter 5 Exercises**

5.24 Penguins diving. A study of king penguins looked for a relationship between how deep the penguins dive to seek food and how long they stay underwater. For all but the shallowest dives, there is a linear relationship that is different for different penguins. The study report gives a scatterplot for one penguin titled “The relation of dive duration (DD) to depth (D).” Duration DD is measured in minutes, and depth D is in meters. The report then says, “The regression equation for this bird is: $DD = 2.69 + 0.0138D$.”

(a) What is the slope of the regression line? Explain in specific language what this slope says about this penguin’s dives.
(b) According to the regression line, how long does a typical dive to a depth of 200 meters last?
(c) The dives varied from 40 meters to 300 meters in depth. Plot the regression line from $x = 40$ to $x = 300$.

5.25 Measuring water quality. Biochemical oxygen demand (BOD) measures organic pollutants in water by measuring the amount of oxygen consumed by microorganisms that break down these compounds. BOD is hard to measure accurately. Total organic carbon (TOC) is easy to measure, so it is common to measure TOC and use regression to predict BOD. A typical regression equation for water entering a municipal treatment plant is:

$$BOD = -55.43 + 1.507 \text{ TOC}$$
Both BOD and TOC are measured in milligrams per liter of water.

(a) What does the slope of this line say about the relationship between BOD and TOC?

(b) What is the predicted BOD when TOC = 0? Values of BOD less than 0 are impossible. Why do you think the prediction gives an impossible value?

5.26 Sisters and brothers. How strongly do physical characteristics of sisters and brothers correlate? Here are data on the heights (in inches) of 11 adult pairs:

<table>
<thead>
<tr>
<th>Brother</th>
<th>71</th>
<th>68</th>
<th>66</th>
<th>67</th>
<th>70</th>
<th>71</th>
<th>70</th>
<th>73</th>
<th>72</th>
<th>65</th>
<th>66</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sister</td>
<td>69</td>
<td>64</td>
<td>65</td>
<td>63</td>
<td>65</td>
<td>62</td>
<td>65</td>
<td>64</td>
<td>66</td>
<td>59</td>
<td>62</td>
</tr>
</tbody>
</table>

(a) Use your calculator or software to find the correlation and to verify that the least-squares line for predicting sister’s height from brother’s height is \( \hat{y} = 27.64 + 0.527x \). Make a scatterplot that includes this line.

(b) Damien is 70 inches tall. Predict the height of his sister Tonya. Based on the scatterplot and the correlation \( r \), do you expect your prediction to be very accurate? Why?

5.27 Heating a home. Exercise 4.27 (page 110) gives data on degree-days and natural gas consumed by the Sanchez home for 16 consecutive months. There is a very strong linear relationship. Mr. Sanchez asks, “If a month averages 20 degree-days per day (that’s 45°F), how much gas will we use?” Use your calculator or software to find the least-squares regression line and answer his question. Based on a scatterplot and \( r^2 \), do you expect your prediction from the regression line to be quite accurate? Why?

5.28 Does social rejection hurt? Exercise 4.40 (page 114) gives data from a study that shows that social exclusion causes “real pain.” That is, activity in an area of the brain that responds to physical pain goes up as distress from social exclusion goes up. A scatterplot shows a moderately strong linear relationship. Figure 5.10 shows regression output from software for these data.
(a) What is the equation of the least-squares regression line for predicting brain activity from social distress score? Use the equation to predict brain activity for social distress score 2.0.

(b) What percent of the variation in brain activity among these subjects is explained by the straight-line relationship with social distress score?

5.29 Merlins breeding. Exercise 4.39 (page 113) gives data on the number of breeding pairs of merlins in an isolated area in each of nine years and the percent of males who returned the next year. The data show that the percent returning is lower after successful breeding seasons and that the relationship is roughly linear. Figure 5.11 shows software regression output for these data.

(a) What is the equation of the least-squares regression line for predicting the percent of males that return from the number of breeding pairs? Use the equation to predict the percent of returning males after a season with 30 breeding pairs.

(b) What percent of the year-to-year variation in percent of returning males is explained by the straight-line relationship with number of breeding pairs the previous year?

