Yellow Dog Productions/Getty Images

CHAPTER 2

Looking at Data—Relationships

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.

In Section 2.2, we focus on graphical descriptions. The scatterplot is our fundamental graphical tool for displaying the relationship between two quantitative variables. Sections 2.3 and 2.4 move on to numerical summaries for these relationships. Cautions about the use of these methods are discussed in Section 2.5. Graphical and numerical methods for describing the relationship between two categorical variables are presented in Section 2.6. We conclude with Section 2.7, a brief overview of issues related to the distinction between associations and causation.

2.1 Relationships

When you complete this section, you will be able to:

- Identify the key characteristics of a data set to be used to explore a relationship between two variables.
- Categorize variables as response variables or explanatory variables.

In Chapter 1 (page 2), we discussed the key characteristics of a data set. Cases are the objects described by a set of data, and a variable is a characteristic of a case. We also learned to categorize variables as categorical or quantitative. For Chapter 2, we focus on data sets that have pairs of variables that we want to study together. Here is an example.

EXAMPLE 2.1

College students cope with stress. Stress is a common problem for college students. Exploring factors that are associated with stress may lead to strategies that will help students to relieve some of the stress that they experience. A recent study found that students who experienced greater stress had less access to resources that would help them to cope with their stress.[1] The two variables involved in the relationship here are perceived stress and resources to cope. The cases are the 97 students who are the subjects for a particular study.

When we study relationships between two variables, it is not sufficient to collect data on the two variables. *A key idea for this chapter is that both variables must be measured on the same cases.*

USE YOUR KNOWLEDGE

2.1 Facebook friends. Do people who have more Facebook friends spend more time on Facebook? In an introductory statistics class of 38 students, there were 32 users of Facebook. Each of these students was asked to report how many Facebook friends they had and the average amount of time that they spent on Facebook per week.

(a) Who are the cases for this study?

(b) What are the variables?

(c) Are the variables quantitative or categorical? Explain your answer.

We use the term *associated* to describe the relationship between two variables, such as stress and access to resources to cope 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 $3.95, $4.45, and $4.95, respectively. There is a clear association between the size of the Mocha Frappuccino and its price.


Anthony Behar/Sipa USA/Sipa via AP Images

ASSOCIATION BETWEEN VARIABLES

Two variables measured on the same cases are **associated** if knowing the values of one of the variables tells you something about the values of the other variable.

AU/DE/PE: this is the only time the ® is used with "Frappucino" —include with all other occurrences for consistency? Yes, I think ok to include for consistency- JA

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. For example, it's likely that some students in Example 2.1 are highly stressed and have a high level of resources to cope. Others experience little stress and have a low level of resources to cope. The association in that example is much weaker than the one in the Mocha Frappuccino example.

Examining relationships

To examine the relationship between two or more variables, we first need to know some basic characteristics of the data.

EXAMPLE 2.3

Stress and resources to cope. Refer to Example 2.1. The study asked 97 first-year college students about their stress (perceived stress) and the availability of resources to deal with stress (resources to cope).[2] Perceived stress is based on responses to 10 questions that are summarized in a single variable. Therefore, we will treat the perceived stress as a quantitative variable. Resources to cope is constructed in a similar way summarizing the responses to 20 questions. We treat resources to cope as a quantitative variable also.

In many situations, we measure a collection of categorical variables and then combine them in a scale that can be viewed as a quantitative variable. The perceived stress and resources to cope are examples. We can also turn the tables in the other direction. Here is an example.

EXAMPLE 2.4

Hemoglobin and anemia. Hemoglobin is a measure of iron in the blood. The units are grams of hemoglobin per deciliter of blood (g/dl). Typical values depend on age and sex. Adult women typically have values between 12 and 16 g/dl.

Anemia is a major problem in developing countries, and many studies have been designed to address the problem. In these studies, computing the mean hemoglobin is not particularly useful. For studies like these, it is more appropriate to use a definition of severe anemia (a hemoglobin of less than 8 g/dl). Thus, for example, researchers can compare the proportions of subjects who are severely anemic for two treatments rather than the difference in the mean hemoglobin levels. In this situation, the categorical variable, severely anemic or not, is much more useful than the quantitative variable, hemoglobin.

When analyzing data to draw conclusions, it is important to carefully consider the best way to summarize the data. *Just because a variable is measured as a quantitative variable, it does not necessarily follow that the best summary is based on the mean (or the median)*. As the previous example illustrates, converting a quantitative variable to a categorical variable is a very useful option to keep in mind.

USE YOUR KNOWLEDGE

2.2 Create a categorical variable from a quantitative variable. Consider the study described in Example 2.3. Suppose that we order the students based on the values of resources to cope from smallest to largest. Then, we define three resource groups: low resources, the first 32 students; medium resources, the next 33 students; and high resources, the remaining 32 students. If we compare the perceived stress of the three resource groups, are we using resource group as a quantitative variable or as a categorical variable? Explain your answer and describe some advantages to using the groups versus the original variable in explaining the results of a study such as this.

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 that 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? In other words, is one of the variables a *response variable* and the other an *explanatory variable*?

RESPONSE VARIABLE, EXPLANATORY VARIABLE

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

EXAMPLE 2.5

Stress and resources to cope. Refer to the study of stress and resources to cope in Example 2.3. Here, the explanatory variable is resources to cope and the response variable is perceived stress.

USE YOUR KNOWLEDGE

2.4 Stress and resources or resources and stress? Consider the scenario described in the previous example. Note that the variable, resources to cope, is constructed by summarizing the responses to 20 questions that include items measuring the skills that the student has developed to reduce stress. Make an argument for treating stress as the explanatory variable and resources to cope as the response variable.

In some studies, it is easy to identify explanatory and response variables. The following example illustrates one situation where this is true: when we actually set values of one variable to see how it affects another variable.

EXAMPLE 2.6

How much calcium do you need? Adolescence is a time when bones are growing very actively. If young people do not have enough calcium, their bones will not grow properly. How much calcium is enough? Research designed to answer this question has been performed for many years at events called "Camp Calcium."[3] At these camps, subjects eat controlled diets that are identical except for the amount of calcium. The amount of calcium retained by the body is the major response variable of interest. Because the amount of calcium consumed is controlled by the researchers, this variable is the explanatory 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 2.7

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.

KEY CHARACTERISTICS OF DATA FOR RELATIONSHIPS

A description of the key characteristics of a data set that will be used to explore a relationship between two variables should include

- **Cases.** Identify the cases and how many there are in the data set.
- **Categorical or quantitative.** Classify each variable as categorical or quantitative.
- **Values.** Identify the possible values for each variable.
- **Explanatory or response.** If appropriate, classify each variable as explanatory or response.
- **Label.** Identify what is used as a label variable if one is present.

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** variable and response variables called **dependent variables**. These terms express

independent variable
dependent variable

mathematical ideas; they are not statistical terms. The concept that underlies this language is that the response *depends* 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.

SECTION 2.1 SUMMARY

- To study relationships between variables, we must measure the variables on the same cases. It is also important to determine the number of cases and to classify each variable as categorical or quantitative.
- Two variables measured on the same cases are **associated** if knowing the values of one variable tells you something about the values of the other variable. 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**.

SECTION 2.1 EXERCISES

For Exercise 2.1, see page 80; for Exercises 2.2 and 2.3, see page 82; and for Exercise 2.4, see page 82.

2.5 High click counts on Twitter. A study was done to identify variables that might produce high click counts on Twitter. You and nine of your friends collect data on all of your tweets for a week. You record the number of click counts, the time of day, the day of the week, the sex of the person posting the tweet, and the length of the tweet.

(a) What are the cases for this study?

(b) Classify each of the variables as categorical or quantitative.

(c) Classify each of the variables as explanatory, response, or neither. Explain your answers.

2.6 Explanatory or response? For each of the following scenarios, classify each of the pair of variables as explanatory or response or neither. Give reasons for your answers.

(a) Whether or not a person likes to sing and whether or not a person likes to dance.

(b) The number of pages in a textbook and the cost of a new copy of the textbook.

(c) The number of alcoholic drinks consumed and the blood alcohol content.

(d) In a study of adolescents, the dose of vitamin D given each day for a year (50, 100, or 200 international units) and the change in total bone mineral content from the beginning of the study to the end of the study.

2.7 Buy and sell prices of used textbooks. Think about a study designed to compare the prices of textbooks for third- and fourth-year college courses in five different majors. For the five majors, you want to examine the relationship between the difference in the price that you pay for a used textbook and the price that the seller gives back to you when you return the textbook. Describe a data set that could be used for this study, and give the key characteristics of the data.

2.8 Protein and fat. Think about a study designed to examine the relationship between protein intake and fat intake in the diets of first-year college students. Describe a data set that could be used for this study, and give the key characteristics of the data.

2.9 Soccer tickets and performance. For the teams in the Big Ten Conference last year, plan a study of the relationship between the average number of tickets sold for home soccer games and the percentage of games won. Give the key characteristics of the data that could be used for your study.

2.2 Scatterplots

When you complete this section, you will be able to:

- Make a scatterplot to examine a relationship between two variables.
- Describe the overall pattern in a scatterplot and any striking deviations from that pattern.
- Use a scatterplot to describe the form, direction, and strength of a relationship.
- Use a scatterplot to identify outliers.
- Identify a linear pattern in a scatterplot.
- Explain the effect of a change of units on a scatterplot.
- Use a log transformation to change a curved relationship into a linear relationship.
- Use different plotting symbols to include information about a categorical variable in a scatterplot.

EXAMPLE 2.8

LAUNDRY

Laundry detergents. Consumers Union provides ratings on a large variety of consumer products. They use sophisticated testing methods as well as surveys of their members to create these ratings. The ratings are published in their magazine, *Consumer Reports*.[4]

© Kristoffer Tripplaar/Alamy

One recent study rated 53 laundry detergents on a scale from 1 to 100. The scale summarizes washing performance under a variety of conditions. Price per load is given in cents.[5] We will examine the relationship between rating and price per load for these laundry detergents. We expect that the higher-priced detergents will tend to have higher ratings.

USE YOUR KNOWLEDGE

LAUNDRY

2.10 Examine the spreadsheet. Examine the spreadsheet that gives the laundry detergent data in the data file LAUNDRY.

(a) How many cases are in the data set?

(b) Describe the labels, variables, and values.

(c) Which columns represent quantitative variables? Which columns give categorical variables.

(d) Is there an explanatory variable? A response variable? Explain your answer.

LAUNDRY

2.11 Use the data set. Using the data set from the previous exercise, create graphical and numerical summaries for the rating and for the price per load.

The most common way to display the relationship between two quantitative variables is a *scatterplot*.

SCATTERPLOT

A **scatterplot** shows the relationship between two quantitative variables measured on the same cases. The values of one variable appear on the horizontal axis, and the values of the other variable appear on the vertical axis. Each case in the data appears as the point in the plot determined by the values of both variables for that case.

EXAMPLE 2.9

LAUNDRY

Laundry detergents. A higher price for a product should be associated with a better product. Therefore, let's treat price per load as the explanatory variable and rating as the response variable in our examination of the relationship between these two variables. We begin with a graphical display.

