Describing Relationships: Regression, Prediction, and Causation

15

Andy Lyons/Getty Images

CASE STUDY Predicting the future course of the stock market could make you rich. No wonder lots of people and lots of computers pore over market data looking for patterns.

There are some surprising methods. The "Super Bowl Indicator" says that the football Super Bowl, played in January or early February, predicts how stocks will behave each year. The current National Football League (NFL) was formed by merging the original NFL with the American Football League (AFL). The current NFL consists of two conferences, the National Football Conference (NFC) and the American Football Conference (AFC). The indicator claims that stocks go up in years when a team from the NFC (or from the old NFL) wins and down when an AFC team wins. The indicator was right in 39 of 49 years between the first Super Bowl in 1967 and 2015. (For purposes of the legend, we will regard the Baltimore Ravens as an old NFL team because they were the Cleveland Browns before the franchise moved to Baltimore in 1996. We will also regard the Tampa Bay Buccaneers as an NFC team, although they were neither a pre-merger team nor an old NFL team and started out as an AFC team.) The indicator is right more than 75% of the time, which seems impressive.

In 2016 (the year in which this was written), the Broncos, an AFC team, won the Super Bowl. According to the Super Bowl Indicator, stocks will fall in 2016. Should I have avoided investing in 2016?

In this chapter, we will study statistical methods to predict one variable from others that go well beyond just counting ups and downs. We will also distinguish between the ability to predict one variable from others and the issue of whether changes in one variable are caused by changes in others. By the end of this chapter, you will be able to critically evaluate the Super Bowl Indicator.

Regression lines

If a scatterplot shows a straight-line relationship between two quantitative variables, we would like to summarize this overall pattern by drawing a line on the graph. A *regression line* summarizes the relationship between two variables, but only in a specific setting: one of the variables helps explain or predict the other. That is, regression describes a relationship between an explanatory variable and a response variable.

Regression line

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

EXAMPLE 1 Fossil bones

In Examples 3 and 4 in Chapter 14, we saw that the lengths of two bones in fossils of the extinct beast archaeopteryx closely follow a straight-line pattern. Figure 15.1 plots the lengths for the five available fossils. The regression line on the plot gives a quick summary of the overall pattern.

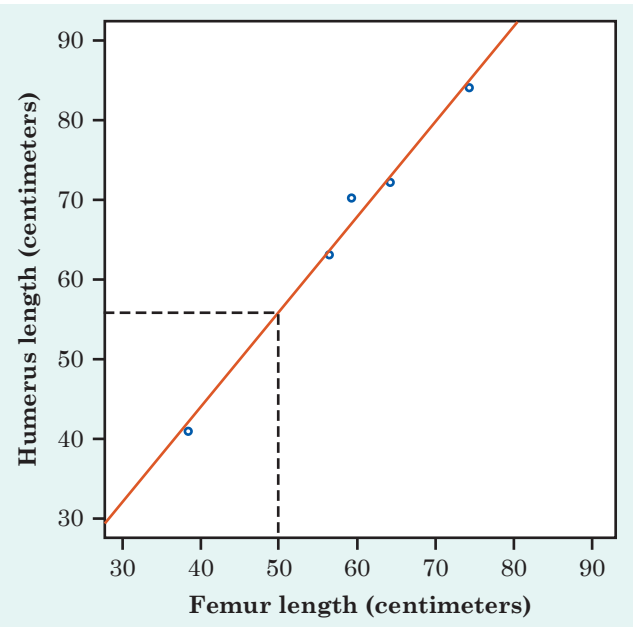


Figure 15.1 Using a straight-line pattern for prediction, Example 1. The data are the lengths of two bones in five fossils of the extinct beast archaeopteryx.

Another archaeopteryx fossil is incomplete. Its femur is 50 centimeters long, but the humerus is missing. Can we predict how long the humerus is? The straight-line pattern connecting humerus length to femur length is so strong that we feel quite safe in using femur length to predict humerus length. Figure 15.1 shows how: starting at the femur length (50 cm), go up to the line, then over to the humerus length axis. We predict a length of about 56 cm. This is the length the humerus would have if this fossil's point lay exactly on the line. All the other points are close to the line, so we think the missing point would also be close to the line. That is, we think this prediction will be quite accurate.

EXAMPLE 2 Presidential elections, the Reagan years

Republican Ronald Reagan was elected president twice, in 1980 and in 1984. His economic policy of tax cuts to stimulate the economy, eventually leading to increases in tax revenue, was still advocated by some Republican presidential candidates in 2015. Figure 15.2 plots the percentage of voters in each state who voted for Reagan's Democratic opponents: Jimmy Carter in 1980 and Walter Mondale in 1984. The plot shows a positive straight-line relationship. We expect this because some states tend to vote Democratic and others tend to vote Republican. There

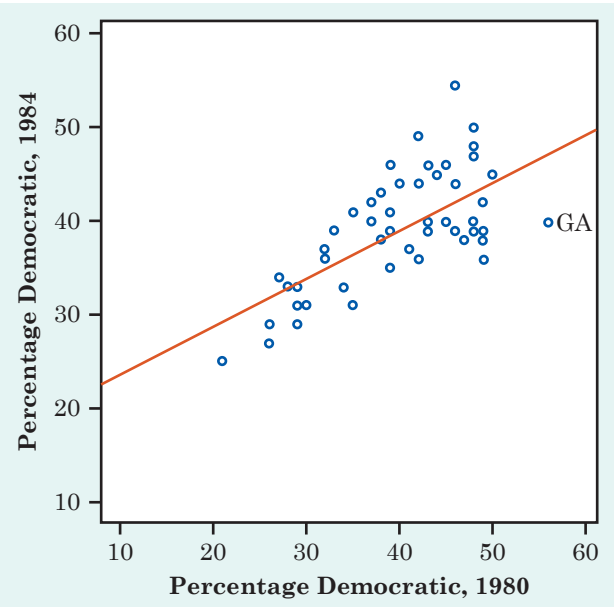


Figure 15.2 A weaker straight-line pattern, Example 2. The data are the percentage in each state who voted Democratic in the two Reagan presidential elections.

is one outlier: Georgia, President Carter's home state, voted 56% for the Democrat Carter in 1980 but only 40% Democratic in 1984.

We could use the regression line drawn in Figure 15.2 to predict a state's 1984 vote from its 1980 vote. The points in this figure are more widely scattered about the line than are the points in the fossil bone plot in Figure 15.1. The correlations, which measure the strength of the straight-line relationships, are $r = 0.994$ for Figure 15.1 and $r = 0.704$ for Figure 15.2. The scatter of the points makes it clear that predictions of voting will be generally less accurate than predictions of bone length.

Regression equations

When a plot shows a straight-line relationship as strong as that in Figure 15.1, it is easy to draw a line close to the points by eye. In Figure 15.2, however, different people might draw quite different lines by eye. Because we want to predict y from x, we want a line that is close to the points in the *vertical* (y) direction. It is hard to concentrate on just the vertical distances when drawing a line by eye. What is more, drawing by eye gives us a line on the graph but not an equation for the line. We need a way to find from the data the equation of the line that comes closest to the points in the vertical direction. There are many ways to make the collection of vertical distances "as small as possible." The most common is the *least-squares* method.