![Simple Linear Regression](image)

**FIGURE 5.11** CrunchIt! regression output for a study of how breeding success affects survival in birds, for Exercise 5.29.

5.30 Husbands and wives. The mean height of American women in their twenties is about 64 inches, and the standard deviation is about 2.7 inches. The mean height of men the same age is about 69.3 inches, with standard deviation about 2.8 inches. If the correlation between the heights of husbands and wives is about \( r = 0.5 \), what is the slope of the regression line of the husband’s height on the wife’s height in young couples? Draw a graph of this regression line. Predict the height of the husband of a woman who is 67 inches tall.

5.31 What’s my grade? In Professor Friedman’s economics course the correlation between the students’ total scores prior to the final examination and their final-examination scores is \( r = 0.6 \). The pre-exam totals for all students in the course have mean 280 and standard deviation 30. The final-exam scores have mean 75 and standard deviation 8. Professor Friedman has lost Julie’s final exam
but knows that her total before the exam was 300. He decides to predict her
final-exam score from her pre-exam total.

(a) What is the slope of the least-squares regression line of final-exam scores on
pre-exam total scores in this course? What is the intercept?

(b) Use the regression line to predict Julie's final-exam score.

(c) Julie doesn't think this method accurately predicts how well she did on the
final exam. Use $r^2$ to argue that her actual score could have been much higher (or
much lower) than the predicted value.

5.32 Going to class. A study of class attendance and grades among first-year students
at a state university showed that in general students who attended a higher
percent of their classes earned higher grades. Class attendance explained 16% of
the variation in grade index among the students. What is the numerical value of
the correlation between percent of classes attended and grade index?

5.33 Keeping water clean. Keeping water supplies clean requires regular
measurement of levels of pollutants. The measurements are indirect—a typical
analysis involves forming a dye by a chemical reaction with the dissolved
pollutant, then passing light through the solution and measuring its "absorbence."
To calibrate such measurements, the laboratory measures known standard
solutions and uses regression to relate absorbence and pollutant concentration.
This is usually done every day. Here is one series of data on the absorbence for
different levels of nitrates. Nitrates are measured in milligrams per liter of water.12

<table>
<thead>
<tr>
<th>Nitrates</th>
<th>50</th>
<th>50</th>
<th>100</th>
<th>200</th>
<th>400</th>
<th>800</th>
<th>1200</th>
<th>1600</th>
<th>2000</th>
<th>2000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Absorbence</td>
<td>7.0</td>
<td>7.5</td>
<td>12.8</td>
<td>24.0</td>
<td>47.0</td>
<td>93.0</td>
<td>138.0</td>
<td>183.0</td>
<td>230.0</td>
<td>226.0</td>
</tr>
</tbody>
</table>

(a) Chemical theory says that these data should lie on a straight line. If the
correlation is not at least 0.997, something went wrong and the calibration
procedure is repeated. Plot the data and find the correlation. Must the calibration
be done again?

(b) The calibration process sets nitrate level and measures absorbence. Once
established, the linear relationship will be used to estimate the nitrate level in
water from a measurement of absorbence. What is the equation of the line used
for estimation? What is the estimated nitrate level in a water specimen with
absorbence 40?

(c) Do you expect estimates of nitrate level from absorbence to be quite accurate?
Why?

5.34 Always plot your data! Table 5.1 presents four sets of data prepared by the
statistician Frank Anscombe to illustrate the dangers of calculating without first
plotting the data.13

(a) Without making scatterplots, find the correlation and the least-squares
regression line for all four data sets. What do you notice? Use the regression line
to predict $y$ for $x = 10$.

