

Looking at Data— Relationships



Do large breeds of dogs have shorter lives? See Example 2.1.

Introduction

In Chapter 1 we learned to use graphical and numerical methods to describe the distribution of a single variable. Many of the interesting examples of the use of statistics involve relationships between pairs of variables. Learning ways to describe relationships with graphical and numerical methods is the focus of this chapter.

2.1 Scatterplots

2.2 Correlation

2.3 Least-Squares Regression

2.4 Cautions about Correlation and Regression

2.5 Data Analysis for Two-Way Tables

2.6 The Question of Causation

EXAMPLE

2.1 Large breeds of dogs have shorter lives. Purebred dogs from breeds that are large tend to have shorter life spans than purebred dogs from breeds that are small. For example, one study found that miniature poodles lived an average of 9.3 years while Great Danes lived an average of only 4.6 years.¹ Irish wolfhounds have sometimes been referred to by the nickname “the heartbreak breed” because of their short life span relative to other breeds.²

We are particularly interested in situations where two variables are related in some way. To study relationships, we measure both variables on the same individuals or cases.

USE YOUR KNOWLEDGE

2.1 Relationship between first test and final exam. You want to study the relationship between the score on the first test and the score on the final exam for the 35 students enrolled in an elementary statistics class. Who are the individuals for your study?

We use the term *associated* to describe the relationship between two variables, such as breed and life span in Example 2.1. Here is another example where two variables are associated.

EXAMPLE

2.2 Size and price of a coffee beverage. You visit a local Starbucks to buy a Mocha Frappuccino[®]. The barista explains that this blended coffee beverage comes in three sizes and asks if you want a Tall, a Grande, or a Venti. The prices are \$3.15, \$3.65, and \$4.15, respectively. There is a clear association between the size of the Mocha Frappuccino and its price.

ASSOCIATION BETWEEN VARIABLES

Two variables measured on the same cases are **associated** if knowing the value of one of the variables tells you something about the values of the other variable that you would not know without this information.

In the Mocha Frappuccino example, knowing the size tells you the exact price, so the association here is very strong. Many statistical associations, however, are simply overall tendencies that allow exceptions. Although smokers on the average die earlier than nonsmokers, some people live to 90 while smoking three packs a day. Knowing that a person smokes tells us that the person is in a group of people who are more likely to die at a younger age than people in the group of nonsmokers. The association here is much weaker than the one in the Mocha Frappuccino example.

Examining relationships

When you examine the relationship between two or more variables, first ask the preliminary questions that are familiar from Chapter 1:

- What *individuals* or *cases* do the data describe?
- What *variables* are present? How are they measured?
- Which variables are *quantitative* and which are *categorical*?

EXAMPLE

2.3 Cases and variable types. In Example 2.1 the cases are dog breeds. The type of dog breed is a categorical variable, and the average life span is a quantitative variable. In Example 2.2 the cases are the containers of coffee. Size is a categorical variable with values Tall, Grande, and Venti. Price is a quantitative variable.

USE YOUR KNOWLEDGE

2.2 Suppose we used breed size? Suppose that for the dog breed example we were able to obtain some measure of average size for each of the breeds. If we replaced type of dog breed with the average breed size, how would this change the explanation in Example 2.3?

2.3 Replace names by ounces. In the Mocha Frappuccino example, the variable size is categorical, with Tall, Grande, and Venti as the possible values. Suppose you converted these values to the number of ounces: Tall is 12 ounces, Grande is 16 ounces, and Venti is 24 ounces. For studying the relationship between ounces and price, describe the cases and the variables, and state whether each is quantitative or categorical.

When you examine the relationship between two variables, a new question becomes important:

- Is your purpose simply to explore the nature of the relationship, or do you hope to show that one of the variables can explain variation in the other? That is, are some of the variables *response variables* and others *explanatory variables*?

RESPONSE VARIABLE, EXPLANATORY VARIABLE

A **response variable** measures an outcome of a study. An **explanatory variable** explains or causes changes in the response variables.

It is easiest to identify explanatory and response variables when we actually set values of one variable in order to see how it affects another variable.

EXAMPLE

2.4 Beer drinking and blood alcohol levels. How does drinking beer affect the level of alcohol in our blood? The legal limit for driving in most states is 0.08%. Student volunteers at Ohio State University drank different numbers of cans of beer. Thirty minutes later, a police officer measured their blood alcohol content. Number of beers consumed is the explanatory variable, and percent of alcohol in the blood is the response variable.

When you don't set the values of either variable but just observe both variables, there may or may not be explanatory and response variables. Whether there are depends on how you plan to use the data.

EXAMPLE

2.5 Student loans. A college student aid officer looks at the findings of the National Student Loan Survey. She notes data on the amount of debt of recent graduates, their current income, and how stressful they feel about college debt. She isn't interested in predictions but is simply trying to understand the situation of recent college graduates.

A sociologist looks at the same data with an eye to using amount of debt and income, along with other variables, to explain the stress caused by college debt. Now amount of debt and income are explanatory variables, and stress level is the response variable.

In many studies, the goal is to show that changes in one or more explanatory variables actually *cause* changes in a response variable. But many explanatory-response relationships do not involve direct causation. The SAT scores of high school students help predict the students' future college grades, but high SAT scores certainly don't cause high college grades.

independent variable
dependent variable

Some of the statistical techniques in this chapter require us to distinguish explanatory from response variables; others make no use of this distinction. You will often see explanatory variables called **independent variables** and response variables called **dependent variables**. The idea behind this language is that response variables depend on explanatory variables. Because the words "independent" and "dependent" have other meanings in statistics that are unrelated to the explanatory-response distinction, we prefer to avoid those words.

Most statistical studies examine data on more than one variable. Fortunately, statistical analysis of several-variable data builds on the tools used for examining individual variables. The principles that guide our work also remain the same:

- Start with a graphical display of the data.
- Look for overall patterns and deviations from those patterns.
- Based on what you see, use numerical summaries to describe specific aspects of the data.

2.1 Scatterplots

The most useful graph for displaying the relationship between two quantitative variables is a *scatterplot*.

SCATTERPLOT

A **scatterplot** shows the relationship between two quantitative variables measured on the same individuals. The values of one variable appear

on the horizontal axis, and the values of the other variable appear on the vertical axis. Each individual in the data appears as the point in the plot fixed by the values of both variables for that individual.

Always plot the explanatory variable, if there is one, on the horizontal axis (the x axis) of a scatterplot. As a reminder, we usually call the explanatory variable x and the response variable y . If there is no explanatory-response distinction, either variable can go on the horizontal axis.

EXAMPLE

2.6 SAT scores. More than a million high school seniors take the SAT college entrance examination each year. We sometimes see the states “rated” by the average SAT scores of their seniors. For example, Illinois students average 1179 on the SAT, which looks better than the 1038 average of Massachusetts students. Rating states by SAT scores makes little sense, however, because average SAT score is largely explained by what percent of a state’s students take the SAT. The scatterplot in Figure 2.1 allows us to see how the mean SAT score in each state is related to the percent of that state’s high school seniors who take the SAT.³

Each point on the plot represents a single individual—that is, a single state. Because we think that the percent taking the exam influences mean score, percent taking is the explanatory variable and we plot it horizontally. For example, 20% of West Virginia high school seniors take the SAT, and their mean score is 1032. West Virginia appears as the point (20, 1032) in the scatterplot, above 20 on the x axis and to the right of 1032 on the y axis.

We see at once that state average SAT score is closely related to the percent of students who take the SAT. Illinois has a high mean score, but only 11% of Illinois seniors take the SAT. In Massachusetts, on the other hand, 82% of seniors take the exam.

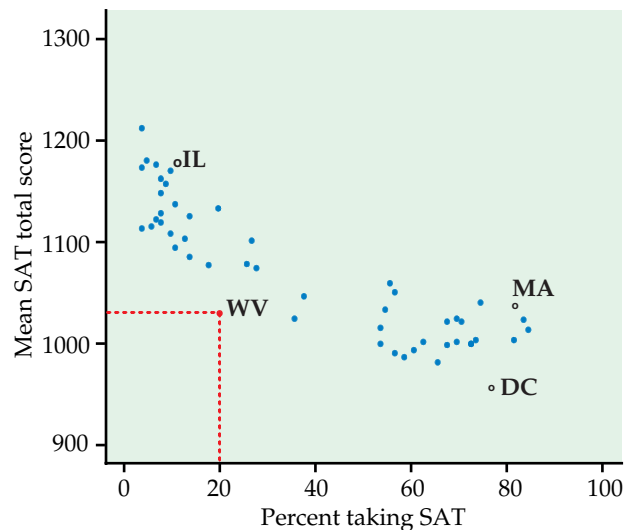


FIGURE 2.1 State mean SAT scores plotted against the percent of high school seniors in each state who take the SAT exams, for Example 2.6. The point for West Virginia (20% take the SAT, mean score 1032) is highlighted.

Interpreting scatterplots

To look more closely at a scatterplot such as Figure 2.1, apply the strategies of exploratory analysis learned in Chapter 1.

EXAMINING A SCATTERPLOT

In any graph of data, look for the **overall pattern** and for striking **deviations** from that pattern.

You can describe the overall pattern of a scatterplot by the **form**, **direction**, and **strength** of the relationship.

An important kind of deviation is an **outlier**, an individual value that falls outside the overall pattern of the relationship.

clusters

Figure 2.1 shows an interesting *form*: there are two distinct **clusters** of states. In one cluster, more than half of high school seniors take the SAT, and the mean scores are low. Fewer than 40% of seniors in states in the other cluster take the SAT—fewer than 20% in most of these states—and these states have higher mean scores.

Clusters in a graph suggest that the data describe several distinct kinds of individuals. The two clusters in Figure 2.1 do in fact describe two distinct sets of states. There are two common college entrance examinations, the SAT and the ACT. Each state tends to prefer one or the other. In ACT states (the left cluster in Figure 2.1), most students who take the SAT are applying to selective out-of-state colleges. This select group performs well. In SAT states (the right cluster), many seniors take the SAT, and this broader group has a lower mean score.

There are no clear *outliers* in Figure 2.1, but each cluster does include a state whose mean SAT score is lower than we would expect from the percent of its students who take the SAT. These points are West Virginia in the cluster of ACT states and the District of Columbia (a city rather than a state) in the cluster of SAT states.

The relationship in Figure 2.1 also has a clear *direction*: states in which a higher percent of students take the SAT tend to have lower mean scores. This is true both between the clusters and within each cluster. This is a *negative association* between the two variables.

POSITIVE ASSOCIATION, NEGATIVE ASSOCIATION

Two variables are **positively associated** when above-average values of one tend to accompany above-average values of the other and below-average values also tend to occur together.

Two variables are **negatively associated** when above-average values of one accompany below-average values of the other, and vice versa.

When a scatterplot shows distinct clusters, it is often useful to describe the overall pattern separately within each cluster. The *form* of the relationship in

linear relationship

the ACT states is roughly **linear**. That is, the points roughly follow a straight line. The *strength* of a relationship in a scatterplot is determined by how closely the points follow a clear form. The linear pattern among the ACT states is moderately strong because the points show only modest scatter about the straight-line pattern. In summary, the ACT states in Figure 2.1 show a moderately strong negative linear relationship. The cluster of SAT states shows a much weaker relationship between percent taking the SAT and mean SAT score.

USE YOUR KNOWLEDGE

2.4 Make a scatterplot. In our Mocha Frappuccino example, the 12-ounce drink costs \$3.15, the 16-ounce drink costs \$3.65, and the 24-ounce drink costs \$4.15. Explain which variable should be used as the explanatory variable and make a scatterplot. Describe the scatterplot and the association between these two variables.

Adding categorical variables to scatterplots

The Census Bureau groups the states into four broad regions, named Midwest, Northeast, South, and West. We might ask about regional patterns in SAT exam scores. Figure 2.2 repeats part of Figure 2.1, with an important difference. We have plotted only the Northeast and Midwest groups of states, using the plot symbol “e” for the northeastern states and the symbol “m” for the midwestern states.