Least-squares regression line

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

Figure 15.3 illustrates the least-squares idea. This figure magnifies the center part of Figure 15.1 to focus on three of the points. We see the vertical distances of these three points from the regression line. To find the least-squares line, look at these vertical distances (all five for the fossil data), square them, and move the line until the sum of the squares is the smallest it can be for any line. The lines drawn on the scatterplots in Figures 15.1 and 15.2 are the least-squares regression lines. We won't give the formula for finding the least-squares line from data—that's a job for a calculator or computer. You should, however, be able to use the equation that the machine produces.

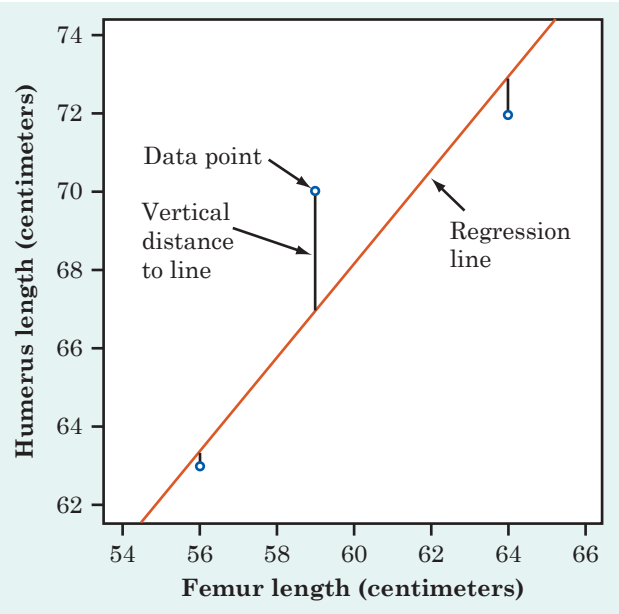


Figure 15.3 A regression line aims to predict *y* from *x*. So, a good regression line makes the vertical distances from the data points to the line small.

In writing the equation of a line, x stands as usual for the explanatory variable and y for the response variable. The equation of a line has the form

$$y = a + bx$$

The number b is the **slope** of the line, the amount by which y changes when x increases by one unit. The number a is the **intercept**, the value of y when $x = 0$. To use the equation for prediction, just substitute your x-value into the equation and calculate the resulting y-value.

EXAMPLE 3 Using a regression equation

In Example 1, we used the "up-and-over" method in Figure 15.1 to predict the humerus length for a fossil whose femur length is 50 cm. The equation of the least-squares line is

$$\text{humerus length} = -3.66 + (1.197 \times \text{femur length})$$

The *slope* of this line is $b = 1.197$. This means that for these fossils, humerus length goes up by 1.197 cm when femur length goes up 1 cm. The slope of a regression line is usually important for understanding the data. The slope is the rate of change, the amount of change in the predicted y when x increases by 1.

Statistics in Your World

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

The *intercept* of the least-squares line is $a = -3.66$. This is the value of the predicted y when $x = 0$. Although we need the intercept to draw the line, it is statistically meaningful only when x can actually take values close to zero. Here, femur length 0 is impossible (recall that the femur is a bone in the leg), so the intercept has no statistical meaning.

To use the equation for *prediction*, substitute the value of x and calculate y. The predicted humerus length for a fossil with a femur 50 cm long is

$$\begin{aligned}\text{humerus length} &= -3.66 + (1.197)(50)\\ &= 56.2 \text{ cm}\end{aligned}$$

To *draw the line* on the scatterplot, predict y for two different values of x. This gives two points. Plot them and draw the line through them.

NOW IT'S YOUR TURN

15.1 Fossil bones. Use the equation of the least-squares line

$$\text{humerus length} = -3.66 + (1.197 \times \text{femur length})$$

to predict the humerus length for a fossil with a femur 70 cm long.

Understanding prediction

Computers make prediction easy and automatic, even from very large sets of data. Anything that can be done automatically is often done thoughtlessly. Regression software will happily fit a straight line to a curved relationship, for example. Also, the computer cannot decide which is the explanatory variable and which is the response variable. This is important because the same data give two different lines depending on which is the explanatory variable.

In practice, we often use several explanatory variables to predict a response. As part of its admissions process, a college might use SAT Math and Verbal scores and high school grades in English, math, and science (five explanatory variables) to predict first-year college grades. Although the details are messy, all statistical methods of predicting a response share some basic properties of least-squares regression lines.

- **Prediction is based on fitting some "model" to a set of data.** In Figures 15.1 and 15.2, our model is a straight line that we draw through the points in a scatterplot. Other prediction methods use more elaborate models.

- **Prediction works best when the model fits the data closely.** Compare again Figure 15.1, where the data closely follow a line, with Figure 15.2, where they do not. Prediction is more trustworthy in Figure 15.1. Also, it is not so easy to see patterns when there are many variables, but if the data do not have strong patterns, prediction may be very inaccurate.
- **Prediction outside the range of the available data is risky.** Suppose that you have data on a child's growth between three and eight years of age. You find a strong straight-line relationship between age x and height y. If you fit a regression line to these data and use it to predict height at age 25 years, you will predict that the child will be 8 feet tall. Growth slows down and stops at maturity, so extending the straight line to adult ages is foolish. No one would make this mistake in predicting height. But almost all economic predictions try to tell us what will happen next quarter or next year. No wonder economic predictions are often wrong. Prediction outside the range of available data is referred to as extrapolation. Beware of **extrapolation**!

EXAMPLE 4 Predicting the national deficit

The Congressional Budget Office is required to submit annual reports that predict the federal budget and its deficit or surplus for the next five years. These forecasts depend on future economic trends (unknown) and on what Congress will decide about taxes and spending (also unknown). Even the prediction of the state of the budget if current policies are not changed has been wildly inaccurate. The forecast made in January 2008 for 2012, for example, underestimated the deficit by nearly $1000 billion! The January 2009 forecast for 2013 underestimated the deficit by $423 billion, but the January 2010 forecast for 2014 underestimated the deficit by only $8 billion. As Senator Everett Dirksen once said, "A billion here and a billion there and pretty soon you are talking real money." In 1999, the Budget Office was predicting a surplus (ignoring Social Security) of $996 billion over the following 10 years. Politicians debated what to do with the money, but no one else believed the prediction (correctly, as it turned out). In 2012, there was a $1087 billion deficit; in 2013, a $680 billion deficit; and in 2014, a $483 billion deficit. The forecast in January 2015 is for a $652 billion deficit in 2019. Time will tell how accurate this forecast is.

Correlation and regression

Correlation measures the direction and strength of a straight-line relationship. Regression draws a line to describe the relationship. Correlation and regression are closely connected, even though regression requires choosing an explanatory variable and correlation does not.