(b) Make a scatterplot for each of the data sets and add the regression line to
each plot.
**TABLE 5.1** Four data sets for exploring correlation and regression

<table>
<thead>
<tr>
<th>Data Set A</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>10</td>
<td>8</td>
<td>13</td>
<td>9</td>
<td>11</td>
<td>14</td>
<td>6</td>
<td>4</td>
<td>12</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>y</td>
<td>8.04</td>
<td>6.95</td>
<td>7.58</td>
<td>8.81</td>
<td>8.33</td>
<td>9.96</td>
<td>7.24</td>
<td>4.26</td>
<td>10.84</td>
<td>4.82</td>
<td>5.68</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data Set B</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>10</td>
<td>8</td>
<td>13</td>
<td>9</td>
<td>11</td>
<td>14</td>
<td>6</td>
<td>4</td>
<td>12</td>
<td>7</td>
<td>5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data Set C</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>10</td>
<td>8</td>
<td>13</td>
<td>9</td>
<td>11</td>
<td>14</td>
<td>6</td>
<td>4</td>
<td>12</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>y</td>
<td>7.46</td>
<td>6.77</td>
<td>12.74</td>
<td>7.11</td>
<td>7.81</td>
<td>8.84</td>
<td>6.08</td>
<td>5.39</td>
<td>8.15</td>
<td>6.42</td>
<td>5.73</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data Set D</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>19</td>
</tr>
<tr>
<td>y</td>
<td>6.58</td>
<td>5.76</td>
<td>7.71</td>
<td>8.84</td>
<td>8.47</td>
<td>7.04</td>
<td>5.25</td>
<td>5.56</td>
<td>7.91</td>
<td>6.89</td>
<td>12.50</td>
</tr>
</tbody>
</table>

(c) In which of the four cases would you be willing to use the regression line to describe the dependence of $y$ on $x$? Explain your answer in each case.

5.35 Drilling into the past. Drilling down beneath a lake in Alaska yields chemical evidence of past changes in climate. Biological silicon, left by the skeletons of single-celled creatures called diatoms, is a measure of the abundance of life in the lake. A rather complex variable based on the ratio of certain isotopes relative to ocean water gives an indirect measure of moisture, mostly from snow. As we drill down, we look further into the past. Here are data from 2300 to 12,000 years ago:  

<table>
<thead>
<tr>
<th>Isotope (%)</th>
<th>Silicon (mg/g)</th>
<th>Isotope (%)</th>
<th>Silicon (mg/g)</th>
<th>Isotope (%)</th>
<th>Silicon (mg/g)</th>
</tr>
</thead>
<tbody>
<tr>
<td>−19.90</td>
<td>97</td>
<td>−20.71</td>
<td>154</td>
<td>−21.63</td>
<td>224</td>
</tr>
<tr>
<td>−19.84</td>
<td>106</td>
<td>−20.80</td>
<td>265</td>
<td>−21.63</td>
<td>237</td>
</tr>
<tr>
<td>−19.46</td>
<td>118</td>
<td>−20.86</td>
<td>267</td>
<td>−21.19</td>
<td>188</td>
</tr>
<tr>
<td>−20.20</td>
<td>141</td>
<td>−21.28</td>
<td>296</td>
<td>−19.37</td>
<td>337</td>
</tr>
</tbody>
</table>

(a) Make a scatterplot of silicon (response) against isotope (explanatory). Ignoring the outlier, describe the direction, form, and strength of the relationship. The researchers say that this and relationships among other variables they
measured are evidence for cyclic changes in climate that are linked to changes in the sun's activity.

(b) The researchers single out one point: "The open circle in the plot is an outlier that was excluded in the correlation analysis." Circle this outlier on your graph. What is the correlation with and without this point? The point strongly influences the correlation. Explain why the outlier moves \( r \) in the direction revealed by your calculations.

5.36 Managing diabetes. People with diabetes must manage their blood sugar levels carefully. They measure their fasting plasma glucose (FPG) several times a day with a glucose meter. Another measurement, made at regular medical checkups, is called HbA. This is roughly the percent of red blood cells that have a glucose molecule attached. It measures average exposure to glucose over a period of several months. Table 5.2 gives data on both HbA and FPG for 18 diabetics five months after they had completed a diabetes education class.15