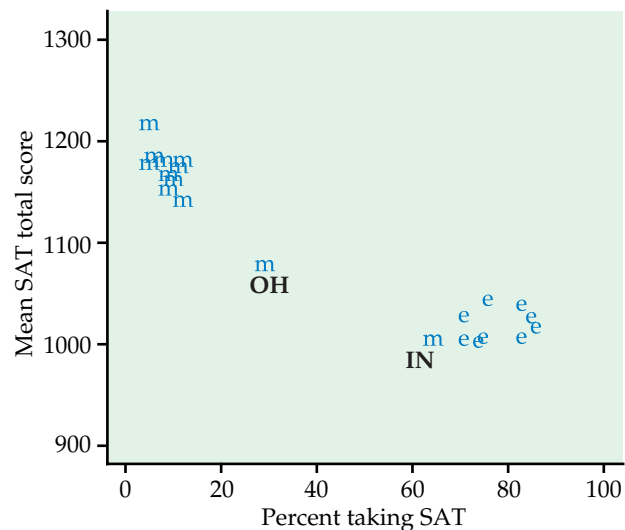


FIGURE 2.2 State mean SAT scores and percent taking the SAT for the northeastern states (plot symbol “e”) and the midwestern states (plot symbol “m”).

The regional comparison is striking. The 9 northeastern states are all SAT states—in fact, at least 70% of high school graduates in each of these states take the SAT. The 12 midwestern states are mostly ACT states. In 10 of these states, the percent taking the SAT is between 4% and 11%. One midwestern state is clearly an outlier within the region. Indiana is an SAT state (63% take the SAT) that falls close to the northeastern cluster. Ohio, where 28% take the SAT, also lies outside the midwestern cluster.

In dividing the states into regions, we introduced a third variable into the scatterplot. “Region” is a categorical variable that has four values, although we plotted data from only two of the four regions. The two regions are displayed by the two different plotting symbols.⁴

CATEGORICAL VARIABLES IN SCATTERPLOTS

To add a categorical variable to a scatterplot, use a different plot color or symbol for each category.

More examples of scatterplots

Experience in examining scatterplots is the foundation for more detailed study of relationships among quantitative variables. Here is an example with a pattern different from that in Figure 2.1.



EXAMPLE

2.7 The Trans-Alaska Oil Pipeline. The Trans-Alaska Oil Pipeline is a tube formed from 1/2-inch-thick steel that carries oil across 800 miles of sensitive arctic and subarctic terrain. The pipe and the welds that join pipe segments were carefully examined before installation. How accurate are field measurements of the depth of small defects? Figure 2.3 compares the results of measurements on 100 defects made in the field with measurements of the same defects made in the laboratory.⁵ We plot the laboratory results on the x axis because they are a standard against which we compare the field results.

What is the overall pattern of this scatterplot? There is a positive linear association between the two variables. This is what we expect from two measurements of the same quantity. If field and laboratory measurements agree, the points will all fall on the $y = x$ line drawn on the plot, except for small random variations in the measurements. In fact, we see that the points for larger

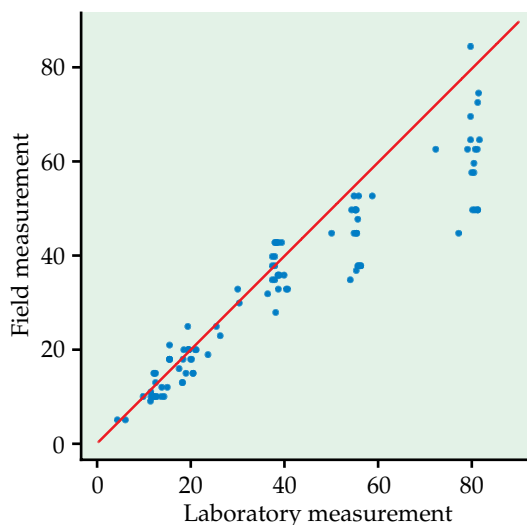


FIGURE 2.3 Depths of small defects in pipe for the Trans-Alaska Oil Pipeline, measured in the field and in the laboratory, for Example 2.7. If the two measurements were the same, the points would lie on the $y = x$ line that is drawn on the graph.

defects fall systematically below this line. That is, the field measurements are too small compared with the laboratory results as a standard. This is an important finding that can be used to adjust future field measurements.

Another part of the overall pattern is that the strength of the linear relationship decreases as the size of the defects increases. Field data show more variation (vertical spread in the plot) for large defect sizes than for small sizes. An increase in the spread in a response variable as the size of the response increases is a common pattern. It implies that predictions of the response based on the overall pattern will be less accurate for large responses.

Did you notice a fine point of graphing technique? Because both x and y measure the same thing, the graph is square and the same scales appear on both axes.

Some scatterplots appear quite different from the cloud of points in Figure 2.1 and the linear pattern in Figure 2.3. This is true, for example, in experiments in which measurements of a response variable are taken at a few selected levels of the explanatory variable. The following example illustrates the use of scatterplots in this setting.



EXAMPLE

2.8 Predators and prey. Here is one way in which nature regulates the size of animal populations: high population density attracts predators, who remove a higher proportion of the population than when the density of the prey is low. One study looked at kelp perch and their common predator, the kelp bass. The researcher set up four large circular pens on sandy ocean bottom in southern California. He chose young perch at random from a large group and placed 10, 20, 40, and 60 perch in the four pens. Then he dropped the nets protecting the pens, allowing bass to swarm in, and counted the perch left after 2 hours. Here are data on the proportions of perch eaten in four repetitions of this setup:⁶

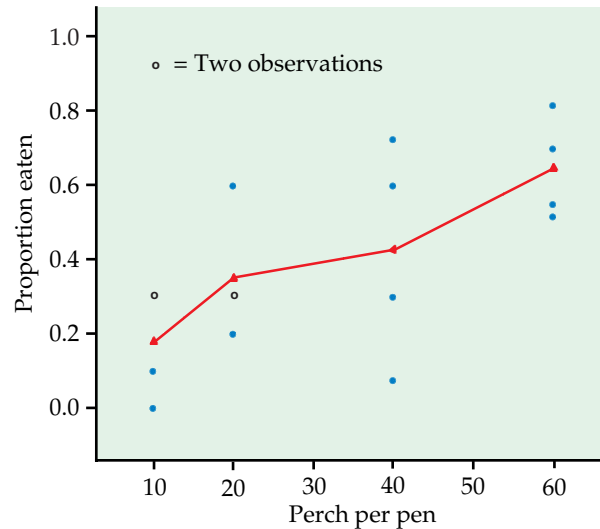
Perch	Proportion killed			
10	0.0	0.1	0.3	0.3
20	0.2	0.3	0.3	0.6
40	0.075	0.3	0.6	0.725
60	0.517	0.55	0.7	0.817

The scatterplot in Figure 2.4 displays the results of this experiment. Because number of perch in a pen is the explanatory variable, we plot it horizontally as the x variable. The proportion of perch eaten by bass is the response variable y . Notice that there are two identical responses in the 10-perch group and also in the 20-perch group. These pairs of observations occupy the same points on the plot, so we use a different symbol for points that represent two observations. Most software does not alert you to repeated values in your data when making scatterplots. This can affect the impression the plot creates, especially when there are just a few points.

The vertical spread of points above each pen size shows the variation in proportions of perch eaten by bass. To see the overall pattern behind this



FIGURE 2.4 Data from an experiment in ecology; proportion of perch eaten by bass plotted against the number of perch present, for Example 2.8. The lines connect the mean responses (triangles) for each group.



variation, plot the mean response for each pen size. In Figure 2.4, these means are marked by triangles and joined by line segments. There is a clear positive association between number of prey present and proportion eaten by predators. Moreover, the relationship is not far from linear.

BEYOND THE BASICS

Scatterplot Smoothers

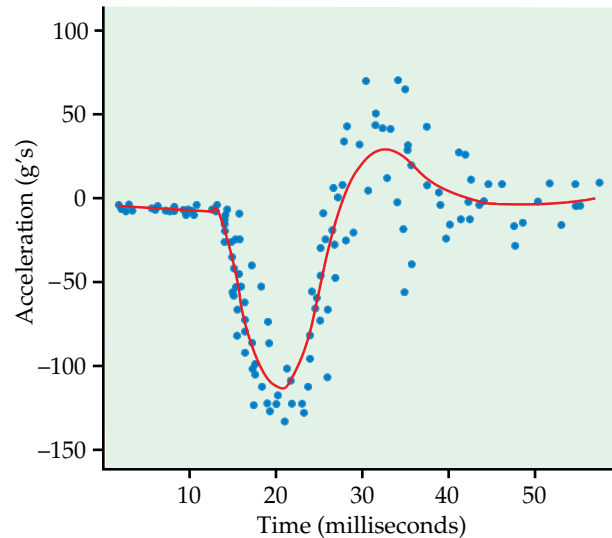
smoothing

A scatterplot provides a complete picture of the relationship between two quantitative variables. A complete picture is often too detailed for easy interpretation, so we try to describe the plot in terms of an overall pattern and deviations from that pattern. Though we can often do this by eye, more systematic methods of extracting the overall pattern are helpful. This is called **smoothing** a scatterplot. Example 2.9 suggests how to proceed when we are plotting a response variable y against an explanatory variable x . We smoothed Figure 2.4 by averaging the y -values separately for each x -value. Though not all scatterplots have many y -values at the same value of x , as did Figure 2.4, modern software provides scatterplot smoothers that form an overall pattern by looking at the y -values for points in the neighborhood of each x -value. Smoothers use *resistant* calculations, so they are not affected by outliers in the plot.

EXAMPLE

2.9 Dummies in motorcycle crashes. Crash a motorcycle into a wall. The rider, fortunately, is a dummy with an instrument to measure acceleration (change of velocity) mounted in its head. Figure 2.5 plots the acceleration of the dummy's head against time in milliseconds.⁷ Acceleration is measured in g 's, or multiples of the acceleration due to gravity at the earth's surface. The motorcycle approaches the wall at a constant speed (acceleration near 0). As it hits, the dummy's head snaps forward and decelerates violently (negative

FIGURE 2.5 Smoothing a scatterplot, for Example 2.9. Time plot of the acceleration of the head of a crash dummy as a motorcycle hits a wall, with the overall pattern calculated by a scatterplot smoother.



acceleration reaching more than 100 g's), then snaps back again (up to 75 g's) and wobbles a bit before coming to rest.

The scatterplot has a clear overall pattern, but it does not obey a simple form such as linear. Moreover, the strength of the pattern varies, from quite strong at the left of the plot to weaker (much more scatter) at the right. A scatterplot smoother deals with this complexity quite effectively and draws a line on the plot to represent the overall pattern.

Categorical explanatory variables

Scatterplots display the association between two quantitative variables. To display a relationship between a categorical explanatory variable and a quantitative response variable, make a side-by-side comparison of the distributions of the response for each category. We have already met some tools for such comparisons:

- A back-to-back stemplot compares two distributions. See the comparison of literacy rates (the quantitative response) for females and males (two categories) on page 11.
- Side-by-side boxplots compare any number of distributions. See the comparison of gas mileage (the quantitative response) for minicompact and two-seater cars on the highway and in the city (four categories) in Figure 1.19 (page 37).

You can also use a type of scatterplot to display the association between a categorical explanatory variable and a quantitative response. Suppose, for example, that the prey-predator study of Example 2.8 had compared four species of prey rather than four densities of prey. The plot in Figure 2.4 remains helpful if we mark the prey species as A, B, C, and D at equal intervals on the horizontal axis in place of the count of perch per pen. A graph of the

mean or median responses at the four locations still shows the overall nature of the relationship.

Many categorical variables, like prey species or type of car, have no natural order from smallest to largest. In such situations we cannot speak of a positive or negative association with the response variable. If the mean responses in our plot increase as we go from left to right, we could make them decrease by writing the categories in the opposite order. The plot simply presents a side-by-side comparison of several distributions. The categorical variable labels the distributions. Some categorical variables do have a least-to-most order, however. We can then speak of the direction of the association between the categorical explanatory variable and the quantitative response. Look again at the boxplots of income by level of education in Figure 1.23, on page 52. Although the Census Bureau records education in categories, such as “did not graduate from high school,” the categories have an order from less education to more education. The boxes in Figure 1.23 are arranged in order of increasing education. They show a positive association between education and income: people with more education tend to have higher incomes.

SECTION 2.1 Summary

To study relationships between variables, we must measure the variables on the same group of individuals or cases.

If we think that a variable x may explain or even cause changes in another variable y , we call x an **explanatory variable** and y a **response variable**.

A **scatterplot** displays the relationship between two quantitative variables. Mark values of one variable on the horizontal axis (x axis) and values of the other variable on the vertical axis (y axis). Plot each individual’s data as a point on the graph.

Always plot the explanatory variable, if there is one, on the x axis of a scatterplot. Plot the response variable on the y axis.