Figure 2.1 gives a scatterplot that displays the relationship between the response variable, rating, and the explanatory variable, price per load. The most striking feature that we see in the plot is a case that appears to be very different from the others. One of the laundry detergents has a rating that is about average (51), but the price per load (56 cents) is almost double that of the other products.

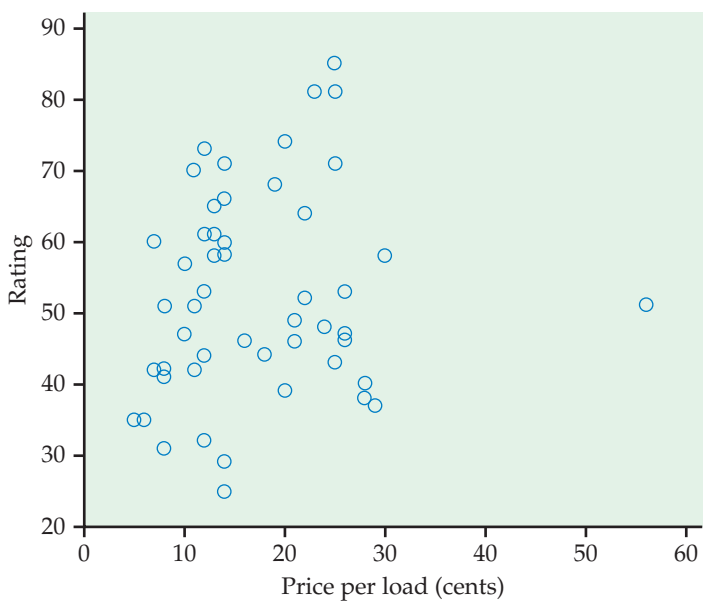


FIGURE 2.1 Scatterplot of price per load (in cents) versus rating for 53 laundry detergents, Example 2.9.

Cases that fall well outside the general pattern of the relationship are called outliers. We provide a more detailed description of these in Section 2.5. For now, we remove this case and focus on the relationship of the remaining data.

Figure 2.2 gives the scatterplot with the outlier removed. The relationship is weak. Paying a high price for your laundry detergent will not guarantee that you have selected a highly rated product.

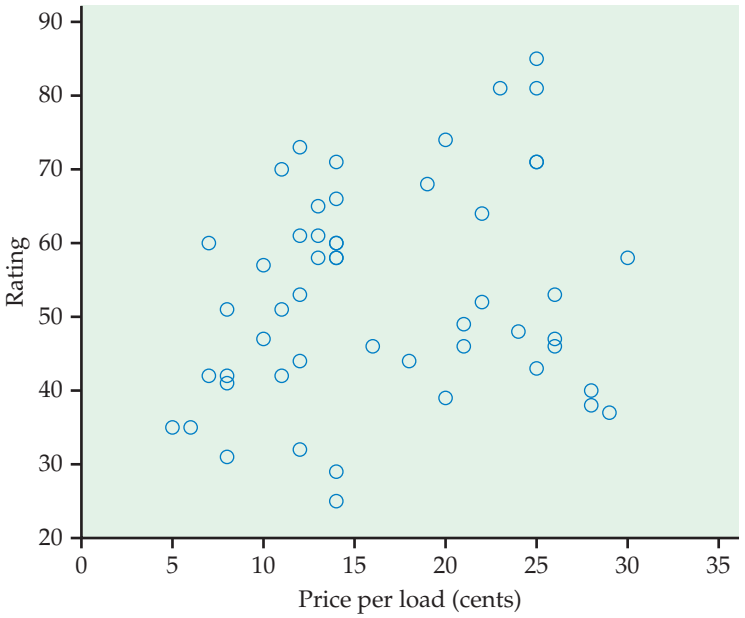


FIGURE 2.2 Scatterplot of price per load (in cents) versus rating for 52 laundry detergents (with the outlier removed), Example 2.9.

Always plot the explanatory variable, if there is one, on the horizontal axis (the x axis) of a scatterplot. 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. Time plots such as the one in Figure 1.12 (page 22) are special scatterplots where the explanatory variable x is a measure of time.

USE YOUR KNOWLEDGE

LAUND

2.12 Make a scatterplot. Let's consider the laundry data with the outlier removed.

(a) Make a scatterplot similar to Figure 2.2.

(b) Two of the laundry detergents cost 14 cents per load with a rating of 60. Mark the location of these items on your plot.

(c) Cases with identical values for both variables are generally indistinguishable in a scatterplot. To what extent do you think that this could give a distorted picture of the relationship between two variables for a data set that has a large number of duplicate values? Explain your answer.

jitter

(d) An option called **jitter** ~~jitter~~ is available with some statistical software that will add a little noise to each point so that points with identical values will appear to be different. If you have software that includes this option, apply it to your plot and summarize the effect of the jittering.

LAUNDRY

2.13 Change the units. Refer to the laundry data set with the outlier.

(a) Create a spreadsheet with the price per load expressed in dollars.

(b) Make a scatterplot for the data in your spreadsheet.

(c) Describe how this scatterplot differs from Figure 2.2.

Interpreting scatterplots

To look more closely at a scatterplot such as Figure 2.2, 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.

linear relationship

The relationship in Figure 2.2 is difficult to see. Looking at it carefully suggests that its *form* is approximately **linear**. In other words, it may be appropriate to summarize the **relationship** with a straight line. To explore this possibility, we can use software to put a straight line through the data. We will see more details about how this is done in Section 2.4.

EXAMPLE 2.10

LAUND

Scatterplot with a straight line. Figure 2.3 plots the laundry detergent data with a straight line. The line helps us to see and to evaluate the linear form of the relationship. In Section 2.4 (page 107), we will learn how to determine this line.

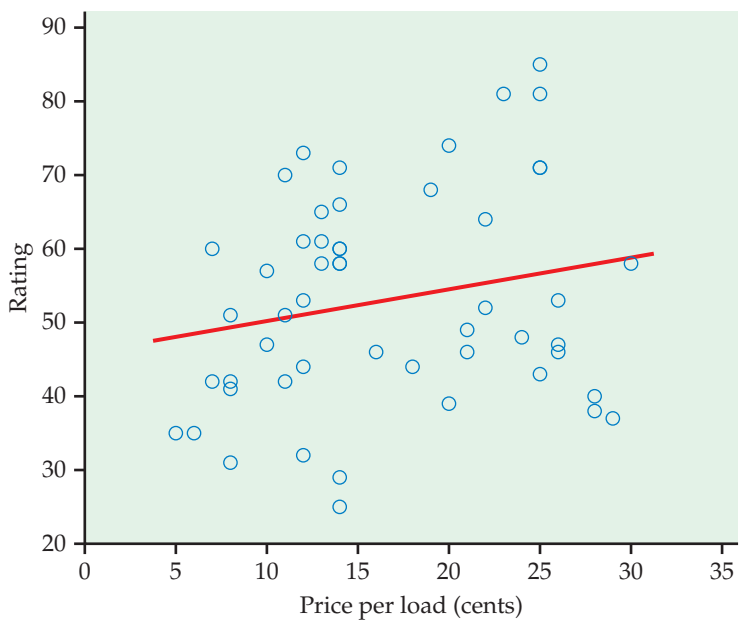


FIGURE 2.3 Scatterplot of rating versus price per load (in cents), with a fitted straight line, Example 2.10.

There is a large amount of scatter about the line. We see that there are eight laundry detergents with a price of 14 cents per load. For these products, the variation in ratings is substantial, from 25 to 71. We do not see any additional outliers in this plot.

Although it is very weak, the relationship in Figure 2.3 has a *direction*, laundry detergents that cost more have somewhat higher ratings. This is a *positive 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 tend to accompany below-average values of the other, and vice versa.

The *strength* of a relationship in a scatterplot is determined by how closely the points follow a clear form. The overall relationship in Figure 2.3 is weak. Here is an example of a stronger linear relationship.

EXAMPLE 2.11

EDSPEND

benchmarking

Education spending and population: Benchmarking. We expect that states with larger populations would spend more on education than states with smaller populations.[6] What is the nature of this relationship? Can we use this relationship to evaluate whether some states are spending more than we expect or less than we expect? This type of exercise is called **benchmarking**. The basic idea is to compare processes or procedures of an organization with those of similar organizations.

Figure 2.4 is a spreadsheet giving the education spending and the populations of the 50 U.S. states for 2015. Figure 2.5 is a scatterplot of the education spending versus the population with a straight line. The scatterplot shows a strong positive relationship between these two variables.

FIGURE 2.4 State spending (in billions of dollars) and population (in millions) for the 50 U.S. states, Example 2.11.

Excel

	A	B	C
1	State	Spending	Population
2	Alabama	14.9	4.9
3	Alaska	3.8	0.7
4	Arizona	14.8	6.8
5	Arkansas	8.5	3.0
6	California	110.7	39.2
7	Colorado	14.3	5.4
8	Connecticut	13.1	3.6
9	Delaware	4.0	0.9
10	Florida	40.1	20.2
11	Georgia	26.4	10.2
12	Hawaii	3.3	1.4
13	Idaho	3.4	1.7
14	Illinois	38.0	12.9
15	Indiana	17.5	6.6
16	Iowa	10.8	3.1
17	Kansas	8.6	2.9
18	Kentucky	12.8	4.4
19	Louisiana	13.1	4.7
20	Maine	4.2	1.3
21	Maryland	18.6	6.0
22	Massachusetts	21.2	6.8
23	Michigan	24.0	9.9
24	Minnesota	17.7	5.5
25	Mississippi	8.1	3.0
26	Missouri	16.3	6.1

Excel

	A	B	C
27	Montana	2.7	1.0
28	Nebraska	6.5	1.9
29	Nevada	5.7	2.9
30	New Hampshire	4.2	1.3
31	New Jersey	36.6	9.0
32	New Mexico	6.6	2.1
33	New York	74.3	19.8
34	North Carolina	30.2	10.0
35	North Dakota	3.0	0.8
36	Ohio	32.1	11.6
37	Oklahoma	10.8	3.9
38	Oregon	11.3	4.0
39	Pennsylvania	37.5	12.8
40	Rhode Island	3.5	1.1
41	South Carolina	10.8	4.9
42	South Dakota	2.1	0.9
43	Tennessee	15.9	6.6
44	Texas	89.4	27.4
45	Utah	10.0	3.0
46	Vermont	2.5	0.6
47	Virginia	24.9	8.4
48	Washington	21.2	7.2
49	West Virginia	6.0	1.8
50	Wisconsin	18.9	5.8
51	Wyoming	2.7	0.6
52			

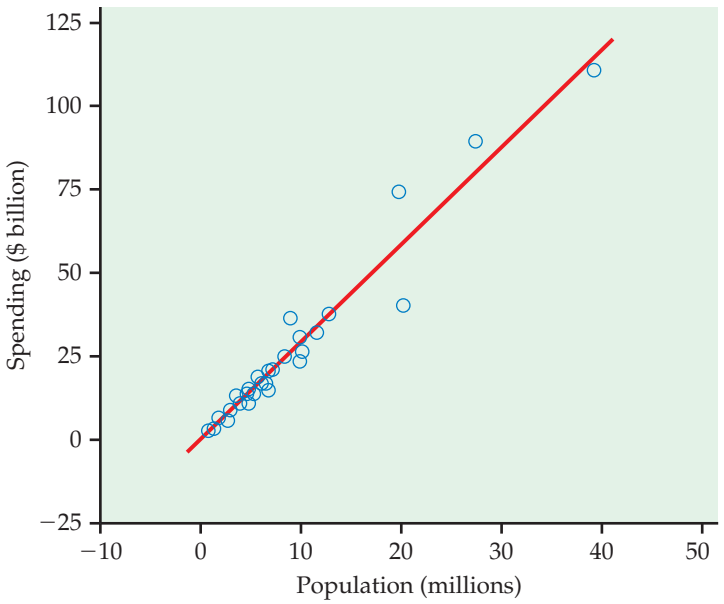


FIGURE 2.5 Scatterplot of state spending (in billions of dollars) versus population for the 50 U.S. states, with a fitted straight line, Example 2.11.