<table>
<thead>
<tr>
<th>Subject</th>
<th>HbA (%)</th>
<th>FPG (mg/ml)</th>
<th>Subject</th>
<th>HbA (%)</th>
<th>FPG (mg/ml)</th>
<th>Subject</th>
<th>HbA (%)</th>
<th>FPG (mg/ml)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.1</td>
<td>141</td>
<td>7</td>
<td>7.5</td>
<td>96</td>
<td>13</td>
<td>10.6</td>
<td>103</td>
</tr>
<tr>
<td>2</td>
<td>6.3</td>
<td>158</td>
<td>8</td>
<td>7.7</td>
<td>78</td>
<td>14</td>
<td>10.7</td>
<td>172</td>
</tr>
<tr>
<td>3</td>
<td>6.4</td>
<td>112</td>
<td>9</td>
<td>7.9</td>
<td>148</td>
<td>15</td>
<td>10.7</td>
<td>359</td>
</tr>
<tr>
<td>4</td>
<td>6.8</td>
<td>153</td>
<td>10</td>
<td>8.7</td>
<td>172</td>
<td>16</td>
<td>11.2</td>
<td>145</td>
</tr>
<tr>
<td>5</td>
<td>7.0</td>
<td>134</td>
<td>11</td>
<td>9.4</td>
<td>200</td>
<td>17</td>
<td>13.7</td>
<td>147</td>
</tr>
<tr>
<td>6</td>
<td>7.1</td>
<td>95</td>
<td>12</td>
<td>10.4</td>
<td>271</td>
<td>18</td>
<td>19.3</td>
<td>255</td>
</tr>
</tbody>
</table>

(a) Make a scatterplot with HbA as the explanatory variable. There is a positive linear relationship, but it is surprisingly weak.

(b) Subject 15 is an outlier in the \( y \) direction. Subject 18 is an outlier in the \( x \) direction. Find the correlation for all 18 subjects, for all except Subject 15, and for all except Subject 18. Are either or both of these subjects influential for the correlation? Explain in simple language why \( r \) changes in opposite directions when we remove each of these points.

5.37 Drilling into the past, continued. Is the outlier in Exercise 5.35 also strongly influential for the regression line? Calculate and draw on your graph two regression lines, and discuss what you see. Explain why adding the outlier moves the regression line in the direction shown on your graph.

5.38 Managing diabetes, continued. Add three regression lines for predicting FPG from HbA to your scatterplot from Exercise 5.36: for all 18 subjects, for all except Subject 15, and for all except Subject 18. Is either Subject 15 or Subject 18 strongly influential for the least-squares line? Explain in simple language what features of the scatterplot explain the degree of influence.

5.39 Influence in regression. The Correlation and Regression applet allows you to animate Figure 5.8. Click to create a group of 10 points in the lower-left corner of the scatterplot with a strong straight-line pattern (correlation about 0.9). Click the "Show least-squares line" box to display the regression line.
(a) Add one point at the upper right that is far from the other 10 points but exactly on the regression line. Why does this outlier have no effect on the line even though it changes the correlation?

(b) Now use the mouse to drag this last point straight down. You see that one end of the least-squares line chases this single point, while the other end remains near the middle of the original group of 10. What makes the last point so influential?

5.40 Beavers and beetles. Ecologists sometimes find rather strange relationships in our environment. For example, do beavers benefit beetles? Researchers laid out 23 circular plots, each 4 meters in diameter, in an area where beavers were cutting down cottonwood trees. In each plot, they counted the number of stumps from trees cut by beavers and the number of clusters of beetle larvae. Ecologists think that the new sprouts from stumps are more tender than other cottonwood growth, so that beetles prefer them. If so, more stumps should produce more beetle larvae. Here are the data:

<table>
<thead>
<tr>
<th>Stumps</th>
<th>Beetle larvae</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>30</td>
</tr>
<tr>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>3</td>
<td>24</td>
</tr>
<tr>
<td>3</td>
<td>36</td>
</tr>
<tr>
<td>4</td>
<td>40</td>
</tr>
<tr>
<td>3</td>
<td>43</td>
</tr>
<tr>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>2</td>
<td>27</td>
</tr>
<tr>
<td>5</td>
<td>56</td>
</tr>
<tr>
<td>1</td>
<td>18</td>
</tr>
<tr>
<td>3</td>
<td>40</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Stumps</th>
<th>Beetle larvae</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>25</td>
</tr>
<tr>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>21</td>
</tr>
<tr>
<td>2</td>
<td>14</td>
</tr>
<tr>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>54</td>
</tr>
<tr>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>13</td>
</tr>
<tr>
<td>1</td>
<td>50</td>
</tr>
</tbody>
</table>

Analyze these data to see if they support the “beavers benefit beetles” idea. Follow the four-step process (page 53) in reporting your work.