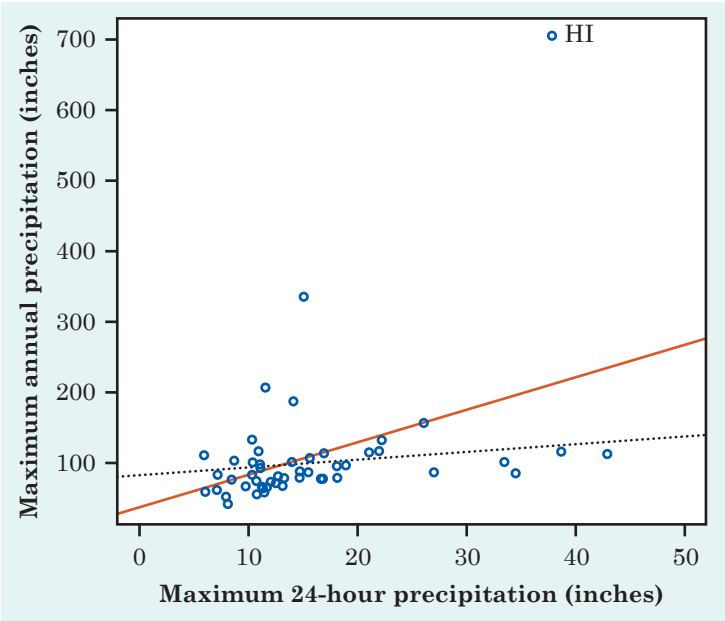


Figure 15.4 Least-squares regression lines are strongly influenced by outliers. The solid line is based on all 50 data points. The dotted line leaves out Hawaii.

Both correlation and regression are strongly affected by outliers. Be wary if your scatterplot shows strong outliers. Figure 15.4 plots the record-high yearly precipitation in each state against that state's record-high 24-hour precipitation. Hawaii is a high outlier, with a yearly record of 704.83 inches of rain recorded at Kukui in 1982. The correlation for all 50 states in Figure 15.4 is 0.510. If we leave out Hawaii, the correlation drops to $r = 0.248$. The solid line in the figure is the least-squares line for predicting the annual record from the 24-hour record. If we leave out Hawaii, the least-squares line drops down to the dotted line. This line is nearly flat—there is little relation between yearly and 24-hour record precipitation once we decide to ignore Hawaii.

The usefulness of the regression line for prediction depends on the strength of the association. That is, the usefulness of a regression line depends on the correlation between the variables. It turns out that the *square* of the correlation is the right measure.

r^2 in regression

The **square of the correlation, r^2**, is the proportion of the variation in the values of y that is explained by the least-squares regression of y on x.

The idea is that when there is a straight-line relationship, some of the variation in y is accounted for by the fact that as x changes, it pulls y along with it.

EXAMPLE 5 Using r^2

Look again at Figure 15.1. There is a lot of variation in the humerus lengths of these five fossils, from a low of 41 cm to a high of 84 cm. The scatterplot shows that we can explain almost all of this variation by looking at femur length and at the regression line. As femur length increases, it pulls humerus length up with it along the line. There is very little leftover variation in humerus length, which appears in the scatter of points about the line. Because $r = 0.994$ for these data, $r^2 = (0.994)^2 = 0.988$. So, the variation "along the line" as femur length pulls humerus length with it accounts for 98.8% of all the variation in humerus length. The scatter of the points about the line accounts for only the remaining 1.2%. Little leftover scatter says that prediction will be accurate.

Contrast the voting data in Figure 15.2. There is still a straight-line relationship between the 1980 and 1984 Democratic votes, but there is also much more scatter of points about the regression line. Here, $r = 0.704$ and so $r^2 = 0.496$. Only about half the observed variation in the 1984 Democratic vote is explained by the straight-line pattern. You would still guess a higher 1984 Democratic vote for a state that was 45% Democratic in 1980 than for a state that was only 30% Democratic in 1980. But lots of variation remains in the 1984 votes of states with the same 1980 vote. That is the other half of the total variation among the states in 1984. It is due to other factors, such as differences in the main issues in the two elections and the fact that President Reagan's two Democratic opponents came from different parts of the country.

In reporting a regression, it is usual to give r^2 as a measure of how successful the regression was in explaining the response. When you see a correlation, square it to get a better feel for the strength of the association. Perfect correlation ($r = -1$ or $r = 1$) means the points lie exactly on a line. Then, $r^2 = 1$ and all of the variation in one variable is accounted for by the straight-line relationship with the other variable. If $r = -0.7$ or $r = 0.7$, $r^2 = 0.49$ and about

Did the vote counters cheat? Republican Bruce Marks was ahead of Democrat William Stinson when the voting-machines results were tallied in their 1993 Pennsylvania election. But Stinson was ahead after absentee ballots were counted by the Democrats, who controlled the election board. A court fight followed. The court called in a statistician, who used regression with data from past elections to predict the counts of absentee ballots from the voting-machines results. Marks's lead of 564 votes from the machines predicted that he would get 133 more absentee votes than Stinson. In fact, Stinson got 1025 more absentee votes than Marks. Did the vote counters cheat?

half the variation is accounted for by the straight-line relationship. In the r^2 scale, correlation ± 0.7 is about halfway between 0 and ± 1.

NOW IT'S YOUR TURN

15.2 At the ballpark. Table 14.2 (page 334) gives data on the prices charged for for beer (per ounce) and for a hot dog at Major League Baseball stadiums. The correlation between the prices is $r = 0.36$. What proportion of the variation in hot dog prices is explained by the least-squares regression of hot dog prices on beer prices (per ounce)?

Norm Betts/Bloomberg via Getty Images

The question of causation

There is a strong relationship between cigarette smoking and death rate from lung cancer. Does smoking cigarettes *cause* lung cancer? There is a strong association between the availability of handguns in a nation and that nation's homicide rate from guns. Does easy access to handguns *cause* more murders? It says right on the pack that cigarettes cause cancer. Whether more guns cause more murders is hotly debated. Why is the evidence for cigarettes and cancer better than the evidence for guns and homicide?

We already know three big facts about statistical evidence for cause and effect.

Statistics and causation

1. A strong relationship between two variables does not always mean that changes in one variable cause changes in the other.

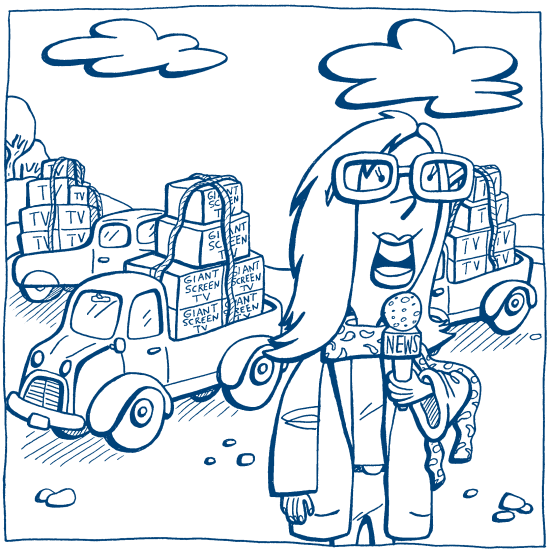
2. The relationship between two variables is often influenced by other variables lurking in the background.

3. The best evidence for causation comes from randomized comparative experiments.

EXAMPLE 6 Does television extend life?

Measure the number of television sets per person x and the life expectancy y for the world's nations. There is a high positive correlation: nations with many TV sets have higher life expectancies.

The basic meaning of causation is that by changing x we can bring about a change in y. Could we lengthen the lives of people in Botswana by shipping them TV sets? No. Rich nations have more TV sets than poor nations. Rich nations also have longer life expectancies because they offer better nutrition, clean water, and better health care. There is no cause-and-effect tie between TV sets and length of life.



"In a new attack on third-world poverty, aid organizations today began delivery of 100,000 television sets."