USE YOUR KNOWLEDGE

AU: note, changed prices to match previous example. Okay?

2.14 Make a scatterplot. In our Mocha Frappuccino example, the 12-ounce drink costs $3.95, the 16-ounce drink costs $4.45, and the 24-ounce drink costs $4.95. Explain which variable should be used as the explanatory variable, and make a scatterplot and include the fitted straight line if your software includes this option. Describe the scatterplot and the association between these two variables.

Of course, not all relationships are linear. Here is an example where a relationship is described by a curve.

EXAMPLE 2.12

CALCIUM

Calcium retention. Our bodies need calcium to build strong bones. How much calcium do we need? Does the amount that we need depend on our age? Questions like these are studied by nutrition researchers. One series of studies used the amount of calcium retained by the body as a response variable and the amount of calcium consumed as an explanatory variable.[7]

Figure 2.6 is a scatterplot of calcium retention in milligrams per day (mg/d) versus calcium intake (mg/d) for 56 children aged 11 to 15 years.

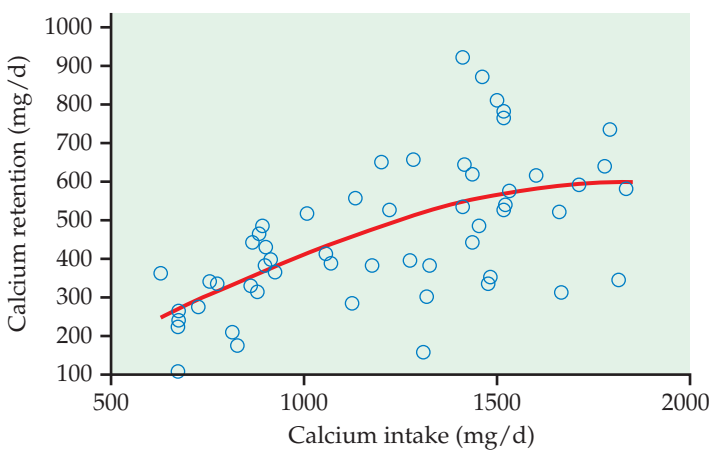


FIGURE 2.6 Scatterplot of calcium retention (mg/d) versus calcium intake (mg/d) for 56 children, with a fitted curve, Example 2.12. There is a positive relationship between these two variables, but it is not linear.

A smooth curve generated by software helps us see the relationship between the two variables.

There is clearly a relationship here. As calcium intake increases, the body retains more calcium. However, the relationship is not linear. The curve is approximately linear for low values of intake, but then the line curves more and becomes almost level.

transformation

There are many kinds of curved relationships like that in Figure 2.6. For some of these, we can apply a **transformation** to the data that will make the relationship approximately linear. To do this, we replace the original values with the transformed values and then use the transformed values for our analysis.

Transforming data is common in statistical practice. There are systematic principles that describe how transformations behave and guide the search for transformations that will, for example, make a distribution more Normal or a curved relationship more linear.

The log transformation

log transformation

The most important transformation that we will use is the **log transformation**. This transformation can be used for variables that have positive values only. Occasionally, we use it when there are zeros, but in this case we first replace the zero values by some small value, often one-half of the smallest positive value in the data set.

You have probably encountered logarithms in one of your high school mathematics courses as a way to do certain kinds of arithmetic. Logarithms are a powerful tool when used in statistical analyses. We will use natural logarithms. Statistical software and statistical calculators generally provide easy ways to perform this transformation.

Let's try a log transformation on our calcium retention data. Here are the details.

EXAMPLE 2.13

CALCIUM

Calcium retention with logarithms. Figure 2.7 is a scatterplot of the log of calcium retention versus calcium intake. The plot includes a fitted straight line to help us see the relationship. We see that the transformation has worked. Our relationship is now approximately linear.

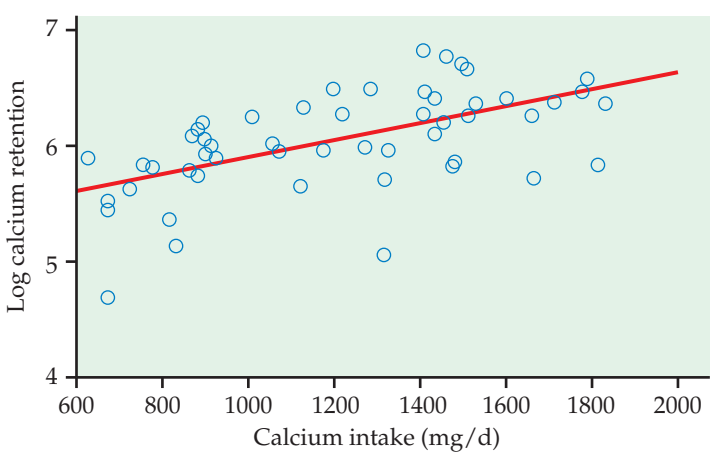


FIGURE 2.7 Scatterplot of log calcium retention versus calcium intake, with a fitted line, for 56 children, Example 2.13. The relationship is approximately linear.

Our analysis of the calcium retention data in Examples 2.12 and 2.13 reminds us of an important issue when describing relationships. In Example 2.12, we noted that the relationship appeared to become approximately flat. Biological processes are consistent with this observation. There is probably a point where additional intake does not result in any additional retention. With our transformed relationship in Figure 2.7, however, there is no leveling off as we saw in Figure 2.6, even though we appear to have a good fit to the data. The relationship and fit apply to the range of data that are analyzed. *We cannot assume that the relationship extends beyond the range of the data*.

For the calcium data, we used a log transformation to describe the curved relationship in Figure 2.6 as the linear relationship in Figure 2.7. Here is another application of a log transformation.

EXAMPLE 2.14

EDSPEND

Education spending and population with logarithms. Let's examine the relationship between spending and population using logs for both variables. Figure 2.8 gives the plot with the fitted line.

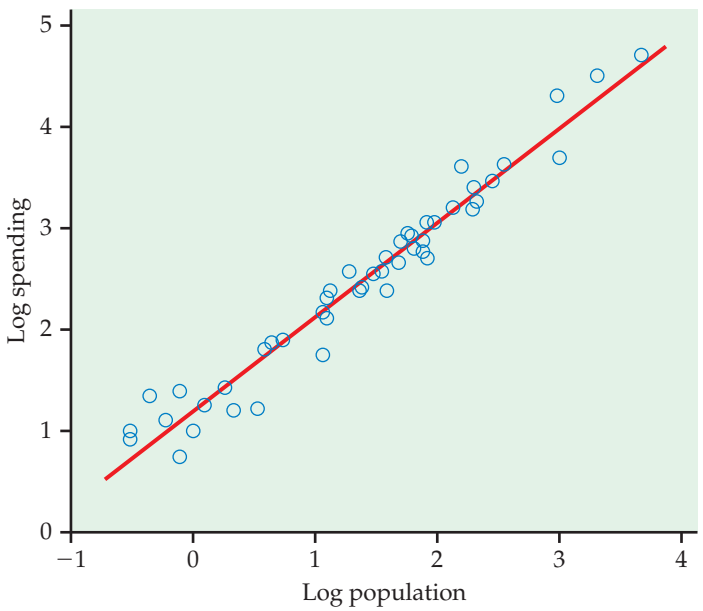


FIGURE 2.8 Scatterplot of log spending versus log population for the 50 U.S. states, with a fitted line, Example 2.14. The relationship is approximately linear.

USE YOUR KNOWLEDGE

2.15 Compare the plots. Compare the plot in Figure 2.8 with the one in Figure 2.5 (page 90). Which one do you prefer? Give reasons for your answer.

EDSPEND

Use of transformations and the interpretation of scatterplots are an art that requires judgment and knowledge about the variables that we are studying. *Always ask yourself if the relationship that you see makes sense*. If it does not, then additional analyses are needed to understand the data.

Adding categorical variables to scatterplots

In Figure 2.3 (page 88), we looked at the relationship between the rating and the price per load for 52 laundry detergents. A more detailed look at the data shows that there are two different types of laundry detergent included in this data set, liquid and powder. Let's examine where these two types of laundry detergents are in our plot.

CATEGORICAL VARIABLES IN SCATTERPLOTS

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

EXAMPLE 2.15

LAUND

Rating versus price and type of laundry detergent. In our scatterplot, we use the color blue for liquids and the color red for powders. The scatterplot is given in Figure 2.9. Separate lines are given for each type of laundry detergent. Most of the laundry detergents are liquids. There are three powders with somewhat low prices and four powders with relatively high prices. The prices of the powders are similar to the prices of the liquids.

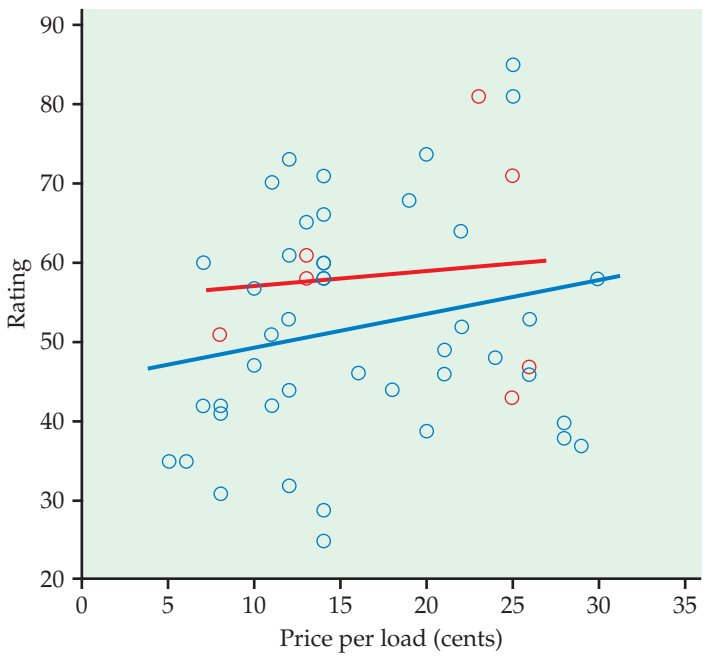


FIGURE 2.9 Scatterplot of rating versus price per load (in cents), with fitted straight lines, for 52 laundry detergents, Example 2.15. The type of detergent is indicated by the color: blue for liquid and red for powder.