5.41 Climate change. Global warming has many indirect effects on climate. For example, the summer monsoon winds in the Arabian Sea bring rain to India and are critical for agriculture. As the climate warms and winter snow cover in the vast landmass of Europe and Asia decreases, the land heats more rapidly in the summer. This may increase the strength of the monsoon. Here are data on snow cover (in millions of square kilometers) and summer wind stress (in newtons per square meter):

<table>
<thead>
<tr>
<th>Snow cover</th>
<th>Wind stress</th>
<th>Snow cover</th>
<th>Wind stress</th>
<th>Snow cover</th>
<th>Wind stress</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.6</td>
<td>0.125</td>
<td>16.6</td>
<td>0.111</td>
<td>26.6</td>
<td>0.062</td>
</tr>
<tr>
<td>5.9</td>
<td>0.160</td>
<td>18.2</td>
<td>0.106</td>
<td>27.1</td>
<td>0.051</td>
</tr>
<tr>
<td>6.8</td>
<td>0.158</td>
<td>15.2</td>
<td>0.143</td>
<td>27.5</td>
<td>0.068</td>
</tr>
<tr>
<td>7.7</td>
<td>0.155</td>
<td>16.2</td>
<td>0.153</td>
<td>28.4</td>
<td>0.055</td>
</tr>
<tr>
<td>7.9</td>
<td>0.169</td>
<td>17.1</td>
<td>0.155</td>
<td>28.6</td>
<td>0.033</td>
</tr>
<tr>
<td>7.8</td>
<td>0.173</td>
<td>17.3</td>
<td>0.133</td>
<td>29.6</td>
<td>0.029</td>
</tr>
<tr>
<td>8.1</td>
<td>0.196</td>
<td>18.1</td>
<td>0.130</td>
<td>29.4</td>
<td>0.024</td>
</tr>
</tbody>
</table>

Analyze these data to uncover the nature and strength of the effect of decreasing snow cover on wind stress. Follow the four-step process (page 53) in reporting your work.
### TABLE 5.3 Reaction times in a computer game

