4
Regression Analysis: Exploring Relationships Between Variables

LEARNING OBJECTIVES

After completing this chapter, you should be able to:

1. Write a description of the relationship between two numerical variables based on a scatterplot.
2. Define, calculate, and interpret the correlation coefficient between two variables.
3. Summarize the linear relationship between two variables with a regression line.
4. Interpret the intercept and the slope of a regression line.
5. Define, calculate, and interpret the coefficient of determination.
6. Avoid errors that are common when applying regression.

The online real estate value calculator Zoopraisal (launched by Zoocasa.com) can estimate the market value of any home in Canada. You need merely type in an address. Zoopraisal estimates the value of a home even if the home has not been on the market for many years. How can this tool come up with an estimate for something that is not for sale? The answer is that Zoopraisal takes advantage of relationships between the value of a home and other easily observed variables—size of the home, selling price of nearby homes, number of bedrooms, and so on.

What role does genetics play in determining basic physical characteristics, such as height? This question fascinated nineteenth-century statistician Francis Galton (1822–1911). He examined the heights of thousands of father-son pairs to determine the nature of the relationship between these heights. If a father is 15 centimetres taller than average, how much taller than average will his son be? How certain can we be of the answer? Will there be much variability? If there’s a lot of variability, then perhaps factors other than the father’s genetic material play a role in determining height.

Relationships between variables can be used to predict as yet unseen observations. You might think that estimating the value of a piece of real estate and understanding the role of genetics in determining height are unrelated. However, both take advantage of relationships between two numerical variables. They use a technique called regression, developed by Galton, to analyze the extent of such relationships.

As in previous chapters, graphs play a major role in revealing patterns in data, and graphs become even more important when we have two variables, not just one. For this reason, we’ll start by using graphs to visualize relationships between two numerical variables, and then we’ll talk about quantifying these relationships.
CASE STUDY

Interactive Video Game Exercise

Physical activity is essential to staying healthy. As stated by the Public Health Agency of Canada, people who are physically active are more productive, more likely to avoid illness and injury, and live longer and healthier lives.

Physical activity tends to decline steadily between high school and young adulthood. Given the appeal of television and video games in that age group, researchers have been investigating different ways in which active gaming can be used to improve physical activity in young adults.

One such study was conducted by researchers at the University of British Columbia. Fourteen low-active young males were randomly assigned to either the intervention group or the control group. The participants in the intervention group were required to exercise on a GameBike interactive video gaming system that was linked to a Sony Playstation 2 and a television monitor. The system reads the participant’s speed and steering, giving the participant the opportunity to play a variety of video games. The participants in the control group were required to exercise on a standard stationary bicycle. All participants were given a recommended exercise training regime consisting of moderate-intensity exercise, three days a week for 30 minutes a day for six weeks. Despite the recommendations, all participants were allowed complete freedom to choose the exercise intensity and frequency with which they participated in training.

Researchers observed significant differences in the attendance of the interactive video game and the traditional cycling groups. They also observed a significant change in aerobic fitness in the interactive video game group, while there was no improvement in aerobic fitness in the control group. At the end of this chapter you will see what the linear relationship between attendance and improvement in aerobic fitness tells us about the possible health benefits of interactive video game exercise for young males.

SECTION 4.1

Visualizing Variability with a Scatterplot

Canada’s life expectancy—the number of years a person is expected to live—ranks among the top in the world. Nevertheless, life expectancy varies greatly within Canada, reflecting the disparities in socio-economic characteristics and living habits of people in different regions. As an example, let us examine how life expectancy varies across the nation in relation to the smoking habits of the population. The primary tool for examining relationships between numerical variables such as these is the scatterplot. In a scatterplot, each point represents one observation. The location of the point depends on the values of the two variables. Figure 4.1 shows a scatterplot of these data, taken from the 2007 Canadian Community Health Survey. Each point represents a health region, showing us the percentage of adults who smoke and the life expectancy in that region. As you might have expected, the picture shows that health regions with the lowest life expectancies tend to have the highest rates of smoking. The point in the lower right corner represents the 10 largest communities in Nunavut.

When examining histograms (or other pictures of distributions for a single variable), we look for centre, spread, and shape. When studying scatterplots, we look for trend (which is like centre), strength (which is like spread), and shape (which is like, well, shape). Let’s take a closer look at these characteristics.
Recognizing Trend

The trend of a relationship is the general tendency of the scatterplot as you scan from left to right. Usually trends are either increasing (uphill, /) or decreasing (downhill, \\), but other possibilities exist. Increasing trends are called positive relationships (or positive trends) and decreasing trends are called negative relationships (or negative trends).

Figure 4.2 shows examples of positive and negative trends. Figure 4.2a reveals a positive trend between the age of a used car and the kilometres it was driven (mileage). The positive trend matches our common sense: we expect older cars to have been driven farther, because generally, the longer a car is owned, the more kilometres it travels. Figure 4.2b shows a negative trend—the birthrate of a country against that country’s literacy rate. The negative trend suggests that countries with higher literacy rates tend to have a lower rate of childbirth.

Figure 4.3a shows an example of a scatterplot with no trend. The variables plotted are age and weight of road cyclists who participated in the 2012 Olympics. Even though athletes stay fit throughout their competitive career, we might still expect that the weight of an athlete is related to his or her age, so that older athletes tend to weigh more than younger ones. At the 2012 Olympics, that was indeed the case for most sports. However, cycling (mountain bike, track, and road), basketball, rowing, and volleyball were sports that did not show such a trend. The lack of trend means that, on average, athletes in these disciplines tended to weigh the same, irrespective of their age.

Figure 4.3b shows simulated data of a relationship between two variables that cannot be easily characterized as positive or negative—for smaller \( x \)-values the trend is negative (\( \backslash \)), and for larger \( x \)-values it is positive (\( / \)).
4.1 VISUALIZING VARIABILITY WITH A SCATTERPLOT

CHAPTER 4

FIGURE 4.3 Scatterplots with (a) no trend and (b) a changing trend. (Sources: (a) Could you be an athlete? Olympics 2012 by age, weight and height, www.theguardian.com; (b) simulated data)

Seeing Strength of a Relationship

Weak relationships result in a large amount of scatter in the scatterplot. A large amount of scatter means that points have a great deal of spread in the vertical direction. This vertical spread makes it somewhat harder to detect a trend. Strong relationships have little vertical variation.

Figure 4.4 allows us to compare the strengths of two relationships. Figure 4.4a shows the relationship between height and weight for a sample of active adults. Figure 4.4b involves the same group of adults, but this time we examine the relationship between waist size (typically given in inches, even in Canada) and weight. Which relationship is stronger?

The relationship between waist size and weight is the stronger one (Figure 4.4b). To see this, in Figure 4.4a, consider the data for people who are 165 centimetres tall. Their weights vary anywhere from about 50 kilograms to 105 kilograms, a range of 55 kilograms. If you were using height to predict weight, you could be off by quite a bit. Compare this with the data in Figure 4.4b for people with a waist size of 30 inches. Their weights vary from about 50 kilograms to 80 kilograms, only a 30-kilogram range. The relationship between waist size and weight is stronger than that between height and weight because there is less vertical spread, so better predictions can be made. If you had to guess someone’s weight, and could ask only one question before guessing, you’d do a better job if you asked about the person’s waist size than about his or her height.

Labeling a trend as strong, very strong, or weak is a subjective judgment. Different statisticians might have different opinions. Later in this section, we’ll see how we can measure strength with a number.

FIGURE 4.4 (a) This graph shows a relatively weak relationship. (b) This graph shows a stronger relationship, because the points have less vertical spread. (Source: Heinz et al. 2003)
Identifying Shape

The simplest shape for a trend is linear. Fortunately, linear trends are quite common in real-life situations. Linear trends always increase (or decrease) at the same rate. They are called linear because the trend can be summarized with a straight line. Scatterplots of linear trends often resemble a football, as shown in Figure 4.4a, particularly if there is some scatter and a large number of observations. Figure 4.5 shows a linear trend from data provided in Nielsen’s Global E-Commerce Report, a study aimed at understanding consumers’ online shopping and purchasing intentions in 60 countries (Nielsen 2014). Each point represents one of the e-commerce categories included in the study. The variables measured are the percentage of global respondents who plan to browse for products in that category online in the next six months, and the percentage who plan to buy products in that category online in the next six months. Figure 4.5 shows that a positive, linear relationship exists between online searching and online shopping intentions: those who plan to browse online also tend to buy online, and the more they plan to browse, the more they tend to plan to buy. (Categories with low browsing and purchasing intentions include flowers, alcohol, and baby products, whereas those with high browsing and purchasing intentions include clothing and accessories, event tickets, and hotel reservations.) We’ve added a straight line to the scatterplot to highlight the linear trend.

Not all trends are linear; in fact, a great variety of shapes can occur. But don’t worry about that for now: all we want to do is classify trends as either linear or not linear.

Figure 4.6 shows the relationship between the Life Satisfaction Index and the income per capita (in U.S. dollars) for countries that are members of the Organisation for Economic Co-operation and Development (OECD 2013). You can see that there is a tendency for countries with a higher income per capita to have a higher Life Satisfaction Index. However, the relationship is not linear, as is made clear by the curved line superimposed on the graph.

Nonlinear trends are more difficult to summarize than linear trends. This book does not cover nonlinear trends. Although our focus is on linear trends, it is very important that you first examine a scatterplot to be sure that the trend is linear. If you apply the
techniques in this chapter to a nonlinear trend, you might reach disastrously incorrect conclusions!

**KEY POINT**

When examining relationships between two variables, look for the trend, the strength of the trend, and the shape of the trend.

**Writing Clear Descriptions of Relationships**

Good communication skills are vital for success in general, and being able to clearly describe patterns in data is an important goal of this book. Here are some tips to help you describe relationships between two variables.

- A written description should always include (1) trend, (2) shape, and (3) strength (not necessarily in that order) and should explain what all of these mean in (4) the context of the data. You should also mention any observations that do not fit the general trend.

Example 1 demonstrates how to write a clear, precise description of the relationship between numerical variables.

**EXAMPLE 1 Age and Mileage of Used Cars**

Figure 4.2a on page 132 displays the relationship between the age and number of kilometres of a sample of used cars.

**QUESTION** Describe the relationship.

**SOLUTION** The relationship between the age and number of kilometres of used cars is positive and linear. This means that older cars tend to have greater mileage. The relationship is moderately strong: some scatter is present, but not enough to hide the shape of the relationship. There is one exceptional point: one car is only about six years old but has been driven many kilometres.

**TRY THIS!** Exercise 4.7

The description in Example 1 is good because it mentions trend (a “positive” relationship), shape (“linear”), and strength (“moderately strong”) and does so in context (“older cars tend to have greater mileage”).

- It is very important that your descriptions be precise. For example, it would be wrong to say that older cars have higher kilometres. This statement is not true of every car in the data set. The one exceptional car (upper left corner of the plot) is relatively new (about six years old) but has a high number of kilometres driven (about 400,000 kilometres). Some older cars have relatively few kilometres on them. To be precise, you could say that older cars *tend* to have been driven for more kilometres. The word *tend* indicates that you are describing a trend that has variability, so the trend you describe is not true of all individuals but instead is a characteristic of the entire group.

- When writing a description of a relationship, you should also mention unusual features, such as outliers, small clusters of points, or anything else that does not seem to be part of the general pattern. Figure 4.7 includes an outlier. These data are from a group of adults who reported their weights and heights. One person clearly wrote the wrong height.
SECTION 4.2

Measuring the Strength of a Relationship with Correlation

The correlation coefficient is a number that measures the strength of the linear relationship between two numerical variables—for example, the relationship between people's heights and weights. We can't emphasize enough that the correlation coefficient makes sense only if the trend is linear and if both variables are numerical.

The correlation coefficient, represented by the letter $r$, is a statistic whose value is always between $-1$ and $+1$. Both the value and the sign (positive or negative) of $r$ have information we can use. If the value of $r$ is close to $-1$ or $+1$, then the relationship is very strong; if $r$ is close to 0, the relationship is weak. If the value of the correlation coefficient is positive, then the trend is positive; if the value is negative, the trend is negative.

**Visualizing the Correlation Coefficient**

Figure 4.8 on the next page presents a series of scatterplots that show relationships of gradually decreasing strength. The strongest linear relationship appears in Figure 4.8a; the points fall exactly along a line. Because the trend is positive and perfectly linear, the correlation coefficient is equal to 1.

The next scatterplot, Figure 4.8b, is slightly weaker. The points are more spread out vertically. We can see a linear trend, but the points do not fall exactly along a line. The trend is still positive, so the correlation coefficient is also positive. However, the value of the correlation coefficient is less than 1 (it is 0.98).

The remaining scatterplots show weaker and weaker linear relationships, and their correlation coefficients gradually decrease. In the last scatterplot, Figure 4.8f, there is no linear relationship between the two variables, and the correlation coefficient has a value of 0.0.

The next set of scatterplots (Figure 4.9 on page 138) starts with that same Olympic cyclists’ data (having a correlation of 0.0), and the negative correlations gradually get stronger. The last figure has a correlation of $-1.00$.

**The Correlation Coefficient in Context**

The correlation between online browsing intentions and buying intentions is $r = 0.89$. If we are told, or already know, that the relationship between these variables is linear, then we know that the trend is positive and strong. The fact that the correlation is high and close to 1 means that there is little scatter in the scatterplot.

University admission offices sometimes report correlations between students’ Grade 12 average marks and their first-year grades. If the relationship is linear and the correlation is high, this justifies using Grade 12 marks to make admissions decisions, because a high correlation would indicate a strong relationship between Grade 12 marks and academic...
4.2 Measuring the Strength of a Relationship with Correlation

Chapter 4

FIGURE 4.8 Scatterplots with decreasing positive correlation coefficients. (Sources: (a) simulated data; (b) World Health Organization, www.who.int; (c) Federation of Canadian Municipalities 2012; (d) Boué 2010; (e) www.nhl.com; (f) www.theguardian.com)

A positive correlation means that students who have above-average Grade 12 marks tend to get above-average grades in first year. Conversely, those who have below-average Grade 12 marks tend to get below-average grades in first year. Note that we’re careful to say “tend to.” Certainly, some students with low Grade 12 marks do very well, and some with high Grade 12 marks struggle to pass their courses. The correlation coefficient does not tell us about individual students; it tells us about the overall trend.

More Context: Correlation Does Not Mean Causation!

Quite often, you’ll hear someone use the correlation coefficient to support a claim of cause and effect. For example, one of the authors once read that a politician wanted to close liquor stores in a city because there was a positive correlation between the number of liquor stores in a neighbourhood and the amount of crime.
As you learned in Chapter 1, we can’t form cause-and-effect conclusions from observational studies. If your data came from an observational study, it doesn’t matter how strong the correlation is. Even a correlation near 1 is not enough to conclude that changing one variable (closing down liquor stores) will lead to a change in the other variable (crime rate).

A positive correlation also exists between the number of toques sold in Canada per week and the number of brush fires in Australia per week. Are brush fires in Australia caused by cold Canadians? Probably not. The correlation is likely to be the result of weather. When it is winter in Canada, people buy toques. When winter is happening in Canada, summer is happening in Australia (which is located in the southern hemisphere), and summer is brush-fire season.
What, then, can we conclude from the fact that the number of liquor stores in a
neighbourhood is positively correlated with the crime rate in that neighbourhood? Only
that neighbourhoods with a higher-than-average number of liquor stores typically (but
not always) have a higher-than-average crime rate.

If you learn nothing else from this book, remember this: no matter how tempting, do
not conclude that a cause-and-effect relationship between two variables exists just because
there is a correlation, no matter how close to $+1$ or $-1$ that correlation might be!