Plot points with different colors or symbols to see the effect of a categorical variable in a scatterplot.

In examining a scatterplot, look for an overall pattern showing the **form**, **direction**, and **strength** of the relationship, and then for **outliers** or other deviations from this pattern.

Form: Linear relationships, where the points show a straight-line pattern, are an important form of relationship between two variables. Curved relationships and **clusters** are other forms to watch for.

Direction: If the relationship has a clear direction, we speak of either **positive association** (high values of the two variables tend to occur together) or **negative association** (high values of one variable tend to occur with low values of the other variable).

Strength: The **strength** of a relationship is determined by how close the points in the scatterplot lie to a simple form such as a line.

To display the relationship between a categorical explanatory variable and a quantitative response variable, make a graph that compares the distributions of the response for each category of the explanatory variable.

SECTION 2.1 Exercises

For Exercise 2.1, see page 84; for Exercises 2.2 and 2.3, see page 85; and for Exercise 2.4, see page 89.

2.5 Average temperatures. Here are the average temperatures in degrees for Lafayette, Indiana, during the months of February through May:

Month	February	March	April	May
Temperature (degrees F)	30	41	51	62

- (a) Explain why month should be the explanatory variable for examining this relationship.
- (b) Make a scatterplot and describe the relationship.

2.6 Relationship between first test and final exam. How strong is the relationship between the score on the first test and the score on the the final exam in an elementary statistics course? Here are data for eight students from such a course:

First-test score	153	144	162	149	127	118	158	153
Final-exam score	145	140	145	170	145	175	170	160

- (a) Which variable should play the role of the explanatory variable in describing this relationship?
- (b) Make a scatterplot and describe the relationship.
- (c) Give some possible reasons why this relationship is so weak.

2.7 Relationship between second test and final exam. Refer to the previous exercise. Here are the data for the second test and the final exam for the same students:

Second-test score	158	162	144	162	136	158	175	153
Final-exam score	145	140	145	170	145	175	170	160

- (a) Explain why you should use the second-test score as the explanatory variable.
- (b) Make a scatterplot and describe the relationship.
- (c) Why do you think the relationship between the second-test score and the final-exam score is stronger than the relationship between the first-test score and the final-exam score?

2.8 Add an outlier to the plot. Refer to the previous exercise. Add a ninth student whose scores on the second test and final exam would lead you to classify the additional data point as an outlier. Highlight the outlier on your scatterplot and describe the performance of the student on the second exam and

final exam and why that leads to the conclusion that the result is an outlier. Give a possible reason for the performance of this student.

2.9 Explanatory and response variables. In each of the following situations, is it more reasonable to simply explore the relationship between the two variables or to view one of the variables as an explanatory variable and the other as a response variable? In the latter case, which is the explanatory variable and which is the response variable?

- (a) The weight of a child and the age of the child from birth to 10 years.
- (b) High school English grades and high school math grades.
- (c) The rental price of apartments and the number of bedrooms in the apartment.
- (d) The amount of sugar added to a cup of coffee and how sweet the coffee tastes.
- (e) The student evaluation scores for an instructor and the student evaluation scores for the course.

2.10 Parents' income and student loans. How well does the income of a college student's parents predict how much the student will borrow to pay for college? We have data on parents' income and college debt for a sample of 1200 recent college graduates. What are the explanatory and response variables? Are these variables categorical or quantitative? Do you expect a positive or negative association between these variables? Why?

2.11 Reading ability and IQ. A study of reading ability in schoolchildren chose 60 fifth-grade children at random from a school. The researchers had the children's scores on an IQ test and on a test of reading ability.⁸ Figure 2.6 (on page 96) plots reading test score (response) against IQ score (explanatory).

- (a) Explain why we should expect a positive association between IQ and reading score for children in the same grade. Does the scatterplot show a positive association?
- (b) A group of four points appear to be outliers. In what way do these children's IQ and reading scores deviate from the overall pattern?
- (c) Ignoring the outliers, is the association between IQ and reading scores roughly linear? Is it very strong? Explain your answers.

2.12 Treasury bills and common stocks. What is the relationship between returns from buying Treasury bills and returns from buying common stocks? The stemplots in Figure 1.22 (page 44) show the two

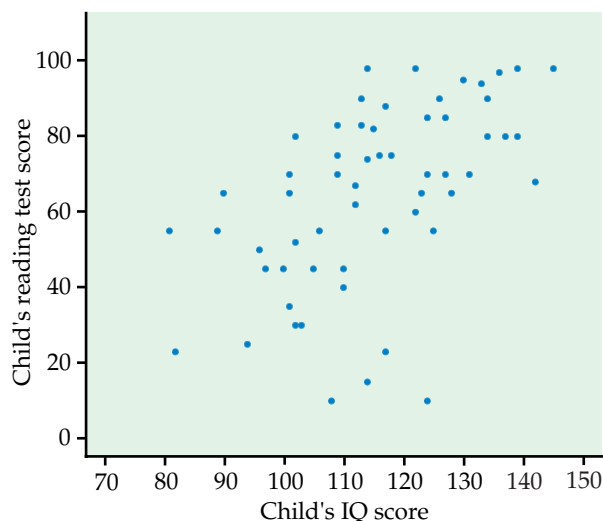


FIGURE 2.6 IQ and reading test scores for 60 fifth-grade children, for Exercise 2.11.

individual distributions of percent returns. To see the relationship, we need a scatterplot. Figure 2.7 plots the annual returns on stocks for the years 1950 to 2003 against the returns on Treasury bills for the same years.

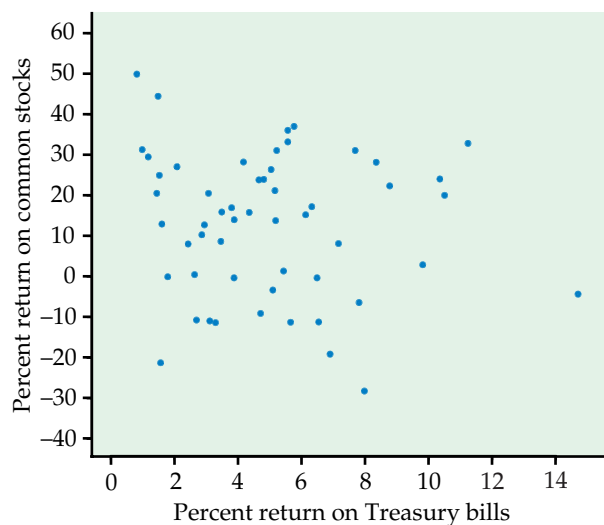


FIGURE 2.7 Percent return on Treasury bills and common stocks for the years 1950 to 2003, for Exercise 2.12.

- (a) The best year for stocks during this period was 1954. The worst year was 1974. About what were the returns on stocks in those two years?
- (b) Treasury bills are a measure of the general level of interest rates. The years around 1980 saw very

high interest rates. Treasury bill returns peaked in 1981. About what was the percent return that year?

(c) Some people say that high Treasury bill returns tend to go with low returns on stocks. Does such a pattern appear clearly in Figure 2.7? Does the plot have any clear pattern?

2.13 Can children estimate their reading ability? The main purpose of the study cited in Exercise 2.11 was to ask whether schoolchildren can estimate their own reading ability. The researchers had the children's scores on a test of reading ability. They asked each child to estimate his or her reading level, on a scale from 1 (low) to 5 (high). Figure 2.8 is a scatterplot of the children's estimates (response) against their reading scores (explanatory).

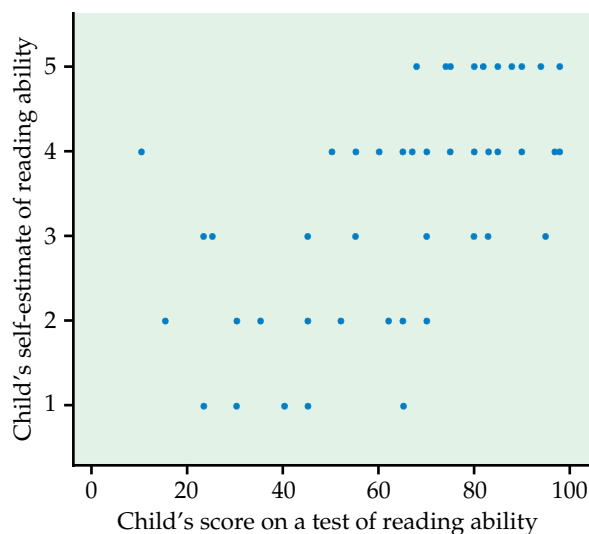


FIGURE 2.8 Reading test scores for 60 fifth-grade children and the children's estimates of their own reading levels, for Exercise 2.13.

- (a) What explains the "stair-step" pattern in the plot?
- (b) Is there an overall positive association between reading score and self-estimate?
- (c) There is one clear outlier. What is this child's self-estimated reading level? Does this appear to over- or underestimate the level as measured by the test?

2.14 Literacy of men and women. Table 1.2 (page 10) shows the percent of men and women at least 15 years old who were literate in 2002 in the major Islamic nations for which data were available. Make a scatterplot of these data, taking male literacy as the explanatory variable. Describe the direction,

form, and strength of the relationship. Are there any identical observations that plot as the same point? Are there any clear outliers?

2.15 Small falcons in Sweden. Often the percent of an animal species in the wild that survive to breed again is lower following a successful breeding season. This is part of nature's self-regulation, tending to keep population size stable. A study of merlins (small falcons) in northern Sweden observed the number of breeding pairs in an isolated area and the percent of males (banded for identification) who returned the next breeding season. Here are data for nine years:⁹

Pairs	28	29	29	29	30	32	33	38	38
Percent	82	83	70	61	69	58	43	50	47

- (a) Why is the response variable the *percent* of males that return rather than the *number* of males that return?
- (b) Make a scatterplot. To emphasize the pattern, also plot the mean response for years with 29 and 38 breeding pairs and draw lines connecting the mean responses for the six values of the explanatory variable.
- (c) Describe the pattern. Do the data support the theory that a smaller percent of birds survive following a successful breeding season?

2.16 City and highway gas mileage. Table 1.10 (page 31) gives the city and highway gas mileages for minicompact and two-seater cars. We expect a positive association between the city and highway mileages of a group of vehicles. We have already seen that the Honda Insight is a different type of car, so omit it as you work with these data.

(a) Make a scatterplot that shows the relationship between city and highway mileage, using city mileage as the explanatory variable. Use different plotting symbols for the two types of cars.

(b) Interpret the plot. Is there a positive association? Is the form of the plot roughly linear? Is the form of the relationship similar for the two types of car? What is the most important difference between the two types?

2.17 Social rejection and pain. We often describe our emotional reaction to social rejection as “pain.” A clever study asked whether social rejection causes activity in areas of the brain that are known to be activated by physical pain. If it does, we really do experience social and physical pain in similar ways. Subjects were first included and then deliberately

excluded from a social activity while increases in blood flow in their brains were measured. After each activity, the subjects filled out questionnaires that assessed how excluded they felt.

Below are data for 13 subjects.¹⁰ The explanatory variable is “social distress” measured by each subject’s questionnaire score after exclusion relative to the score after inclusion. (So values greater than 1 show the degree of distress caused by exclusion.) The response variable is activity in the anterior cingulate cortex, a region of the brain that is activated by physical pain.

Subject	Social distress	Brain activity	Subject	Social distress	Brain activity
1	1.26	−0.055	8	2.18	0.025
2	1.85	−0.040	9	2.58	0.027
3	1.10	−0.026	10	2.75	0.033
4	2.50	−0.017	11	2.75	0.064
5	2.17	−0.017	12	3.33	0.077
6	2.67	0.017	13	3.65	0.124
7	2.01	0.021			

Plot brain activity against social distress. Describe the direction, form, and strength of the relationship, as well as any outliers. Do the data suggest that brain activity in the “pain” region is directly related to the distress from social exclusion?

2.18 Biological clocks. Many plants and animals have “biological clocks” that coordinate activities with the time of day. When researchers looked at the length of the biological cycles in the plant *Arabidopsis* by measuring leaf movements, they found that the length of the cycle is not always 24 hours. The researchers suspected that the plants adapt their clocks to their north-south position. Plants don’t know geography, but they do respond to light, so the researchers looked at the relationship between the plants’ cycle lengths and the length of the day on June 21 at their locations. The data file has data on cycle length and day length, both in hours, for 146 plants.¹¹ Plot cycle length as the response variable against day length as the explanatory variable. Does there appear to be a positive association? Is it a strong association? Explain your answers.