<table>
<thead>
<tr>
<th>Time</th>
<th>Distance</th>
<th>Hand</th>
<th>Time</th>
<th>Distance</th>
<th>Hand</th>
</tr>
</thead>
<tbody>
<tr>
<td>115</td>
<td>190.70</td>
<td>right</td>
<td>240</td>
<td>190.70</td>
<td>left</td>
</tr>
<tr>
<td>96</td>
<td>138.52</td>
<td>right</td>
<td>190</td>
<td>138.52</td>
<td>left</td>
</tr>
<tr>
<td>110</td>
<td>165.08</td>
<td>right</td>
<td>170</td>
<td>165.08</td>
<td>left</td>
</tr>
<tr>
<td>100</td>
<td>126.19</td>
<td>right</td>
<td>125</td>
<td>126.19</td>
<td>left</td>
</tr>
<tr>
<td>111</td>
<td>163.19</td>
<td>right</td>
<td>315</td>
<td>163.19</td>
<td>left</td>
</tr>
<tr>
<td>101</td>
<td>305.66</td>
<td>right</td>
<td>240</td>
<td>305.66</td>
<td>left</td>
</tr>
<tr>
<td>111</td>
<td>176.15</td>
<td>right</td>
<td>141</td>
<td>176.15</td>
<td>left</td>
</tr>
<tr>
<td>106</td>
<td>162.78</td>
<td>right</td>
<td>210</td>
<td>162.78</td>
<td>left</td>
</tr>
<tr>
<td>96</td>
<td>147.87</td>
<td>right</td>
<td>200</td>
<td>147.87</td>
<td>left</td>
</tr>
<tr>
<td>96</td>
<td>271.46</td>
<td>right</td>
<td>401</td>
<td>271.46</td>
<td>left</td>
</tr>
<tr>
<td>95</td>
<td>40.25</td>
<td>right</td>
<td>320</td>
<td>40.25</td>
<td>left</td>
</tr>
<tr>
<td>96</td>
<td>24.76</td>
<td>right</td>
<td>113</td>
<td>24.76</td>
<td>left</td>
</tr>
<tr>
<td>96</td>
<td>104.80</td>
<td>right</td>
<td>176</td>
<td>104.80</td>
<td>left</td>
</tr>
<tr>
<td>106</td>
<td>136.80</td>
<td>right</td>
<td>211</td>
<td>136.80</td>
<td>left</td>
</tr>
<tr>
<td>100</td>
<td>308.60</td>
<td>right</td>
<td>238</td>
<td>308.60</td>
<td>left</td>
</tr>
<tr>
<td>113</td>
<td>279.80</td>
<td>right</td>
<td>316</td>
<td>279.80</td>
<td>left</td>
</tr>
<tr>
<td>123</td>
<td>125.51</td>
<td>right</td>
<td>176</td>
<td>125.51</td>
<td>left</td>
</tr>
<tr>
<td>111</td>
<td>329.80</td>
<td>right</td>
<td>173</td>
<td>329.80</td>
<td>left</td>
</tr>
<tr>
<td>95</td>
<td>51.66</td>
<td>right</td>
<td>210</td>
<td>51.66</td>
<td>left</td>
</tr>
<tr>
<td>108</td>
<td>201.95</td>
<td>right</td>
<td>170</td>
<td>201.95</td>
<td>left</td>
</tr>
</tbody>
</table>

5.42 **A computer game.** A multimedia statistics learning system includes a test of skill in using the computer’s mouse. The software displays a circle at a random location on the computer screen. The subject clicks in the circle with the mouse as quickly as possible. A new circle appears as soon as the subject clicks the old one. Table 5.3 gives data for one subject’s trials, 20 with each hand. Distance is the distance from the cursor location to the center of the new circle, in units whose actual size depends on the size of the screen. Time is the time required to click in the new circle, in milliseconds.18

(a) We suspect that time depends on distance. Make a scatterplot of time against distance, using separate symbols for each hand.

(b) Describe the pattern. How can you tell that the subject is right-handed?

(c) Find the regression line of time on distance separately for each hand. Draw these lines on your plot. Which regression does a better job of predicting time from distance? Give numerical measures that describe the success of the two regressions.

5.43 **Climate change: look more closely.** The report from which the data in Exercise 5.41 were taken is not clear about the time period that the data describe. Your work for Exercise 5.41 should include a scatterplot. That plot shows an odd pattern that correlation and regression don’t describe. What is this pattern? On
5.44 **Using residuals.** It is possible that the subject in Exercise 5.42 got better in later trials due to learning. It is also possible that he got worse due to fatigue. Plot the residuals from each regression against the time order of the trials (down the columns in Table 5.3). Is either of these systematic effects of time visible in the data?

5.45 **How residuals behave.** Return to the merlin data of Exercise 4.39 (page 113). Figure 5.11 shows basic regression output.

(a) Use the regression equation from the output to obtain the residuals step-by-step. That is, find the predicted percent $\hat{y}$ of returning males for each number of breeding pairs, then find the residuals $y - \hat{y}$.

(b) The residuals are the part of the response left over after the straight-line tie to the explanatory variable is removed. Find the correlation between the residuals and the explanatory variable. Your result should not be a surprise.

5.46 **Using residuals.** Make a residual plot (residual against explanatory variable) for the merlin regression of the previous exercise. Use a $y$-scale from $-20$ to $20$ or wider to better see the pattern. Add a horizontal line at $y = 0$, the mean of the residuals.

(a) Describe the pattern if we ignore the two years with $x = 38$. Do the $x = 38$ years fit this pattern?