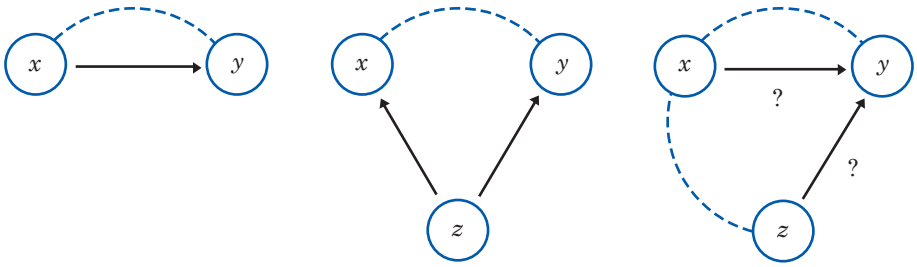
Example 6 illustrates our first two big facts. Correlations such as this are sometimes called "nonsense correlations." The correlation is real. What is nonsense is the conclusion that changing one of the variables causes changes in the other. A lurking variable—such as national wealth in Example 6—that influences both x and y can create a high correlation even though there is no direct connection between x and y. We might call this *common response*: both the explanatory and the response variable are responding to some lurking variable.

EXAMPLE 7 Obesity in mothers and daughters

What causes obesity in children? Inheritance from parents, overeating, lack of physical activity, and too much television have all been named as explanatory variables.

The results of a study of Mexican American girls aged 9 to 12 years are typical. Researchers measured body mass index (BMI), a measure of weight relative to height, for both the girls and their mothers. People with high BMI are overweight or obese. They also measured hours of television watched, minutes of physical activity, and intake of several kinds of food. The result: the girls' BMIs were weakly correlated with physical activity ($r = -0.18$), diet, and television. The strongest correlation ($r = 0.506$) was between the BMI of daughters and the BMI of their mothers.

Body type is in part determined by heredity. Daughters inherit half their genes from their mothers. There is, therefore, a direct causal link between the BMI of mothers and daughters. Of course, the causal link is far from perfect. The mothers' BMIs explain only 25.6% (that's r^2 again) of the variation among the daughters' BMIs. Other factors, some measured



(a) Causation (b) Common response (c) Confounding

Figure 15.5 Some explanations for an observed association. A dashed line shows an association. An arrow shows a cause-and-effect link. Variable *x* is explanatory, *y* is a response variable, and *z* is a lurking variable.

in the study and others not measured, also influence BMI. *Even when direct causation is present, it is rarely a complete explanation of an association between two variables.*

Can we use r or r^2 from Example 7 to say how much inheritance contributes to the daughters' BMIs? No. Remember *confounding*. It may well be that mothers who are overweight also set an example of little exercise, poor eating habits, and lots of television. Their daughters pick up these habits to some extent, so the influence of heredity is mixed up with influences from the girls' environment. We can't say how much of the correlation between mother and daughter BMIs is due to inheritance.

Figure 15.5 shows in outline form how a variety of underlying links between variables can explain association. The dashed line represents an observed association between the variables x and y. Some associations are explained by a *direct cause-and-effect* link between the variables. The first diagram in Figure 15.5 shows "x causes y" by an arrow running from x to y. The second diagram illustrates *common response*. The observed association between the variables x and y is explained by a lurking variable z. Both x and y change in response to changes in z. This common response creates an association even though there may be no direct causal link between x and y. The third diagram in Figure 15.5 illustrates *confounding*. Both the explanatory variable x and the lurking variable z may influence the response variable y. Variables x and z are themselves associated, so we cannot distinguish the influence of x from the influence of z. We cannot say how strong the direct effect of x on y is. In fact, it can be hard to say if x influences y at all.

In Example 7, there is a causal link between the BMI of mothers and daughters. However, other factors, some measured in the study and some not measured, also influence the BMI of daughters. This is an example of confounding, illustrated in Figure 15.5(c). The x in the figure corresponds

to the BMI of the mother, the z to one of the other factors, and the y to the BMI of the daughter.

Both common response and confounding involve the influence of a lurking variable or variables z on the response variable y. We won't belabor the distinction between the two kinds of relationships. Just remember that "beware the lurking variable" is good advice in thinking about relationships between variables. Here is another example of common response, in a setting where we want to do prediction.

EXAMPLE 8 SAT scores and college grades

High scores on the SAT examinations in high school certainly do not *cause* high grades in college. The moderate association (r^2 is about 27%) is no doubt explained by common response variables such as academic ability, study habits, and staying sober. Figure 15.5(b) illustrates this. In the figure, z might correspond to academic ability, x to SAT scores, and y to grades in college.

The ability of SAT scores to partly predict college performance doesn't depend on causation. We need only believe that the relationship between SAT scores and college grades that we see in past years will continue to hold for this year's high school graduates. Think once more of our fossils, where femur length predicts humerus length very well. The strong relationship is explained by common response to the overall age and size of the beasts whose fossils we now examine. *Prediction doesn't require causation.*

Discussion of these examples has brought to light two more big facts about causation:

STATISTICAL CONTROVERSIES

Gun Control and Crime

Do strict controls on guns, especially handguns, reduce crime? To many people, the answer must be "Yes." More than half of all murders in the United States are committed with handguns. The U.S. murder rate (per 100,000 population) is 1.7 times that of Canada, and the rate of murders with handguns is 15 times higher. Surely guns help bad things happen. Then John Lott, a University of Chicago economist, did an elaborate statistical study using data from all 3054 counties in the United States over the 18-year period from 1977 to 1994. Lott found that as states relaxed gun laws to allow adults to carry guns, the crime rate dropped. He argued that guns reduce crime by allowing citizens to defend themselves and by making criminals hesitate.

Lott used regression methods to determine the relationship between crime and many explanatory variables and to isolate the effect of permits to carry concealed guns after adjusting for other explanatory variables. You can find a link to a copy of Lott's study at www2.lib.uchicago.edu/~llou/guns.html.

The resulting debate, still going on, has been loud. People feel strongly about gun control. Most reacted to Lott's work based on whether or not they liked his conclusion. Gun supporters painted Lott as Moses revealing truth at last; opponents knew he must be both wrong and evil.

Is Lott right? What do you see as the weaknesses of his study based on what you have learned about statistics?

More about statistics and causation

4. The observed relationship between two variables may be due to **direct causation, common response,** or **confounding**. Two or more of these factors may be present together.

5. An observed relationship can, however, be used for prediction without worrying about causation as long as the patterns found in past data continue to hold true.

NOW IT'S YOUR TURN

15.3 At the ballpark. Table 14.2 (page 334) gives data on the prices charged for beer (per ounce) and for a hot dog at Major League Baseball stadiums. The correlation between the prices is $r = 0.36$. Do you think the observed relationship is due to direct causation, common response, confounding, or some combination of these? Explain your answer.

Evidence for causation

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

Doctors had long observed that most lung cancer patients were smokers. Observational studies comparing smokers and "similar" (in the sense of characteristics such as age, gender, and overall health) nonsmokers showed a strong association between smoking and death from lung cancer. Could the association be explained by lurking variables that the studies could not measure? Might there be, for example, a genetic factor that

predisposes people both to nicotine addiction and to lung cancer? Smoking and lung cancer would then be positively associated even if smoking had no direct effect on the lungs. How were these objections overcome?