2.19 Business revenue and team value in the NBA. Management theory says that the value of a business should depend on its operating income, the income produced by the business after taxes. (Operating income excludes income from sales of assets and investments, which don’t reflect the actual business.) Total revenue, which ignores costs, should be less

TABLE 2.1
NBA teams as businesses

Team	Value (\$millions)	Revenue (\$millions)	Income (\$millions)
Los Angeles Lakers	447	149	22.8
New York Knicks	401	160	13.5
Chicago Bulls	356	119	49.0
Dallas Mavericks	338	117	−17.7
Philadelphia 76ers	328	109	2.0
Boston Celtics	290	97	25.6
Detroit Pistons	284	102	23.5
San Antonio Spurs	283	105	18.5
Phoenix Suns	282	109	21.5
Indiana Pacers	280	94	10.1
Houston Rockets	278	82	15.2
Sacramento Kings	275	102	−16.8
Washington Wizards	274	98	28.5
Portland Trail Blazers	272	97	−85.1
Cleveland Cavaliers	258	72	3.8
Toronto Raptors	249	96	10.6
New Jersey Nets	244	94	−1.6
Utah Jazz	239	85	13.8
Miami Heat	236	91	7.9
Minnesota Timberwolves	230	85	6.9
Memphis Grizzlies	227	63	−19.7
Denver Nuggets	218	75	7.9
New Orleans Hornets	216	80	21.9
Los Angeles Clippers	208	72	15.9
Atlanta Hawks	202	78	−8.4
Orlando Magic	199	80	13.1
Seattle Supersonics	196	70	2.4
Golden State Warriors	188	70	7.8
Milwaukee Bucks	174	70	−15.1

important. Table 2.1 shows the values, operating incomes, and revenues of an unusual group of businesses: the teams in the National Basketball Association (NBA).¹² Professional sports teams are generally privately owned, often by very wealthy individuals who may treat their team as a source of prestige rather than as a business.

(a) Plot team value against revenue. There are several outliers. Which teams are these, and in what way are they outliers? Is there a positive association between value and revenue? Is the pattern roughly linear?

(b) Now plot value against operating income. Are the same teams outliers? Does revenue or operating income better predict the value of an NBA team?

2.20 Two problems with feet. Metatarsus adductus (call it MA) is a turning in of the front part of the foot that is common in adolescents and usually corrects itself. Hallux abducto valgus (call it HAV) is a deformation of the big toe that is not common in youth and often requires surgery. Perhaps the severity of MA can help predict the severity of HAV. Table 2.2 gives data

TABLE 2.2
Two measurements of foot deformities

HAV angle	MA angle	HAV angle	MA angle	HAV angle	MA angle
28	18	21	15	16	10
32	16	17	16	30	12
25	22	16	10	30	10
34	17	21	7	20	10
38	33	23	11	50	12
26	10	14	15	25	25
25	18	32	12	26	30
18	13	25	16	28	22
30	19	21	16	31	24
26	10	22	18	38	20
28	17	20	10	32	37
13	14	18	15	21	23
20	20	26	16		

on 38 consecutive patients who came to a medical center for HAV surgery.¹³ Using X-rays, doctors measured the angle of deformity for both MA and HAV. They speculated that there is a positive association—more serious MA is associated with more serious HAV.

(a) Make a scatterplot of the data in Table 2.2. (Which is the explanatory variable?)

(b) Describe the form, direction, and strength of the relationship between MA angle and HAV angle. Are there any clear outliers in your graph?

(c) Do you think the data confirm the doctors' speculation?

2.21 Body mass and metabolic rate. Metabolic rate, the rate at which the body consumes energy, is important in studies of weight gain, dieting, and exercise. The table below gives data on the lean body mass and resting metabolic rate for 12 women and 7 men who are subjects in a study of dieting. Lean body mass, given in kilograms, is a person's weight leaving out all fat. Metabolic rate is measured in calories burned per 24 hours, the same calories

Subject	Sex	Mass	Rate	Subject	Sex	Mass	Rate
1	M	62.0	1792	11	F	40.3	1189
2	M	62.9	1666	12	F	33.1	913
3	F	36.1	995	13	M	51.9	1460
4	F	54.6	1425	14	F	42.4	1124
5	F	48.5	1396	15	F	34.5	1052
6	F	42.0	1418	16	F	51.1	1347
7	M	47.4	1362	17	F	41.2	1204
8	F	50.6	1502	18	M	51.9	1867
9	F	42.0	1256	19	M	46.9	1439
10	M	48.7	1614				

TABLE 2.3

World record times for the 10,000-meter run

Men				Women	
Record year	Time (seconds)	Record year	Time (seconds)	Record year	Time (seconds)
1912	1880.8	1962	1698.2	1967	2286.4
1921	1840.2	1963	1695.6	1970	2130.5
1924	1835.4	1965	1659.3	1975	2100.4
1924	1823.2	1972	1658.4	1975	2041.4
1924	1806.2	1973	1650.8	1977	1995.1
1937	1805.6	1977	1650.5	1979	1972.5
1938	1802.0	1978	1642.4	1981	1950.8
1939	1792.6	1984	1633.8	1981	1937.2
1944	1775.4	1989	1628.2	1982	1895.3
1949	1768.2	1993	1627.9	1983	1895.0
1949	1767.2	1993	1618.4	1983	1887.6
1949	1761.2	1994	1612.2	1984	1873.8
1950	1742.6	1995	1603.5	1985	1859.4
1953	1741.6	1996	1598.1	1986	1813.7
1954	1734.2	1997	1591.3	1993	1771.8
1956	1722.8	1997	1587.8		
1956	1710.4	1998	1582.7		
1960	1698.8	2004	1580.3		

used to describe the energy content of foods. The researchers believe that lean body mass is an important influence on metabolic rate.

(a) Make a scatterplot of the data, using different symbols or colors for men and women.

(b) Is the association between these variables positive or negative? What is the form of the relationship? How strong is the relationship? Does the pattern of the relationship differ for women and men? How do the male subjects as a group differ from the female subjects as a group?

2.22 Fuel consumption and speed. How does the fuel consumption of a car change as its speed increases? Below are data for a British Ford Escort. Speed is measured in kilometers per hour, and fuel consumption is measured in liters of gasoline used per 100 kilometers traveled.¹⁴

Speed (km/h)	Fuel used (liters/100 km)	Speed (km/h)	Fuel used (liter/100 km)
10	21.00	90	7.57
20	13.00	100	8.27
30	10.00	110	9.03
40	8.00	120	9.87
50	7.00	130	10.79
60	5.90	140	11.77
70	6.30	150	12.83
80	6.95		

(a) Make a scatterplot. (Which variable should go on the x axis?)

(b) Describe the form of the relationship. In what way is it not linear? Explain why the form of the relationship makes sense.

(c) It does not make sense to describe the variables as either positively associated or negatively associated. Why not?

(d) Is the relationship reasonably strong or quite weak? Explain your answer.

2.23 World records for the 10K. Table 2.3 shows the progress of world record times (in seconds) for the 10,000-meter run up to mid-2004.¹⁵ Concentrate on the women’s world record times. Make a scatterplot with year as the explanatory variable. Describe the pattern of improvement over time that your plot displays.

2.24 CHALLENGE How do icicles grow? How fast do icicles grow? Japanese researchers measured the growth of icicles in a cold chamber under various conditions of temperature, wind, and water flow.¹⁶ Table 2.4 contains data produced under two sets of conditions. In both cases, there was no wind and the temperature was set at -11°C . Water flowed over the icicle at a higher rate (29.6 milligrams per second) in Run 8905 and at a slower rate (11.9 mg/s) in Run 8903.

TABLE 2.4
Growth of icicles over time

Run 8903				Run 8905			
Time (min)	Length (cm)	Time (min)	Length (cm)	Time (min)	Length (cm)	Time (min)	Length (cm)
10	0.6	130	18.1	10	0.3	130	10.4
20	1.8	140	19.9	20	0.6	140	11.0
30	2.9	150	21.0	30	1.0	150	11.9
40	4.0	160	23.4	40	1.3	160	12.7
50	5.0	170	24.7	50	3.2	170	13.9
60	6.1	180	27.8	60	4.0	180	14.6
70	7.9			70	5.3	190	15.8
80	10.1			80	6.0	200	16.2
90	10.9			90	6.9	210	17.9
100	12.7			100	7.8	220	18.8
110	14.4			110	8.3	230	19.9
120	16.6			120	9.6	240	21.1

- (a) Make a scatterplot of the length of the icicle in centimeters versus time in minutes, using separate symbols for the two runs.
- (b) Write a careful explanation of what your plot shows about the growth of icicles.

2.25 **Records for men and women in the 10K.** Table 2.3 shows the progress of world record times (in seconds) for the 10,000-meter run for both men and women.

- (a) Make a scatterplot of world record time against year, using separate symbols for men and women. Describe the pattern for each sex. Then compare the progress of men and women.
- (b) Women began running this long distance later than men, so we might expect their improvement to be more rapid. Moreover, it is often said that men have little advantage over women in distance running as opposed to sprints, where muscular strength plays a greater role. Do the data appear to support these claims?

2.26 **Worms and plant growth.** To demonstrate the effect of nematodes (microscopic worms) on plant growth, a botanist introduces different numbers of nematodes into 16 planting pots. He then transplants a tomato seedling into each pot. Here are data on the increase in height of the seedlings (in centimeters) 14 days after planting:¹⁷

Nematodes	Seedling growth			
0	10.8	9.1	13.5	9.2
1,000	11.1	11.1	8.2	11.3
5,000	5.4	4.6	7.4	5.0
10,000	5.8	5.3	3.2	7.5

- (a) Make a scatterplot of the response variable (growth) against the explanatory variable (nematode count). Then compute the mean growth for each group of seedlings, plot the means against the nematode counts, and connect these four points with line segments.


- (b) Briefly describe the conclusions about the effects of nematodes on plant growth that these data suggest.

2.27 **Mutual funds.** Fidelity Investments, like other large mutual funds companies, offers many “sector funds” that concentrate their investments in narrow segments of the stock market. These funds often rise or fall by much more than the market as a whole. We can group them by broader market sector to compare returns. Here are percent total returns for 23 Fidelity “Select Portfolios” funds for the year 2003, a year in which stocks rose sharply:¹⁸

Market sector	Fund returns (percent)							
Consumer	23.9	14.1	41.8	43.9	31.1			
Financial services	32.3	36.5	30.6	36.9	27.5			
Technology	26.1	62.7	68.1	71.9	57.0	35.0	59.4	
Natural resources	22.9	7.6	32.1	28.7	29.5	19.1		

- (a) Make a plot of total return against market sector (space the four market sectors equally on the horizontal axis). Compute the mean return for each sector, add the means to your plot, and connect the means with line segments.
- (b) Based on the data, which of these market sectors were the best places to invest in 2003? Hindsight is wonderful.

(c) Does it make sense to speak of a positive or negative association between market sector and total return?

2.28  **Mutual funds in another year.** The data for 2003 in the previous exercise make sector funds look attractive. Stocks rose sharply in 2003, after falling sharply in 2002 (and also in 2001 and 2000). Let's look at the percent returns for both 2003 and 2002 for these same 23 funds. Here they are:

2002 return	2003 return	2002 return	2003 return	2002 return	2003 return
-17.1	23.9	-0.7	36.9	-37.8	59.4
-6.7	14.1	-5.6	27.5	-11.5	22.9
-21.1	41.8	-26.9	26.1	-0.7	36.9
-12.8	43.9	-42.0	62.7	64.3	32.1
-18.9	31.1	-47.8	68.1	-9.6	28.7
-7.7	32.3	-50.5	71.9	-11.7	29.5
-17.2	36.5	-49.5	57.0	-2.3	19.1
-11.4	30.6	-23.4	35.0		

Do a careful graphical analysis of these data: side-by-side comparison of the distributions of returns in 2002 and 2003 and also a description of the relationship between the returns of the same funds in these two years. What are your most important findings? (The outlier is Fidelity Gold Fund.)

2.2 Correlation

A scatterplot displays the form, direction, and strength of the relationship between two quantitative variables. Linear (straight-line) relations are particularly important because a straight line is a simple pattern that is quite common. We say a linear relationship is strong if the points lie close to a straight line, and weak if they are widely scattered about a line. Our eyes are not good judges of how strong a relationship is. The two scatterplots in Figure 2.9 depict exactly the same data, but the plot on the right is drawn smaller in a large field. The plot on the left seems to show a stronger relationship. Our eyes can be fooled by changing the plotting scales or the amount of white space around the cloud of points in a scatterplot.¹⁹ We need to follow our strategy for data analysis by using a numerical measure to supplement the graph. *Correlation* is the measure we use.