In this example, we used a categorical variable, type, to distinguish the two types of laundry detergents in our plot. Suppose that the additional variable that we want to investigate is quantitative. In this situation, we sometimes can combine the values into ranges of the quantitative variable—such as high, medium, and low—to create a categorical variable.

Careful judgment is needed in using this graphical method. Don't be discouraged if your first attempt is not very successful. In performing a good data analysis, you will often produce several plots before you find the one that you believe to be the most effective in describing the data.[8]

Scatterplot smoothers

In Figure 2.6 (page 90), we added a curve to our scatterplot to better understand the relationship between calcium retention and calcium intake. This curve helped us to see that the amount of calcium retained tends to level off as the intake increases. The method that we used to construct the curve is called **smoothing**.

smoothing

Today, most statistical software includes options to perform the calculations needed for smoothing. The technical details vary, but the basic idea is that there is a smoothing parameter that controls the degree to which the relationship is smoothed. Here is another example.

EXAMPLE 2.16

Laundry rating versus price with a smooth fit. Figure 2.2 (page 87) gives the scatterplot for rating versus price for the remaining 52 laundry detergents that we studied in Example 2.9. In Figure 2.3 (page 88), we added a straight line to the plot to help us see the relationship. Figure 2.10 shows the laundry detergent with two different smooth curves. The first (a) used a relatively small value of the smoothing parameter. The second (b) used a larger value, making the curve smoother. Overall, the relationship is very weak and there is no clear pattern in the plot.

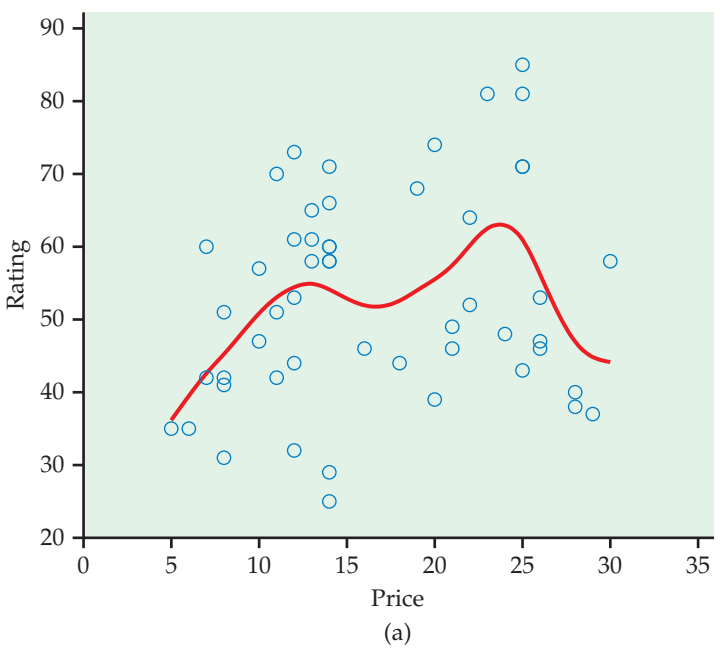
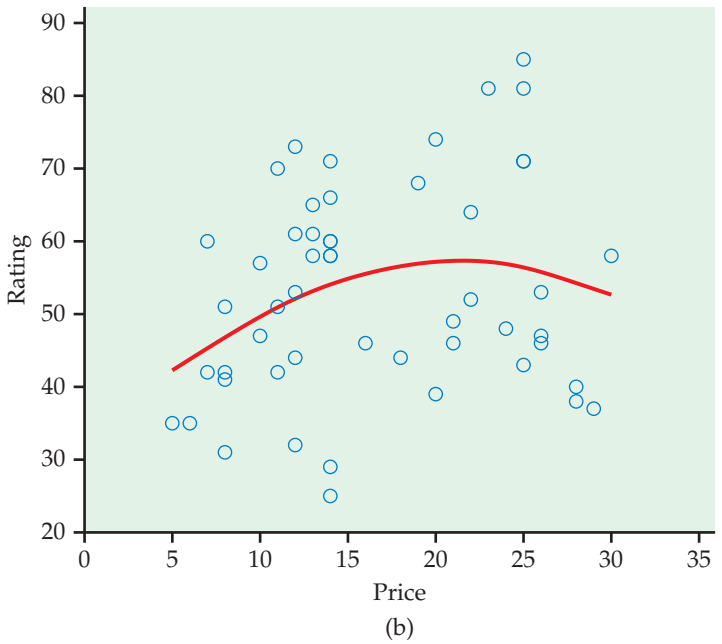


FIGURE 2.10 Scatterplot of rating versus price per load (in cents), with smooth curves, Example 2.16: (a) with a small value of the smoothing parameter; (b) with a higher value of the smoothing parameter.



DE/PE: continued line style OK? not included in design

FIGURE 2.10 *Continued*

Scatterplot smoothers can help you to learn about relationships between two quantitative variables. They can confirm that there is a linear relationship, or they can suggest other features that are not evident in a casual look at the scatterplot. Here is an example of the latter scenario.

EXAMPLE 2.17

EDSPEND

A smooth fit for education spending and population with logs. Figure 2.11 gives the scatterplot of log education spending versus log population with a smooth curve. The curve suggests that the relationship is approximately linear except for states with relatively small populations. For these, the spending is relatively flat.

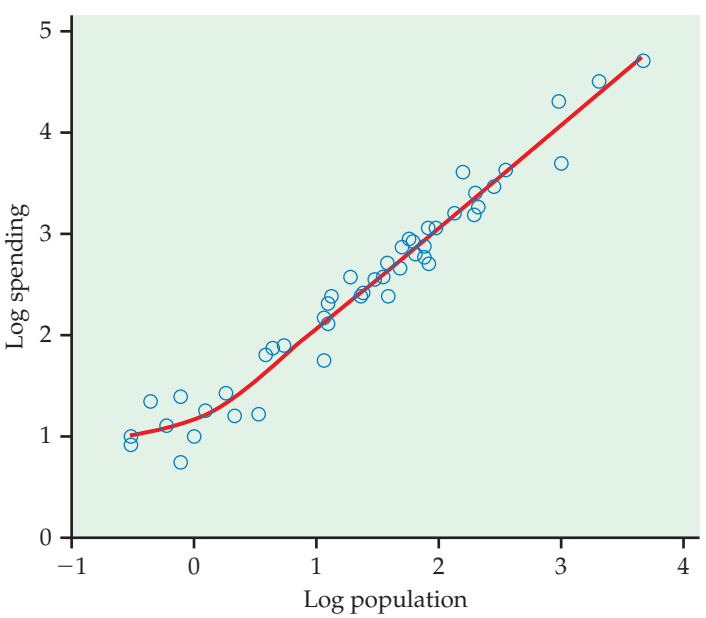


FIGURE 2.11 Scatterplot of log spending versus log population, with a smooth curve fitted to the data, for 50 U.S. states, Example 2.17. This smooth curve fits the data very well and suggests that the relationship is generally linear except for states with small populations.

Categorical explanatory variables

Scatterplots display the association between two quantitative variables. To display a relationship between a categorical variable and a quantitative variable, make a side-by-side comparison of the distributions of the response for each category. Back-to-back stemplots (page 12) and side-by-side boxplots (page 37) are useful tools for this purpose.

We will study methods for describing the association between two categorical variables in Section 2.6 (page 136).

SECTION 2.2 SUMMARY

- 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.
- 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 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. Plot points with different colors or symbols to see the effect of a categorical variable in a scatterplot.
- 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.
- A **log transformation** of one or both variables in a scatterplot can help us to understand the relationship between two quantitative variables.
- A **scatterplot smoother** is a tool to examine the relationship between two quantitative variables by fitting a smooth curve to the data. The amount of smoothing can be varied using a **smoothing parameter**.

SECTION 2.2 EXERCISES

For Exercises 2.10 and 2.11, see page 85; for Exercises 2.12 and 2.13, see page 87; for Exercise 2.14, see page 90; and for Exercise 2.15, see page 92.

2.16 Make some sketches. For each of the following situations, make a scatterplot that illustrates the given relationship between two variables.

(a) No apparent relationship.

(b) A strong negative linear relationship.

(c) A weak positive relationship that is not linear.

(d) A more complicated relationship. Explain the relationship.

2.17 What's wrong? Explain what is wrong with each of the following:

(a) If two variables are negatively associated, then low values of one variable are associated with low values of the other variable.

(b) A stemplot can be used to examine the relationship between two variables.

(c) In a scatterplot, we put the response variable on the x axis and the explanatory variable on the y axis.

2.18 Blueberries and anthocyanins. Anthocyanins are compounds that have been associated with health benefits associated with the heart, bones, and the brain. Blueberries are a good source of many different anthocyanins. Researchers at the Piedmont Research Station of North Carolina State University have assembled a database giving the concentrations of 18 different anthocyanins for 267 varieties of blueberries.[9] Four of the anthocyanins measured are delphinidin-3-arabinoside, malvidin-3-arabinoside, cyanidin-3-galactoside, and delphinidin-3-glucoside, all measured in units of mg per 100g of berries (dry weight). In the data file, we have simplified the names of these anthocyanins to Antho1, Antho2, Antho3, and Antho4. In Exercises 1.167 and 1.168 (page 77), you examined the distributions of each Antho3 and Antho4. BERRIES

(a) Make a scatterplot of the data with Antho3 on the x axis and Antho4 on the y axis.

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

(c) Are there any outliers or unusual observations?

(d) Is it useful to add a straight line to your scatterplot? Explain your answer.

(e) If you have access to the appropriate software, explore the use of a scatterplot smoother to understand this relationship. Summarize what you have found using this method.

2.19 Blueberries and anthocyanins with logs. Refer to the previous exercise. Transform each of the variables with a log, make a scatterplot and answer the questions in the previous exercise for the transformed data. BERRIES

2.20 Blueberries and anthocyanins: Raw data or logs. Refer to Exercises 2.18 and 2.19.

(a) Compare your results from the two exercises.

(b) For exploring and explaining the relationship between Antho4 and Antho3, do you prefer the analysis you performed in Exercise 2.18 or the one you performed in Exercise 2.19? Give reasons for your answer. BERRIES

2.21 Fuel consumption. Natural Resources Canada tests new vehicles each year and reports several variables related to fuel consumption for vehicles in different classes.[10] For 2015 they provide data for 527 vehicles that use regular fuel. Two variables reported are carbon dioxide (CO_2) emissions and highway fuel consumption. CO_2 is measured in grams per kilometer (g/km) and highway fuel consumption measured in liters per 100 kilometers (L/100km). CANFREG

(a) Make a scatterplot of the data with highway fuel consumption on the x axis and CO_2 emissions on the y axis.

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

(c) Are there any outliers or unusual observations?

(d) Is it useful to add a straight line to your scatterplot? Explain your answer.

(e) If you have access to the appropriate software, explore the use of a scatterplot smoother to understand this relationship. Summarize what you find using this method.