Let's answer this question in general terms. What are the criteria for establishing causation when we cannot do an experiment?

- **The association is strong.** The association between smoking and lung cancer is very strong.
- **The association is consistent.** Many studies of different kinds of people in many countries link smoking to lung cancer. That reduces the chance that a lurking variable specific to one group or one study explains the association.
- **Higher doses are associated with stronger responses.** People who smoke more cigarettes per day or who smoke over a longer period get lung cancer more often. People who stop smoking reduce their risk.
- **The alleged cause precedes the effect in time.** Lung cancer develops after years of smoking. The number of men dying of lung cancer rose as smoking became more common, with a lag of about 30 years. Lung cancer kills more men than any other form of cancer. Lung cancer was rare among women until women began to smoke. Lung cancer in women rose along with smoking, again with a lag of about 30 years, and has now passed breast cancer as the leading cause of cancer death among women.
- **The alleged cause is plausible.** Experiments with animals show that tars from cigarette smoke do cause cancer.

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

Correlation, prediction, and big data

In 2008, researchers at Google were able to track the spread of influenza across the United States much faster than the Centers for Disease Control and Prevention (CDC). By using computer algorithms to explore millions of online Internet searches, the researchers discovered a correlation between what people searched for online and whether they had flu symptoms. The researchers used this correlation to make their surprisingly accurate predictions.

Massive databases, or "big data," that are collected by Google, Facebook, credit card companies, and others contain petabytes—or 10^{15} bytes—

of data and continue to grow in size. Big data allow researchers, businesses, and industry to search for correlations and patterns in data that will enable them to make accurate predictions about public health, economic trends, or consumer behavior. Using big data to make predictions is increasingly common. Big data explored with clever algorithms open exciting possibilities. Will the experience of Google become the norm?

Proponents for big data often make the following claims for its value. First, big data include all members of a population, eliminating the need for statistical sampling. Second, there is no need to worry about causation because correlations are all we need to know for making accurate predictions. Third, scientific and statistical theory is unnecessary because, with enough data, the numbers speak for themselves.

Are these claims correct? First, as we saw in Chapter 3, it is true that sampling variability is reduced by increasing the sample size and will become negligible with a sufficiently large sample. It is also true that there is no sampling variability when one has information on the entire population of interest. However, sampling variability is not the only source of error in statistics computed from data. Bias is another source of error and is not eliminated because the sample size is extremely large. Big data are often enormous convenience samples, the result of recording huge numbers of web searches, credit card purchases, or mobile phones pinging the nearest phone tower. This is not equivalent to having information about the entire population of interest. For example, in principle, it is possible to record every message on Twitter and use these data to draw conclusions about public opinion. However, Twitter users are not representative of the population as a whole. According to the Pew Research Internet Project, in 2013, U.S.-based users were disproportionally young, urban or suburban, and black. In other words, the large amount of data generated by Twitter users is biased when the goal is to draw conclusions about public opinion of all adults in the United States.

Second, it is true that correlation can be exploited for purposes of prediction even if there is no causal relation between explanatory and response variables. However, if you have no idea what is behind a correlation, you have no idea what might cause prediction to fail, especially when one exploits the correlation to extrapolate to new situations. For a few winters after their success in 2008, Google Flu Trends continued to accurately track the spread of influenza using the correlations they discovered. But during the 2012–2013 flu season, data from the CDC showed that Google's estimate of the spread of flu-like illnesses was overstated by almost a factor of two. A possible explanation was that the news was full of stories about the flu, and this provoked Internet searches by people who were otherwise healthy. The failure to understand why search terms were correlated with the spread of flu resulted in incorrectly assuming previous correlations extrapolated into the future.

Adding to the perception of the infallibility of big data are news reports touting successes, with few reports of the failures. The claim that theory is unnecessary because the numbers speak for themselves is misleading when all the numbers concerning successes and failures of big data are not reported. Statistical theory has much to say that can prevent data analysts from making serious errors. Providing examples of where mistakes have been made and explaining how, with proper statistical understanding and tools, those mistakes could have been avoided is an important contribution.

The era of big data is exciting and challenging and has opened incredible opportunities for researchers, businesses, and industry. But simply being big does not exempt big data from statistical pitfalls such as bias and extrapolation.

STATISTICS IN SUMMARY

Chapter Specifics

- **Regression** is the name for statistical methods that fit some model to data in order to predict a response variable from one or more explanatory variables.
- The simplest kind of regression fits a straight line on a scatterplot for use in predicting y from x. The most common way to fit a line is the **least-squares** method, which finds the line that makes the sum of the squared vertical distances of the data points from the line as small as possible.
- The **squared correlation r^2** tells us what fraction of the variation in the responses is explained by the straight-line tie between y and x.
- **Extrapolation**, or prediction outside the range of the data, is risky because the pattern may be different there. Beware of extrapolation!
- A strong relationship between two variables is not always evidence that changes in one variable **cause** changes in the other. Lurking variables can create relationships through **common response** or **confounding**.
- If we cannot do experiments, it is often difficult to get convincing evidence for causation.

In Chapter 14, we used scatterplots and the correlation to explore and describe the relationship between two quantitative variables. In this chapter, we looked carefully at fitting a straight line to data in a scatterplot when there appears to be a straight-line trend, and then

we used this line to predict the response from the explanatory variable. In doing this, we have used data to draw conclusions. We assume that the straight line that we fit to our data describes the actual relationship between the response and the explanatory variable and, thus, that conclusions (predictions) about additional values of the response based on other values of the explanatory variable are valid.

Are these conclusions (predictions) justified? The squared correlation provides information about the likelihood of a successful prediction. Small values of the squared correlation suggest that our predictions are not likely to be accurate. Extrapolation is another setting in which our predictions are not likely to be accurate.

Finally, when there is a strong relationship between two variables, it is tempting to draw an additional conclusion: namely, that changes in one variable cause changes in another. However, the case for causation requires more than a strong relationship. Unless our data are produced by a proper experiment, the case for causation is difficult to prove.

CASE STUDY EVALUATED What should we conclude about the Super Bowl Indicator described in the Case Study at the beginning of this chapter? To evaluate the Super Bowl Indicator, answer the following questions.

1. We wrote this Case Study on March 4, 2016, the year in which the Broncos won the Super Bowl. The Super Bowl Indicator predicts stocks should go down in 2016. Did they go down?
2. Stocks went down only 12 times in the 49 years between 1967 and 2015. If you simply predicted "up" every year, how would you have performed?
3. There are 19 original NFL and NFC teams and only 13 AFC teams. How often would you expect "NFL wins" to occur if one assumes that the chance of winning is proportional to the number of teams? How does this compare with simply predicting "up" every year?
4. Write a paragraph, in language that someone who knows no statistics would understand, explaining why the association between the Super Bowl Indicator and stock prices is not surprising and why it would be incorrect to conclude that the Super Bowl outcome causes changes in stock prices.

LaunchPad macmillan learning Online Resources

video

- The StatClips ~~Video~~ *Regression—Introduction and Motivation* describes many of the topics in this chapter in the context of an example about hair growth.

StatBoards video

- The ~~StatBoard Video~~ *Beware Extrapolation!* discusses the dangers of extrapolation in the context of several examples.