**Finding the Correlation Coefficient**

The correlation coefficient is a messy quantity to compute, so it is best to use technology to
calculate its value. Nevertheless, it is useful to know how it is computed. Without technol-
ogy, we compute the value of $r$ by first multiplying the $x$ and $y$ values for each observation
together. We then add up all these products and from this we subtract $n \bar{x} \bar{y}$, where $n$ is the
number of two-variable data points you have, $\bar{x}$ is the mean of all the $x$-values, and $\bar{y}$ is the
mean of all the $y$-values. A positive difference means that as one variable increases in value,
so does the other and there is a positive linear relationship between the $x$-variable and the
$y$-variable. A negative difference means that as one variable increases in value, the other vari-
able decreases in value (or vice versa) and the linear relationship between the two variables is
negative. If there is no relationship between the two variables, then this difference should
be zero (or very close to zero). The last step is to divide this difference by the product of
$n - 1$, $s_x$, and $s_y$, where $s_x$ is the standard deviation of all the $x$-values and $s_y$ is the standard
deviation of all the $y$-values.

This description for computing the correlation coefficient is summarized by the
following formula:

$$
Formula 4.1: \ r = \frac{\sum xy - n \bar{x} \bar{y}}{(n - 1)s_x s_y}
$$

The next example illustrates how to use Formula 4.1 in a calculation.

**EXAMPLE 2 Heights and Weights of Six Women**

Figure 4.10a shows the scatterplot for heights and weights of six women.

**QUESTION** Using the data provided, find the correlation coefficient between height
(in centimetres) and weight (in kilograms) for these six women.

<table>
<thead>
<tr>
<th>Height</th>
<th>155</th>
<th>158</th>
<th>160</th>
<th>163</th>
<th>168</th>
<th>174</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight</td>
<td>47</td>
<td>50</td>
<td>64</td>
<td>57</td>
<td>75</td>
<td>73</td>
</tr>
</tbody>
</table>

![FIGURE 4.10a Scatterplot showing heights and weights of six women.](image-url)
Before proceeding, we verify that the condition of linearity holds. Figure 4.10a suggests that the trend’s shape can be considered linear; a straight line through the data might summarize the trend, although this is hard to see with so few points.

Next, we calculate the correlation coefficient. Ordinarily, we use technology to do this, and Figure 4.10b shows the output from StatCrunch, which gives us the value \( r = 0.88 \).

Because the sample size is small, we can confirm this output using Formula 4.1. It is helpful to go through the steps of this calculation to better understand how the correlation coefficient measures linear relationships between variables.

The first step is to multiply each of the \( x \)-values by its corresponding \( y \)-value, and then add up the products. We have:

\[
\sum_{i=1}^{n} x_i y_i = (155 \times 47) + (158 \times 50) + (160 \times 64) + (163 \times 57) + (168 \times 75) + (174 \times 73)
\]
\[
= 7285 + 7900 + 10240 + 9291 + 12600 + 12702
\]
\[
= 60018
\]

Next we calculate average values of height and weight and then determine the standard deviation for each.

For the height: \( \bar{x} = 163 \) and \( s_x = 6.986 \)

For the weight: \( \bar{y} = 61 \) and \( s_y = 11.679 \)

Being reminded that there are six pairs of heights and weights (\( n = 6 \)), we now use Formula 4.1 to find the value of the correlation coefficient. We get

\[
r = \frac{\sum_{i=1}^{n} x_i y_i - n \bar{x} \bar{y}}{(n - 1) s_x s_y}
\]
\[
= \frac{60018 - (6 \times 163 \times 61)}{(6 - 1) \times 6.986 \times 11.679}
\]
\[
= \frac{360}{407.95}
\]
\[
= 0.8825
\]

The correlation between height and weight for these six women comes out to be about 0.88.

The correlation coefficient for the linear association of weights and heights of these six women is \( r = 0.88 \). Thus, there is a strong positive correlation between height and weight for these women. Taller women tend to weigh more.

Exercise 4.19a
Understanding the Correlation Coefficient

The correlation coefficient has a few features you should know about when interpreting the value of \( r \) or deciding whether you should even compute the value.

- Changing the order of the variables does not change \( r \). Note that in the equation for \( r \) it doesn’t matter which variable is called \( x \) and which is called \( y \). In practice, this means that if the correlation between life expectancy for men and women is 0.977, then the correlation between life expectancy for women and men is also 0.977. This makes sense because the correlation measures the strength of the linear relationship between \( x \) and \( y \), and that strength will be the same no matter which variable gets plotted on the horizontal axis and which on the vertical.

  Figure 4.11a and b have the same correlation; we’ve just swapped axes.

- Adding or multiplying by a constant does not affect \( r \). The correlation between the heights and weights of the six women in Example 2 was 0.88. What would happen if all six women in the sample had been asked to wear 8-centimetre high platform heels when their heights were measured? Everyone would have been 8 centimetres taller. Would this have changed the value of \( r \)? Intuitively, you should sense that it wouldn’t. Figure 4.12a shows a scatterplot of the original data, and Figure 4.12b shows the data with the women in 8-centimetre heels.

We haven’t changed the strength of the relationship. All we’ve done is shift the points on the scatterplot 8 centimetres to the right. But shifting the points doesn’t change the relationship between height and weight. We can verify that the correlation is unchanged by looking at the formula. Each height is increased by 8 cm, so \( \Sigma xy \) becomes \( \Sigma xy + 8 \Sigma y \), and \( n\bar{x}\bar{y} \) becomes \( n\bar{y} + 8 \Sigma y \). When subtracting in
the numerator of Formula 4.1, the new terms cancel out and we obtain the same value for the correlation. As another example, if science found a way to add five years to the life expectancy of men in all countries in the world, the correlation between life expectancies for men and women would still be the same.

More generally, we can add a constant (a fixed value) to all of the values of one variable, or of both variables, and not affect the correlation coefficient. For the very same reason, we can multiply either or both variables by positive constants without changing \( r \). For example, to convert the women’s heights from centimetres to metres, we divide their heights by 100. Doing this does not change how strong the relationship is; it merely changes the units we’re using to measure height. Because the strength of the relationship does not change, the correlation coefficient does not change.

- The correlation coefficient is unitless. Height is measured in centimetres and weight in kilograms, but \( r \) has no units because the numerator and the denominator in Formula 4.1 have the same units, so they cancel out. This means that we will get the same value for correlation whether we measure height in centimetres, inches, or fathoms.

- Linear, linear, linear. We’ve said it before, but we’ll say it again: We’re talking only about linear relationships here. The correlation can be misleading if you do not have a linear relationship. Figures 4.13a through d illustrate the fact that different nonlinear patterns can have the same correlation. All of these graphs have \( r = 0.817 \), but the graphs have very different shapes. The take-home message is that the correlation alone does not tell us much about the shape of a graph. We must also know that the relationship is linear to make sense of the correlation.

Remember: always make a graph of your data. If the trend is nonlinear, the correlation (and, as you’ll see in the next section, other statistics) can be very misleading.

![FIGURE 4.13 (a-d)](a) Four scatterplots with the same correlation of 0.817 have very different shapes. The correlation coefficient is meaningful only if the trend is linear. (Source: Anscombe, F. Anscombe’s Quartet)
The correlation coefficient does not tell you whether a relationship is linear. However, if you already know that the relationship is linear, then the correlation coefficient tells you how strong the relationship is.

**Key Point**

**What is it?**
- A quantity with values between $-1$ and $1$.

**What does it do?**
- Measures the strength of a linear relationship between two numerical variables.

**How does it do it?**
- Through a formula that uses the sum of products of the two observations for each point, and the mean and standard deviation for each variable.

**How is it used?**
- The sign tells us whether the trend is positive (+) or negative (−). The value tells us the strength. If the value is close to 1 or −1, then the points are tightly clustered about a line; if the value is close to 0, then there is no linear relationship.

**When is it used?**
- The correlation coefficient can be interpreted only when a scatterplot shows a linear relationship between the two variables.

**Section 4.3**

**Modelling Linear Trends**

How much more do people tend to weigh for each additional centimetre in height? How much value do cars lose each year as they age? Do NHL teams with a higher team-payroll do better through winning more points in the regular season standings? Can we predict how much space a book will take on a bookshelf just by knowing how many pages there are in the book? To answer these types of questions, it is not enough to remark that a trend exists. We need to measure the trend and the strength of the trend.

To measure the trend, we will summarize the relationship between two variables with a model. The model consists of an equation and a set of conditions that describe when the model will be appropriate. Ideally, this equation is a very concise and accurate description of the data; if so, then the model is a good fit of the data. When this happens, any understanding we gain about the model accurately applies to our understanding of the real world. If the model is a bad fit, however, then the model does not imitate “real-life” situations and should not be used.

**The Regression Line**

The **regression line** is a tool for making predictions about one variable based on the values of another variable. It also provides us with a useful way of summarizing a linear relationship. Recall from Chapter 3 that we could summarize a sample distribution with a mean and a standard deviation. The regression line works the same way: it reduces a linear relationship to its bare essentials and allows us to analyze a relationship without being distracted by small details.
Review: Equation of a Line

The regression line is given by an equation for a straight line. Recall from algebra that equations for straight lines contain a **y-intercept** and a **slope**. The equation for a straight line is

\[ y = mx + b \]

The letter \( m \) represents the slope, which tells how steep the line is, and the letter \( b \) represents the \( y \)-intercept, which is the value of \( y \) when \( x = 0 \).

Statisticians write the equation of a line slightly differently and put the intercept first. For now, we will use the notation \( b_0 \) to denote the \( y \)-intercept, and \( b_1 \) to denote the slope (we will see some different notation in Chapter 14). We then write the regression line as

\[ y = b_0 + b_1 x \]

We often use the names of variables in place of \( x \) and \( y \) to emphasize that the regression line is a model about two real-world variables. We will sometimes write the word **predicted** in front of the \( y \)-variable to emphasize that the line consists of predictions for the \( y \)-variable, not actual values. A few examples should make this clear.

**Visualizing the Regression Line**

Can we know how wide a book is on the basis of the number of pages in the book? A student took a random sample of books from his shelf, measured the width of the spine (in millimetres, mm), and recorded the number of pages. Figure 4.14 illustrates how the regression line captures the basic trend of a linear relationship between width of the book and the number of pages for this sample. The equation for this line is

\[ \text{Predicted Width} = 6.22 + 0.0366 \text{ Pages} \]

One way of comparing the performance of NHL hockey teams is through the total number of points scored during the regular season. If we consider the amount of money paid by an NHL team to its players, we might expect that teams that pay their players more tend to have better regular season performance and therefore a higher total number of points scored. Figure 4.15 shows the relationship between the total number of points scored during the regular season and the team’s total payroll (in millions of dollars) for the 2011–2012 season. The relationship seems fairly linear (although weak), and the regression line can be used to predict how many regular season points a team scored, given the team’s payroll. The data suggest that teams that pay their players more do tend to score more points on average.

\[ \text{Predicted Regular Season Points} = 61.5 + 0.538 \text{ Payroll} \]
Regression in Context

Suppose you have a 10-year-old car and want to estimate how much it is worth. One of the more important uses of the regression line is to make predictions about what \( y \)-values can be obtained for particular \( x \)-values. Figure 4.16 suggests that the relationship between age and value is linear, and the regression line that summarizes this relationship is

\[
\text{Predicted Value} = 20771 - 1355 \text{ Age}
\]

We can use this equation to predict approximately how much a 10-year-old car is worth:

\[
\begin{align*}
\text{Predicted Value} &= 20771 - (1355 \times 10) \\
&= 20771 - 13550 \\
&= 7221
\end{align*}
\]

The regression line predicts that a 10-year-old car will be valued at about $7221. As we know, many factors other than age affect the value of a car, and perhaps with more information we might make a better prediction. However, if the only thing we know about a car is its age, this may be the best prediction we can get. It is also important to keep in mind that this sample is not representative of all used cars on the market in Canada.
Using the regression line to make predictions requires certain assumptions. We’ll go
into more detail about these assumptions later, but for now, just use common sense. This
predicted value of $7221 is useful only for situations similar to those we observed in the
data set. For instance, if all the cars in our data set are Toyotas, and our 10-year-old car is
a Chevrolet, then the prediction is probably not useful.

EXAMPLE 3 Book Width

A college instructor with far too many books on his shelf is wondering whether he has
room for one more. He has about 20 mm of space left on his shelf, and he can tell from
the online bookstore that the book he wants has 598 pages. The regression line obtained
from a small sample of books is

\[
\text{Predicted Width} = 6.22 + 0.0366 \text{ Pages}
\]

**QUESTION** Will the book fit on his shelf?

**SOLUTION** Assuming that the data used to fit this regression line are representative of
all books, we would predict the width of the book corresponding to 598 pages
to be

\[
\text{Predicted Width} = 6.22 + (0.0366 \times 598)
\]

\[
= 6.22 + 21.8868
\]

\[
= 28.1068 \text{ mm}
\]

**CONCLUSION** The book is predicted to be 28 mm wide.

Even though the actual book width is likely to differ somewhat from 28 mm, it seems that the book will probably not
fit on the shelf.

**TRY THIS!** Exercise 4.21

Common sense tells us that not all books with 598 pages are exactly 28 mm wide.
There is a lot of variation in the width of a book for a given number of pages. In
Chapter 14, we’ll see how to place upper and lower limits on this predicted width so
that we can take into account the uncertainty caused by the variation in books.

Finding the Regression Line

In almost every case, we’ll use technology to find the regression line. However, it is
important to know how the technology works, and to be able to calculate the equation
when we have access only to summary statistics and not to the complete data set.

To understand how technology finds the regression line, imagine trying to draw a
line through the scatterplots in Figures 4.14 through 4.16 to best capture the linear trend.
We could have drawn almost any line, and some of them would have looked like pretty
good summaries of the trend. What makes the regression line special? How do we find
the intercept and slope of the regression line?

The regression line is chosen because it is the line that comes closest to most of
the points. More precisely, the square of the vertical distances between the points and
the line, on average, is bigger for any other line we might draw than for the regression
To find this best fit line, we need to find the slope and the intercept. The slope, $b_1$, of the regression line is the ratio of the standard deviations of the two variables, multiplied by the correlation coefficient:

**Formula 4.2a:** The slope, $b_1 = r \frac{s_y}{s_x}$

Once we have found the slope, we can find the intercept. Finding the intercept, $b_0$, requires that we first find the means of the variables, $\bar{x}$ and $\bar{y}$:

**Formula 4.2b:** The intercept, $b_0 = \bar{y} - b_1 \bar{x}$

Now put these quantities into the equation of a line, and the regression line is given by

**Formula 4.2c:** The regression line, Predicted $y = b_0 + b_1 x$

---

**EXAMPLE 4 Term and Final Exam Marks**

A professor wants to predict the final exam mark for students in an introductory statistics class based on their going-into-the-final-exam mark, or term mark. Figure 4.18a shows a scatterplot of the term mark and the final exam mark for a sample of 211 students. The scatterplot suggests a moderately strong, positive linear relationship: students with high term marks tend to get high final exam marks. The average term mark of this sample was 70.19 with standard deviation 13.30. The average final exam mark was 66.82 with standard deviation 16.28. The correlation between term mark and final exam mark was 0.745.

**QUESTIONS**

a. Find the equation of the regression line that best summarizes this relationship. Note that the $x$-variable is the term mark and the $y$-variable is the final exam mark.

b. Using the equation, find the predicted final exam mark for a student in this class with a term mark of 74%.
The scatterplot suggests a moderate linear positive relationship: students with higher term marks tend to have higher final exam marks. (Source: Stallard 2013)

**SOLUTIONS**

a. With access to the full data set on the text's companion website, we can use technology to find (and plot) the regression line. Still, when summary statistics are provided (means and standard deviations of the two variables as well as their correlation), it is not time-consuming to use Formula 4.2 to find the regression line.

Figure 4.18b shows StatCrunch output for the regression line. (StatCrunch provides quite a bit more information than we will use in this chapter.)

According to technology, the equation of the regression line is

\[ \text{Predicted FinalExamMark} = 2.81 + 0.91 \text{ TermMark} \]

We now check this calculation by hand, using Formula 4.2.