2.22 Fuel consumption with a line. Refer to the previous exercise. CANFREG

(a) Add a line to the plot. To what extent to you think that the line does a good job of summarizing the relationship?

(b) If your have the appropriate software, use smooth curves to examine the relationship. Does your analysis support the idea of using a straight line to summarize the relationship? Explain your answer.

2.23 Fuel consumption for different types of vehicles. Refer to the previous two exercises. Those exercises examined data for vehicles that used regular fuel. Data are also available for vehicles that use several other types of fuel. There are 1067 vehicles in total. The variable Fuel has four different possible values: X, for regular fuel; Z, for premium fuel; D, for diesel; and E, for ethanol. CANFUEL

(a) Make a scatterplot of all of the data using different symbols or colors for the different fuel types.

(b) Does the relationship between CO_2 and highway fuel consumption depend upon the type of fuel that the vehicle uses? Explain your answer.

2.24 Bone strength. Osteoporosis is a condition where bones become weak. It affects more than 200 million people worldwide. Exercise is one way to produce strong bones and to prevent osteoporosis. Because we use our dominant arm (the right arm for most people) more than our nondominant arm, we expect the bone in our dominant arm to be stronger than the bone in our nondominant arm. By comparing the strengths, we can get an idea of the effect that exercise

can have on bone strength. Here are some data on the strength of bones, measured in Newton meters divided by 1000 (Nm/1000), for the arms of 15 young men:[11]
ARMSTR

ID	Nondominant	Dominant	ID	Nondominant	Dominant
1	15.7	16.3	9	15.9	20.1
2	25.2	26.9	10	13.7	18.7
3	17.9	18.7	11	17.7	18.7
4	19.1	22.0	12	15.5	15.2
5	12.0	14.8	13	14.4	16.2
6	20.0	19.8	14	14.1	15.0
7	12.3	13.1	15	12.3	12.9
8	14.4	17.5			

Before attempting to compare the arm strengths of the nondominant and dominant arms, let's take a careful look at the data for these two variables.

(a) Make a scatterplot of the data with the nondominant arm strength on the x axis and the dominant arm strength on the y axis.

(b) Describe the overall pattern in the scatterplot and any striking deviations from the pattern.

(c) Describe the form, direction, and strength of the relationship.

(d) Identify any outliers.

(e) Is the relationship approximately linear?

2.25 Bone strength for baseball players. Refer to the previous exercise. The study collected arm bone strength information for two groups of young men. The data in the previous exercise were for a control group. The second group in the study comprised men who played baseball. We know that these baseball players use their dominant arm in throwing (those who throw with their nondominant arm were excluded), so they get more arm exercise than the controls. Here are the data for the baseball players:

ID	Nondominant	Dominant	ID	Nondominant	Dominant
16	17.0	19.3	24	15.1	19.4
17	16.9	19.0	25	13.5	20.4
18	17.7	25.2	26	13.6	17.1
19	21.2	37.7	27	20.3	26.5
20	21.0	40.3	28	17.3	30.3
21	14.6	20.8	29	14.6	17.4
22	31.5	36.9	30	22.6	35.0
23	14.9	21.2			

Answer the questions in the previous exercise for the baseball players. ARMSTR

2.26 Compare the baseball players with the controls. Refer to the previous two exercises. ARMSTR

(a) Plot the data for the two groups on the same graph using different symbols for the baseball players and the controls.

(b) Use your plot to describe and compare the relationships for the two variables. Write a short paragraph summarizing what you have found.

2.27 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?

AU/DE/PE: data file needed?

2.28 What's in the beer? The website **beer100.com** advertises itself as "Your Place for All Things Beer." One of their "things" is a list of 159 domestic beer brands with the percent alcohol, calories per 12 ounces, and carbohydrates per 12 ounces (in grams).[12]
BEERD

(a) Figure 2.12 gives a scatterplot of calories versus percent alcohol. Give a short summary of what can be learned from the plot.

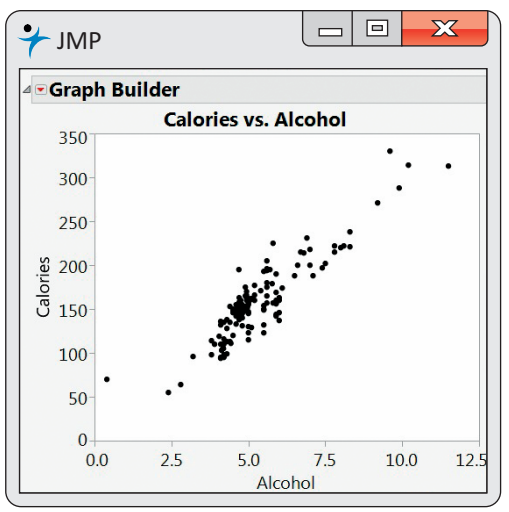


FIGURE 2.12 Scatterplot of calories versus percent alcohol for 159 domestic brands of beer, Exercise 2.28.

(b) One of the points is an outlier. Find the brand of the outlier. How is this brand of beer different from the other brands?

(c) Remove the outlier from the data set and generate a scatterplot of the remaining data.

(d) Describe the relationship between calories and percent alcohol based on what you see in your scatterplot.

2.29 More beer. Refer to the previous exercise. BEERD

(a) Make a scatterplot of calories versus percent alcohol using the data set without the outlier.

(b) Describe the relationship between these two variables. If your software is capable, use a line and smoothers to explore the relationship.

2.30 Imported beer. The beer100 website also gives data for imported beers. Describe the relationship between calories and percent alcohol for these imported beers. BEERI

2.31 Compare domestic with imported. Plot calories versus percent alcohol for domestic and imported beers on the same scatterplot. Use different colors or symbols for the two types of beers. Summarize what this plot tells you about the relationship and the difference between the two types of beer. In particular, note any characteristics that are better shown in this plot relative to what was learned in Exercises 2.28, 2.29, and 2.30. BEER

2.32 Decay of a radioactive element. Barium-137m is a radioactive form of the element barium that decays very rapidly. It is easy and safe to use for lab experiments in schools and colleges.[13] In a typical experiment, the radioactivity of a sample of barium-137m is measured for one minute. It is then measured for three additional one-minute periods, separated by two minutes. So data are recorded at one, three, five, and seven minutes after the start of the first counting period. The measurement units are counts. Here are the data for one of these experiments:[14] DECAY

Time	1	3	5	7
Count	578	317	203	118

(a) Make a scatterplot of the data. Give reasons for the choice of which variables to use on the x and y axes.

(b) Describe the overall pattern in the scatterplot and any striking deviations from the pattern.

(c) Describe the form, direction, and strength of the relationship.

(d) Identify any outliers.

(e) Is the relationship approximately linear?

2.33 Use a log for the radioactive decay. Refer to the previous exercise. Transform the counts using a log transformation. Then repeat parts (a) through (e) for the transformed data and compare your results with those from the previous exercise. DECAY

2.34 Internet use and babies. The World Bank collects data on many variables related to world development for countries throughout the world. Two of these are Internet use, in number of users per 100 people, and birthrate, in births per 1000 people.[15] Figure 2.13 is a scatterplot of birthrate versus Internet use for the 106 countries that have data available for both variables. INBIRTH

(a) Describe the relationship between these two variables.

(b) A friend looks at this plot and concludes that using the Internet will decrease the number of babies born. Write a short paragraph explaining why the association seen in the scatterplot does not provide a reason to draw this conclusion.

2.35 Try a log. Refer to the previous exercise. INBIRTH

(a) Make a scatterplot of the log of births per 1000 people versus Internet users per 100 people.

(b) Describe the relationship that you see in this plot and compare it with Figure 2.13.

(c) Which plot do you prefer? Give reasons for your answer.

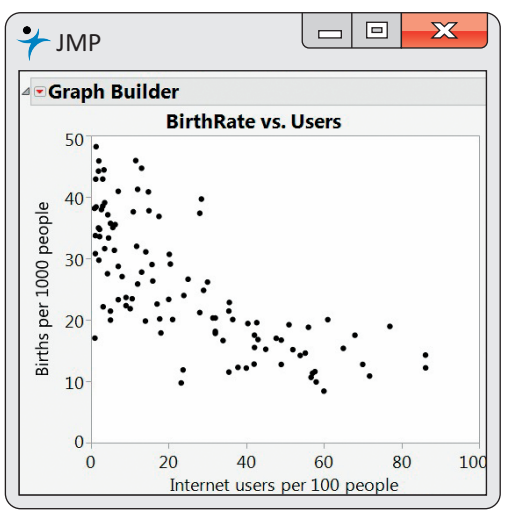


FIGURE 2.13 Scatterplot of births (per 1000 people) versus Internet users (per 100 people) for 106 countries, Exercise 2.34.

AU: Please confirm x-ref page

2.36 Make another plot. Refer to Exercise 2.34. INBIRTH

(a) Make a new data set that has Internet users expressed as users per 10,000 people and births as births per 10,000 people.

(b) Explain why these transformations to give new variables are linear transformations. (*Hint:* See linear transformations on page 44.)

(c) Make a scatterplot using the transformed variables.

(d) Compare your new plot with the one in Figure 2.13.

(e) Why do you think that the analysts at the World Bank chose to express births as births per 1000 people and Internet users as users per 100 people?

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

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				

(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?

(c) 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?

2.3 Correlation

When you complete this section, you will be able to:

- Use a correlation to describe the direction and strength of a linear relationship between two quantitative variables.
- Interpret the sign of a correlation.
- Identify situations in which the correlation is not a good measure of association between two quantitative variables.
- Identify a linear pattern in a scatterplot.
- For describing the relationship between two quantitative variables, identify the roles of the correlation, a numerical summary, and the scatterplot, a graphical summary.

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 2.14 depict exactly the same data, but the plot on the right is drawn smaller in a large field. The plot on the right seems to show a stronger relationship.

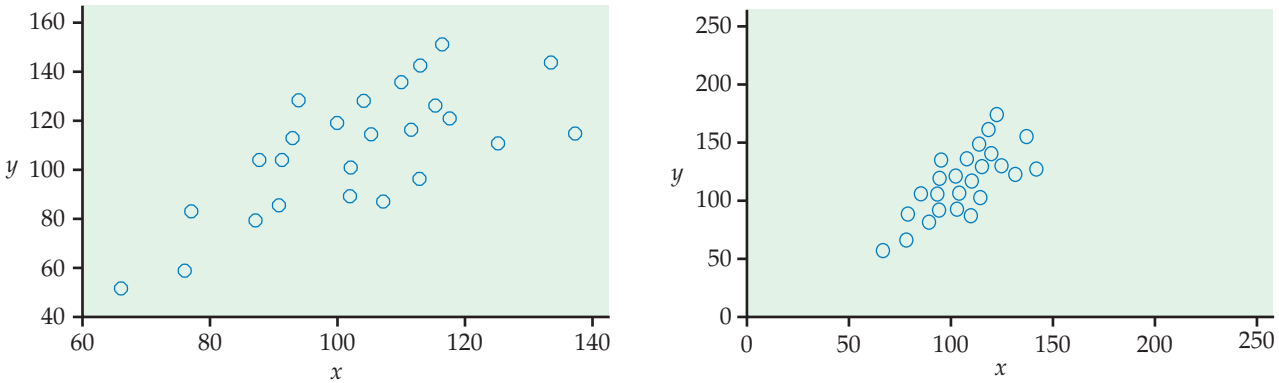


FIGURE 2.14 Two scatterplots of the same data. The linear pattern in the plot on the right appears stronger because of the surrounding space.