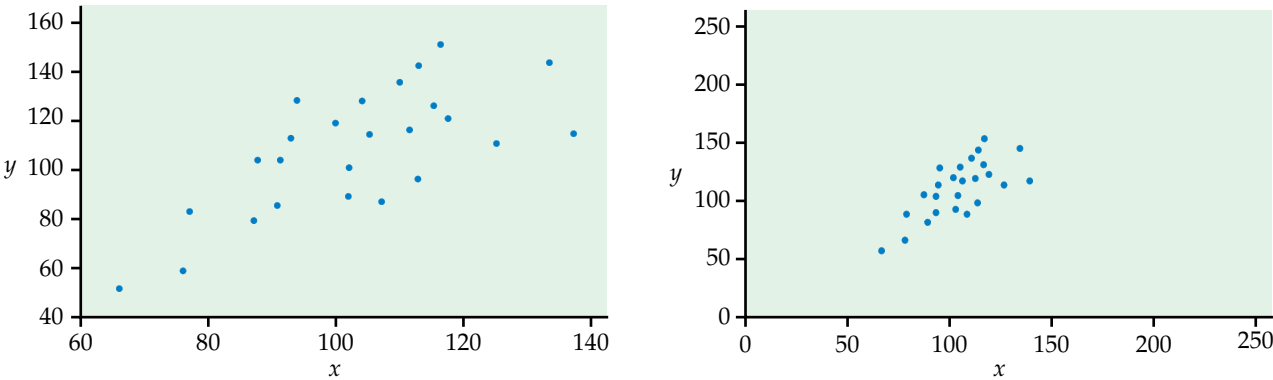


FIGURE 2.9 Two scatterplots of the same data. The linear pattern in the plot on the right appears stronger because of the surrounding space.

The correlation r

We have data on variables x and y for n individuals. Think, for example, of measuring height and weight for n people. Then x_1 and y_1 are your height and your weight, x_2 and y_2 are my height and my weight, and so on. For the i th individual, height x_i goes with weight y_i . Here is the definition of correlation.

CORRELATION

The **correlation** measures the direction and strength of the linear relationship between two quantitative variables. Correlation is usually written as r .

Suppose that we have data on variables x and y for n individuals. The means and standard deviations of the two variables are \bar{x} and s_x for the x -values, and \bar{y} and s_y for the y -values. The correlation r between x and y is

$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

As always, the summation sign \sum means “add these terms for all the individuals.” The formula for the correlation r is a bit complex. It helps us see what correlation is but is not convenient for actually calculating r . In practice you should use software or a calculator that finds r from keyed-in values of two variables x and y . Exercise 2.29 asks you to calculate a correlation step-by-step from the definition to solidify its meaning.

The formula for r begins by standardizing the observations. Suppose, for example, that x is height in centimeters and y is weight in kilograms and that we have height and weight measurements for n people. Then \bar{x} and s_x are the mean and standard deviation of the n heights, both in centimeters. The value

$$\frac{x_i - \bar{x}}{s_x}$$

is the standardized height of the i th person, familiar from Chapter 1. The standardized height says how many standard deviations above or below the mean a person’s height lies. Standardized values have no units—in this example, they are no longer measured in centimeters. Standardize the weights also. The correlation r is an average of the products of the standardized height and the standardized weight for the n people.

Properties of correlation

The formula for correlation helps us see that r is positive when there is a positive association between the variables. Height and weight, for example, have a positive association. People who are above average in height tend to also be above average in weight. Both the standardized height and the standardized weight for such a person are positive. People who are below average in height tend also to have below-average weight. Then both standardized height and standardized weight are negative. In both cases, the products in the formula for r are mostly positive and so r is positive. In the same way, we can see that

r is negative when the association between x and y is negative. More detailed study of the formula gives more detailed properties of r . Here is what you need to know in order to interpret correlation:



- Correlation makes no use of the distinction between explanatory and response variables. It makes no difference which variable you call x and which you call y in calculating the correlation.
- *Correlation requires that both variables be quantitative, so that it makes sense to do the arithmetic indicated by the formula for r .* We cannot calculate a correlation between the incomes of a group of people and what city they live in, because city is a categorical variable.
- Because r uses the standardized values of the observations, r does not change when we change the units of measurement of x , y , or both. Measuring height in inches rather than centimeters and weight in pounds rather than kilograms does not change the correlation between height and weight. The correlation r itself has no unit of measurement; it is just a number.
- Positive r indicates positive association between the variables, and negative r indicates negative association.
- The correlation r is always a number between -1 and 1 . Values of r near 0 indicate a very weak linear relationship. The strength of the relationship increases as r moves away from 0 toward either -1 or 1 . Values of r close to -1 or 1 indicate that the points lie close to a straight line. The extreme values $r = -1$ and $r = 1$ occur only when the points in a scatterplot lie exactly along a straight line.



- Correlation measures the strength of only the linear relationship between two variables. *Correlation does not describe curved relationships between variables, no matter how strong they are.*
- *Like the mean and standard deviation, the correlation is not resistant: r is strongly affected by a few outlying observations.* Use r with caution when outliers appear in the scatterplot.



The scatterplots in Figure 2.10 illustrate how values of r closer to 1 or -1 correspond to stronger linear relationships. To make the essential meaning of r clear, the standard deviations of both variables in these plots are equal and the horizontal and vertical scales are the same. In general, it is not so easy to guess the value of r from the appearance of a scatterplot. Remember that changing the plotting scales in a scatterplot may mislead our eyes, but it does not change the standardized values of the variables and therefore cannot change the correlation. To explore how extreme observations can influence r , use the *Correlation and Regression* applet available on the text CD and Web site.

EXAMPLE

2.10 Scatterplots and correlations. The real data we have examined also illustrate the behavior of correlation.

Figure 2.1 (page 87), despite the clusters, shows a quite strong negative linear association between the percent of a state's high school seniors who take the SAT exam and their mean SAT score. The correlation is $r = -0.877$.

Figure 2.3 (page 90) shows a strong positive linear association between the two measurements of defect depth. The correlation is $r = 0.944$. That the

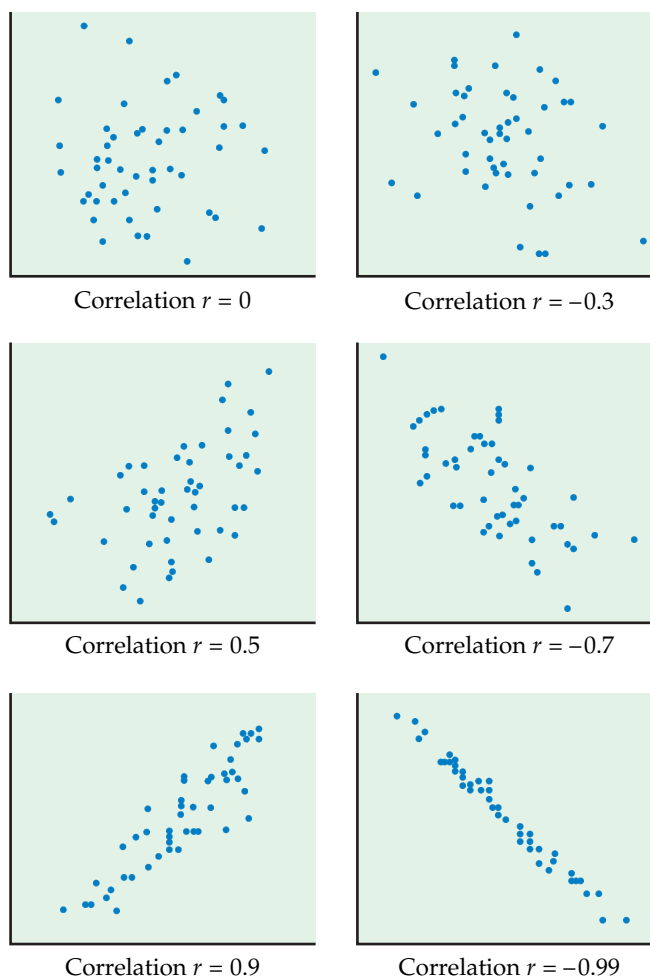


FIGURE 2.10 How the correlation r measures the direction and strength of a linear association.



pattern doesn't follow the $y = x$ line drawn on the graph doesn't matter—correlation measures closeness to whatever line describes the data, not to a line that we specify in advance.

Figure 2.7 (page 96) shows a very weak relationship between returns on Treasury bills and on common stocks. We expect a small negative r , and calculation gives $r = -0.113$.

The correlation between time and acceleration for the motorcycle crash data graphed in Figure 2.5 (page 93) is $r = 0.296$. Because the relationship is not at all linear, r provides no useful information. *Always plot your data before calculating common statistical measures such as correlation.*

Finally, remember that **correlation is not a complete description of two-variable data**, even when the relationship between the variables is linear. You should give the means and standard deviations of both x and y along with the correlation. (Because the formula for correlation uses the means and standard deviations, these measures are the proper choices to accompany a correlation.) Conclusions based on correlations alone may require rethinking in the light of a more complete description of the data.



EXAMPLE

2.11 Scoring of figure skating in the Olympics. Until a scandal at the 2002 Olympics brought change, figure skating was scored by judges on a scale from 0.0 to 6.0. The scores were often controversial. We have the scores awarded by two judges, Pierre and Elena, to many skaters. How well do they agree? We calculate that the correlation between their scores is $r = 0.9$. But the mean of Pierre's scores is 0.8 point lower than Elena's mean.

These facts do not contradict each other. They are simply different kinds of information. The mean scores show that Pierre awards lower scores than Elena. But because Pierre gives *every* skater a score about 0.8 point lower than Elena, the correlation remains high. Adding the same number to all values of either x or y does not change the correlation. If both judges score the same skaters, the competition is scored consistently because Pierre and Elena agree on which performances are better than others. The high r shows their agreement. But if Pierre scores some skaters and Elena others, we must add 0.8 points to Pierre's scores to arrive at a fair comparison.

SECTION 2.2 Summary

The **correlation** r measures the direction and strength of the linear (straight line) association between two quantitative variables x and y . Although you can calculate a correlation for any scatterplot, r measures only linear relationships.

Correlation indicates the direction of a linear relationship by its sign: $r > 0$ for a positive association and $r < 0$ for a negative association.

Correlation always satisfies $-1 \leq r \leq 1$ and indicates the strength of a relationship by how close it is to -1 or 1 . Perfect correlation, $r = \pm 1$, occurs only when the points lie exactly on a straight line.

Correlation ignores the distinction between explanatory and response variables. The value of r is not affected by changes in the unit of measurement of either variable. Correlation is not resistant, so outliers can greatly change the value of r .

SECTION 2.2 Exercises

2.29 Coffee prices and deforestation. Coffee is a leading export from several developing countries. When coffee prices are high, farmers often clear forest to plant more coffee trees. Here are data for five years on prices paid to coffee growers in Indonesia and the rate of deforestation in a national park that lies in a coffee-producing region:²⁰

Price (cents per pound)	Deforestation (percent)
29	0.49
40	1.59
54	1.69
55	1.82
72	3.10

- Make a scatterplot. Which is the explanatory variable? What kind of pattern does your plot show?
- Find the correlation r step-by-step. That is, find the mean and standard deviation of the two variables. Then find the five standardized values for each variable and use the formula for r . Explain how your value for r matches your graph in (a).
- Now enter these data into your calculator or software and use the correlation function to find r . Check that you get the same result as in (b).

2.30 First test and final exam. In Exercise 2.6 you looked at the relationship between the score on the first test and the score on the final exam in an elementary statistics course. The data for eight students from such a course are presented in the following table.

First-test score	153	144	162	149	127	118	158	153
Final-exam score	145	140	145	170	145	175	170	160

- (a) Find the correlation between these two variables.
- (b) In Exercise 2.6 we noted that the relationship between these two variables is weak. Does your calculation of the correlation support this statement? Explain your answer.

2.31 Second test and final exam. Refer to the previous exercise. Here are the data for the second test and the final exam for the same students:

Second-test score	158	162	144	162	136	158	175	153
Final-exam score	145	140	145	170	145	175	170	160

- (a) Find the correlation between these two variables.
- (b) In Exercise 2.7 we noted that the relationship between these two variables is stronger than the relationship between the two variables in the previous exercise. How do the values of the correlations that you calculated support this statement? Explain your answer.