We are given that

For term marks: \( \bar{x} = 70.19, s_x = 13.30 \)

For final exam marks: \( \bar{y} = 66.82, s_y = 16.28 \)

and \( r = 0.745 \)

First we must find the slope:

\[ b_1 = r \frac{s_y}{s_x} = 0.745 \times \frac{16.28}{13.30} = 0.912 \]
Now we can use the slope to find the intercept:

\[ b_0 = \bar{y} - b_1 \bar{x} = 66.82 - (0.912 \times 70.19) = 66.82 - 64.01 = 2.81 \]

We can then write:

- Predicted FinalExamMark = 2.81 + 0.912 TermMark
- Predicted FinalExamMark = 2.81 + 0.912 TermMark
  = 2.81 + (0.912 \times 74)
  = 2.81 + 67.488
  = 70.3

**CONCLUSION**

We would expect someone with a term mark of 74% to have a final exam mark of 70.3%.

**TRY THIS!** Exercise 4.23

Different software packages present the intercept and slope differently. Therefore, you need to learn how to read the output of the software you are using. Example 5 shows output from several packages.

**EXAMPLE 5 Technology Output for Regression**

Figure 4.19 shows outputs from Minitab, StatCrunch, the TI-83/84, Excel, and SPSS for finding the regression equation in Example 4 for final exam marks and term marks.

**QUESTION** For each software package, explain how to find the equation of the regression line from the given output.

**CONCLUSION**

- **Figure 4.19a:** Minitab gives us a simple equation directly:
  Final Exam Mark = 2.81 + 0.912 Term Mark.
  However the more statistically correct format would be
  Predicted Final Exam Mark = 2.81 + 0.912 Term Mark

- **Figure 4.19b:** StatCrunch gives the equation directly near the top, but it also lists the intercept and slope separately in the table near the bottom.

- **Figure 4.19c:** TI-83/84 gives us the coefficients to put together. The “a” value is the intercept, and the “b” value is the slope. If the diagnostics are on (use the CATALOG button), the TI-83/84 also gives the correlation.

- **Figure 4.19d:** Excel shows the coefficients in the column labelled “Coefficients.” The intercept is in the first row, labelled “Intercept,” and the slope is in the row labelled “Variable 1.”
An important use of the regression line is to make predictions about the value of $y$ that we will see for a given value of $x$. However, the regression line provides more information than just a predicted $y$-value. The regression line can also tell us about the rate of change of the mean value of $y$ with respect to $x$ and can help us understand the underlying theory behind cause-and-effect relationships.
Choosing $x$ and $y$: Order Matters

In Section 4.2, you saw that the correlation coefficient is the same no matter which variable you choose for $x$ and which you choose for $y$. With regression, however, order matters.

Consider the collection of data about book widths. We used it earlier to predict the width of a book, given the number of pages. The equation of the regression line for this prediction problem (shown in Figure 4.20a) is

\[ \text{Predicted Width} = 6.22 + 0.037 \text{ Pages} \]

But what if we instead wanted to predict how many \textit{Pages} there are in a book on the basis of the width of the book?

To do this, we would switch the order of the variables, and use \textit{Pages} as our $y$-variable and \textit{Width} as our $x$-variable. Then the slope is calculated to be 19.6 (Figure 4.20b).

It is tempting to think that because we are flipping the graph over when we switch $x$ and $y$, we can just flip the slope over to get the new slope. If this were true, then we could find the new slope simply by calculating $1/(\text{old slope})$. However, that approach doesn’t work. That would give us a slope of

\[ \frac{1}{0.037} = 27.0, \text{ which is not the same as the correct value of 19.6.} \]

How, then, do we know which variable goes where?

We use the variable plotted on the horizontal axis to make predictions about the variable plotted on the vertical axis. For this reason, the $x$-variable is called the \textbf{explanatory variable}, the \textbf{predictor variable}, or the \textbf{independent variable}. The $y$-variable is called the \textbf{response variable}, the \textbf{predicted variable}, or the \textbf{dependent variable}. These names reflect the different roles played by the $x$- and $y$-variables in regression. Which variable is which depends on what the regression line will be used to predict.

You’ll see many pairs of terms used for the $x$- and $y$-variables in regression; some are shown in Table 4.1.

![FIGURE 4.20](a) Predicting width from the number of pages. (b) Predicting the number of pages from width.

<table>
<thead>
<tr>
<th>$x$-Variable</th>
<th>$y$-Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predictor variable</td>
<td>Predicted variable</td>
</tr>
<tr>
<td>Explanatory variable</td>
<td>Response variable</td>
</tr>
<tr>
<td>Independent variable</td>
<td>Dependent variable</td>
</tr>
</tbody>
</table>

\[\text{TABLE 4.1} \quad \text{Terms used for the } x\text{- and } y\text{-variables.}\]
EXAMPLE 6 Bedridden

It is hard to measure the height of people who are bedridden, and for many medical reasons it is often important to know a bedridden patient’s height. However, it is not so difficult to measure the length of the ulna (the bone that runs from the elbow to the wrist). Data collected on non-bedridden people shows a strong linear relationship between ulnar length and height.

QUESTION When making a scatterplot to predict height from ulnar length, which variable should be plotted on the $x$-axis and which on the $y$-axis?

CONCLUSION We are measuring ulnar length to predict a person’s height. Therefore, ulnar length is the predictor (independent variable) and is plotted on the $x$-axis, while height is the response (dependent variable) and is plotted on the $y$-axis.

TRY THIS! Exercise 4.35

The Regression Line Is a Line of Averages

Figure 4.21 shows a histogram of the weights (in kilograms) of a sample of 507 active people. What is the typical weight?

One way of answering this question is to calculate the mean of the sample. The distribution of weights is a little right-skewed but not terribly so, and so the mean will probably give us a good idea of what is typical. The average weight of this group is 69.2 kilograms.

Now, we know that a relationship exists between height and weight and that shorter people tend to weigh less. If we know someone’s height, then, what weight would we guess for that person? Surely not the average weight of the whole group! We can do better than that. For instance, what’s the typical weight of someone 170 centimetres tall? To answer this, it makes sense to look only at the weights of those people in our sample who are about 170 centimetres tall. To make sure we have enough 170-centimetres-tall people in our sample, let’s include everyone who is approximately 170 centimetres tall. So let’s look at a slice of people whose height is between 169 centimetres and 171 centimetres. We come up with 37 people. Some of their weights are

63, 66, 53, 57, 67 . . . kilograms
Figure 4.22a shows this slice, which is centred at 170 centimetres.

The mean of these numbers is 66.3 kilograms. We put a special symbol on the plot to record this point—a triangle at the point (170, 66.3), shown in Figure 4.22b.

The reason for marking this point is that if we wanted to predict the weights of those who were 170 centimetres tall, one good answer would be 66.3 kilograms.

What if we wanted to predict the weight of someone who is 178 centimetres tall? We could take a slice of the sample and look at those people who are between 177 centimetres and 179 centimetres tall. Typically, they’re heavier than the people who are about 170 centimetres tall. Here are some of their weights:

83, 94, 73, 72, 60 . . . kilograms

Their mean weight is 79.3 kg. Let’s put another special triangle symbol at (178, 79.3) to record this. We can continue in this fashion, and Figure 4.22c shows where the mean weights are for a few more heights.

Note that the means fall (nearly) on a straight line. What could be the equation of this line? Figure 4.22d shows the regression line superimposed on the scatterplot with the means. They’re nearly identical.

In fact, if we knew the distribution of all weights at every height, the means would lie exactly on the regression line. However, because we are working with only a sample of real data, the fit is only approximate.

The series of graphs in Figure 4.22 illustrates a fundamental feature of the regression line: it is a line of means. You plug in a value for $x$, and the regression line “predicts” the mean $y$-value for everyone in the population with that $x$-value.
EXAMPLE 7 An Alternative to Weighing a Bear?

A bear's body mass is of interest to researchers who study bears in the field. However, weighing bears in the field is difficult, requiring a net, a heavy spring scale, and the effort of two people to suspend the animal from the scale. Chest girth, on the other hand, can be measured much more easily with a soft measuring tape. Because of this, researchers from the Ministry of National Resources wanted to determine if a black bear's mass could be predicted from its chest girth. As you can see in Figure 4.23, the scatterplot of body mass (response variable) versus chest girth (predictor) for a sample of 50 bears shows a reasonably linear relationship.

The regression line is

\[
\text{Predicted Body Mass} = -125 + 2.33 \times \text{Chest Girth}
\]

**Question** What is the predicted mean body mass of a black bear with chest girth of 84 centimetres?

**Solution** We predict by estimating the mean body mass of a bear with chest girth of 84 centimetres. To do this, we substitute 84 into the regression line:

\[
\text{Predicted Body Mass} = -125 + (2.33 \times 84) = 70.7 \text{ kilograms}
\]
4.3 MODELLING LINEAR TRENDS

CONCLUSION We predict that the mean body mass of black bears with a chest girth of 84 centimetres is about 71 kilograms.

TRY THIS! Exercise 4.37

Interpreting the Slope

The slope tells us how to compare the mean $y$-values for objects that are 1 unit apart on the $x$-variable. For example, how different is the mean body mass of black bears whose chest girths differ by 1 centimetre? The slope in Example 7 tells us that the means differ by 2.33 kilograms. What if the bears’ chest girths differ by 10 centimetres? Then the difference in mean body mass is $10 \times 2.33 = 23.3$ kilograms.

We should pay attention to whether the slope is 0 or very close to 0. When the slope is 0, we say that no (linear) relationship exists between $x$ and $y$. The reason is that a 0 slope means that no matter what value of $x$ you consider, the predicted value of $y$ is always the same. The slope is 0 whenever the correlation coefficient is 0.

The slope is close to 0 whenever the correlation coefficient is close to 0. For example, the slope for Olympic road cyclists’ ages and their weights (see Figure 4.3a) is very close to 0 ($b_1 = 0.025$).

**EXAMPLE 8 Math and Reading EQAO Scores**

Each year EQAO (Education Quality Accountability Office) publishes student achievement results for elementary schools in Ontario. To study the relationship between achievement in reading and achievement in mathematics for students in Grade 3, a regression analysis relating the percentage of students at or above provincial standard in reading with the percentage of students at or above provincial standard in mathematics resulted in the following model:

\[
\text{Predicted Reading} = 32.4 + 0.548 \text{ Math}
\]

This model was based on data from 61 schools belonging to the same school board.

The scatterplot in Figure 4.24 shows a linear relationship with a moderately strong positive trend.
Interpreting the Intercept
The intercept tells us the predicted mean $y$-value when the $x$-value is 0. Quite often, this is not terribly helpful. Sometimes it has no applicable meaning. For example, the regression line to predict weight, given someone’s height, tells us that if a person is 0 centimetres tall, then his or her predicted weight is negative 200.5 kilograms!

Before interpreting the intercept, ask yourself whether it makes sense to talk about the $x$-variable taking on a 0 value. For example, suppose that a law school has found a regression line that predicts a student’s first year average grade from the student’s LSAT scaled score. You might think it makes sense to talk about getting a 0 on the LSAT score (no questions right). However, the lowest possible scaled LSAT score is 120, so it is not possible to get a score of 0. (One lesson statisticians learn early is that you must know something about the data you analyze—knowing only the numbers is not enough!)

EXAMPLE 9 Predicting Bears’ Body Mass
What can we learn about the relationship between the chest girth and the body mass of a bear by looking at the intercept and slope of the regression line? The regression model is

$$\text{Predicted Body Mass} = -125 + 2.33 \text{ Chest Girth}$$

**QUESTION** Interpret the intercept and the slope.

**CONCLUSION** The intercept is not meaningful. It tells us that the mean body mass for bears with chest girth of 0 cm is negative 125 kilograms.
The slope is meaningful. The positive sign indicates that there is a positive trend: bears with larger chest girths tend to weigh more. In particular, for each additional centimetre of chest girth, a bear weighs on average 2.33 kilograms more.

**TRY THIS!** Exercise 4.59

---

**SNAPSHOT** THE INTERCEPT OF A REGRESSION LINE

**WHAT IS IT?**
- In the regression line $y = b_0 + b_1x$, the intercept is the value of $b_0$.

**WHAT DOES IT DO?**
- It tells us the average $y$-value when the $x$-value is 0.

**HOW DOES IT DO IT?**
- The intercept is the $y$-value where the regression line crosses the $y$-axis.

**HOW IS IT USED?**
- It is always used to graph the line and to obtain predicted values. However, its value does not always have a meaningful interpretation.

**WHEN IS IT USED?**
- When a scatterplot has a linear trend. It can be interpreted only if the value of $x = 0$ makes sense for the observed data.

---

**EXAMPLE 10** Age and Value of Cars

Figure 4.25 shows the relationship between age and VMR Canada values for a sample of cars.

![Figure 4.25](image)

The regression line is

\[
\text{Predicted Value} = 20771 - 1355 \text{ Age}
\]

**QUESTION** Interpret the slope and the intercept.

**CONCLUSION** The intercept estimates that the average value of a new car (0 years old) in this sample is $20,771. The slope tells us that, on average, cars lost $1355 in value per year.

**TRY THIS!** Exercise 4.61

---

**WHAT IS IT?**
- In the regression line $y = b_0 + b_1x$, the intercept is the value of $b_0$.

**WHAT DOES IT DO?**
- It tells us the average $y$-value when the $x$-value is 0.

**HOW DOES IT DO IT?**
- The intercept is the $y$-value where the regression line crosses the $y$-axis.

**HOW IS IT USED?**
- It is always used to graph the line and to obtain predicted values. However, its value does not always have a meaningful interpretation.

**WHEN IS IT USED?**
- When a scatterplot has a linear trend. It can be interpreted only if the value of $x = 0$ makes sense for the observed data.
SECTION 4.4

Evaluating the Linear Model

Regression is a very powerful tool for understanding linear relationships between numerical variables. However, we need to be aware of several potential interpretation pitfalls so that we can avoid them. We will discuss some of them in this section. We will also discuss methods for determining just how well the regression model fits the data.

Pitfalls to Avoid

You can avoid most pitfalls by simply making a graph of your data and examining it closely before you begin interpreting your linear model. This section will offer some advice for sidestepping a few subtle complications that might arise.

Don’t Fit Linear Models to Nonlinear Relationships

Regression models are useful only for linear relationships. If the relationship is not linear, a regression model can be misleading and deceiving. For this reason, before you fit a regression model, you should always make a scatterplot to verify that the relationship seems linear. Chapter 14 will provide a slightly more sophisticated (and discerning) method for checking linearity.

Figure 4.26 shows an example of a bad regression model. The relationship between the life expectancy at birth of people in a country and the country’s per-person wine consumption is nonlinear. The regression model is

\[
\text{Predicted Life Expectancy} = 65.4 + 2.13 \text{ Wine Consumption}
\]

but it provides a poor fit. The regression model suggests that countries with middle values of wine consumption should have lower mortality rates than they actually do.

Correlation Is Not Causation

One important goal of science and business is to discover relationships between variables such that if you make a change to one variable, you know the other variable will change in a reliably predictable way. This is what is meant by “x causes y”: make a change in x, and a change in y will usually follow. For example, the distance it takes to stop your car depends on how fast you were travelling when you first applied the brakes (among other things); the amount of memory an mp3 file takes on your hard drive depends on the length of the song; and the size of your phone bill depends on how many minutes you talked. In these cases, a strong causal relationship exists between two variables, and if you were to collect data and make a scatterplot, you would see a relationship between the variables.
In statistics, however, we are often faced with the reverse situation, in which we see a relationship between two variables and wonder whether there is a cause-and-effect relationship. The correlation coefficient for the linear relationship could be quite strong, but as we saw earlier, correlation does not mean cause and effect. A strong correlation or a good-fitting regression line is not evidence of a cause-and-effect relationship.