CHECK THE BASICS

For Exercise 15.1, see page 344; for Exercise 15.2, see page 348; for Exercise 15.3, see page 352.

15.4 Least-squares. The least-squares regression line

(a) is the line that makes the sum of the vertical distances of the data points from the line as small as possible.

(b) is the line that makes the sum of the vertical distances of the data points from the line as large as possible.

(c) is the line that makes the sum of the squared vertical distances of the data points from the line as small as possible.

(d) is the line that makes the sum of the squared vertical distances of the data points from the line as large as possible.

15.5 Correlation. The quantity that tells us what fraction of the variation in the responses is explained by the straight-line tie between the response and explanatory variables is

(a) the correlation.

(b) the absolute value of the correlation.

(c) the square root of the correlation.

(d) the square of the correlation.

15.6 Extrapolation. Extrapolation, or prediction outside the range of the data, is risky

(a) because the pattern observed in the data may be different outside the range of the data.

(b) because correlation does not necessarily imply causation.

(c) unless the correlation is very close to 1.

(d) unless the square of the correlation is very close to 1.

15.7 Prediction. An observed relationship between two variables can be used for prediction

(a) as long as we know the relationship is due to direct causation.

(b) as long as the relationship is a straight-line relationship.

(c) as long as the patterns found in past data continue to hold true.

(d) in all of the above instances.

15.8 Causation. The best evidence that changes in one variable cause changes in another comes from

(a) randomized comparative experiments.

(b) data for which the square of the correlation is near 1.

(c) higher values of the explanatory variable are associated with stronger responses.

(d) a plausible theory for theory for causation.

CHAPTER 15 EXERCISES

15.9 Obesity in mothers and daughters. The study in Example 7 found that the correlation between the body mass index of young girls and their hours of physical activity in a day was $r = -0.18$. Why might we expect this correlation to be negative? What percentage of the variation in BMI among the girls in the study can be explained by the

straight-line relationship with hours of activity?

15.10 State SAT scores. Figure 14.9 (page 329) plots the average SAT Mathematics score of each state's high school seniors against the percentage of each state's seniors who took the exam. In addition to two clusters, the plot shows an overall roughly straight-line pattern. The least-squares regression line for predicting average SAT Math score from percentage taking is

$$\text{average Math SAT score} = 588.4 - (1.228 \times \text{percentage taking})$$

(a) What does the slope $b = -1.228$ tell us about the relationship between these variables?

(b) In New York State, the percentage of high school seniors who took the SAT was 76%. Predict their average score. (The actual average score in New York was 502.)

(c) On page 345, we mention that using least-squares regression to do prediction outside the range of available data is risky. For what range of data is it reasonable to use the least-squares regression line for predicting average SAT Math score from percentage taking?

15.11 The endangered manatee. Figure 14.10 (page 331) plots the number of manatee deaths by boats versus the number of boats registered in Florida (in thousands). There is a clear straight-line pattern with a modest amount of scatter. The correlation between these variables is $r = 0.953$. What percentage of the observed variation among the manatees deaths by boats is explained by the straight-line relationship between manatee deaths and number of boats registered?

15.12 State SAT scores. The correlation between the average SAT Mathematics score in the states and the percent of high school seniors who take the SAT is $r = -0.89$.

(a) The correlation is negative. What does that tell us?

(b) How well does proportion taking predict average score? (Use r^2 in your answer.)

15.13 The endangered manatee. The least-squares line for predicting manatee deaths by boats from number of boats registered in Florida, based on the data plotted in Figure 14.10 (page 331), is

$$\text{number of manatee deaths by boats} = -44.59 + (0.132 \times \text{number of boats registered in Florida})$$

Explain in words the meaning of the slope $b = 0.132$. Then predict the number of manatee deaths by boats when the number of boats registered in Florida is 1000.

15.14 Global warming. Here are annual average global temperatures for the last 21 years in degrees Celsius:

Year	1994	1995	1996
Temperature	14.23	14.35	14.22
Year	1997	1998	1999
Temperature	14.42	14.54	14.36
Year	2000	2001	2002
Temperature	14.33	14.45	14.51
Year	2003	2004	2005
Temperature	14.52	14.48	14.55

Year	2006	2007	2008
Temperature	14.50	14.49	14.41
Year	2009	2010	2011
Temperature	14.50	14.56	14.43
Year	2012	2013	2014
Temperature	14.48	14.52	14.59

You made a scatterplot of these data in Exercise 14.18 (page 332). The least-squares regression line is

$$\text{temperature} = -7.8 + (0.0111 \times \text{year})$$

What would you predict for the annual average temperature for 2014 based on this line? How accurate is your prediction?

15.15 Wine and heart disease. Drinking moderate amounts of wine may help prevent heart attacks. Let's look at data for entire nations. Table 15.1 gives data on yearly wine consumption (liters of alcohol from drinking wine, per person) and yearly deaths from heart disease (deaths per 100,000 people) in 19 developed countries in 2001.

(a) Make a scatterplot that shows how national wine consumption helps explain heart disease death rates.

(b) Describe in words the direction, form, and strength of the relationship.

(c) The correlation for these variables is $r = -0.645$. Why does this value agree with your description in part (b)?

15.16 The 2008 and 2012 presidential elections. Democrat Barack Obama was elected president in 2008 and 2012. Figure 15.6 plots the percentage who voted for Obama in 2008 and 2012 for each of the 50 states and the District of Columbia.

(a) Describe in words the direction, form, and strength of the relationship between the percentage of votes for Obama in 2008 and the percentage

TABLE 15.1 Wine consumption and heart disease

Country	Alcohol from wine[a]	Heart disease death rate[b]	Country	Alcohol from wine[a]	Heart disease death rate[b]
Australia	3.25	80	Italy	7.50	60
Austria	4.75	100	Netherlands	2.75	70
Belgium	2.75	60	New Zealand	2.50	100
Canada	1.50	80	Norway	1.75	80
Denmark	4.50	90	Spain	5.00	50
Finland	3.00	120	Sweden	2.50	90
France	8.50	40	Switzerland	6.00	70
Germany	3.75	90	United Kingdom	2.75	120
Iceland	1.25	110	United States	1.25	120
Ireland	2.00	130			

[a]Liters of alcohol from drinking wine, per person.

[b]Deaths per 100,000 people, ischemic heart disease.

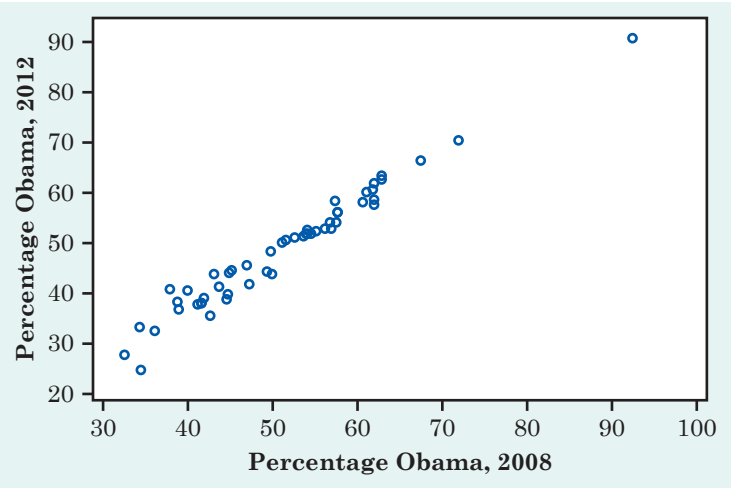


Figure 15.6 Scatterplot of the percentage who voted for Obama in 2008 and 2012 for each of the 50 states and the District of Columbia, Exercise 15.16.

in 2012. Are there any unusual features in the plot?