2.32 The effect of an outlier. Refer to the previous exercise. Add a ninth student whose scores on the second test and final exam would lead you to classify the additional data point as an outlier. Recalculate the correlation with this additional case and summarize the effect it has on the value of the correlation.

2.33 The effect of a different point. Examine the data in Exercise 2.31 and add a ninth student who has low scores on the second test and the final exam, and fits the overall pattern of the other scores in the data set. Calculate the correlation and compare it with the correlation that you calculated in Exercise 2.31. Write a short summary of your findings.

2.34 Perch and bass. Figure 2.4 (page 92) displays the positive association between number of prey (perch) present in an area and the proportion eaten by predators (bass).

- (a) Do you think the correlation between these variables is closest to $r = 0.1$, $r = 0.6$, or $r = 0.9$? Explain the reason for your guess.
- (b) Calculate the correlation. Was your guess correct?

2.35 IQ and reading scores. Figure 2.6 (page 96) displays the positive association between the IQ scores of fifth-grade students and their reading scores. Do you think the correlation between these variables is

closest to $r = 0.1$, $r = 0.6$, or $r = 0.9$? Explain the reason for your guess.

2.36



Mutual funds. Mutual fund reports often give correlations to describe how the prices of different investments are related. You look at the correlations between three Fidelity funds and the Standard & Poor's 500 stock index, which describes stocks of large U.S. companies. The three funds are Dividend Growth (stocks of large U.S. companies), Small Cap Stock (stocks of small U.S. companies), and Emerging Markets (stocks in developing countries). For 2003, the three correlations are $r = 0.35$, $r = 0.81$, and $r = 0.98$.²¹

- (a) Which correlation goes with each fund? Explain your answer.
- (b) The correlations of the three funds with the index are all positive. Does this tell you that stocks went up in 2003? Explain your answer.

2.37 Coffee prices in dollars or euros. Coffee is currently priced in dollars. If it were priced in euros, and the dollar prices in Exercise 2.29 were translated into the equivalent prices in euros, would the correlation between coffee price and percent deforestation change? Explain your answer.

2.38 Mutual funds. Exercise 2.28 (page 101) gives data on the returns from 23 Fidelity “sector funds” in 2002 (a down-year for stocks) and 2003 (an up-year).

- (a) Make a scatterplot if you did not do so in Exercise 2.28. Fidelity Gold Fund, the only fund with a positive return in both years, is an extreme outlier.
- (b) To demonstrate that correlation is not resistant, find r for all 23 funds and then find r for the 22 funds other than Gold. Explain from Gold's position in your plot why omitting this point makes r more negative.

2.39 NBA teams. Table 2.1 (page 98) gives the values of the 29 teams in the National Basketball Association, along with their total revenues and operating incomes. You made scatterplots of value against both explanatory variables in Exercise 2.19.

- (a) Find the correlations of team value with revenue and with operating income. Do you think that the two values of r provide a good first comparison of what the plots show about predicting value?
- (b) Portland is an outlier in the plot of value against income. How does r change when you remove Portland? Explain from the position of this point why the change has the direction it does.

2.40 Correlations measure strong and weak linear associations. Your scatterplots for Exercises 2.18 (page 97) and 2.24 (Table 2.4, page 100) illustrate a quite weak linear association and a very strong linear association. Find the correlations that go with these plots. It isn't surprising that a laboratory experiment on physical behavior (the icicles) gives a much stronger correlation than field data on living things (the biological clock). How strong a correlation must be to interest scientists depends on the field of study.

2.41 Heights of people who date each other. A student wonders if tall women tend to date taller men than do short women. She measures herself, her dormitory roommate, and the women in the adjoining rooms; then she measures the next man each woman dates. Here are the data (heights in inches):



Women (x)	66	64	66	65	70	65
Men (y)	72	68	70	68	71	65

- Make a scatterplot of these data. Based on the scatterplot, do you expect the correlation to be positive or negative? Near ± 1 or not?
- Find the correlation r between the heights of the men and women.
- How would r change if all the men were 6 inches shorter than the heights given in the table? Does the correlation tell us whether women tend to date men taller than themselves?
- If heights were measured in centimeters rather than inches, how would the correlation change? (There are 2.54 centimeters in an inch.)
- If every woman dated a man exactly 3 inches taller than herself, what would be the correlation between male and female heights?

2.42 An interesting set of data. Make a scatterplot of the following data.

x	1	2	3	4	10	10
y	1	3	3	5	1	11

Use your calculator to show that the correlation is about 0.5. What feature of the data is responsible for reducing the correlation to this value despite a strong straight-line association between x and y in most of the observations?

2.43   **Use the applet.** You are going to use the *Correlation and Regression*



applet to make different scatterplots with 10 points that have correlation close to 0.8. *Many patterns can have the same correlation. Always plot your data before you trust a correlation.*

- Stop after adding the first 2 points. What is the value of the correlation? Why does it have this value no matter where the 2 points are located?
- Make a lower-left to upper-right pattern of 10 points with correlation about $r = 0.8$. (You can drag points up or down to adjust r after you have 10 points.) Make a rough sketch of your scatterplot.
- Make another scatterplot, this time with 9 points in a vertical stack at the left of the plot. Add one point far to the right and move it until the correlation is close to 0.8. Make a rough sketch of your scatterplot.
- Make yet another scatterplot, this time with 10 points in a curved pattern that starts at the lower left, rises to the right, then falls again at the far right. Adjust the points up or down until you have a quite smooth curve with correlation close to 0.8. Make a rough sketch of this scatterplot also.

2.44 Gas mileage and speed. Exercise 2.22 (page 99) gives data on gas mileage against speed for a small car. Make a scatterplot if you have not already done so, then find the correlation r . Explain why r is close to zero despite a strong relationship between speed and gas used.

2.45 City and highway gas mileage. Table 1.10 (page 31) gives the city and highway gas mileages for 21 two-seater cars, including the Honda Insight gas-electric hybrid car.

- Make a scatterplot of highway mileage y against city mileage x for all 21 cars. There is a strong positive linear association. The Insight lies far from the other points. Does the Insight extend the linear pattern of the other cars, or is it far from the line they form?
- Find the correlation between city and highway mileages both without and with the Insight. Based on your answer to (a), explain why r changes in this direction when you add the Insight.

2.46   **Use the applet.** Go to the *Correlation and Regression* applet. Click on the scatterplot to create a group of 10 points in the lower-left corner of the scatterplot with a strong straight-line negative pattern (correlation about -0.9).

- Add one point at the upper left that is in line with the first 10. How does the correlation change?

(b) Drag this last point down until it is opposite the group of 10 points. How small can you make the correlation? Can you make the correlation positive? *A single outlier can greatly strengthen or weaken a correlation. Always plot your data to check for outlying points.*

2.47 What is the correlation? Suppose that women always married men 2 years older than themselves. Draw a scatterplot of the ages of 5 married couples, with the wife's age as the explanatory variable. What is the correlation r for your data? Why?

2.48 CHALLENGE High correlation does not mean that the values are the same. Investment reports often include correlations. Following a table of correlations among mutual funds, a report adds, "Two funds can have perfect correlation, yet different levels of risk. For example, Fund A and Fund B may be perfectly correlated, yet Fund A moves 20% whenever Fund B moves 10%." Write a brief explanation, for someone who knows no statistics, of how this can happen. Include a sketch to illustrate your explanation.

2.49 Student ratings of teachers. A college newspaper interviews a psychologist about student ratings of the teaching of faculty members. The psychologist says, "The evidence indicates that the correlation between the research productivity and teaching rating of faculty members is close to zero." The paper reports this as "Professor McDaniel said that good researchers tend to be poor teachers, and vice versa." Explain why the paper's report is wrong. Write a statement in plain language (don't use the word "correlation") to explain the psychologist's meaning.

2.50 What's wrong? Each of the following statements contains a blunder. Explain in each case what is wrong.

(a) "There is a high correlation between the gender of American workers and their income."

(b) "We found a high correlation ($r = 1.09$) between students' ratings of faculty teaching and ratings made by other faculty members."

(c) "The correlation between planting rate and yield of corn was found to be $r = 0.23$ bushel."

2.51 CHALLENGE IQ and GPA. Table 1.9 (page 29) reports data on 78 seventh-grade students. We expect a positive association between IQ and GPA. Moreover, some people think that self-concept is related to school performance. Examine in detail the relationships between GPA and the two explanatory variables IQ and self-concept. Are the relationships roughly linear? How strong are they? Are there unusual points? What is the effect of removing these points?

2.52 CHALLENGE Effect of a change in units. Consider again the correlation r between the speed of a car and its gas consumption, from the data in Exercise 2.22 (page 99).

(a) Transform the data so that speed is measured in miles per hour and fuel consumption in gallons per mile. (There are 1.609 kilometers in a mile and 3.785 liters in a gallon.) Make a scatterplot and find the correlation for both the original and the transformed data. How did the change of units affect your results?

(b) Now express fuel consumption in miles per gallon. (So each value is $1/x$ if x is gallons per mile.) Again make a scatterplot and find the correlation. How did this change of units affect your results?

(Lesson: The effects of a linear transformation of the form $x_{\text{new}} = a + bx$ are simple. The effects of a nonlinear transformation are more complex.)

2.3 Least-Squares Regression

Correlation measures the direction and strength of the linear (straight-line) relationship between two quantitative variables. If a scatterplot shows a linear relationship, we would like to summarize this overall pattern by drawing a line on the scatterplot. A *regression line* summarizes the relationship between two variables, but only in a specific setting: when 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 . Regression, unlike correlation, requires that we have an explanatory variable and a response variable.

EXAMPLE

2.12 Fidgeting and fat gain. Does fidgeting keep you slim? Some people don't gain weight even when they overeat. Perhaps fidgeting and other “nonexercise activity” (NEA) explains why—the body might spontaneously increase nonexercise activity when fed more. Researchers deliberately overfed 16 healthy young adults for 8 weeks. They measured fat gain (in kilograms) and, as an explanatory variable, increase in energy use (in calories) from activity other than deliberate exercise—fidgeting, daily living, and the like. Here are the data:²²

NEA increase (cal)	−94	−57	−29	135	143	151	245	355
Fat gain (kg)	4.2	3.0	3.7	2.7	3.2	3.6	2.4	1.3
NEA increase (cal)	392	473	486	535	571	580	620	690
Fat gain (kg)	3.8	1.7	1.6	2.2	1.0	0.4	2.3	1.1

Figure 2.11 is a scatterplot of these data. The plot shows a moderately strong negative linear association with no outliers. The correlation is $r = -0.7786$. People with larger increases in nonexercise activity do indeed gain less fat. A line drawn through the points will describe the overall pattern well.

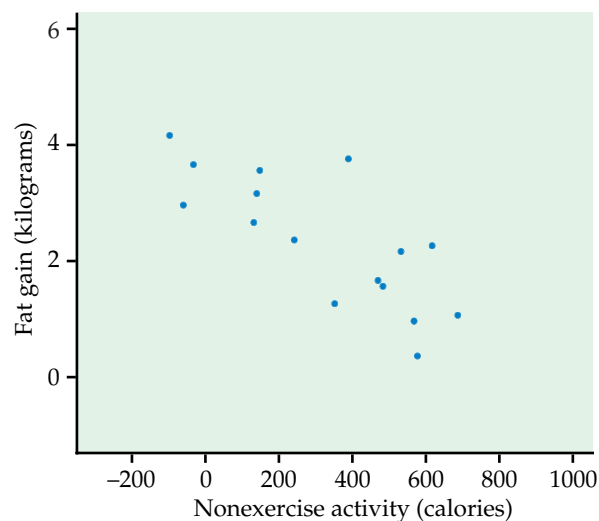


FIGURE 2.11 Fat gain after 8 weeks of overeating, plotted against the increase in nonexercise activity over the same period, for Example 2.12.

Fitting a line to data

fitting a line

When a scatterplot displays a linear pattern, we can describe the overall pattern by drawing a straight line through the points. Of course, no straight line passes exactly through all of the points. **Fitting a line** to data means drawing a line that comes as close as possible to the points. The equation of a line fitted to the data gives a compact description of the dependence of the response variable y on the explanatory variable x .

STRAIGHT LINES

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

$$y = b_0 + b_1x$$

In this equation, b_1 is the **slope**, the amount by which y changes when x increases by one unit. The number b_0 is the **intercept**, the value of y when $x = 0$.