Be particularly careful about drawing cause-and-effect conclusions when interpreting the slope of a regression line. For example, for the EQAO Grade 3 data,

\[
\text{Predicted Reading} = 32.4 + 0.548 \times \text{Math}
\]

Even if this regression line fits the relationship very well, it does not give us sufficient evidence to conclude that if a school were to improve its math score by 10 percentage points (so that 10% more students are at or above the provincial standard), its critical reading score will go up by about 5.5 percentage points. As you learned in Chapter 1, because these data were not collected from a controlled experiment, the presence of confounding factors could prevent you from making a causal interpretation.

When can we conclude that a relationship between two variables means a cause-and-effect relationship is present? Strictly speaking, never from an observational study and only when the data were collected from a controlled experiment. (Even in a controlled experiment, care must be taken that the experiment was designed correctly.) However, for many important questions, controlled experiments are not possible. In these cases, we can sometimes make conclusions about causality after a number of observational studies have been collected and examined, and if there is a strong theoretical explanation for why the variables should be related in a cause-and-effect fashion. For instance, it took many years of observational studies to conclude that smoking causes lung cancer, including studies that compared twins—one twin who smoked and one who did not—and numerous controlled experiments on lab animals.

Moreover, beware of the algebra trap. In algebra, you were taught to interpret the slope to mean that "as \( x \) increases by 1 unit, \( y \) goes up by \( b_1 \) units." However, quite often with data, the phrase "as \( x \) increases" doesn't make sense. When looking at the height and weight data, where \( x \) is height and \( y \) is weight, to say "\( x \) increases" means that people are growing taller! This is not accurate. It is much more accurate to interpret the slope as making comparisons between groups. For example, when comparing people of a certain height with those who are 1 centimetre taller, you can see that the taller individuals tend to weigh, on average, \( b_1 \) kilograms more.

**Beware of Outliers**

Recall that when calculating sample means, we were warned that outliers can have a big effect. Because the regression line is a line of means, you might think that outliers would have a big effect on the regression line. And you’d be right. You should always check a scatterplot before performing a regression analysis to be sure there are no outliers.

The graphs in Figure 4.27 illustrate this effect. Both graphs in Figure 4.27 show yearly expenditure for meals and snacks purchased from restaurants (restaurants include refreshment stands, snack bars, vending machines, mobile canteens, caterers, and coffee wagons) versus total income before taxes for a small random sample of respondents to the 2009 Survey of Household Spending. Figure 4.27a includes the complete sample of 100 respondents. The respondent with an income of $250,000 is an outlier and has a strong influence on the regression line. Figure 4.27b excludes this observation.

There is a positive linear relationship between expenditure for restaurants and income. However, compare the correlation coefficients and note how the inclusion of the
outlier makes that relationship seem stronger than it actually is. These types of observations are called **influential points** because their presence or absence has a big effect on conclusions. When you have influential points in your data, it is good practice to try the regression and correlation with and without these points (as we did) and to comment on the difference.

**Regressions of Aggregate Data**

Researchers sometimes do regression analysis based on what we call **aggregate data**. Aggregate data are those for which each plotted point is from a summary of a group of individuals. For example, in a study to examine the relationship between EQAO mathematics and reading results for Grade 3 students, we might use the combined results for each of the school boards in the province rather than the scores of individual schools.

There is nothing wrong with using aggregate data, as long as you don’t assume that relationships that hold for the aggregate data will also hold for the individual data. For example, Figure 4.28 shows scatterplots of 2012–2013 EQAO results for Grade 3 students in Ontario. The variables represent the percentage of students in Grade 3 at or above provincial standard in reading and in mathematics. In Figure 4.28a, each point represents the percentages at or above standard for a single school. In Figure 4.28b, each point represents the percentages at or above standard for all schools in a school board. As a consequence of the lower variability of the aggregate data, the scatterplot in Figure 4.28b seems to show a much stronger relationship between the two variables.

We can still interpret Figure 4.28b, as long as we’re careful to remember that we are talking about school boards, and not about individual schools. We can say that a strong...
correlation exists between a school board’s percentage of students at or above provincial standard in reading and a school board’s percentage of students at or above provincial standard in mathematics. We cannot say that there is the same correlation between a school’s percentage of students at or above provincial standard in reading and a school’s percentage of students at or above provincial standard in mathematics.

Don’t Extrapolate!

Extrapolation means that we use the regression line to make predictions beyond the range of our data. This practice can be dangerous, because although the relationship may have a linear shape for the range we’re observing, that might not be true over a larger range. This means that our predictions might be wrong outside the range of observed x-values.

Figure 4.29a shows a graph of height versus age for children between 2 and 9 years old from a large study. We’ve superimposed the regression line on the graph, and it looks like a very good fit. However, although the regression model provides a good fit for children aged 2 through 9, it fails when we use the same model to predict heights for older individuals.

The regression line of the data shown in Figure 4.29a is

$$\text{Predicted Height} = 80.6 + 6.22 \text{ Age}$$

However, we observed only children between the ages of 2 and 9. Can we use this line to predict the height of a 20-year-old?

The regression model predicts that the mean height of 20-year-olds is 205 cm:

$$\text{Predicted Height} = 80.6 + 6.22 \text{ Age} = 80.6 + (6.22 \times 20) = 205$$

We can see from Figure 4.29b that the regression model provides a poor fit if we include people over the age of 9. Beyond that age, the trend is no longer linear, so we get bad predictions from the model.

It is often tempting to use the regression model beyond the range of data used to build the model. Don’t. Unless you are very confident that the linear trend continues without change beyond the range of observed data, you must collect new data to cover the new range of values.
The Origin of the Word Regression (Regression Toward the Mean)

The term regression was coined by Francis Galton, who used the regression model to study genetic relationships. He noticed that even though taller-than-average fathers tended to have taller-than-average sons, the sons were somewhat closer to average than the fathers were. Also, shorter-than-average fathers tended to have sons who were closer to the average than their fathers. He called this phenomenon regression toward mediocrity, applying the term regression in the sense of a backward movement. Later on it came to be known as regression toward the mean.

You can see how regression toward the mean works by examining the formula for the slope of the regression line:

\[ b_1 = r \frac{s_y}{s_x} \]

This formula tells us that fathers who are one standard deviation taller than average (\( s_x \) centimetres above average) have sons who are not one standard deviation taller than average (\( s_y \) but are instead \( r \) times \( s_y \) centimetres taller than average. Because \( r \) is a number between \(-1\) and 1, \( r \) times \( s_y \) is usually smaller than \( s_y \). Thus the “rise” will be less than the “run” in terms of standard deviations.

The Sports Illustrated jinx is an example of regression toward the mean. According to the jinx, athletes who make the cover of Sports Illustrated end up having a really bad year after appearing. Some professional athletes have refused to appear on the cover of Sports Illustrated. (Once, the editors published a picture of a black cat in that place of honour, because no athlete would agree to grace the cover.) However, if an athlete’s performance in the first year is several standard deviations above average, the second year is likely to be closer to average. This is an example of regression toward the mean. For a star athlete, closer to average can seem disastrous.

The Coefficient of Determination

If we are convinced that the relationship we are examining is linear, then the regression line provides the best numerical summary of the relationship. But how good is “best”? The correlation coefficient, which measures the strength of linear relationships, can also be used to measure how well the regression line summarizes the data.

The coefficient of determination is simply the correlation coefficient squared: \( r^2 \). In fact, this quantity is often called \( r \)-squared. When reporting \( r \)-squared, we will multiply by 100% to convert it to a percentage. Because \( r \) is always between \(-1\) and 1, \( r \)-squared will be between 0% and 100%. A value of 100% means the relationship is perfectly linear and the regression line perfectly predicts the observations. A value of 0% means there is no linear relationship and the regression line does a horrible job.

For example, when we predicted the width of a book from the number of pages in the book, we found the correlation between these variables to be \( r = 0.9202 \). So the coefficient of determination is \( 0.9202^2 = 0.8468 \), which we report as 84.7%.

What does this value of 84.7% mean? A useful interpretation of \( r \)-squared is that it measures how much of the variation in the response variable is explained by the linear relationship with the explanatory variable. For example, 84.7% of the variation in book widths was explained by the number of pages. What does this mean?
Figure 4.30 shows a scatterplot (simulated data) with a constant value for $y$ ($y = 6240$) no matter what the $x$-value is. You can see that there is no variation in $y$, so there is also nothing to explain.

Figure 4.31 shows the height in inches and in centimetres for several people. Here, variation in the $y$-variable does occur. Height as measured in centimetres varies from about 150 cm to about 190 cm. However, the points are perfectly linear and have a correlation of 1.000. That means that if you are given an $x$-value (a person’s height in inches), then you know the $y$-value (the person’s height in centimetres) precisely. Thus all the variation in $y$ is explained by the regression model. In this case, the coefficient of determination is 100%; all variation in $y$ is perfectly explained by the best-fit line.

Real data are messier. Figure 4.32 shows a plot of the age and value of some cars. The regression line has been superimposed to remind us that there is, in fact, a linear trend and that the regression line does capture it. The regression model explains some of the variation in $y$, but as we can see, it’s not perfect; plugging the value of $x$ into the regression line gives us an imperfect prediction of what $y$ will be. In fact, for these data, $r = -0.837$, so we’ve explained $(-0.837)^2 = 0.701$, or about 70.1%, of the variation in $y$ with this regression line.

The practical implication of $r$-squared is that it helps determine which explanatory variable would be best for making predictions about the response variable. For example, is
waist size or height a better predictor of weight? We can see the answer to this question from the scatterplots in Figure 4.33, which show that the linear relationship is stronger (has less scatter) for waist size.

The \( r \)-squared for predicting weight from height is 51.4\% (Figure 4.33a), and the \( r \)-squared for predicting weight from waist size is 81.7\% (Figure 4.33b). We can explain more of the variation in these people’s weights by using their waist sizes than by using their heights, and therefore, we can make better (more precise) predictions using waist size as the predictor.

If the relationship is linear, the larger the coefficient of determination \( (r \)-squared), the smaller the variation or scatter about the regression line, and the more accurate the predictions tend to be.
As we mentioned at the beginning of the chapter, researchers observed a significant difference in the attendance of participants who cycled using an interactive video game system and those who cycled on a regular stationary bike. There was also an improvement in the aerobic fitness of the interactive video game group, with no improvement in the control group. We will now examine the relationship that exists between these two outcomes.

The scatterplot in Figure 4.34 shows the relationship between attendance (measured as the percentage of recommended training sessions attended) and change in maximal aerobic power ($VO_{2max}$, an indicator of aerobic fitness) for all participants. There is a strong, positive, linear trend between attendance and change in $VO_{2max}$. The positive trend indicates that higher attendance rates are related to larger improvements in aerobic fitness. The regression line is

$$\text{Change in } VO_{2max} = -5.02 + 0.1935 \text{ Attendance}$$

The coefficient of determination of the regression line is $r^2 = 0.69$, meaning that 69% of the variation in change in $VO_{2max}$ is explained by attendance.

This model allows us to conclude that the improvement in aerobic fitness for the active gaming group is related to attendance. A higher attendance rate implies a higher volume of exercise, and volume of exercise is of key importance for health benefits. By promoting high attendance levels, active gaming ultimately led to improved health in this group of young males. Although these are encouraging findings, note that we cannot generalize them to other populations (young females, for example).
EXPLORING
STATISTICS

CLASS ACTIVITY
Guessing the Age
of Famous People

GOALS
In this activity, you will learn how to interpret the slope and intercept of a
regression line, using data you collect in class.

MATERIALS
Graph paper and a calculator or computer.

ACTIVITY
Your instructor will give you a list of names of famous people. Beside each name,
write your guess of the person’s age, in years. Even if you don’t know the age or don’t
know who the person is, give your best guess. If you work in a group, your group
should discuss the guessed ages and record the best guess of the group. After you’ve
finished, your instructor will give you a list of the actual ages of these people.

To examine the relationship between the actual ages and the ages you guessed,
make a scatterplot with actual age on the x-axis and guessed age on the y-axis. Use
technology to find the equation for the regression line and insert the line in the graph.
Calculate the correlation.

AFTER THE ACTIVITY
1. How would you describe the relationship between the ages you guessed and the
actual ages?
2. Is a regression line appropriate for your data? Explain why or why not.
3. What does it mean if a point falls above your regression line? Below your line?
4. What is the intercept of your line? What is the slope? Interpret the slope and
intercept. Explain what these tell you about your ability to guess the ages of these
people.

BEFORE THE ACTIVITY
1. Suppose you guessed every age correctly. What would be the equation of the
regression line? What would be the correlation?
2. What correlation do you think you will actually get? Why?
3. Suppose you consistently guess that people are older than they actually are. How
will the intercept of the regression line compare with your intercept in Question 1?
How about the slope?
Summary of Learning Objectives

1. What are the important features of a scatterplot that need to be discussed when describing the relationship between two numerical variables? You should discuss the trend, the strength of the trend, the shape of the trend, and any unusual observations, always in the context of the data.

2. What is the correlation coefficient, how do I calculate it, and how can I interpret it? The correlation coefficient is a number between −1 and 1 that measures the strength of the linear relationship between two variables.

**Formula 4.1:** Correlation Coefficient \( r \) = \( \frac{\sum (x - \overline{x})(y - \overline{y})}{n - 1}s_x s_y \)

It can only be interpreted if a scatterplot shows that the relationship between the variables is linear. In that case, the sign of \( r \) tells us whether the linear trend is positive or negative. Values of \( r \) close to −1 and 1 mean that the linear relationship is strong, whereas values close to 0 indicate that there is no linear relationship.

3. What is a regression line and how do I obtain it? The regression line tells us the average \( y \)-values for any value of \( x \). First obtain the slope:

**Formula 4.2a:** Slope \( \beta_1 = r \frac{s_y}{s_x} \)

Then find the intercept:

**Formula 4.2b:** Intercept \( \beta_0 = \overline{y} - \beta_1 \overline{x} \)

Write the regression line:

**Formula 4.2c:** Predicted \( y = \beta_0 + \beta_1 x \)

4. How do I interpret the slope and the intercept of the regression line? The regression line can only be interpreted if a scatterplot shows a linear relationship. In that case, the slope tells us how different the mean \( y \)-value is for observations that are 1 unit apart on the \( x \)-variable. The intercept is the mean value of \( y \) when \( x = 0 \), but it does not always have a useful interpretation.

5. What is the coefficient of determination, how do I calculate it, and how do I interpret it? The coefficient of determination (or \( r \)-squared) gives us the percentage of variation in the \( y \)-variable that is explained by the regression line. When a relationship is linear, it is calculated as the correlation coefficient squared (\( r^2 \)). The linear relationship is very strong if its value is close to 1, or 100%.

6. What are some common errors to avoid when using regression? Correlation does not imply causation: do not make cause-and-effect conclusions if the data are observational. Don’t extrapolate. Beware of outliers and influential points. Beware of conclusions you make from aggregate data.

Important Terms

| Scatterplot, 134 | Negative relationship (negative trend), 134 |
| Trend, 134     | Linear, 136 |
| Strength, 134  | Correlation coefficient, 138 |
| Shape, 134     | Regression line, 146 |
| Positive relation (positive trend), 134 | Intercept, 147 |
|                | Slope, 147 |
|                | Explanatory variable, predictor variable, independent variable, 154 |
|                | Response variable, predicted variable, dependent variable, 154 |
|                | Influential point, 163 |
|                | Aggregate data, 163 |
|                | Extrapolation, 164 |
|                | Regression toward the mean, 165 |
|                | Coefficient of determination, \( r^2 \), 165 |

Sources

SECTION 4.1

4.1 Predicting Land Value  Both figures concern the assessed value of land (with homes on the land), and both use the same data set.

a. Which do you think has a stronger relationship with value of the land—
   the number of acres of land or the number of rooms in the homes?
   Why?

b. If you were trying to predict the value of a parcel of land in this area
   (on which there is a home), would you be able to make a better
   prediction by knowing the acreage or the number of rooms in the
   house? Explain.