Our eyes can be fooled by changing the plotting scales or the amount of white space around the cloud of points in a scatterplot.[16] We need to follow our strategy for data analysis by using a numerical measure to supplement the graph. *Correlation* is the measure we use.

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 ith 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 Σ 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 computes r from the values of x and y pairs.

LOOK BACK
standardize, p. 59

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 ith person. 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. You can 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.

USE YOUR KNOWLEDGE

LAUNDRY

LAUNDRY

2.38 Laundry detergents. Example 2.8 (page 85) describes data on the rating and price per load for 53 laundry detergents. Use these data to compute the correlation between rating and the price per load.

2.39 Change the units. Refer to the previous exercise. Express the price per load in dollars.

(a) Is the transformation from cents to dollars a linear transformation? Explain your answer.

(b) Compute the correlation between rating and price per load expressed in dollars.

(c) How does the correlation that you computed in part (b) compare with the one you computed in the previous exercise?

(d) What can you say in general about the effect of changing units using linear transformations on the size of the correlation?

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, 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 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*. For example, 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 (a linear transformation) 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*.

standardize, p. 30

- *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 2.15 illustrate how values of r closer to 1 or -1 correspond to stronger linear relationships. To make the essential meaning of

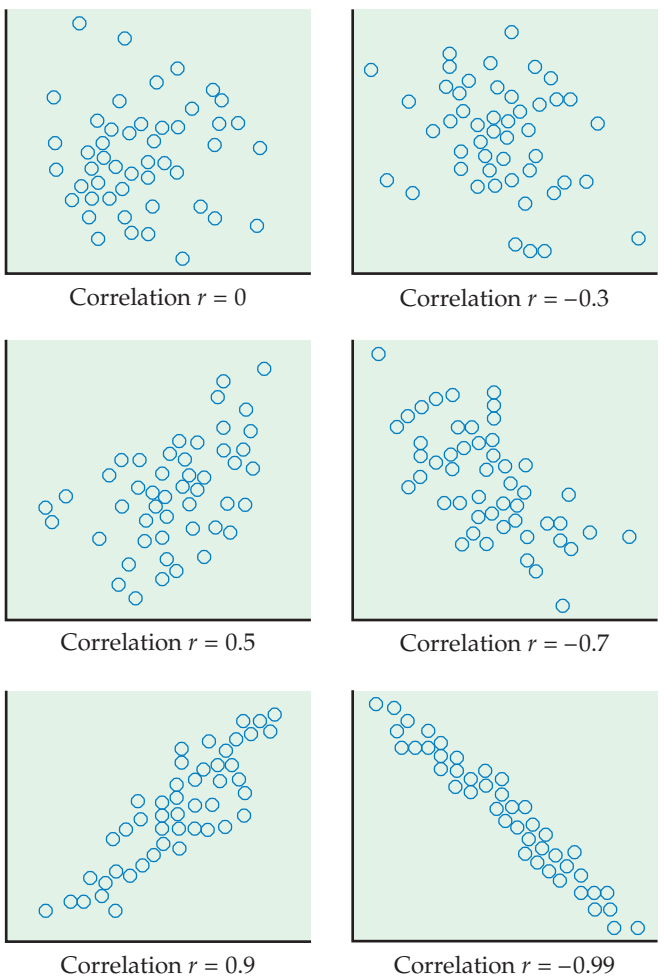


FIGURE 2.15 How the correlation r measures the direction and strength of a linear association.

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 website. Also, see Exercises 2.56 and 2.57 (page 106).

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 2.18

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.

AU/DE/PE: data file needed?

These facts in the example above 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 point to Pierre's scores to arrive at a fair comparison.

SECTION 2.3 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 $-1 \le r \le 1$ 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 2.3 EXERCISES

For Exercises 2.38 and 2.39, see page 102.

2.40 Correlations and scatterplots. Explain why you should always look at a scatterplot when you want to use a correlation to describe the relationship between two quantitative variables.

2.41 Interpret some correlations. For each of the following correlations, describe the relationship between the two quantitative variables in terms of the direction and the strength of the linear relationship.

(a) $r = 0.9$.

(b) $r = -0.9$.

(c) $r = -0.3$.

(d) $r = 0.0$.

2.42 Blueberries and anthocyanins. In Exercise 2.18 (page 97), you examined the relationship between Antho4 and Antho3, two anthocyanins found in blueberries. BERRIES

(a) Find the correlation between these two anthocyanins.

(b) Look at the scatterplot for these data that you made in part (a) of Exercise 2.18 (or make one if you did not do that exercise). Is the correlation a good numerical summary of the graphical display in the scatterplot? Explain your answer.

(c) Does the size of the correlation suggest that the amounts of these two anthocyanins is approximately equal in these blueberries? Explain why or why not.

2.43 Blueberries and anthocyanins with logs. In Exercise 2.19 (page 97), you examined the relationship between Antho4 and Antho3, two anthocyanins found in blueberries, using logs for both variables. Answer the questions in the previous exercise for the variables transformed in this way. BERRIES

2.44 Fuel consumption. In Exercise 2.21 (page 97), you examined the relationship between CO_2 emissions and highway fuel consumption for 527 vehicles that use regular fuel. Find the correlation between these two variables. Write a short paragraph describing the relationship using the scatterplot and the correlation. CANFREG

2.45 Fuel consumption for different types of vehicles. In Exercise 2.23 (page 97), you examined the relationship between CO_2 emissions and highway fuel consumption for 1067 vehicles that use four different types of fuel. Find the correlations between CO_2 and highway fuel consumption for each of these four categories of vehicle. Summarize your results explaining similarities and differences in the relationships among the four types of fuel. CANFUEL

2.46 Strong association but no correlation. Here is a data set that illustrates an important point about correlation: CORR

X	25	35	45	55	65
Y	10	30	50	30	10

(a) Make a scatterplot of Y versus X.

(b) Describe the relationship between Y and X. Is it weak or strong? Is it linear?

(c) Find the correlation between Y and X.

(d) What important point about correlation does this exercise illustrate?

2.47 Bone strength. Exercise 2.24 (page 97) gives the bone strengths of the dominant and the nondominant arms for 15 men who were controls in a study. ARMSTR

(a) Find the correlation between the bone strength of the dominant arm and the bone strength of the nondominant arm.

(b) Look at the scatterplot for these data that you made in part (a) of Exercise 2.24 (or make one if you did not do that exercise). Is the correlation a good numerical summary of the graphical display in the scatterplot? Explain your answer.

2.48 Bone strength for baseball players. Refer to the previous exercise. Similar data for baseball players are given in Exercise 2.25 (page 98). Answer parts (a) and (b) of the previous exercise for these data. ARMSTR

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 Decay of a radioactive element. Data for an experiment on the decay of barium-137m is given in Exercise 2.32 (page 99). DECAY

(a) Find the correlation between the radioactive counts and the time after the start of the first counting period.

(b) Does the correlation give a good numerical summary of the relationship between these two variables? Explain your answer.

2.51 Decay in the log scale. Refer to the previous exercise and to Exercise 2.33 (page 99), where the counts were transformed by a log. DECAY

(a) Find the correlation between the log counts and the time after the start of the first counting period.

(b) Does the correlation give a good numerical summary of the relationship between these two variables? Explain your answer.

Is there extra space here?

your results for this exercise with those vious exercise.

2.52 Brand names and generic products.

(a) If a store always prices its generic "store brand" products at 80% of the brand name products' prices, what would be the correlation between the prices of the brand name products and the store brand products? (*Hint:* Draw a scatterplot for several prices.)

(b) If the store always prices its generic products $2 less than the corresponding brand name products, then what would be the correlation between the prices of the brand name products and the store brand products?

2.53 Alcohol and calories in beer. Figure 2.12 (page 98) gives a scatterplot of the calories versus percent alcohol for 159 brands of domestic beer. BEERD

(a) Compute the correlation for these data.

(b) Does the correlation do a good job of describing the direction and strength of this relationship? Explain your answer.

2.54 Alcohol and calories in beer revisited. Refer to the previous exercise. The data that you used to compute the correlation includes an outlier. BEERD

(a) Remove the outlier and recompute the correlation.

(b) Write a short paragraph about the possible effects of outliers on a correlation using this example to illustrate your ideas.

2.55 Compare domestic with imported. In Exercise 2.31 (page 99), you compared domestic beers with imported beers with respect to the relationship between calories and percent alcohol. In that exercise, you used scatterplots to make the comparison. Compute the correlations for these two categories of beer and write a new summary of the comparison using correlations in addition to the scatterplots. BEERD, BEERI

2.56 Use the applet. Go to the *Correlation and Regression* applet. Click on the scatterplot to create a group of 12 points in the lower-right corner of the scatterplot with a strong straight-line negative pattern (correlation about −0.9).

(a) Add one point at the upper left that is in line with the first 12. How does the correlation change?

(b) Drag this last point down until it is opposite the group of 12 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.*

2.57 Use the applet. You are going to use the *Correlation and Regression* applet to make different scatterplots with 12 points that have correlation close to 0.8. *Many patterns can have the same correlation. Always plot your data before you trust a correlation.*

(a) Stop after adding the first two points. What is the value of the correlation? Why does it have this value no matter where the two points are located?

(b) Make a lower-left to upper-right pattern of 12 points with correlation about $r = 0.8$. (You can drag points up or down to adjust r after you have 12 points.) Make a rough sketch of your scatterplot.

(c) Make another scatterplot, this time with 11 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.

(d) Make yet another scatterplot, this time with 12 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.

2.58 An interesting set of data. Make a scatterplot of the following data: INTER

X	1	2	3	4	10	10
Y	1	3	3	5	1	10

Verify 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?

2.59 Internet use and babies. Figure 2.13 (page 99) is a scatterplot of the number of births per 1000 people rate versus Internet users per 100 people for 106 countries. In Exercise 2.34 (page 99), you described this relationship. INBIRTH

(a) Make a plot of the data similar to Figure 2.13 and report the correlation.

(b) Is the correlation a good numerical summary for this relationship? Explain your answer.

2.60 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 age of American workers and their occupation.

(b) We found a high correlation ($r = 1.19$) between students' ratings of faculty teaching and ratings made by other faculty members.

(c) The correlation between the sex of a group of students and the color of their cell phone was $r = 0.23$.

2.4 Least-Squares Regression

When you complete this section, you will be able to:

- Draw a straight line on a scatterplot of a set of data, given the equation of the line.
- Predict a value of the response variable y for a given value of the explanatory variable x using a regression equation.
- Explain the meaning of the term *least squares*.
- Calculate the equation of a least-squares regression line from the means and standard deviations of the explanatory and response variables and their correlation.
- Read the output of statistical software to find the equation of the least-squares regression line and the value of r^2.
- Explain the meaning of r^2 in the regression setting.