EXAMPLE

2.13 Regression line for fat gain. Any straight line describing the nonexercise activity data has the form

$$\text{fat gain} = b_0 + (b_1 \times \text{NEA increase})$$

In Figure 2.12 we have drawn the regression line with the equation

$$\text{fat gain} = 3.505 - (0.00344 \times \text{NEA increase})$$

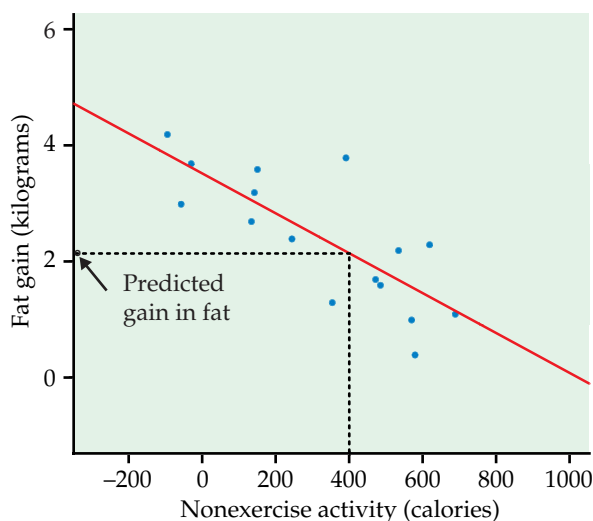


FIGURE 2.12 A regression line fitted to the nonexercise activity data and used to predict fat gain for an NEA increase of 400 calories.

The figure shows that this line fits the data well. The slope $b_1 = -0.00344$ tells us that fat gained goes down by 0.00344 kilogram for each added calorie of NEA.

The slope b_1 of a line $y = b_0 + b_1x$ is the *rate of change* in the response y as the explanatory variable x changes. The slope of a regression line is an important numerical description of the relationship between the two variables. For Example 2.13, the intercept, $b_0 = 3.505$ kilograms, is the estimated fat gain if NEA does not change when a person overeats.

USE YOUR KNOWLEDGE

2.53 Plot the data with the line. Make a sketch of the data in Example 2.12 and plot the line

$$\text{fat gain} = 4.505 - (0.00344 \times \text{NEA increase})$$

on your sketch. Explain why this line does not give a good fit to the data.

Prediction

prediction We can use a regression line to **predict** the response y for a specific value of the explanatory variable x .

EXAMPLE

2.14 Prediction for fat gain. Based on the linear pattern, we want to predict the fat gain for an individual whose NEA increases by 400 calories when she overeats. To use the fitted line to predict fat gain, go “up and over” on the graph in Figure 2.12. From 400 calories on the x axis, go up to the fitted line and over to the y axis. The graph shows that the predicted gain in fat is a bit more than 2 kilograms.

If we have the equation of the line, it is faster and more accurate to substitute $x = 400$ in the equation. The predicted fat gain is

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

The accuracy of predictions from a regression line depends on how much scatter about the line the data show. In Figure 2.12, fat gains for similar increases in NEA show a spread of 1 or 2 kilograms. The regression line summarizes the pattern but gives only roughly accurate predictions.

USE YOUR KNOWLEDGE

2.54 Predict the fat gain. Use the regression equation in Example 2.13 to predict the fat gain for a person whose NEA increases by 600 calories.

EXAMPLE

2.15 Is this prediction reasonable? Can we predict the fat gain for someone whose nonexercise activity increases by 1500 calories when she overeats? We can certainly substitute 1500 calories into the equation of the line. The prediction is

$$\text{fat gain} = 3.505 - (0.00344 \times 1500) = -1.66 \text{ kilograms}$$

That is, we predict that this individual loses fat when she overeats. This prediction is not trustworthy. Look again at Figure 2.12. An NEA increase of 1500 calories is far outside the range of our data. We can't say whether increases this large ever occur, or whether the relationship remains linear at such extreme values. Predicting fat gain when NEA increases by 1500 calories *extrapolates* the relationship beyond what the data show.

EXTRAPOLATION

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

USE YOUR KNOWLEDGE

2.55 Would you use the regression equation to predict? Consider the following values for NEA increase: -400 , 200 , 500 , 1000 . For each, decide whether you would use the regression equation in Example 2.13 to predict fat gain or whether you would be concerned that the prediction would not be trustworthy because of extrapolation. Give reasons for your answers.

Least-squares regression

Different people might draw different lines by eye on a scatterplot. This is especially true when the points are widely scattered. We need a way to draw a regression line that doesn't depend on our guess as to where the line should go. No line will pass exactly through all the points, but we want one that is as close as possible. We will use the line to predict y from x , so we want a line that is as close as possible to the points in the *vertical* direction. That's because the prediction errors we make are errors in y , which is the vertical direction in the scatterplot.

The line in Figure 2.12 predicts 2.13 kilograms of fat gain for an increase in nonexercise activity of 400 calories. If the actual fat gain turns out to be 2.3 kilograms, the error is

$$\begin{aligned} \text{error} &= \text{observed gain} - \text{predicted gain} \\ &= 2.3 - 2.13 = 0.17 \text{ kilograms} \end{aligned}$$

Errors are positive if the observed response lies above the line, and negative if the response lies below the line. We want a regression line that makes these prediction errors as small as possible. Figure 2.13 illustrates the idea. For clarity, the plot shows only three of the points from Figure 2.12, along with the line, on an expanded scale. The line passes below two of the points and above one of them. The vertical distances of the data points from the line appear as vertical line segments. A “good” regression line makes these distances as small as possible. There are many ways to make “as small as possible” precise. The most common is the *least-squares* idea. The line in Figures 2.12 and 2.13 is in fact the least-squares regression line.

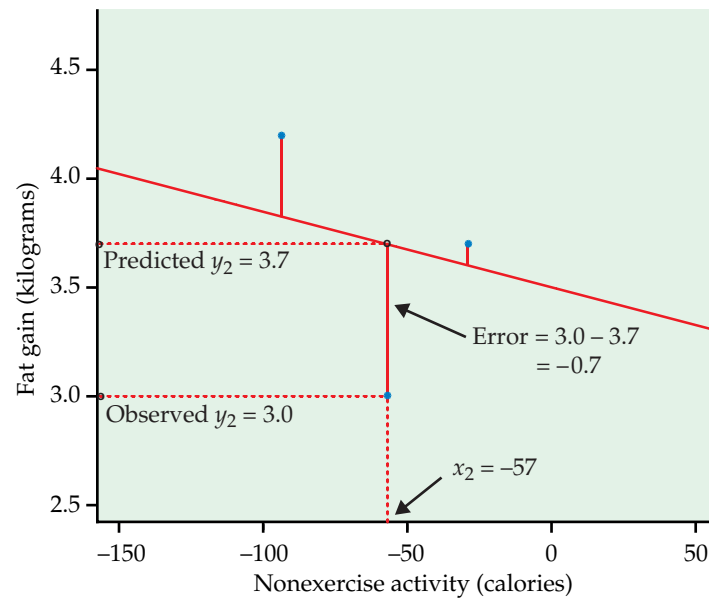


FIGURE 2.13 The least-squares idea: make the errors in predicting y as small as possible by minimizing the sum of their squares.

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.

Here is the least-squares idea expressed as a mathematical problem. We represent n observations on two variables x and y as

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

If we draw a line $y = b_0 + b_1x$ through the scatterplot of these observations, the line predicts the value of y corresponding to x_i as $\hat{y}_i = b_0 + b_1x_i$. We write \hat{y} (read “y-hat”) in the equation of a regression line to emphasize that the line gives a *predicted* response \hat{y} for any x . The predicted response will usually not be exactly the same as the actually *observed* response y . The method of least squares chooses the line that makes the sum of the squares of these errors as small as possible. To find this line, we must find the values of the intercept b_0

and the slope b_1 that minimize

$$\sum (\text{error})^2 = \sum (y_i - b_0 - b_1 x_i)^2$$

for the given observations x_i and y_i . For the NEA data, for example, we must find the b_0 and b_1 that minimize

$$(-94 - b_0 - 4.2b_1)^2 + (-57 - b_0 - 3.0b_1)^2 + \cdots + (690 - b_0 - 1.1b_1)^2$$

These values are the intercept and slope of the least-squares line.

You will use software or a calculator with a regression function to find the equation of the least-squares regression line from data on x and y . We will therefore give the equation of the least-squares line in a form that helps our understanding but is not efficient for calculation.

EQUATION OF THE LEAST-SQUARES REGRESSION LINE

We have data on an explanatory variable x and a response variable y for n individuals. The means and standard deviations of the sample data are \bar{x} and s_x for x and \bar{y} and s_y for y , and the correlation between x and y is r . The equation of the least-squares regression line of y on x is

$$\hat{y} = b_0 + b_1 x$$

with **slope**

$$b_1 = r \frac{s_y}{s_x}$$

and **intercept**

$$b_0 = \bar{y} - b_1 \bar{x}$$

EXAMPLE

2.16 Check the calculations. Verify from the data in Example 2.12 that the mean and standard deviation of the 16 increases in NEA are

$$\bar{x} = 324.8 \text{ calories} \quad \text{and} \quad s_x = 257.66 \text{ calories}$$

The mean and standard deviation of the 16 fat gains are

$$\bar{y} = 2.388 \text{ kg} \quad \text{and} \quad s_y = 1.1389 \text{ kg}$$

The correlation between fat gain and NEA increase is $r = -0.7786$. The least-squares regression line of fat gain y on NEA increase x therefore has slope

$$\begin{aligned} b_1 &= r \frac{s_y}{s_x} = -0.7786 \frac{1.1389}{257.66} \\ &= -0.00344 \text{ kg per calorie} \end{aligned}$$

and intercept

$$\begin{aligned} b_0 &= \bar{y} - b_1 \bar{x} = 2.388 - (-0.00344)(324.8) \\ &= 3.505 \text{ kg} \end{aligned}$$

The equation of the least-squares line is

$$\hat{y} = 3.505 - 0.00344x$$



When doing calculations like this by hand, you may need to carry extra decimal places in the preliminary calculations to get accurate values of the slope and intercept. Using software or a calculator with a regression function eliminates this worry.

Interpreting the regression line

The slope $b_1 = -0.00344$ kilograms per calorie in Example 2.16 is the change in fat gain as NEA increases. The units “kilograms of fat gained per calorie of NEA” come from the units of y (kilograms) and x (calories). Although the correlation does not change when we change the units of measurement, the equation of the least-squares line does change. The slope in grams per calorie would be 1000 times as large as the slope in kilograms per calorie, because there are 1000 grams in a kilogram. The small value of the slope, $b_1 = -0.00344$, does not mean that the effect of increased NEA on fat gain is small—it just reflects the choice of kilograms as the unit for fat gain. *The slope and intercept of the least-squares line depend on the units of measurement—you can’t conclude anything from their size.*



The expression $b_1 = rs_y/s_x$ for the slope says that, along the regression line, **a change of one standard deviation in x corresponds to a change of r standard deviations in y .** When the variables are perfectly correlated ($r = 1$ or $r = -1$), the change in the predicted response \hat{y} is the same (in standard deviation units) as the change in x . Otherwise, when $-1 < r < 1$, the change in \hat{y} is less than the change in x . As the correlation grows less strong, the prediction \hat{y} moves less in response to changes in x .

The least-squares regression line always passes through the point (\bar{x}, \bar{y}) on the graph of y against x . Check that when you substitute $\bar{x} = 324.8$ into the equation of the regression line in Example 2.16, the result is $\hat{y} = 2.388$, equal to the mean of y . So the least-squares regression line of y on x is the line with slope rs_y/s_x that passes through the point (\bar{x}, \bar{y}) . We can describe regression entirely in terms of the basic descriptive measures \bar{x} , s_x , \bar{y} , s_y , and r . If both x and y are standardized variables, so that their means are 0 and their standard deviations are 1, then the regression line has slope r and passes through the origin.

Figure 2.14 displays the basic regression output for the nonexercise activity data from two statistical software packages. Other software produces very similar output. You can find the slope and intercept of the least-squares line, calculated to more decimal places than we need, in both outputs. The software also provides information that we do not yet need, including some that we trimmed from Figure 2.14. Part of the art of using software is to ignore the extra information that is almost always present. Look for the results that you need. Once you understand a statistical method, you can read output from almost any software.

Correlation and regression

Least-squares regression looks at the distances of the data points from the line only in the y direction. So the two variables x and y play different roles in regression.