(Source: Minitab File, Student 12, "Assess")

4.3 Car Value and Age of Student  The figure shows the age of students and the value of their cars according to VMR Canada. Does it show an increasing trend, a decreasing trend, or very little trend? Explain.

4.4 Shoe Size and Grades  The figure shows a scatterplot of shoe size and grade on a statistics exam for some students. Does it show an increasing trend, a decreasing trend, or no trend? Is there a strong relationship?
4.5 Weight Loss  The scatterplot shows the actual weight and desired weight change of some students. Thus, if they weighed 70 kg and wanted to weigh 60, the desired weight change would be negative 10 kg.

Explain what you see. In particular, what does it mean that the trend is negative?

4.6 Comparing Salaries  The figure shows the salary and year of first employment for some professors at a university. Explain, in context, what the negative trend shows. Who makes the most and who makes the least? (Source: Minitab, Student 12, “Salary”)

4.7 Residential and Non-Residential Building Permits  (Example 1) The figure shows the residential and non-residential building permits issued in 2012, by metropolitan area. Explain what the trend shows. (Source: Canada Mortgage and Housing Corporation)

4.8 Household Income and Food Expenditure  The scatterplot shows data from a random sample of respondents to the 2009 Survey of Household Spending—the annual household income and the annual household expenditure for food. Describe and interpret the trend. (Source: Statistics Canada)

4.9 Adult Smoking and Youth Smoking  The figure shows the percentage of adults who smoke and the percentage of youth aged 12 to 19 who smoke. Each point represents a province (data on youth smoking were unavailable for Prince Edward Island). Describe what you see. Is the trend positive or negative? What does that mean? (Source: Statistics Canada)

4.10 Food Expenditures and Household Size  The figure shows a scatterplot of household size and yearly food expenditure for a sample of respondents to the Survey of Household Spending 2009. Describe what you see. Is the trend positive, negative, or near zero? Explain.
SECTION 4.2

4.11 Children’s Ages and Heights  The figure shows information about the ages and heights of several children. Why would it not be appropriate to find the correlation or to perform linear regression with this data set? Explain.

4.12 Blackjack Tips  The figure shows the amount of money won by people playing blackjack and the amount of tips they gave to the dealer, in dollars. Would it be appropriate to find a correlation for this data set? Explain.

4.13 University Education and Median Income  The scatterplot shows data from all provinces and territories taken from the Labour Force Survey—the percentage of the population aged 25 to 64 with university education and the median family income. Would it be appropriate to find the correlation for this data set? Explain.

4.14 Ages of Women Who Give Birth  The figure shows a scatterplot of birthrate (live births per 1000 women) and age of the mother in Canada. Would it make sense to find the correlation for this data set? Explain. According to this graph, at approximately what age is the highest fertility rate? (Source: Statistics Canada)

4.15 Do Older Students Get Better Grades?  On the basis of the scatterplot, do you think the correlation coefficient between age and average grade in a statistics class for this figure is positive, negative, or near zero?

4.16 Handspans  Refer to the figure. Is the correlation coefficient between the handspan of the dominant hand and that of the nondominant hand positive, negative, or near zero?
4.17 Matching Pick the letter of the graph that goes with each numerical value listed below for the correlation. Correlations:

- 0.767
- 0.299
- -0.980

4.18 Matching Pick the letter of the graph that goes with each numerical value listed below for the correlation. Correlations:

- -0.903
- 0.374
- 0.777

4.19 Cost of Flights (Example 2) The table for part a shows approximate distances between selected cities and the approximate cost of flights between those cities in November 2013.

a. Calculate the correlation of the numbers shown in the table by using a computer or statistical calculator.

<table>
<thead>
<tr>
<th>Cost ($)</th>
<th>Distance (kilometres)</th>
</tr>
</thead>
<tbody>
<tr>
<td>230</td>
<td>350</td>
</tr>
<tr>
<td>250</td>
<td>550</td>
</tr>
<tr>
<td>290</td>
<td>700</td>
</tr>
<tr>
<td>250</td>
<td>730</td>
</tr>
<tr>
<td>290</td>
<td>1050</td>
</tr>
<tr>
<td>380</td>
<td>1300</td>
</tr>
</tbody>
</table>

b. The table for part b shows the same information, except that the distance was converted to miles by dividing the number of kilometres by 1.609. What happens to the correlation when numbers are multiplied by a constant (we multiplied by 1/1.609)?

<table>
<thead>
<tr>
<th>Cost ($)</th>
<th>Distance (miles)</th>
</tr>
</thead>
<tbody>
<tr>
<td>230</td>
<td>217.5</td>
</tr>
<tr>
<td>250</td>
<td>341.8</td>
</tr>
<tr>
<td>290</td>
<td>435.1</td>
</tr>
<tr>
<td>250</td>
<td>453.7</td>
</tr>
<tr>
<td>290</td>
<td>652.6</td>
</tr>
<tr>
<td>380</td>
<td>808.0</td>
</tr>
</tbody>
</table>
c. Suppose that Transport Canada adds a tax to each flight. Fifty dollars are added to every flight, no matter how long it is. The table for part c shows the new data. What happens to the correlation when a constant is added to each number?

<table>
<thead>
<tr>
<th>Cost ($)</th>
<th>Distance (miles)</th>
</tr>
</thead>
<tbody>
<tr>
<td>280</td>
<td>217.5</td>
</tr>
<tr>
<td>300</td>
<td>341.8</td>
</tr>
<tr>
<td>340</td>
<td>435.1</td>
</tr>
<tr>
<td>300</td>
<td>453.7</td>
</tr>
<tr>
<td>340</td>
<td>652.6</td>
</tr>
<tr>
<td>430</td>
<td>808.0</td>
</tr>
</tbody>
</table>

b. Use the regression equation above the graph to get a more precise estimate of the median hourly pay for women for an occupation that has a median pay of 25 dollars per hour for men.

### 4.22 Number of Births and Population

The figure shows the number of births and the populations (in thousands) for the 13 provinces and territories in 2012, according to Statistics Canada. The correlation is 0.996.

- **a.** Find a rough estimate (by using the scatterplot) of the number of births in a province or territory with a population of about 4 million (4000 thousand).
- **b.** Use the regression equation above the graph to get a more precise estimate of the number of births for a province or territory with a population of 4 million.

<table>
<thead>
<tr>
<th>Predicted Births (K) = 0.85 + 0.0107 Population (K)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted Births (K) = 0.85 + 0.0107 Population (K)</td>
</tr>
</tbody>
</table>

### 4.23–4.26

Refer to the following table for Exercises 4.23–4.26. It provides various body measurements for some university women. The height is self-reported. All measurements are in centimetres.

- **Ht** is the height.
- **Head** is the circumference of the head.
- **HandL** is the length of the hand.
- **FootL** is the length of the foot.
- **HandW** is the width of the hand with the fingers spread wide.
- **Armspan** is the armspan with the arms spread wide.

<table>
<thead>
<tr>
<th>Ht</th>
<th>Head</th>
<th>HandL</th>
<th>FootL</th>
<th>HandW</th>
<th>Armspan</th>
</tr>
</thead>
<tbody>
<tr>
<td>157.5</td>
<td>56.0</td>
<td>17.0</td>
<td>23.0</td>
<td>20.0</td>
<td>159.5</td>
</tr>
<tr>
<td>165</td>
<td>54.0</td>
<td>17.5</td>
<td>23.5</td>
<td>18.0</td>
<td>161.0</td>
</tr>
<tr>
<td>168</td>
<td>55.0</td>
<td>18.0</td>
<td>24.0</td>
<td>16.5</td>
<td>161.5</td>
</tr>
<tr>
<td>160</td>
<td>58.0</td>
<td>18.5</td>
<td>23.5</td>
<td>19.5</td>
<td>160.5</td>
</tr>
<tr>
<td>173</td>
<td>55.0</td>
<td>19.0</td>
<td>25.0</td>
<td>18.0</td>
<td>173.0</td>
</tr>
<tr>
<td>142</td>
<td>54.5</td>
<td>16.5</td>
<td>22.0</td>
<td>17.0</td>
<td>142.0</td>
</tr>
<tr>
<td>155</td>
<td>50.0</td>
<td>16.0</td>
<td>22.0</td>
<td>18.0</td>
<td>155.0</td>
</tr>
<tr>
<td>170</td>
<td>53.5</td>
<td>18.0</td>
<td>23.5</td>
<td>15.0</td>
<td>165.0</td>
</tr>
<tr>
<td>168</td>
<td>55.5</td>
<td>16.0</td>
<td>21.0</td>
<td>20.0</td>
<td>166.0</td>
</tr>
<tr>
<td>162.5</td>
<td>57.0</td>
<td>19.0</td>
<td>24.5</td>
<td>20.0</td>
<td>163.0</td>
</tr>
<tr>
<td>156</td>
<td>55.0</td>
<td>19.0</td>
<td>24.5</td>
<td>21.0</td>
<td>152.0</td>
</tr>
</tbody>
</table>
**TRY 4.23 Height and Armspan for Women (Example 4)**

TI-83/84 output from a linear model for predicting armspan from height (both in cm) is given in the figure. Summary statistics are also provided.

![Regression equation and statistics](image)

To do parts a–c, assume that the relationship between armspan and height is linear.

a. Report the regression equation, using the words “Height” and “Armspan,” not x and y, using the output given.

b. Verify the slope by using the formula \( b_1 = \frac{s_y}{s_x} \).

c. Verify the y-intercept by using the formula \( b_0 = \bar{y} - b_1 \bar{x} \).

d. Using the regression equation, predict the armspan (in cm) for someone 163 cm tall.

**4.24 Hand and Foot Length for Women** Refer to the data shown in the table for Exercises 4.23–4.26. Some computer output is shown in the figure. Assume the trend is linear. Summary statistics for these data are shown below.

<table>
<thead>
<tr>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hand, x</td>
<td>17.682</td>
</tr>
<tr>
<td>Foot, y</td>
<td>23.318</td>
</tr>
</tbody>
</table>

The regression equation is \( Y = 5.67 + 0.998 X \)

Pearson correlation of HandL and FootL = 0.948

a. Report the regression equation, using the words “Hand” and “Foot,” not x and y.

b. Verify the slope by using the formula \( b_1 = \frac{s_y}{s_x} \).

c. Verify the y-intercept by using the formula \( b_0 = \bar{y} - b_1 \bar{x} \).

d. Using the regression equation, predict the foot length (in cm) for someone who has a hand length of 18 cm.

**4.25 Hand Width and Armspan for Women** Refer to the data shown in the table for Exercises 4.23–4.26. Use hand width as the predictor and armspan as the response.

a. Make a scatterplot of the data.

b. Explain why linear regression is probably not appropriate for these variables.

**4.26 Head Circumference and Hand Width for Women** Refer to the data shown in the table for Exercises 4.23–4.26. Use head circumference as x and hand width as y.

a. Make a scatterplot of the data.

b. Explain why linear regression is probably not appropriate for these variables.

**TRY 4.27 Height and Armspan for Men (Example 5)** Measurements were made for a sample of adult men. A regression line was fit to predict the men’s armspan from their height. The output from several different statistical technologies is provided. The scatterplot confirms that the relationship between armspan and height is linear.

a. Report the equation for predicting armspan from height. Use words such as armspan, not just x and y.

b. Report the slope and intercept from each technology, using all the digits given.
4.28 Hand Length and Foot Length for Men Measurements were made for a sample of adult men. Assume that the relationship between their hand length and foot length is linear. Output for predicting foot length from hand length is provided from several different statistical technologies.

a. Report the equation for predicting foot length from hand length. Use words like foot or foot length, not just $x$ and $y$.

b. Report the slope and intercept from each technology, using all the digits given.

4.29 Height and Head Circumference for Men Explain what makes this scatterplot hard to interpret. What should have been done differently?

4.30 Elbow-to-Wrist and Knee-to-Ankle Measurements The scatterplot shows measurements for the distance from elbow to wrist and from knee to ankle for some male students. Explain what makes this scatterplot difficult to interpret.

4.31 Comparing Correlation for Armspan and Height The correlation between height and armspan in a sample of adult women was found (in Exercise 4.23) to be $r = 0.950$. The correlation between armspan and height in a sample of adult men was found (in Exercise 4.27) to be $r = 0.829$. Which relationship—the relationship between height and armspan for women, or the relationship between height and armspan for men—is stronger? Explain.

4.32 Age and Weight for Men and Women The scatterplot shows a solid blue line for predicting weight from age of men; the dotted red line is for predicting weight from age of women. The data were collected from a large statistics class.

a. Which line is higher and what does that mean?

b. Which line has a steeper slope and what does that mean?
4.33 Social Insurance Number and Age The figure shows a scatterplot of the last two digits of some students’ Social Insurance numbers and their ages.

a. If a regression line were drawn on this graph, would it have a positive slope, a negative slope, or a slope near 0?
b. Give an estimate of the numerical value of the correlation between age and Social Insurance number.
c. Explain what this graph tells us about the relation between Social Insurance number and age.

4.34 Teeter-Totter The figure shows a scatterplot of the height of the left seat of a teeter-totter and the height of the right seat of the same teeter-totter. Estimate the numerical value of the correlation and explain the reason for your estimate.

4.35 Choosing the Predictor and Response (Example 6) Pick out which variable you think should be the predictor (x) and which variable should be the response (y). Explain your choices.

a. You collect data on the number of litres of gas it takes to fill up the tank after driving a certain number of kilometres. You wish to know how many kilometres you’ve driven based on the number of litres it took to fill up the tank.
b. You have data on laying date and clutch size (the number of eggs in a single nesting) for a certain species of owl. You want to estimate the clutch size for eggs laid on April 2.
c. You wish to buy a belt for a friend and know only his weight. You have data on the weight and waist sizes for a large sample of adult men.

4.36 Choosing the Predictor and Response Pick out which variable you think should be the predictor (x) and which variable should be the response (y). Explain your choices.

a. Weights of nuggets of gold (in grams) and their market value over the last few days are provided, and you wish to use this to estimate the value of a gold ring that weighs 115 grams.

b. You have data collected on the amount of time since chlorine was added to the public swimming pool and the concentration of chlorine still in the pool. (Chlorine evaporates over time.) Chlorine was added to the pool at 8 a.m., and you wish to know what the concentration is now, at 3 p.m.
c. You have data on the circumference of oak trees (measured 30 cm from the ground) and their age (in years). An oak tree in the park has a circumference of 90 cm, and you wish to know approximately how old it is.

4.37 Fertility Rate and Contraceptive Prevalence Rate (Example 7) The figure shows a scatterplot with the regression line. The data are for 145 countries. The predictor is the contraceptive prevalence rate (the percentage of women who are practising, or whose partners are practising, any form of contraception). The response is the fertility rate (in number of births per woman). The data came from the World Databank.

a. Explain what the trend shows.
b. Use the regression equation to predict the fertility rate of a country with a contraceptive prevalence rate of 70%.

4.38 Daily Newspaper Circulation The figure shows a fitted line plot. The data represent a random sample of 30 Canadian paid daily newspapers. The predictor is the population of the newspaper’s local market. The response is the newspaper’s average daily circulation. (Source: www.newspaperscanada.ca)

a. Explain what the trend shows.
b. Use the regression equation to predict the circulation of the Brandon Sun, given that the population of Brandon is 46 thousand people.
4.39 Drivers’ Fatalities and Ages The figure shows a graph of the death rate in motor vehicle traffic collisions and the age of the driver. The numbers came from Transport Canada.

a. Explain what the graph tells us about drivers at different ages; state which ages show the safest drivers and which show the most dangerous drivers.

b. Explain why it would not be appropriate to use these data for linear regression.