(b) The least-squares regression line is

$$\text{percentage in 2012} = -4.75 + (1.05 \times \text{percentage in 2008})$$

Draw this line on a separate sheet of paper. (To draw the line, use the equation to predict y for $x = 90$ and for $x = 50$. Plot the two (x, y) points and draw the line through them.)

(c) The correlation between these variables is $r = 0.983$. What percentage of the observed variation in 2012 percentages can be explained by straight-line dependence on 2008 percentages?

15.17 Beavers and beetles. Ecologists sometimes find rather strange relationships in our environment. One study seems to show that beavers benefit beetles. The researchers laid out 23 circular plots, each 4 meters in diameter, in an area where beavers were cutting down cottonwood trees. In each plot, they counted the number of stumps from trees cut by beavers and the number of clusters of beetle larvae. Here are the data:

Stumps:	2	2	1	3	3
Larvae clusters:	10	30	12	24	36
Stumps:	4	3	1	2	5
Larvae clusters:	40	43	11	27	56
Stumps:	1	3	2	1	2
Larvae clusters:	18	40	25	8	21
Stumps:	2	1	1	4	1
Larvae clusters:	14	16	6	54	9
Stumps:	2	1	4		
Larvae clusters:	13	14	50		

(a) Make a scatterplot that shows how the number of beaver-caused stumps influences the number of beetle larvae clusters. What does your plot show? (Ecologists think that the new sprouts from stumps are more

tender than other cottonwood growth so that beetles prefer them.)

(b) The least-squares regression line is

$$\text{larvae clusters} = -1.286 + (11.894 \times \text{stumps})$$

Draw this line on your plot. (To draw the line, use the equation to predict y for $x = 1$ and for $x = 5$. Plot the two (x, y) points and draw the line through them.)

(c) The correlation between these variables is $r = 0.916$. What percentage of the observed variation in beetle larvae counts can be explained by straight-line dependence on stump counts?

(d) Based on your work in parts (a), (b), and (c), do you think that counting stumps offers a quick and reliable way to predict beetle larvae clusters?

15.18 Wine and heart disease. Table 15.1 gives data on wine consumption and heart disease death rates in 19 countries in 2001. A scatterplot (Exercise 15.15) shows a moderately strong relationship. The least-squares regression line for predicting heart disease death rate from wine consumption, calculated from the data in Table 15.1, is

$$y = 115.86 - 8.05x$$

Use this equation to predict the heart disease death rate in a country where adults average 1 liter of alcohol from wine each year and in a country that averages 8 liters per year. Use these two results to draw the least-squares line on your scatterplot.

15.19 Strong association but no correlation. Exercise 14.24 gives these data on the speed (miles per hour) and mileage (miles per gallon) of a car:

Speed:	25	35	45	55	65
Mileage:	20	24	26	24	20

The least-squares line for predicting mileage from speed is

$$\text{mileage} = 22.8 + (0 \times \text{speed})$$

(a) Make a scatterplot of the data and draw this line on the plot.

(b) The correlation between mileage and speed is $r = 0$. What does this say about the usefulness of the regression line in predicting mileage?

15.20 Wine and heart disease. In Exercises 15.15 and 15.18, you examined data on wine consumption and heart disease deaths from Table 15.1. Suggest some differences among nations that may be confounded with wine-drinking habits. (*Note*: What is more, data about nations may tell us little about individual people. So these data alone are not evidence that you can lower your risk of heart disease by drinking more wine.)

15.21 Correlation and regression. If the correlation between two variables x and y is $r = 0$, there is no straight-line relationship between the variables. It turns out that the correlation is 0 exactly when the slope of the least-squares regression line is 0. Explain why slope 0 means that there is no straight-line relationship between x and y. Start by drawing a line with slope 0 and explaining why in this situation x has no value for predicting y.

15.22 Acid rain. Researchers studying acid rain measured the acidity of precipitation in a Colorado wilderness area for 150 consecutive weeks. Acidity is measured by pH. Lower pH

values show higher acidity. The acid rain researchers observed a straight-line pattern over time. They reported that the least-squares regression line

$$\text{pH} = 5.43 - (0.0053 \times \text{weeks})$$

fit the data well.

(a) Draw a graph of this line. Is the association positive or negative? Explain in plain language what this association means.

(b) According to the regression line, what was the pH at the beginning of the study (weeks = 1)? At the end (weeks = 150)?

(c) What is the slope of the regression line? Explain clearly what this slope says about the change in the pH of the precipitation in this wilderness area.

(d) Is it reasonable to use this least-squares regression line to predict the pH of precipitation after 200 weeks? Explain your answer.

15.23 Review of straight lines. Fred keeps his savings in his mattress. He began with $1000 from his mother and adds $250 each year. His total savings y after x years are given by the equation

$$y = 1000 + 250x$$

(a) Draw a graph of this equation. (Choose two values of x, such as 0 and 10. Compute the corresponding values of y from the equation. Plot these two points on graph paper and draw the straight line joining them.)

(b) After 20 years, how much will Fred have in his mattress?

(c) If Fred had added $300 instead of $250 each year to his initial $1000, what is the equation that describes his savings after x years?

15.24 Review of straight lines. During the period after birth, a male white rat gains exactly 39 grams (g) per week. (This rat is unusually regular in his growth, but 39 g per week is a realistic rate.)

(a) If the rat weighed 110 g at birth, give an equation for his weight after x weeks. What is the slope of this line?

(b) Draw a graph of this line between birth and 10 weeks of age.

(c) Would you be willing to use this line to predict the rat's weight at age two years? Do the prediction and think about the reasonableness of the result. (There are 454 grams in a pound. A large cat weighs about 10 pounds.)

15.25 More on correlation and regression. In Exercises 15.11 and 15.13, the correlation and the slope of the least-squares line for the number of boats registered in Florida and the number of manatee deaths by boats are both positive. In Exercises 15.15 and 15.18, both the correlation and the slope for wine consumption and heart disease deaths are negative. Is it possible for these two quantities (the correlation and the slope) to have opposite signs? Explain your answer.

15.26 Always plot your data! Table 15.2 presents four sets of data prepared by the statistician Frank Anscombe to illustrate the dangers of calculating without first plotting the data. *All four sets have the same correlation and the same least-squares regression line* to several decimal places. The regression equation is

$$y = 3 + 0.5x$$

TABLE 15.2 Four data sets for exploring correlation and regression

Data Set A											
x	10	8	13	9	11	14	6	4	12	7	5
y	8.04	6.95	7.58	8.81	8.33	9.96	7.24	4.26	10.84	4.82	5.68
Data Set B											
x	10	8	13	9	11	14	6	4	12	7	5
y	9.14	8.14	8.74	8.77	9.26	8.10	6.13	3.10	9.13	7.26	4.74
Data Set C											
x	10	8	13	9	11	14	6	4	12	7	5
y	7.46	6.77	12.74	7.11	7.81	8.84	6.08	5.39	8.15	6.42	5.73
Data Set D											
x	8	8	8	8	8	8	8	8	8	8	19
y	6.58	5.76	7.71	8.84	8.47	7.04	5.25	5.56	7.91	6.89	12.50

Source: Frank J. Anscombe, "Graphs in statistical analysis," *The American Statistician*, 27 (1973), pp. 17–21.