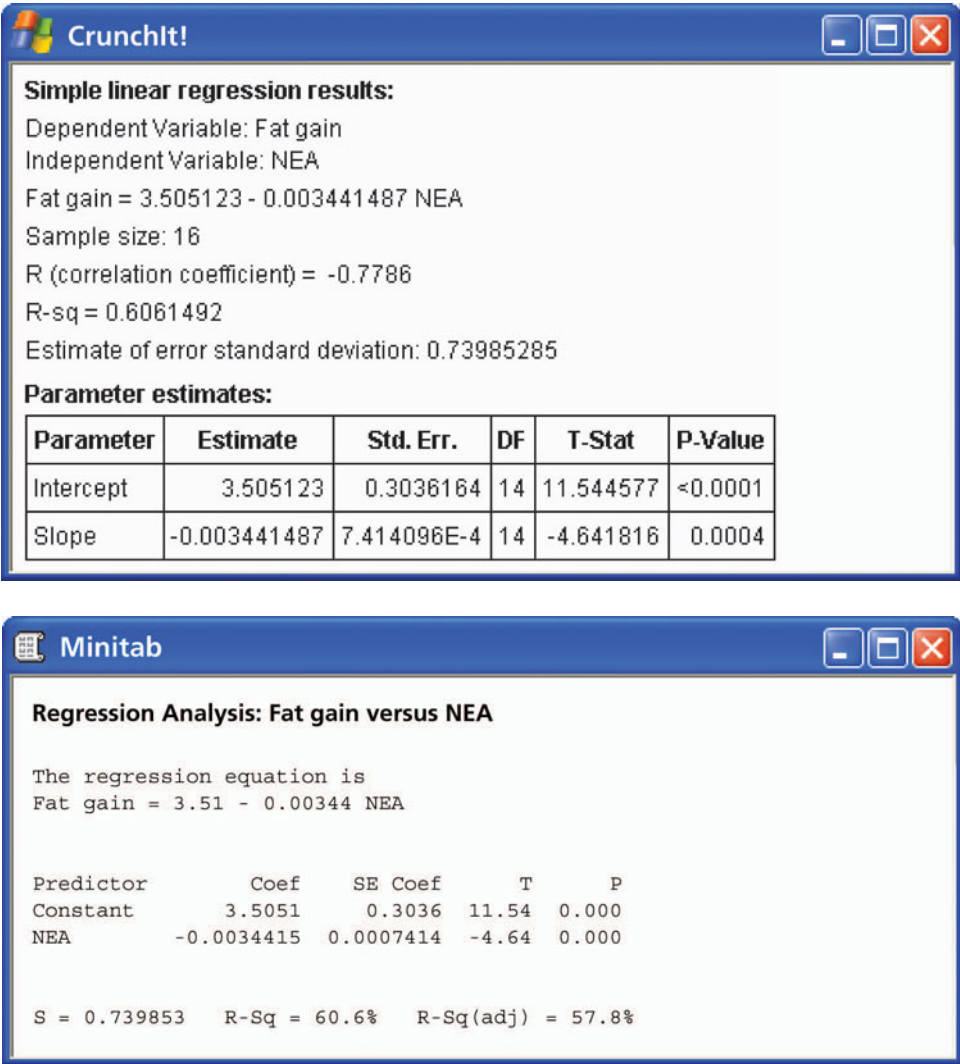


FIGURE 2.14 Regression results for the nonexercise activity data from two statistical software packages. Other software produces similar output.

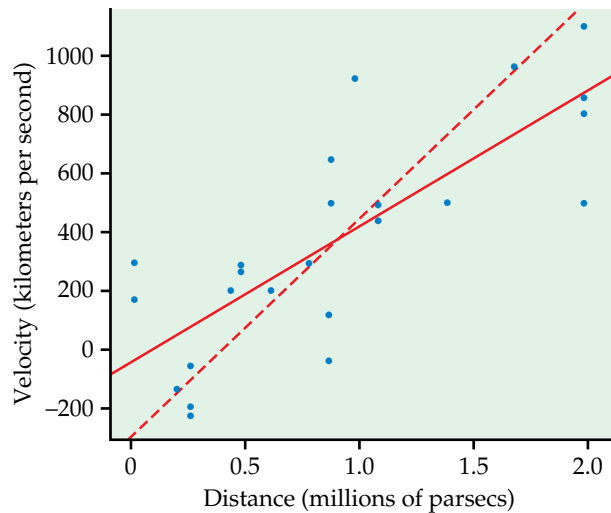


EXAMPLE

2.17 The universe is expanding. Figure 2.15 is a scatterplot of data that played a central role in the discovery that the universe is expanding. They are the distances from the earth of 24 spiral galaxies and the speed at which these galaxies are moving away from us, reported by the astronomer Edwin Hubble in 1929.²³ There is a positive linear relationship, $r = 0.7842$. More distant galaxies are moving away more rapidly. Astronomers believe that there is in fact a perfect linear relationship, and that the scatter is caused by imperfect measurements.

The two lines on the plot are the two least-squares regression lines. The regression line of velocity on distance is solid. The regression line of distance on velocity is dashed. *Although there is only one correlation between velocity and distance, regression of velocity on distance and regression of distance on velocity give different lines. In doing regression, you must choose which variable is explanatory.*

FIGURE 2.15 Hubble's data on the velocity and distance of 24 galaxies, for Example 2.17. The lines are the least-squares regression lines of velocity on distance (solid) and of distance on velocity (dashed).



Even though the correlation r ignores the distinction between explanatory and response variables, there is a close connection between correlation and regression. We saw that the slope of the least-squares line involves r . Another connection between correlation and regression is even more important. In fact, the numerical value of r as a measure of the strength of a linear relationship is best interpreted by thinking about regression. Here is the fact we need.

r^2 IN REGRESSION

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

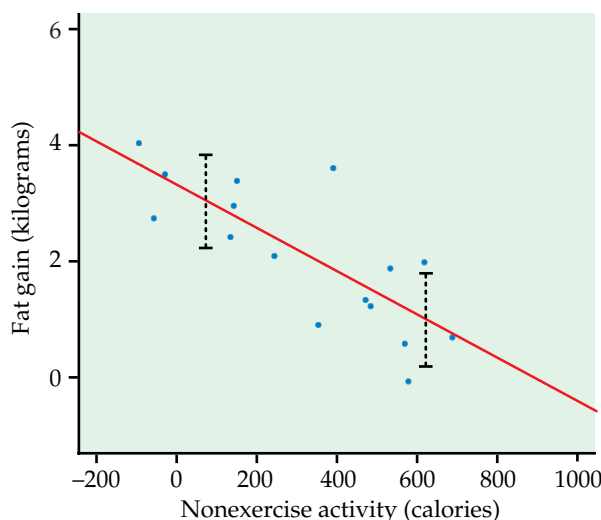
The correlation between NEA increase and fat gain for the 16 subjects in Example 2.12 is $r = -0.7786$. Because $r^2 = 0.606$, the straight-line relationship between NEA and fat gain explains about 61% of the vertical scatter in fat gains in Figure 2.12. When you report a regression, give r^2 as a measure of how successfully the regression explains the response. Both software outputs in Figure 2.14 include r^2 , either in decimal form or as a percent. When you see a correlation, square it to get a better feel for the strength of the association. Perfect correlation ($r = -1$ or $r = 1$) means the points lie exactly on a line. Then $r^2 = 1$ and all of the variation in one variable is accounted for by the linear relationship with the other variable. If $r = -0.7$ or $r = 0.7$, $r^2 = 0.49$ and about half the variation is accounted for by the linear relationship. In the r^2 scale, correlation ± 0.7 is about halfway between 0 and ± 1 .

USE YOUR KNOWLEDGE

2.56 What fraction of the variation is explained? Consider the following correlations: -0.9 , -0.5 , -0.3 , 0 , 0.3 , 0.5 , and 0.9 . For each, give the fraction of the variation in y that is explained by the least-squares regression of y on x . Summarize what you have found from performing these calculations.

The use of r^2 to describe the success of regression in explaining the response y is very common. It rests on the fact that there are two sources of variation in the responses y in a regression setting. Figure 2.16 gives a rough visual picture of the two sources. The first reason for the variation in fat gains is that there is a relationship between fat gain y and increase in NEA x . As x increases from -94 calories to 690 calories among the 16 subjects, it pulls fat gain y with it along the regression line in the figure. The linear relationship explains this part of the variation in fat gains.

FIGURE 2.16 Explained and unexplained variation in regression. As x increases, it pulls y with it along the line. That is the variation explained by the regression. The scatter of the data points above and below the line, suggested by the dashed segments, is not explained by the regression.



The fat gains do not lie exactly on the line, however, but are scattered above and below it. This is the second source of variation in y , and the regression line tells us nothing about how large it is. The vertical dashed lines in Figure 2.16 show a rough average for the spread in y when we fix a value of x . We use r^2 to measure variation along the line as a fraction of the total variation in the fat gains. In Figure 2.16, about 61% of the variation in fat gains among the 16 subjects is due to the straight-line tie between y and x . The remaining 39% is vertical scatter in the observed responses remaining after the line has fixed the predicted responses.

*Understanding r^2

Here is a more specific interpretation of r^2 . The fat gains y in Figure 2.16 range from 0.4 kilograms to 4.2 kilograms. The variance of these responses, a measure of how variable they are, is

$$\text{variance of observed values } y = 1.297$$

Much of this variability is due to the fact that as x increases from -94 calories to 690 calories it pulls height y along with it. If the only variability in the observed responses were due to the straight-line dependence of fat gain on NEA, the observed gains would lie exactly on the regression line. That is, they would be the same as the predicted gains \hat{y} . We can compute the predicted gains by substitut-

*This explanation is optional reading.

ing the NEA values for each subject into the equation of the least-squares line. Their variance describes the variability in the predicted responses. The result is

$$\text{variance of predicted values } \hat{y} = 0.786$$

This is what the variance would be if the responses fell exactly on the line, that is, if the linear relationship explained 100% of the observed variation in y . Because the responses don't fall exactly on the line, the variance of the predicted values is smaller than the variance of the observed values. Here is the fact we need:

$$\begin{aligned} r^2 &= \frac{\text{variance of predicted values } \hat{y}}{\text{variance of observed values } y} \\ &= \frac{0.786}{1.297} = 0.606 \end{aligned}$$

This fact is always true. The squared correlation gives the variance the responses would have if there were no scatter about the least-squares line as a fraction of the variance of the actual responses. This is the exact meaning of “fraction of variation explained” as an interpretation of r^2 .

These connections with correlation are special properties of least-squares regression. They are not true for other methods of fitting a line to data. One reason that least squares is the most common method for fitting a regression line to data is that it has many convenient special properties.

BEYOND THE BASICS

Transforming Relationships

How is the weight of an animal's brain related to the weight of its body? Figure 2.17 is a scatterplot of brain weight against body weight for 96 species of mam-

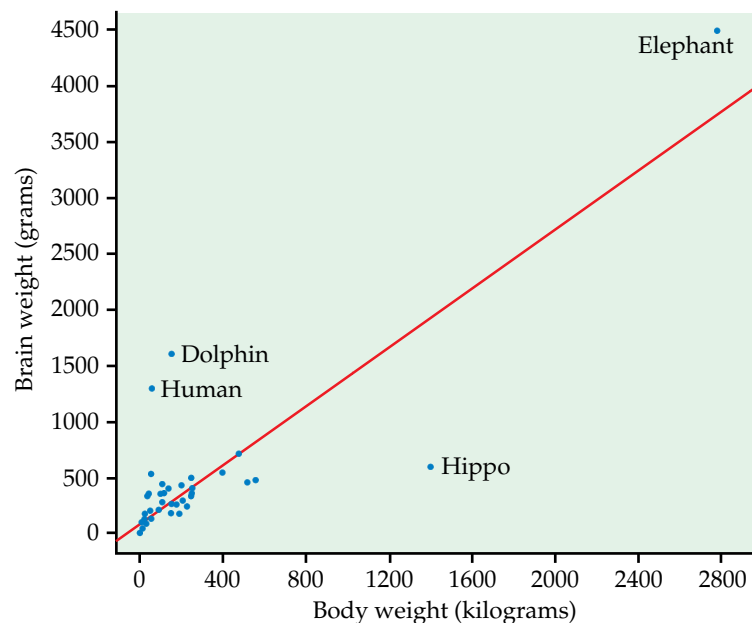


FIGURE 2.17 Scatterplot of brain weight against body weight for 96 species of mammals.