4.40 Do Women Tend to Live Longer Than Men? The figure shows life expectancy versus age for males and females in Canada (2009–2011), up to age 110. Females are represented by the blue circles and males by the red squares. These figures were reported by Statistics Canada.

a. Find your own age on the graph and estimate your life expectancy from the appropriate graph.

b. Would it make sense to find the best straight line for this graph? Why or why not?

c. Is it reasonable to predict the life expectancy for a person who is 120 from the regression line for these data? Why or why not?

d. Explain what it means that nearly all of the blue circles (for women) are above the red squares (for men). (Above the age of 100, the red squares cover the blue circles because both are in the same place.)

4.41 Does the Cost of a Flight Depend on the Distance? The table gives the round-trip fare for direct flights between Toronto and some other cities on major airlines. These were the lowest prices shown on www.ca.kayak.com on November 28, 2013. The airlines varied. (The travel dates were all the same, in mid-January.) See page 185 for guidance.

a. Without doing any calculations, predict whether the correlation and slope will be positive or negative. Explain your prediction.

b. Make a scatterplot with the number of nights (in thousands) on the x-axis and the amount of money spent (in millions of dollars) on the y-axis. Was your prediction correct?

Round-Trip Air Fares from Toronto

<table>
<thead>
<tr>
<th>City</th>
<th>Cost ($)</th>
<th>Distance (kilometres)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Montreal</td>
<td>227</td>
<td>506</td>
</tr>
<tr>
<td>Vancouver</td>
<td>513</td>
<td>3366</td>
</tr>
<tr>
<td>Halifax</td>
<td>374</td>
<td>1268</td>
</tr>
<tr>
<td>Fredericton</td>
<td>286</td>
<td>1041</td>
</tr>
<tr>
<td>Calgary</td>
<td>487</td>
<td>2716</td>
</tr>
<tr>
<td>Ottawa</td>
<td>227</td>
<td>354</td>
</tr>
<tr>
<td>Regina</td>
<td>454</td>
<td>2045</td>
</tr>
<tr>
<td>St. John’s</td>
<td>456</td>
<td>2117</td>
</tr>
<tr>
<td>Winnipeg</td>
<td>403</td>
<td>1516</td>
</tr>
<tr>
<td>Quebec City</td>
<td>249</td>
<td>729</td>
</tr>
<tr>
<td>Saskatoon</td>
<td>465</td>
<td>2227</td>
</tr>
<tr>
<td>Charlottetown</td>
<td>399</td>
<td>1314</td>
</tr>
<tr>
<td>Boston</td>
<td>286</td>
<td>696</td>
</tr>
<tr>
<td>New York</td>
<td>250</td>
<td>554</td>
</tr>
<tr>
<td>Orlando</td>
<td>296</td>
<td>1688</td>
</tr>
<tr>
<td>Mexico City</td>
<td>602</td>
<td>3256</td>
</tr>
<tr>
<td>Chicago</td>
<td>286</td>
<td>703</td>
</tr>
</tbody>
</table>

How much would it cost, on average, to fly to Washington, D.C., which is 566 kilometres from Toronto? To answer this question, perform a complete regression analysis, including a scatterplot with a regression line.

4.42 Travelling Abroad Canadians like to travel abroad, taking more than 30 million overnight trips in 2012. The table gives the number of nights spent (in thousands) and the amount of money spent (in millions of Canadian dollars) for the top 13 countries visited by Canadians outside North America in 2012. (Source: Statistics Canada, Tourism and the Centre for Education Statistics)

<table>
<thead>
<tr>
<th>Country</th>
<th>Nights</th>
<th>Spending in Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cuba</td>
<td>8947</td>
<td>748</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>10955</td>
<td>1,056</td>
</tr>
<tr>
<td>Dominican Republic</td>
<td>6704</td>
<td>674</td>
</tr>
<tr>
<td>France</td>
<td>9068</td>
<td>942</td>
</tr>
<tr>
<td>Italy</td>
<td>3897</td>
<td>480</td>
</tr>
<tr>
<td>Germany</td>
<td>3501</td>
<td>311</td>
</tr>
<tr>
<td>Mainland China</td>
<td>6445</td>
<td>521</td>
</tr>
<tr>
<td>Spain</td>
<td>2615</td>
<td>284</td>
</tr>
<tr>
<td>Jamaica</td>
<td>2189</td>
<td>248</td>
</tr>
<tr>
<td>Netherlands</td>
<td>1579</td>
<td>158</td>
</tr>
<tr>
<td>Hong Kong</td>
<td>2709</td>
<td>211</td>
</tr>
<tr>
<td>Republic of Ireland</td>
<td>2161</td>
<td>205</td>
</tr>
<tr>
<td>Australia</td>
<td>3808</td>
<td>329</td>
</tr>
</tbody>
</table>

a. Without doing any calculations, predict whether the correlation and slope will be positive or negative. Explain your prediction.

b. Make a scatterplot with the number of nights (in thousands) on the x-axis and the amount of money spent (in millions of dollars) on the y-axis. Was your prediction correct?
c. Find the numerical value of the correlation.
d. Find the value of the slope and explain what it means in context. Be careful with the units.
e. Explain why giving and interpreting the value for the intercept does not make sense in this situation.

4.43 Pumpkinseed Sunfish  Characterizing the growth pattern of fish is a crucial aspect of fisheries research and management. One widely used approach is based on the linear relationship between body length and scale length established for a sample of fish of a given species and population caught at the same time of year. For instance, consider the data on body length and scale length for a sample of pumpkinseed sunfish caught in the Otonobee River in Ontario in late September. (Source: Michael Fox, Trent University, unpublished data)

<table>
<thead>
<tr>
<th>Body Length (mm)</th>
<th>Scale Length (mm)</th>
<th>Body Length (mm)</th>
<th>Scale Length (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>47</td>
<td>0.458</td>
<td>109</td>
<td>2.063</td>
</tr>
<tr>
<td>48</td>
<td>0.500</td>
<td>122</td>
<td>2.500</td>
</tr>
<tr>
<td>45</td>
<td>0.417</td>
<td>116</td>
<td>2.458</td>
</tr>
<tr>
<td>75</td>
<td>1.617</td>
<td>125</td>
<td>2.708</td>
</tr>
<tr>
<td>81</td>
<td>1.250</td>
<td>135</td>
<td>2.958</td>
</tr>
<tr>
<td>78</td>
<td>1.542</td>
<td>133</td>
<td>2.771</td>
</tr>
<tr>
<td>64</td>
<td>1.125</td>
<td>135</td>
<td>2.917</td>
</tr>
<tr>
<td>85</td>
<td>1.792</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. Make a scatterplot of the data, with scale length on the x-axis and body length on the y-axis.
b. Find the numerical value for the correlation.
c. Find the equation of the regression line and insert the line in the scatterplot, or use technology to insert the line in your plot.
d. Report the value of the slope and explain what it means in context.
e. Report the value of the intercept, known as the “body-scale intercept.” This value is used by fish researchers to estimate the length of a fish at each age of its life before it was captured.

4.45 Calories and Fat  The table gives the calories and the total fat content (in grams) in a sample of Tim Hortons breakfast items. (Source: timhortons.com)
a. Do you expect the correlation between calories and fat content to be positive or negative? Why?
b. Make a scatterplot of the data.
c. Report the correlation.
d. Report the equation of the regression line and insert the line in the scatterplot, or use technology to insert the line in your plot.
e. Report the slope of the regression line and explain what it shows.

<table>
<thead>
<tr>
<th>Breakfast Item</th>
<th>Calories</th>
<th>Fat (g)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biscuit, ham, egg, cheese sandwich</td>
<td>400</td>
<td>21</td>
</tr>
<tr>
<td>Biscuit, sausage, egg, cheese sandwich</td>
<td>530</td>
<td>34</td>
</tr>
<tr>
<td>English muffin, egg, cheese sandwich</td>
<td>280</td>
<td>11</td>
</tr>
<tr>
<td>English muffin, bacon, egg, cheese sandwich</td>
<td>330</td>
<td>15</td>
</tr>
<tr>
<td>English muffin, ham, egg, cheese sandwich</td>
<td>300</td>
<td>11</td>
</tr>
<tr>
<td>English muffin, sausage, egg, cheese sandwich</td>
<td>430</td>
<td>25</td>
</tr>
<tr>
<td>Sausage, egg, cheese wrap</td>
<td>420</td>
<td>28</td>
</tr>
<tr>
<td>Bacon, egg, cheese wrap</td>
<td>320</td>
<td>18</td>
</tr>
<tr>
<td>Bagel BELT</td>
<td>460</td>
<td>15</td>
</tr>
<tr>
<td>Sausage bagel BELT</td>
<td>560</td>
<td>25</td>
</tr>
<tr>
<td>Hash brown</td>
<td>100</td>
<td>5</td>
</tr>
<tr>
<td>Oatmeal – maple</td>
<td>220</td>
<td>2.5</td>
</tr>
</tbody>
</table>
4.52 Coefficient of Determination

Assume that the relationship between height and weight is linear. Assume that height is the predictor and weight is the response, and coefficient of determination (as a percent) and explain what it means. If the correlation between height and weight of a large group of people is 0.67, find the coefficient of determination (as a percent) and explain what it means.

4.54 Test Scores

Suppose that students who scored much lower than the mean on their first statistics test were given special tutoring in the subject. Suppose that they tended to show some improvement on the next test. Explain what might cause the rise in grades other than the tutoring program itself.

TRY

4.55 Salary and Year of Employment (Example 8)

The equation for the regression line relating the salary and the year first employed is given above the figure.

a. Report the slope and explain what it means.

b. Either interpret the intercept ($4,255,000) or explain why it is not appropriate to interpret the intercept.

**SECTION 4.4**

4.47 Answer the questions using complete sentences.

a. What is an influential point?

b. It has been noted that people who go to church frequently tend to have lower blood pressure than people who don’t go to church. Does this mean you can lower your blood pressure by going to church? Why or why not? Explain.

4.48 Answer the questions, using complete sentences.

a. What is extrapolation and why is it a bad idea in regression analysis?

b. How is the coefficient of determination related to the correlation, and what does the coefficient of determination show?

4.49 If there is a positive correlation between number of years studying math and shoe size (for children), does that prove that larger shoes cause more studying of math, or vice versa? Can you think of a hidden variable that might be influencing both of the other variables?

4.50 Suppose that the growth rate of children looks like a straight line if the height of a child is observed at the ages of 24 months, 28 months, 32 months, and 36 months. If you use the regression line if the height of a child is observed at the ages of 24 months, 25 years, and 30 years, you might find that the predicted height is 6 metres. What is wrong with the prediction and the process used?

4.51 Coefficient of Determination

If the correlation between height and weight of a large group of people is 0.67, find the coefficient of determination (as a percent) and explain what it means. Assume that height is the predictor and weight is the response, and assume that the relationship between height and weight is linear.

4.52 Coefficient of Determination

Does a correlation of −0.70 or +0.50 give a larger coefficient of determination? We say that the linear relationship having the larger coefficient of determination is more strongly correlated. Which of the values shows a stronger correlation?

4.53 Decrease in Cholesterol

A doctor is studying cholesterol readings in his patients. After reviewing the cholesterol readings, he calls the patients with the highest cholesterol readings (the top 5% of readings in his office) and asks them to come back to discuss cholesterol-lowering methods. When he tests these patients a second time, the average cholesterol readings tended to have gone down somewhat. Explain what statistical phenomenon might have been partly responsible for this lowering of the readings.

4.54 Test Scores

Suppose that students who scored much lower than the mean on their first statistics test were given special tutoring in the subject. Suppose that they tended to show some improvement on the next test. Explain what might cause the rise in grades other than the tutoring program itself.

Try

4.55 Salary and Year of Employment (Example 8)
The equation for the regression line relating the salary and the year first employed is given above the figure.

a. Report the slope and explain what it means.

b. Either interpret the intercept ($4,255,000) or explain why it is not appropriate to interpret the intercept.

**SECTION 4.4**

4.47 Answer the questions using complete sentences.

a. What is an influential point?

b. It has been noted that people who go to church frequently tend to have lower blood pressure than people who don’t go to church. Does this mean you can lower your blood pressure by going to church? Why or why not? Explain.

4.48 Answer the questions, using complete sentences.

a. What is extrapolation and why is it a bad idea in regression analysis?

b. How is the coefficient of determination related to the correlation, and what does the coefficient of determination show?

4.49 If there is a positive correlation between number of years studying math and shoe size (for children), does that prove that larger shoes cause more studying of math, or vice versa? Can you think of a hidden variable that might be influencing both of the other variables?

4.50 Suppose that the growth rate of children looks like a straight line if the height of a child is observed at the ages of 24 months, 28 months, 32 months, and 36 months. If you use the regression line if the height of a child is observed at the ages of 24 months, 25 years, and 30 years, you might find that the predicted height is 6 metres. What is wrong with the prediction and the process used?

4.51 Coefficient of Determination

If the correlation between height and weight of a large group of people is 0.67, find the coefficient of determination (as a percent) and explain what it means. Assume that height is the predictor and weight is the response, and assume that the relationship between height and weight is linear.

4.52 Coefficient of Determination

Does a correlation of −0.70 or +0.50 give a larger coefficient of determination? We say that the linear relationship having the larger coefficient of determination is more strongly correlated. Which of the values shows a stronger correlation?

4.53 Decrease in Cholesterol

A doctor is studying cholesterol readings in his patients. After reviewing the cholesterol readings, he calls the patients with the highest cholesterol readings (the top 5% of readings in his office) and asks them to come back to discuss cholesterol-lowering methods. When he tests these patients a second time, the average cholesterol readings tended to have gone down somewhat. Explain what statistical phenomenon might have been partly responsible for this lowering of the readings.

**SECTION 4.4**

4.47 Answer the questions using complete sentences.

a. What is an influential point?

b. It has been noted that people who go to church frequently tend to have lower blood pressure than people who don’t go to church. Does this mean you can lower your blood pressure by going to church? Why or why not? Explain.

4.48 Answer the questions, using complete sentences.

a. What is extrapolation and why is it a bad idea in regression analysis?

b. How is the coefficient of determination related to the correlation, and what does the coefficient of determination show?

4.49 If there is a positive correlation between number of years studying math and shoe size (for children), does that prove that larger shoes cause more studying of math, or vice versa? Can you think of a hidden variable that might be influencing both of the other variables?

4.50 Suppose that the growth rate of children looks like a straight line if the height of a child is observed at the ages of 24 months, 28 months, 32 months, and 36 months. If you use the regression line if the height of a child is observed at the ages of 24 months, 25 years, and 30 years, you might find that the predicted height is 6 metres. What is wrong with the prediction and the process used?

4.51 Coefficient of Determination

If the correlation between height and weight of a large group of people is 0.67, find the coefficient of determination (as a percent) and explain what it means. Assume that height is the predictor and weight is the response, and assume that the relationship between height and weight is linear.

4.52 Coefficient of Determination

Does a correlation of −0.70 or +0.50 give a larger coefficient of determination? We say that the linear relationship having the larger coefficient of determination is more strongly correlated. Which of the values shows a stronger correlation?

4.53 Decrease in Cholesterol

A doctor is studying cholesterol readings in his patients. After reviewing the cholesterol readings, he calls the patients with the highest cholesterol readings (the top 5% of readings in his office) and asks them to come back to discuss cholesterol-lowering methods. When he tests these patients a second time, the average cholesterol readings tended to have gone down somewhat. Explain what statistical phenomenon might have been partly responsible for this lowering of the readings.