(b) Return to the original data. Make a scatterplot with two least-squares lines: with all nine years and without the two $x = 38$ years. Although the original regression in Figure 5.11 seemed satisfactory, the two $x = 38$ years are influential. We would like more data for years with $x$ greater than 33.

5.47 **Do artificial sweeteners cause weight gain?** People who use artificial sweeteners in place of sugar tend to be heavier than people who use sugar. Does this mean that artificial sweeteners cause weight gain? Give a more plausible explanation for this association.

5.48 **Learning online.** Many colleges offer online versions of courses that are also taught in the classroom. It often happens that the students who enroll in the online version do better than the classroom students on the course exams. This does not show that online instruction is more effective than classroom teaching, because the people who sign up for online courses are often quite different from the classroom students. Suggest some differences between online and classroom students that might explain why online students do better.

5.49 **What explains grade inflation?** Students at almost all colleges and universities get higher grades than was the case 10 or 20 years ago. Is grade inflation caused by lower grading standards? Suggest some lurking variables that might explain higher grades even if standards have remained the same.

5.50 **Grade inflation and the SAT.** The effect of a lurking variable can be surprising when individuals are divided into groups. In recent years, the mean SAT score of all high school seniors has increased. But the mean SAT score has decreased for students at each level of high school grades (A, B, C, and so on). Explain how grade inflation in high school (the lurking variable) can account for this pattern.
5.51 Workers’ incomes. Here is another example of the group effect cautioned about in the previous exercise. Explain how, as a nation’s population grows older, median income can go down for workers in each age group, yet still go up for all workers.

5.52 Some regression math. Use the equation of the least-squares regression line (box on page 120) to show that the regression line for predicting $y$ from $x$ always passes through the point $(\bar{x}, \bar{y})$. That is, when $x = \bar{x}$, the equation gives $\hat{y} = \bar{y}$.

5.53 Will I bomb the final? We expect that students who do well on the midterm exam in a course will usually also do well on the final exam. Gary Smith of Pomona College looked at the exam scores of all 346 students who took his statistics class over a 10-year period. The least-squares line for predicting final exam score from midterm-exam score was $\hat{y} = 46.6 + 0.41x$.

Octavio scores 10 points above the class mean on the midterm. How many points above the class mean do you predict that he will score on the final? (Hint: Use the fact that the least-squares line passes through the point $(\bar{x}, \bar{y})$ and the fact that Octavio’s midterm score is $x + 10$.) This is an example of the phenomenon that gave “regression” its name: students who do well on the midterm will on the average do less well, but still above average, on the final.

5.54 Is regression useful? In Exercise 4.37 (page 113) you used the Correlation and Regression applet to create three scatterplots having correlation about $r = 0.7$ between the horizontal variable $x$ and the vertical variable $y$. Create three similar scatterplots again, and click the “Show least-squares line” box to display the regression lines. Correlation $r = 0.7$ is considered reasonably strong in many areas of work. Because there is a reasonably strong correlation, we might use a regression line to predict $y$ from $x$. In which of your three scatterplots does it make sense to use a straight line for prediction?

5.55 Guessing a regression line. In the Correlation and Regression applet, click on the scatterplot to create a group of 15 to 20 points from lower left to upper right with a clear positive straight-line pattern (correlation around 0.7). Click the “Draw line” button and use the mouse (right-click and drag) to draw a line through the middle of the cloud of points from lower left to upper right. Note the “thermometer” above the plot. The red portion is the sum of the squared vertical distances from the points in the plot to the least-squares line. The green portion is the “extra” sum of squares for your line—it shows by how much your line misses the smallest possible sum of squares.

(a) You drew a line by eye through the middle of the pattern. Yet the right-hand part of the bar is probably almost entirely green. What does that tell you?

(b) Now click the “Show least-squares line” box. Is the slope of the least-squares line smaller (the new line is less steep) or larger (line is steeper) than that of your line? If you repeat this exercise several times, you will consistently get the same result. The least-squares line minimizes the vertical distances of the points from the line. It is not the line through the “middle” of the cloud of points. This is one reason why it is hard to draw a good regression line by eye.