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 2.19

FIDGET

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 eight 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:[17]

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 2.16 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.

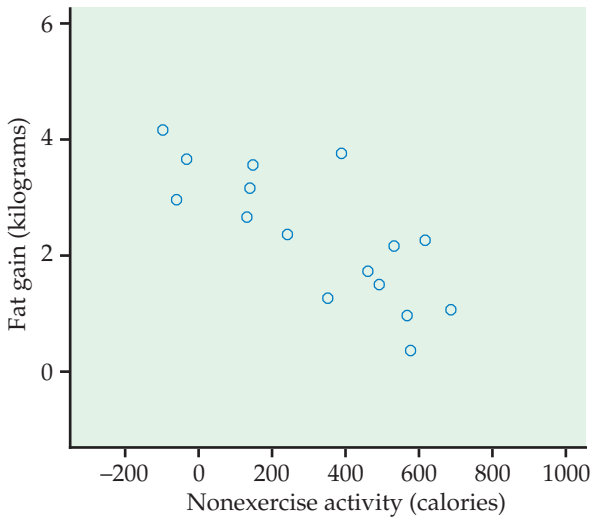


FIGURE 2.16 Fat gain after eight weeks of overeating plotted against the increase in nonexercise activity over the same period, Example 2.19.

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 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 concise description of the relationship between the response variable y and the explanatory variable x. It is the numerical summary that supports the scatterplot, our graphical summary.

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

$$y = b_0 + b_1 x$$

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 $x = 0$.

In practice, we will use software to obtain values of b_0 and b_1 for a given set of data.

EXAMPLE 2.20

FIDGET

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 2.17, we have drawn the regression line with the equation

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

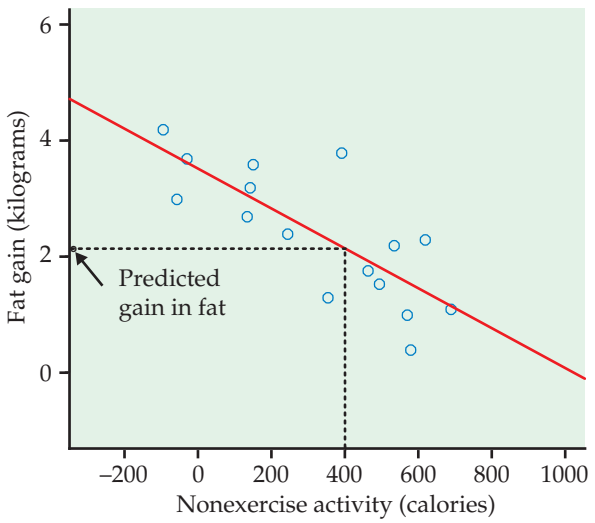


FIGURE 2.17 A regression line fitted to the nonexercise activity data and used to predict fat gain for an NEA increase of 400 calories, Examples 2.20 and 2.21.

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 increase.

The slope b_1 of a line $y = b_0 + b_1x$ is the *average 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 2.20, the intercept, $b_0 = 3.505$ kilograms. This value is the estimated fat gain if NEA does not change. When we substitute the value zero for the NEA increase, the regression equation gives 3.505 (the intercept) as the predicted value of the fat gain.

USE YOUR KNOWLEDGE

FIDGET

2.61 Plot the line. Make a plot of the data in Example 2.19 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. We can interpret the prediction as the *average* value of y corresponding to a collection of cases at the particular value of x or as our best guess at the value of y for an *individual* with the particular value of x.

EXAMPLE 2.21

FIDGET

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 2.17. 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 2.17, 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

2.62 Predict the fat gain. Use the regression equation in Example 2.20 to predict the fat gain for a person whose NEA increases by 250 calories.

EXAMPLE 2.22

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 2.17. 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 and should be avoided.

USE YOUR KNOWLEDGE

2.63 Would you use the regression equation to predict? Consider the following values for NEA increase: −300, 300, 600, 800. For each, decide whether you would use the regression equation in Example 2.20 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 2.17 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

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

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 2.18 illustrates the idea. For clarity, the plot shows only three of the points from Figure 2.17, 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

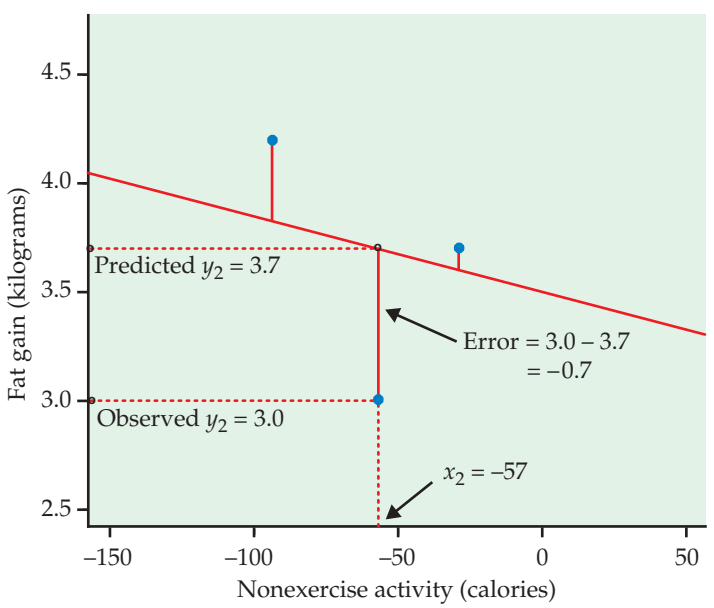


FIGURE 2.18 The least-squares idea: make the errors in predicting y as small as possible by minimizing the sum of their squares.

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 2.17 and 2.18 is, in fact, the least-squares regression line.

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

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

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

$$\sum(\text{error})^2 = \sum(y_i - b_0 - b_1x_i)^2$$

for the given observations x_1 and y_1. For the NEA data, for example, we must find the b_0 and b_1 that minimize

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

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. Therefore, we will 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

$$\hat{y} = b_0 + b_1x$$

with **slope**

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

and **intercept**

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

EXAMPLE 2.23

Check the calculations. Verify from the data in Example 2.19 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 $r = -0.7786$. Therefore, the least-squares regression line of fat gain y on NEA increase x has slope

$$b_1 = r\frac{s_y}{s_x} = -0.7786\,\frac{1.1389}{257.66}$$
$$= -0.00344 \text{ kg per calorie}$$

and intercept

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

The equation of the least-squares line is

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

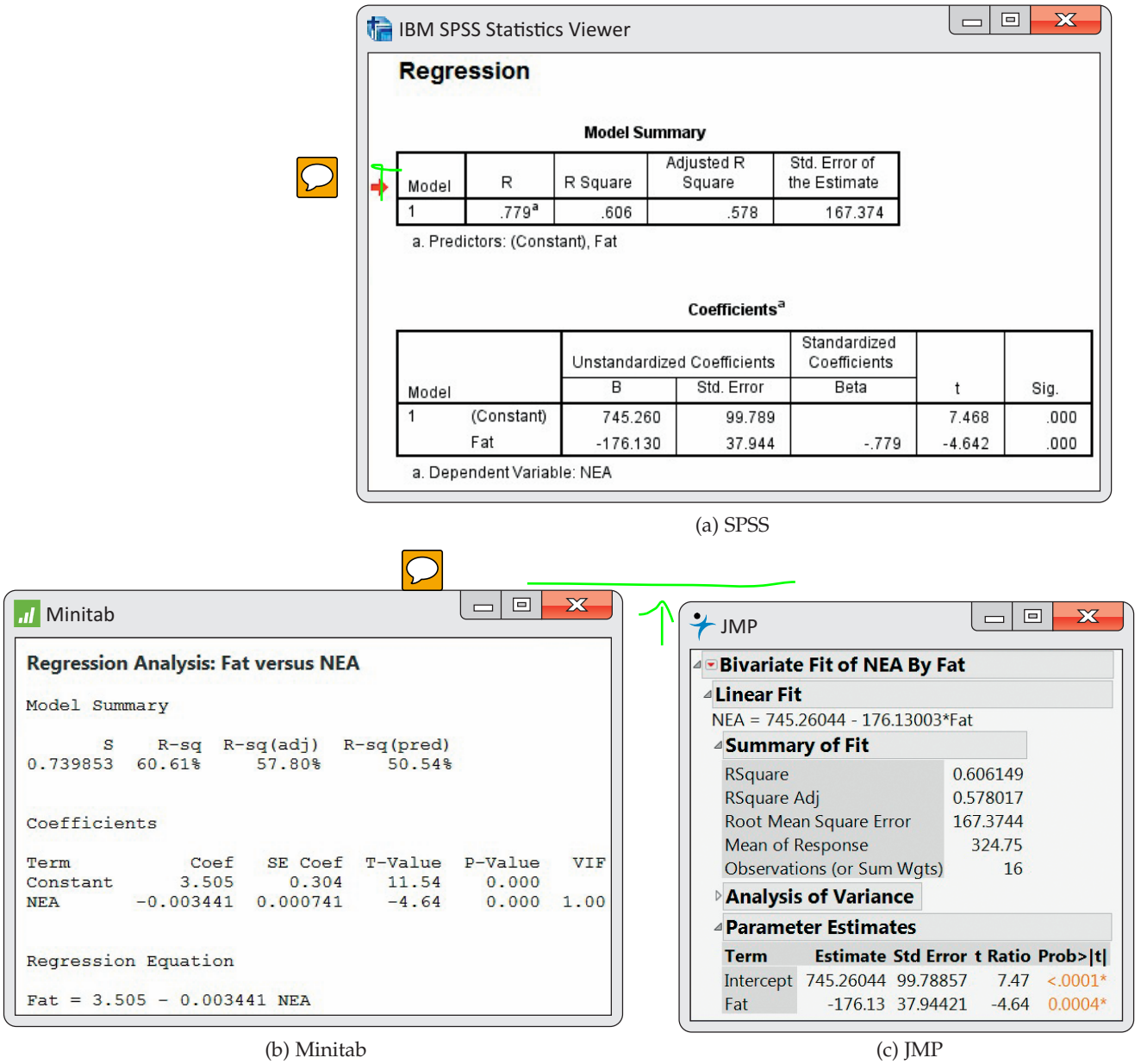
When doing calculations like this by hand, you may need to carry extra decimal places in the preliminary calculations to get accurate values of the slope and intercept. Using software or a calculator with a regression function eliminates this worry.

Interpreting the regression line

The slope $b_1 = -0.00344$ kilograms per calorie in Example 2.23 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*.

EXAMPLE 2.24

Regression using software. Figure 2.19 displays the basic regression output for the nonexercise activity data from three 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 each output. The software also provides information that we do not yet need, including some that we trimmed from Figure 2.19.



(a) SPSS

(b) Minitab

(c) JMP

FIGURE 2.19 Regression results for the nonexercise activity data from three statistical software packages: (a) SPSS; (b) Minitab; (c) JMP. Other software produces similar output.

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.