### Table: Calorie and Fat Content of Donuts and Cookies

<table>
<thead>
<tr>
<th>Item</th>
<th>Calories</th>
<th>Fat (g)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sour cream glazed donut</td>
<td>340</td>
<td>17</td>
</tr>
<tr>
<td>Chocolate dip donut</td>
<td>210</td>
<td>8</td>
</tr>
<tr>
<td>Old fashion plain donut</td>
<td>260</td>
<td>19</td>
</tr>
<tr>
<td>Old fashion glazed donut</td>
<td>320</td>
<td>19</td>
</tr>
<tr>
<td>Double chocolate donut</td>
<td>250</td>
<td>10</td>
</tr>
<tr>
<td>Walnut crunch donut</td>
<td>360</td>
<td>23</td>
</tr>
<tr>
<td>Honey cruller</td>
<td>320</td>
<td>19</td>
</tr>
<tr>
<td>White chocolate macadamia nut cookie</td>
<td>240</td>
<td>12</td>
</tr>
<tr>
<td>Oatmeal raisin spice cookie</td>
<td>220</td>
<td>8</td>
</tr>
<tr>
<td>Chocolate chunk cookie</td>
<td>230</td>
<td>9</td>
</tr>
<tr>
<td>Triple chocolate cookie</td>
<td>250</td>
<td>13</td>
</tr>
<tr>
<td>Caramel chocolate pecan cookie</td>
<td>230</td>
<td>11</td>
</tr>
</tbody>
</table>

**TRY**

4.55 Salary and Year of Employment (Example 8)
The equation for the regression line relating the salary and the year first employed is given above the figure.

a. Report the slope and explain what it means.

b. Either interpret the intercept ($4,255,000) or explain why it is not appropriate to interpret the intercept.

**4.56 Fuel Consumption: Highway and City**
The figure shows the relationship between the fuel consumption (in litres per 100 kilometres) on the highway and in the city for some cars. (Source: Natural Resources Canada, Fuel Consumption Guide 2013)

a. Report the slope and explain what it means.

b. Either interpret the intercept ($4,255,000) or explain why it is not appropriate to interpret the intercept.

### Table: Calories and Fat Content of Donuts and Cookies

<table>
<thead>
<tr>
<th>Item</th>
<th>Calories</th>
<th>Fat (g)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sour cream glazed donut</td>
<td>340</td>
<td>17</td>
</tr>
<tr>
<td>Chocolate dip donut</td>
<td>210</td>
<td>8</td>
</tr>
<tr>
<td>Old fashion plain donut</td>
<td>260</td>
<td>19</td>
</tr>
<tr>
<td>Old fashion glazed donut</td>
<td>320</td>
<td>19</td>
</tr>
<tr>
<td>Double chocolate donut</td>
<td>250</td>
<td>10</td>
</tr>
<tr>
<td>Walnut crunch donut</td>
<td>360</td>
<td>23</td>
</tr>
<tr>
<td>Honey cruller</td>
<td>320</td>
<td>19</td>
</tr>
<tr>
<td>White chocolate macadamia nut cookie</td>
<td>240</td>
<td>12</td>
</tr>
<tr>
<td>Oatmeal raisin spice cookie</td>
<td>220</td>
<td>8</td>
</tr>
<tr>
<td>Chocolate chunk cookie</td>
<td>230</td>
<td>9</td>
</tr>
<tr>
<td>Triple chocolate cookie</td>
<td>250</td>
<td>13</td>
</tr>
<tr>
<td>Caramel chocolate pecan cookie</td>
<td>230</td>
<td>11</td>
</tr>
</tbody>
</table>
f. Add a new point to your data: a 14-kilogram turkey that is free. Give the new value for \( r \) and the new regression equation. Explain what the negative correlation implies. What happened?

<table>
<thead>
<tr>
<th>Weight (kilograms)</th>
<th>Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.58</td>
<td>$17.10</td>
</tr>
<tr>
<td>8.4</td>
<td>$23.87</td>
</tr>
<tr>
<td>9.13</td>
<td>$26.73</td>
</tr>
<tr>
<td>7.58</td>
<td>$19.87</td>
</tr>
<tr>
<td>7.08</td>
<td>$23.24</td>
</tr>
<tr>
<td>4.63</td>
<td>$ 9.08</td>
</tr>
</tbody>
</table>

4.58 Office Space and Parking Rates The figures show the average monthly parking rate downtown and the office space available (in million square feet) downtown for some of Canada’s most populated cities. Comment on the difference in graphs and in the coefficient of determination between (a) the graph that included Toronto, and (b) the one that did not include Toronto. Toronto is the point with 40.1 million square feet of office space.

4.60 Teachers’ Pay and Student–Educator Ratios The figure shows a scatterplot with a regression line for educators’ average pay and the student–educator ratio (in students per educator) for public schooling in each province in 2008, according to Statistics Canada.

4.61 Does Having a Job Affect Students’ Grades? (Example 10) Grades on a political science test and the number of hours of paid work in the week before the test were studied. The instructor was trying to predict the grade on a test from the hours of work. The figure shows a scatterplot and the regression line for these data.

TRY 4.59 Teachers’ Pay and Education Expenditure (Example 9) The figure shows a scatterplot with a regression line for educators’ average pay and the expenditure per student for public schooling in each province in 2008, according to Statistics Canada.

Try

b. Interpret the slope.

c. Interpret the intercept or explain why it should not be interpreted.

TRY
4.62 Waste Disposal and Population  The figure shows a scatterplot with a regression line for amount of waste disposal (in tonnes) and population (in thousands) by province in 2010, according to Statistics Canada.

a. Is the trend positive or negative? What does that mean?

b. Calculate the correlation between waste disposal and population. (Use R-Sq from the figure and take the square root of it.)

c. Report the slope. How many additional tonnes of waste disposal are there, on average, for each additional thousand people?

d. Either report the intercept or explain why it is not appropriate to interpret it.

4.63 Education of Fathers and Mothers  The data shown in the table are the number of years of formal education of the fathers and mothers of a sample of 29 statistics students at a small community college in an area with many recent immigrants. (The means are both about 8, and the standard deviations are both about 4.) The scatterplot (not shown) suggests a linear trend.

<table>
<thead>
<tr>
<th>Father</th>
<th>Mother</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>11</td>
<td>13</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>8</td>
<td>16</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>12</td>
<td>8</td>
</tr>
<tr>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>8</td>
<td>12</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>12</td>
<td>6</td>
</tr>
</tbody>
</table>

4.64 Heights of Fathers and Sons  The table shows some data on a sample of heights of fathers and their sons.

You may want to use the computer or a statistics calculator to verify that

\[ \bar{x}_{\text{father}} = 175.85 \quad \bar{x}_{\text{son}} = 175.4 \]
\[ s_{\text{father}} = 7.166 \quad s_{\text{son}} = 7.003 \]
\[ r = 0.765 \]

The scatterplot (not shown) suggests a linear trend.

a. Find and report the regression equation for predicting the son’s height from the father's height. Then predict the height of a son of a father who is 188 centimetres tall.

b. Find and report the regression equation for predicting the father’s height from the son’s height. Then predict a father’s height from that of a son who is 188 centimetres tall.

c. What phenomenon does this show?
4.65 Test Scores Assume that in a political science class, the teacher gives a midterm exam and a final exam. Assume that the relationship between midterm and final scores is linear. The summary statistics have been simplified for clarity.

Midterm: Mean = 75, Standard deviation = 10
Final: Mean = 75, Standard deviation = 10
Also, \( r = 0.7 \) and \( n = 20 \).

According to the regression equation, for a student who gets a 95 on the midterm, what is the predicted final exam grade? What phenomenon from the chapter does this demonstrate? Explain. See page 186 for guidance.

4.66 Test Scores Assume that in a sociology class, the teacher gives a midterm exam and a final exam. Assume that the relationship between midterm and final scores is linear. Here are the summary statistics:

Midterm: Mean = 72, Standard deviation = 8
Final: Mean = 72, Standard deviation = 8
Also, \( r = 0.75 \) and \( n = 28 \).

a. Find and report the equation of the regression line to predict the final exam score from the midterm score.
b. For a student who gets 55 on the midterm, predict the final exam score.
c. Your answer to part b should be higher than 55. Why?
d. For a student who gets a 100 on the midterm, without doing any calculations, state whether the predicted value would be higher, lower, or the same as 100.

4.67 Heights and Weights of People

The table shows the height and weight of some people. The figure shows that the relationship is linear enough to proceed.

<table>
<thead>
<tr>
<th>Height (inches)</th>
<th>Weight (pounds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td>105</td>
</tr>
<tr>
<td>66</td>
<td>140</td>
</tr>
<tr>
<td>72</td>
<td>185</td>
</tr>
<tr>
<td>70</td>
<td>145</td>
</tr>
<tr>
<td>63</td>
<td>120</td>
</tr>
</tbody>
</table>

a. Calculate the correlation, and find and report the equation of the regression line, using height as the predictor and weight as the response.
b. Change the height to centimetres by multiplying each height in inches by 2.54. Find the weight in kilograms by dividing the weight in pounds by 2.205. Retain at least six digits in each number so that there will be no errors due to rounding.
c. Calculate the correlation between height in centimetres and weight in kilograms, and compare it with the correlation for the heights in inches and the weights in pounds.
d. Find the equation of the regression line for predicting weight from height, using height in cm and weight in kg. Is the equation for weight (in pounds) and height (in inches) the same or different from the equation for weight (in kg) and height (in cm)?

4.68 Heights and Weights of Men

The table shows the heights (in inches) and weights (in pounds) of 14 university men. The figure shows that the relationship is linear enough to proceed.

<table>
<thead>
<tr>
<th>Height (inches)</th>
<th>Weight (pounds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>68</td>
<td>205</td>
</tr>
<tr>
<td>68</td>
<td>168</td>
</tr>
<tr>
<td>74</td>
<td>230</td>
</tr>
<tr>
<td>68</td>
<td>190</td>
</tr>
<tr>
<td>67</td>
<td>185</td>
</tr>
<tr>
<td>69</td>
<td>190</td>
</tr>
<tr>
<td>68</td>
<td>165</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Height (inches)</th>
<th>Weight (pounds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>70</td>
<td>200</td>
</tr>
<tr>
<td>69</td>
<td>175</td>
</tr>
<tr>
<td>72</td>
<td>210</td>
</tr>
<tr>
<td>72</td>
<td>205</td>
</tr>
<tr>
<td>72</td>
<td>185</td>
</tr>
<tr>
<td>73</td>
<td>200</td>
</tr>
<tr>
<td>73</td>
<td>195</td>
</tr>
</tbody>
</table>

a. Find the equation for the regression line with weight (in pounds) as the response and height (in inches) as the predictor. Report the slope and intercept of the regression line and explain what they show. If the intercept is not appropriate to report, explain why.
b. Find the correlation between weight (in pounds) and height (in inches).
c. Find the coefficient of determination and interpret its value.
d. If you changed each height to centimetres by multiplying heights in inches by 2.54, what would the new correlation be? Explain.
e. Find the equation with weight (in pounds) as the response and height (in cm) as the predictor, and interpret the slope.
f. Summarize what you found: Does changing units change the correlation? Does changing units change the regression equation?

4.69 House Price and Square Feet
The figure gives a scatterplot, using data taken from Peterborough This Week Best Homes (November, 2013), of the number of square feet in some houses and their list price for sale.

a. Choose the correct correlation from these choices: +1.00, 0, −1.00, −0.84, +0.84.
b. The equation of the regression line is given above the graph. Report the slope and intercept of the regression line and explain what they show in this context. If the intercept is not appropriate to report, explain why.
c. By looking at the graph or using the equation, predict the average cost of a house that is about 2000 square feet.
d. Find the coefficient of determination from the correlation that you chose in part a and explain what it means in the context of the problem.

4.70 Heights and Weights of Grade 1 and 2 Students
Some statistics students recorded the heights and weights of a sample of Grade 1 and 2 students. Refer to the figure.

a. Choose the correct correlation coefficient from these choices: +1.00, 0, −1.00, −0.84, +0.84.
b. The equation is given above the graph. Report the slope and intercept of the regression line and explain what they show in this context. If the intercept is not appropriate to report, explain why.
c. By looking at the graph or using the equation of the regression line, predict the weight for a Grade 1 or 2 student who is 125 centimetres tall.
d. Find the coefficient of determination from the correlation that you chose in part a and explain what it means in the context of the problem.

4.71 Hours of Study and of TV Viewing
The number of hours of study and the number of hours of TV viewing per week were recorded for some community college students.
a. Make a scatterplot of the data, using hours of study as the predictor.
b. Describe the relationship, if any, between hours of study and hours of TV viewing.
c. Is the correlation strong and positive, strong and negative, or near zero? Do not calculate.

4.72 Salary and Age
The age and salary (dollars per hour) were recorded for some students with jobs.
a. Make a scatterplot of the data, using age as the predictor.
b. Find the equation of the regression line.
c. Report the slope and intercept of the regression line and explain what they show. If the intercept is not appropriate to report, explain why not.

4.73 Age and Weight for Women
The figure shows a scatterplot and regression line for the age and weight of some women in a statistics class.
a. Find the correlation by taking the positive square root of R-Sq shown on the graph and attaching the proper sign.
b. Using the equation given, find the predicted weight for a woman 35 years old.
c. Report the slope and explain what it means. This study did not follow the same people over time.
d. If it is appropriate, report the intercept and explain what it means. Otherwise, explain why it is not appropriate to report.

4.74 Age and Weight for Men
The figure on the next page shows a scatterplot and regression line for the age and weight of some men in a statistics class.
a. Find the correlation by taking the positive square root of R-Sq shown on the graph and attaching the proper sign.
b. Using the equation given, find the predicted weight for a man 35 years old.
c. Report the slope and explain what it means. (Note that this study did not follow the same people over time.)
d. Either report the intercept and explain what it shows or explain why it should not be reported.
4.75 Film Opening Weekend and Total Gross  The figure shows a scatterplot of the opening weekend gross and the total gross in Canada and the United States for some of the top movies of 2013. Describe what you see. Explain the trend and mention any unusual points. (Source: www.boxofficemojo.com)

4.76 Film Opening Weekend and Total Gross  The figure shows a scatterplot of the opening weekend gross and the total gross in Canada and the United States for some of the top movies of 2013. Describe what you see. Explain the trend and mention any unusual points. (Source: www.boxofficemojo.com)

4.77 Student Pulses  A group of students took their pulse (in beats per minute) and a few seconds later took their pulse again. The figure shows the results. Explain what they show.

4.78 Hours of Exercise and Hours of Studying  A group of students reported the number of hours they exercised per week and the number of hours they spent studying statistics per week. The figure shows the results. Explain what they show.

4.79 Tree Heights  Loggers gathered information about some trees. The diameter is in centimetres, the height is in metres, and the volume of the wood is in cubic metres. Loggers are interested in whether they can estimate the volume of the tree given any single dimension. Which is the better predictor of volume: the diameter or the height?

4.80 Salary and Education  Does education pay? The salary per year in dollars, the number of years employed (YrsEm), and the number of years of education after high school (Educ) for the employees of a company were recorded. Determine whether number of years employed or number of years of education after high school is a better predictor of salary. Explain your thinking. (Source: Minitab File)

4.81 Film Budgets and Grosses  Movie studios put a lot of effort trying to predict how much money their movies will make. One possible predictor is the amount of money spent on the production of the movie. In the table on the next page you can see the budget and the amount of money made worldwide for the 15 movies with the highest profit (as of November 2013). The budget and gross are in millions of dollars. Make a scatterplot and comment on what you see. If appropriate, find, report, and interpret the regression line. If it is not appropriate to do so, explain why. (Source: www.boxofficemojo.com)
4.82 Fuel Consumption of Cars  The table gives the fuel consumption (in litres per 100 kilometres) in the city and on the highway for some of the most fuel-efficient cars of 2013, as reported by Consumer Reports. Make a scatterplot, using the city fuel consumption as the predictor. Find the equation of the regression line for predicting the litres per 100 kilometres (L/100 km) on the highway from the litres per 100 kilometres in the city. Use the equation to predict the highway fuel consumption from a city fuel consumption of 4.2 L/100 km. Also find the coefficient of determination and explain what it means.