(a) Make a scatterplot for each of the four data sets and draw the regression line on each of the plots. (To draw the regression line, substitute $x = 5$ and $x = 10$ into the equation. Find the predicted y for each x. Plot these two points and draw the line through them on all four plots.)

(b) In which of the four cases would you be willing to use the regression line to predict y given that $x = 10$? Explain your answer in each case.

15.27 Going to class helps. A study of class attendance and grades among first-year students at a state university showed that in general students who attended a higher percentage of their classes earned higher grades. Class attendance explained 25% of the variation in grade index among the students. What is the numerical value of the correlation between percentage of classes attended and grade index?

15.28 The average age of farm owners. The average age of American farm owners has risen steadily during the last 30 years. Here are data on the average age of farm owners (years) from 1982 to 2012:

Year:	1982	1987	1992	1997
Average age:	50.5	52.0	53.3	54.3

Year:	2002	2007	2012
Average age:	55.3	57.1	58.3

(a) Make a scatterplot of these data. Draw by eye a regression line for predicting a year's farm population.

(b) Extend your line to predict the average age of farm owners in 2100. Is this result reasonable? Why?

15.29 Lots of wine. Exercise 15.18 gives us the least-squares line for predicting heart disease deaths per 100,000 people from liters of alcohol

from wine consumed, per person. The line is based on data from 19 rich countries. The equation is $y = 115.86 - 8.05x$. What is the predicted heart disease death rate for a country where wine consumption is 150 liters of alcohol per person? Explain why this result can't be true. Explain why using the regression line for this prediction is not intelligent.

15.30 Do emergency personnel make injuries worse? Someone says, "There is a strong positive correlation between the number of emergency personnel at the scene of an accident and the extent of injuries of those in the accident. So sending lots of emergency personnel just causes more severe injuries." Explain why this reasoning is wrong.

15.31 Facebook and grades. A September 2010 article on msnbc.com reported on a study that found that college students who are on Facebook while studying or doing homework wind up getting lower grades. Perhaps limiting time on Facebook will improve grades. Can you think of explanations for the association between time on Facebook and grades other than "time on Facebook causes a drop in grades"?

15.32 Freeway exhaust and atherosclerosis. A February 2010 news story on cnet.com reported that the artery walls of people living close to a freeway thicken faster than the walls of those who don't. Researchers correlated changes in artery wall thickness of subjects with estimates of outdoor particulate levels at each subject's home. Does this mean that you can reduce atherosclerosis (the thickening and calcification of arteries) by avoiding living near a freeway? Why?

15.33 Health and wealth. An article entitled "The Health and Wealth of Nations" says, concerning the positive correlation between health and income per capita:

This correlation is commonly thought to reflect a causal link running from income to health… Recently, however, another intriguing possibility has emerged: that the health-income correlation is partly explained by a causal link running the other way—from health to income.

Explain how higher income in a nation can cause better health. Then explain how better health can cause higher income. There is no simple way to determine the direction of the link.

15.34 Is math the key to success in college? A newspaper account of a College Board study of 15,941 high school graduates noted that minority students who take algebra and geometry in high school succeed in college at a rate that is nearly the same as whites. Here is part of the opening of a newspaper account of the study:

The link between high school math and college graduation is "almost magical," says College Board President Donald Stewart, suggesting "math is the gatekeeper for success in college."

"These findings," he says, "justify serious consideration of a national policy to ensure that all students take algebra and geometry."

What lurking variables might explain the association between taking several math courses in high school and success in college? Explain why requiring algebra and geometry may have little effect on who succeeds in college.

15.35 Does low-calorie salad dressing cause weight gain? People who use low-calorie salad dressing in place of regular dressing tend to be heavier than people who use regular dressing. Does this mean that low-calorie salad dressings cause weight gain? Give a more plausible explanation for this association.

15.36 Internet use and school grades. Children who spend many hours on the Internet get lower grades in school, on average, than those who spend less time on the Internet. Suggest some lurking variables that may explain this relationship because they contribute to both heavy Internet use and poor grades.

15.37 Correlation again. The correlation between percentage voting Democrat in 1980 and percentage voting Democrat in 1984 (Example 18.2) is $r = 0.704$. The correlation between percentage of high school seniors taking the SAT and average SAT Mathematics score in the states (Exercise 15.12) is $r = -0.89$. Which of these two correlations indicates a stronger straight-line relationship? Explain your answer.

15.38 Religion is best for lasting joy. An August 2015 article in the *Washington Post* reported a study in which researchers looked at volunteering or working with a charity; taking educational courses; participating in religious organizations; and participating in a political or community organization. Of these, participating in religious organizations was the only social activity associated with "sustained happiness." What do you think of the claim that "joining a religious group" causes "sustained happiness?"

15.39 Living on campus. A February 2, 2008, article in the *Columbus Dispatch* reported a study on the distances students lived from campus and average GPA. Here is a summary of the results:

Residence	Avg. GPA
Residence hall	3.33
Walking distance	3.16
Near campus, long walk or short drive	3.12
Within the county, not near campus	2.97
Outside the county	2.94

Based on these data, the association between the distance a student lives from campus and GPA is negative. Many universities require freshmen to live on campus, but these data have prompted some to suggest that sophomores should also be required to live on campus in order to improve grades. Do these data imply that living closer to campus improves grades? Why?

15.40 Calculating the least-squares line. Like to know the details when you study something? Here is the formula for the least-squares regression line for predicting y from x. Start with

the means $\bar{x}$ and $\bar{y}$ and the standard deviations s_x and s_y of the two variables and the correlation r between them. The least-squares line has equation $y = a + bx$ with

$$\text{slope: } b = r\frac{s_y}{s_x} \qquad \text{intercept: } a = \bar{y} - b\bar{x}$$

Example 4 in Chapter 14 (page 324) gives the means, standard deviations, and correlation for the fossil bone length data. Use these values in the formulas just given to verify the equation of the least-squares line given on page 343:

$$\text{humerus length} = -3.66 + (1.197 \times \text{femur length})$$

The remaining exercises require a two-variable statistics calculator or software that will calculate the least-squares regression line from data.

15.41 Global warming. Return to the global warming data in Exercise 15.14.

(a) Verify the equation given for the least-squares line in that exercise.

(b) Suppose you were told only that the average global temperature was 14.25 degrees Celsius. You now want to "predict" the year in which this occurred. Find the equation of the least-squares regression line that is appropriate for this purpose. What is your prediction?

(c) The two lines in parts (a) and (b) are different. Explain clearly why there are two different regression lines.

15.42 Is wine good for your heart? Table 15.1 gives data on wine consumption and heart disease death rates in 19 countries. Verify the equation of the least-squares line given in Exercise 15.18.

15.43 Always plot your data! A skeptic might wonder if the four very different data sets in Table 15.2 really do have the same correlation and least-squares line. Verify that (to a close approximation) the least-squares line is $y = 3 + 0.5x$, as given in Exercise 15.26.

EXPLORING THE WEB

Follow the QR code to access exercises.