<table>
<thead>
<tr>
<th>Movie</th>
<th>Budget (millions of dollars)</th>
<th>Gross (millions of dollars)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Titanic</td>
<td>200</td>
<td>2185.7</td>
</tr>
<tr>
<td>The Lord of the Rings:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>The Return of the King</td>
<td>94</td>
<td>1141.4</td>
</tr>
<tr>
<td>Jurassic Park</td>
<td>63</td>
<td>1035.3</td>
</tr>
<tr>
<td>Transformers:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dark of the Moon</td>
<td>195</td>
<td>1123.8</td>
</tr>
<tr>
<td>Marvel’s The Avengers</td>
<td>225</td>
<td>1514.3</td>
</tr>
<tr>
<td>The Dark Knight Rises</td>
<td>275</td>
<td>1079.3</td>
</tr>
<tr>
<td>Pirates of the Caribbean:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dead Man’s Chest</td>
<td>225</td>
<td>1060.6</td>
</tr>
<tr>
<td>Iron Man 3</td>
<td>200</td>
<td>1212</td>
</tr>
<tr>
<td>Harry Potter and the</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Deadly Hollows Part 2</td>
<td>125</td>
<td>1328.1</td>
</tr>
<tr>
<td>Toy Story 3</td>
<td>200</td>
<td>1063.8</td>
</tr>
<tr>
<td>Avatar</td>
<td>425</td>
<td>2783.9</td>
</tr>
<tr>
<td>Skyfall</td>
<td>200</td>
<td>1108.7</td>
</tr>
<tr>
<td>Pirates of the Caribbean:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>On Stranger Tides</td>
<td>250</td>
<td>1043.7</td>
</tr>
<tr>
<td>Alice in Wonderland</td>
<td>200</td>
<td>1024.4</td>
</tr>
<tr>
<td>The Hobbit:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>An Unexpected Journey</td>
<td>250</td>
<td>1014.7</td>
</tr>
</tbody>
</table>

4.83 Construct a small set of numbers with at least three points with a perfect positive correlation of 1.00.

4.84 Construct a small set of numbers with at least three points with a perfect negative correlation of $-1.00$.

4.85 Construct a set of numbers (with at least three points) with a strong negative correlation. Then add one point (an influential point) that changes the correlation to positive. Report the data and give the correlation of each set.

4.86 Construct a set of numbers (with at least three points) with a strong positive correlation. Then add one point (an influential point) that changes the correlation to negative. Report the data and give the correlation of each set.

For Exercises 4.83–4.86 show your points in a rough scatterplot and give the coordinates of the points.

<table>
<thead>
<tr>
<th>Vehicle</th>
<th>City (L/100 km)</th>
<th>Highway (L/100 km)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ford Focus</td>
<td>1.7</td>
<td>2.0</td>
</tr>
<tr>
<td>Nissan Leaf</td>
<td>1.8</td>
<td>2.3</td>
</tr>
<tr>
<td>Chevrolet Volt</td>
<td>2.3</td>
<td>2.5</td>
</tr>
<tr>
<td>Toyota Prius Plug-in</td>
<td>2.0</td>
<td>1.9</td>
</tr>
<tr>
<td>Toyota Prius</td>
<td>3.5</td>
<td>4.0</td>
</tr>
<tr>
<td>Lexus CT 200h</td>
<td>4.5</td>
<td>4.8</td>
</tr>
<tr>
<td>Honda Civic Hybrid</td>
<td>4.4</td>
<td>4.2</td>
</tr>
<tr>
<td>Ford Fusion Hybrid</td>
<td>4.0</td>
<td>4.1</td>
</tr>
<tr>
<td>Smart Fortwo Coupe</td>
<td>5.8</td>
<td>4.7</td>
</tr>
</tbody>
</table>

The table gives the budget and gross revenue (in millions of dollars) for some of the highest-grossing movies of all time.
Step 1 ▶ Make a scatterplot
Be sure that distance is the $x$-variable and cost is the $y$-variable. You may want to include the regression line with your scatterplot if you have that option with your technology; refer to Step 4.

Step 2 ▶ Is the linear model appropriate?
In this case the answer is yes, because there is a linear trend. It is hard to see with so few points, but a strong curvature is not present and the cost tends to increase as the distance increases.

Step 3 ▶ Obtain the equation
When finding the regression equation, be sure that you use distance as the $x$-variable and cost as the $y$-variable. For example, if you are using the TI-83/84, the predictor (which you probably put into List 2) because the second column of numbers has the kilometres) has to be entered first and the response (in List 1) is entered second, after the comma. See the TI-83/84 figure.

Step 4 ▶ Add the regression line to the scatterplot
If your technology will make a plot with a line, do so. Refer to the TechTips, which begin on page 187, to see how to do this. If your technology will not draw the line, you can choose two $x$-values (distances), find the corresponding $y$-values (cost), plot these, and then draw the line to connect them. For example, to choose the first point on the line, choose an $x$-value such as $x = 400$ and follow these additional steps:

- **First Predicted Point**
  Pick an arbitrary small distance, such as 400 kilometres, that is still larger than the smallest distance given. Substitute it into the equation you got to find the predicted cost:
  
  \[ \text{Predicted Cost} = 195.02 + 0.113 \times \text{Distance} \]
  
  \[ = 195.02 + 0.113(400) \]
  
  \[ = 240.22 \]
  
  A predicted point is (400, 240.22), so a flight of about 400 kilometres should cost, on average, about $240.20. Put the point on the graph with a symbol you will remember.

- **Second Point**
  To get a second point, pick an arbitrary large distance, such as 3000 kilometres, that is still smaller than the largest distance given. Substitute it into the equation to find the cost and put the point on the graph using the same symbol you chose before.

- **The Line**
  The regression line will be a straight line between these two predicted points (use a ruler to make the line). Extend the line out to the edges of the data, to the left to about 300 kilometres, and to the right to about 3500 kilometres.

Step 5 ▶ Interpret the slope and intercept in context
Predicted \[ y = b_0 + b_1 x \]

Predicted Cost = 195.02 + 0.113 Distance

The slope is $b_1$ (which is the multiplier for Distance), which is 0.113, and the intercept is $b_0$, which is the first number, $195.02$. Fill in the blanks that follow.

For the slope: For every additional kilometre, on average, the price goes up by \( \$ \) dollars.

For the intercept: A trip with zero kilometres should cost about \( \$ \) dollars. Explain why interpreting the $y$-intercept like this is questionable.

Step 6 ▶ Answer the question by using the regression equation
How much would it cost to fly to Washington, D.C., which is 566 kilometres from Toronto?
4.65 Test Scores

Assume that in a political science class, the teacher gives a midterm exam and a final exam. Assume that the association between midterm and final scores is linear. The summary statistics have been simplified for clarity.

Midterm: Mean = 75, Standard deviation = 10
Final: Mean = 75, Standard deviation = 10
Also, \( r = 0.7 \) and \( n = 20 \).

For a student who gets 95 on the midterm, what is the predicted final exam grade? Assume that the graph is linear.

**Step 1**

Find the equation of the line to predict the final exam score from the midterm score: \( y = b_0 + b_1x \)

a. First find the slope: \( b_1 = r \left( \frac{s_{\text{final}}}{s_{\text{midterm}}} \right) \)
b. Then find the \( y \)-intercept, \( b_0 \), from the equation
   \[ b_0 = \bar{y} - b_1 \bar{x} \]
c. Write out the following equation:
   \[ \text{Predicted } y = b_0 + b_1x \]

However, use “Predicted Final” instead of “Predicted \( y \)” and “Midterm” in place of \( x \).

**Step 2**

Use the equation to predict the final exam score for a student who gets 95 on the midterm.

**Step 3**

Your predicted final exam grade should be less than 95. Why?
General Instructions for All Technology

Upload data from the website, or enter data manually using two columns of equal length. Refer to TechTips in Chapter 2 for a review of entering data. Each row represents a single observation, and each column represents a variable. All technologies will use the example that follows.

Example. Analyze the six points in the data table with a scatterplot, correlation, and regression. Use heights (in centimetres) as the $x$-variable and weight (in kilograms) as the $y$-variable.

<table>
<thead>
<tr>
<th>Height</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>155</td>
<td>47</td>
</tr>
<tr>
<td>158</td>
<td>50</td>
</tr>
<tr>
<td>160</td>
<td>64</td>
</tr>
<tr>
<td>163</td>
<td>57</td>
</tr>
<tr>
<td>168</td>
<td>77</td>
</tr>
<tr>
<td>173</td>
<td>73</td>
</tr>
</tbody>
</table>

Making a Scatterplot
1. Press 2ND, STATPLOT (which is the button above 2ND), 4, and ENTER, to turn off plots made previously.
2. Press 2ND, STATPLOT, and 1 (for Plot1).
3. Refer to Figure 4A: Turn on Plot1 by pressing ENTER when On is highlighted.
4. Use the arrows on the keypad to get to the scatterplot (upper left of the six plots) and press ENTER when the scatterplot is highlighted. Be careful with the Xlist and Ylist. To get L1, press 2ND and 1. To get L2, press 2ND and 2.
5. Press ZOOM and 9 (Zoomstat) to create the graph.
6. Press TRACE to see the coordinates of the points, and use the arrows on the keypad to go to other points. Your output will look like Figure 4B, but without the line.

Finding the Correlation and Regression Equation Coefficients
Before finding the correlation, you must turn the diagnostics on, as shown here.

1. Press 2ND, CATALOG, and scroll down to DiagnosticOn and press ENTER twice. The diagnostics will stay on unless you Reset your calculator or change the batteries.
2. Press STAT, choose CALC, and 8 (for LinReg (a + bx)).
3. Press 2ND L1 (or whichever list is X, the predictor), press (comma: the button above the 7), press 2ND L2 (or whichever list is Y, the response), and press ENTER.

Figure 4C shows the output.
TECH TIPS

MINITAB

Making a Scatterplot
1. **Graph > Scatterplot**
2. Leave the default **Simple** and click **OK**.
3. Double click the column containing the weights so that it goes under the **Y Variables**. Then double click the column containing the heights so that it goes under the **X Variables**.
4. Click **OK**. After the graph is made, you can edit the labels by clicking on them.

Finding the Correlation
1. **Stat > Basic Statistics > Correlation**
2. Double click both the predictor column and the response column (in either order).
3. Click **OK**. You will get 0.881.

Finding the Regression Equation Coefficients
1. **Stat > Regression > Regression**
2. Put in the **Response** (y) and **Predictor** (x) columns.
3. Click **OK**. You may need to scroll up to see the regression equation. It will be easier to understand if you have put in labels for the columns, such as “Height” and “Weight.” You will get: Weight = −443 + 9.03 Height.

To Display the Regression Line on a Scatterplot
1. **Stat > Regression > Fitted Line Plot**
2. Double click the **Response** (y) column and then double click the **Predictor** (x) column.
3. Click **OK**. Figure 4D shows the fitted line plot.

EXCEL

Making a Scatterplot
1. Select (highlight) the two columns containing the data, with the predictor column to the left of the response column. You may include the labels at the top or not include them.
2. Click **Insert**, in **Charts** click the picture of a scatterplot, and click the upper left option shown here:

Alternatively, you can add the regression line to an existing scatterplot by right-clicking on the scatterplot to open a graphing menu. Then select **Add > Regression Fit > Linear** (make sure **Fit intercept** is selected) and click **OK**. This instruction will add the regression line, but not the equation of the line.

3. When the chart is active (click on it), **Chart Tools** are shown at the top of screen, right of centre. Click **Layout** (not **Page Layout**), then **Axis Titles**, and **Chart Title** to add appropriate labels. After the labels are added, you can click on them to change the spelling or add words. Delete the labels on the right-hand side, such as **Series 1**, if you see any.

the minimum value for an axis, right-click on the axis numbers, click **Format axis**, in **Axis Options** change the **Minimum to Fixed**, and put in the desired value. You may want to do this twice: once for the x-axis and once for the y-axis. Then click **Close**.

5. If you want to zoom in or out on the data by changing (0, 0). If you want to zoom in or out on the data by changing
Finding the Correlation
1. Click on Data, click on Data Analysis, select Correlation, and click OK.
2. For the Input Range, select (highlight) both columns of data (if you have highlighted the labels as well as the numbers, you must also click on the Labels in first row).
3. Click OK. You will get 0.881638.
   (Alternatively, just click the fx button, for category choose statistical, select CORREL, click OK, and highlight the two columns containing the numbers, one at a time. The correlation will show up on the dialogue screen, and it will appear in the last active cell if you click OK.)

Finding the Coefficients of the Regression Equation
1. Click on Data, Data Analysis, Regression, and OK.
2. For the Input Y Range, select the column of numbers (not words) that represents the response or dependent variable. For the Input X Range, select the column of numbers that represents the predictor or independent variable.
3. Click OK.
   A large summary of the model will be displayed. Look under Coefficients at the bottom. For the Intercept and the slope (next to XVariable1), see Figure 4E, which means the regression line is
\[ y = -200.5 + 1.61x \]

<table>
<thead>
<tr>
<th>Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
</tr>
<tr>
<td>X Variable 1</td>
</tr>
</tbody>
</table>

To Display the Regression Line on a Scatterplot
4. After making the scatterplot, under Chart Tools click Design. In the Chart Layouts group, click the triangle to the right of Quick Layout. Choose Layout 9 (the option in the lower right portion, which shows a line in it and also fx).
   Refer to Figure 4F.
TECH TIPS

STATCRUNCH

Making a Scatterplot
1. Graph > Scatterplot
2. Select an X column and a Y column for the plot.
3. Click Compute! to construct the plot.
4. To copy the graph, click Options and Copy.

Finding the Correlation and Coefficients for the Equation
1. Stat > Regression > Simple Linear
2. Select the X variable and Y variable for the regression.
3. Click Compute! to view the equation and numbers, which are shown in Figure 4G.

<table>
<thead>
<tr>
<th>Options (1 of 2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple linear regression results:</td>
</tr>
<tr>
<td>Dependent Variable: Weight</td>
</tr>
<tr>
<td>Independent Variable: Height</td>
</tr>
<tr>
<td>Weight = 200.51533 + 1.6080778 Height</td>
</tr>
<tr>
<td>Sample size: 6</td>
</tr>
<tr>
<td>R (correlation coefficient) = 0.88163813</td>
</tr>
<tr>
<td>R-sq = 0.7772858</td>
</tr>
<tr>
<td>Estimate of error standard deviation: 6.4246685</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter estimates:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
</tr>
<tr>
<td>Intercept</td>
</tr>
<tr>
<td>Slope</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Analysis of variance table for regression model:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
</tr>
<tr>
<td>Model</td>
</tr>
<tr>
<td>Error</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>

FIGURE 4G StatCrunch regression output.

Plotting the Regression Line on a Scatterplot
1. The instructions for finding the coefficients for the regression equation produce the regression line as part of the output. Simply click on the right arrow at the bottom right-hand corner of the output to move to the next page. Refer to Figure 4H.

FIGURE 4H StatCrunch fitted line plot.
Making a Scatterplot
1. Graphs > Legacy Dialogs > Scatter/Dot
2. Select Simple Scatter and click Define.
3. Highlight the response or dependent variable and click on the right arrow to move it to the Y axis box.
4. Highlight the independent variable and click on the right arrow to move it to the X axis box.
5. Click OK.

Finding the Correlation
1. Analyze > Correlate > Bivariate
2. Highlight and move each variable to the Variable box by clicking on the right arrow.
3. Select Pearson and click OK.

Finding the Coefficients of the Regression Equation
1. Analyze > Regression > Linear
2. Highlight the response or dependent variable and click on the right arrow to move it to the Dependent box.
3. Highlight the independent variable and click on the right arrow to move it to the Independent(s) box.
4. Click OK.
Several tables appear as output. Look at the table of Coefficients at the bottom. The intercept is listed as the (Constant) coefficient, and the slope as the coefficient for the independent variable (in this case Height). See Figure 4I. Once again, the regression line is

\[ y = -201 + 1.61x \]

Plotting the Regression Line on a Scatterplot
1. Follow the steps for making a scatterplot.
2. Double click on the scatterplot to activate the chart editor.
3. Elements > Fit Line at Total > Close (the default is Linear). See Figure 4J.