Facts about least-squares regression

Regression is one of the most common statistical settings, and least squares is the most common method for fitting a regression line to data. Here are some facts about least-squares regression lines.

Fact 1. There is a close connection between correlation and the slope of the least-squares line. The slope is

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

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

Fact 2. The least-squares regression line always passes through the point $(\bar{x}, \bar{y})$ on the graph of y against x. 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.

Fact 3. The distinction between explanatory and response variables is essential in regression. Least-squares regression looks at the distances of the data points from the line only in the y direction. If we reverse the roles of the two variables, we get a different least-squares regression line.

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.

EXAMPLE 2.25

FIDGET

Fidgeting and fat gain. Figure 2.20 is a scatterplot of the fidgeting and fat gain data described in Example 2.19 (page 107). There is a negative linear relationship. The two lines on the plot are the two least-squares regression

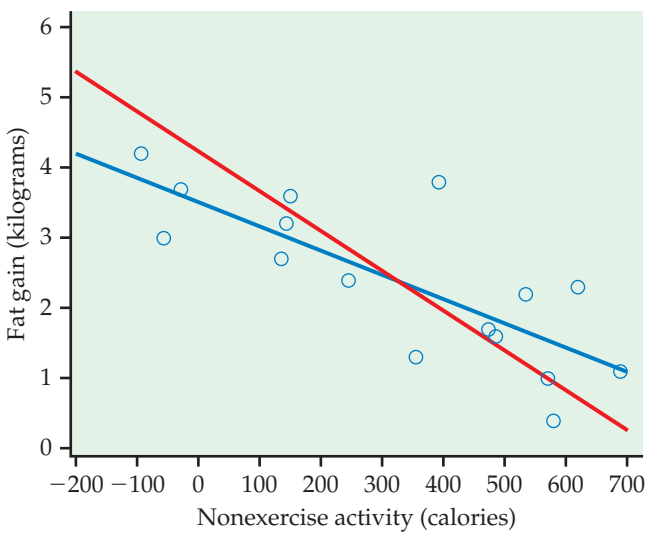


FIGURE 2.20 Scatterplot of fat gain versus nonexercise activity for 16 subjects from Example 2.8. The two lines are the two least-squares regression lines, using nonexercise activity to predict fat gain (blue) and using fat gain to predict nonexercise activity (red), Example 2.25.

lines. The regression line using nonexercise activity to predict fat gain is blue. The regression line using fat gain to predict nonexercise activity is red. *Regression of fat gain on nonexercise activity and regression of nonexercise activity on fat gain give different lines*. In the regression setting, you must decide which variable is explanatory.

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 2.19 (page 107) is $r = -0.7786$. Because $r^2 = 0.6062$, the straight-line relationship between NEA and fat gain explains about 61% of the vertical scatter in fat gains in Figure 2.17 (page 109).

When you report a regression, give r^2 as a measure of how successfully the regression explains the response. All three software outputs in Figure 2.19 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 that the points lie exactly on a line. Then $r^2 = 1$ and all 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

2.64 What fraction of the variation is explained? Consider the following correlations: -0.9, -0.5, -0.2, 0, 0.2, 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 2.17 (page 109) 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 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.

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 dashed lines in Figure 2.17 show a rough average for 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 2.17, about 61% of the variation in fat gains among the 16 subjects is due to the straight-line relationship between y and x. The remaining 39% is vertical scatter in the observed responses remaining after the line has fixed the predicted responses.

Another view of r^2

Here is a more specific interpretation of r^2. The fat gains y in Figure 2.17 range from 0.4 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 to 690 calories, it pulls 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 substituting 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:

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

This fact is always true. The squared correlation gives the variance that 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.

SECTION 2.4 SUMMARY

- A **regression line** is a straight line that describes how a response variable y changes as an explanatory variable x changes.
- The most common method of fitting a line to a scatterplot is least squares. The **least-squares regression line** is the straight line $\hat{y} = b_0 + b_1 x$ that minimizes the sum of the squares of the vertical distances of the observed y-values from the line.

- You can use a regression line to **predict** the value of y for any value of x by substituting this x into the equation of the line. **Extrapolation** beyond the range of x-values spanned by the data is risky.
- The **slope** $b1$ of a regression line $\hat{y} = b_0 + b_1x$ is the rate at which the predicted response $\hat{y}$ changes along the line as the explanatory variable x changes. Specifically, $b1$ is the change in $\hat{y}$ when x increases by 1. The numerical value of the slope depends on the units used to measure x and y.
- The **intercept** $b0$ of a regression line $\hat{y} = b_0 + b_1x$ is the predicted response $\hat{y}$ when the explanatory variable $x = 0$. This prediction is not particularly useful unless x can actually take values near 0.
- The least-squares regression line of y on x is the line with slope $b_1 = rs_y/s_x$ and intercept $b_0 = \bar{y} - b_1\bar{x}$. This line always passes through the point $(\bar{x}, \bar{y})$.
- **Correlation and regression** are closely connected. The correlation r is the slope of the least-squares regression line when we measure both x and y in standardized units. The square of the correlation r^2 is the fraction of the variance of one variable that is explained by least-squares regression on the other variable.

SECTION 2.4 EXERCISES

For Exercise 2.61, see page 109; for Exercise 2.62, see page 110; for Exercise 2.63, see page 111; and for Exercise 2.64, see page 116.

2.65 Blueberries and anthocyanins. In Exercise 2.18 (page 97), you examined the relationship between Antho4 and Antho3, two anthocyanins found in blueberries. In Exercise 2.42 (page 105), you found the correlation between these two variables. BERRIES

(a) Find the equation of the least-squares regression line for predicting Antho4 from Antho3.

(b) Make a scatterplot of the data with the fitted line.

(c) How well does the line fit the data? Explain your answer.

(d) Use the line to predict the value of Antho4 when Antho3 is equal to 1.5.

2.66 Fuel consumption. In Exercise 2.21 (page 97), you examined the relationship between CO_2 emissions and highway fuel consumption for 527 vehicles that use regular fuel. In Exercise 2.44 (page 105), you found the correlation between these two variables. CANFREG

(a) Find the equation of the least-squares regression line for predicting CO_2 emissions from highway fuel consumption.

(b) Make a scatterplot of the data with the fitted line.

(c) How well does the line fit the data? Explain your answer.

(d) Use the line to predict the value of CO_2 for vehicles that consume 8.0 liters per kilometer (L/km).

2.67 Fuel consumption for different types of vehicles. In Exercise 2.23 (page 97), you examined the relationship between CO_2 emissions and highway fuel consumption for 1067 vehicles. You used different plotting symbols for the four different types of fuel used by these vehicles: regular, premium, diesel, and ethanol. CANFREG

AU: should this be CANFUEL?

(a) Find the least-squares equation for predicting CO_2 emissions from highway fuel consumption for all 1067 vehicles.

(b) Make a scatterplot of the data with the fitted line.

(c) Based on what you learned from Example 2.23, do you think that a single least-squares regression line provides a good fit for all four types of vehicles? Explain your answer.

2.68 Bone strength. Exercise 2.24 (page 97), gives the bone strengths of the dominant and the nondominant arms for 15 men who were controls in a study. ARMSTR

(a) Plot the data. Use the bone strength in the nondominant arm as the explanatory variable and bone strength in the dominant arm as the response variable.

(b) The least-squares regression line for these data is

$$\text{dominant} = 2.74 + (0.936 \times \text{nondominant})$$

Add this line to your plot.

(c) Use the scatterplot (a graphical summary), with the least-squares line (a graphical display of a numerical summary) to write a short paragraph describing this relationship.

2.69 Bone strength for baseball players. Refer to the previous exercise. Similar data for baseball players are given in Exercise 2.25 (page 98). Here is the equation of the least-squares line for the baseball players:

$$\text{dominant} = 0.886 + (1.373 \times \text{nondominant})$$

Answer parts (a) and (c) of the previous exercise for these data. ARMSTR

2.70 Predict the bone strength. Refer to Exercise 2.68. A young male who is not a baseball player has a bone strength of 16.0 $cm^4/1000$ in his nondominant arm. Predict the bone strength in the dominant arm for this person. ARMSTR

2.71 Predict the bone strength for a baseball player. Refer to Exercise 2.69. A young male who is a baseball player has a bone strength of 16.0 $cm^4/1000$ in his nondominant arm. Predict the bone strength in the dominant arm for this person. ARMSTR

2.72 Compare the predictions. Refer to the two previous exercises. You have predicted two dominant-arm bone strengths, one for a baseball player and one for a person who is not a baseball player. The nondominant bone strengths are both 16.0 $cm^4/1000$. ARMSTR

(a) Compare the two predictions by computing the difference in means, baseball player minus control.

(b) Explain how the difference in the two predictions is an estimate of the effect of baseball throwing exercise on the strength of arm bones.

(c) For nondominant arm strengths of 12 $cm^4/1000$ and 20 $cm^4/1000$, repeat your predictions and take the differences. Make a table of the results of all three calculations (for 12, 16, and 20 $cm^4/1000$).

(d) Write a short summary of the results of your calculations for the three different nondominant-arm strengths. Be sure to include an explanation of why the differences are not the same for the three nondominant-arm strengths.

2.73 Least-squares regression for radioactive decay. Refer to Exercise 2.32 (page 99) for the data on radioactive decay of barium-137m. Here are the data: DECAY

Time	1	3	5	7
Count	578	317	203	118

(a) Using the least-squares regression equation

$$\text{count} = 602.8 - (74.7 \times \text{time})$$

find the predicted values for the counts.

(b) Compute the differences, observed count minus predicted count. How many of these are positive; how many are negative?

(c) Square and sum the differences that you found in part (b).

(d) Repeat the calculations that you performed in parts (a), (b), and (c) using the equation

$$\text{count} = 500 - (100 \times \text{time})$$

(e) In a short paragraph, explain the least-squares idea using the calculations that you performed in this exercise.

2.74 Least-squares regression for the log counts. Refer to Exercise 2.33 (page 99), where you analyzed the radioactive decay of barium-137m data using log counts. Here are the data: DECAY

Time	1	3	5	7
Log count	6.35957	5.75890	5.31321	4.77068

(a) Using the least-squares regression equation

$$\text{log count} = 6.593 - (0.2606 \times \text{time})$$

find the predicted values for the log counts.

(b) Compute the differences, observed count minus predicted count. How many of these are positive; how many are negative?

(c) Square and sum the differences that you found in part (b).

(d) Repeat the calculations that you performed in parts (a), (b), and (c) using the equation

$$\text{log count} = 7 - (0.2 \times \text{time})$$

(e) In a short paragraph, explain the least-squares idea using the calculations that you performed in this exercise.

2.75 College students by state. How well does the population of a state predict the number of undergraduates? The National Center for Educational

Statistics collects data for each of the 50 U.S. states that we can use to address this question.[18] COLLEGE

(a) Make a scatterplot with population on the x axis and number of undergraduates on the y axis.

(b) Describe the form, direction, and strength of the relationship. Are there any outliers?

(c) For the number of undergraduates, the mean is 302,136 and the standard deviation is 358,460, and for population, the mean is 5,955,551 and the standard deviation is 6,620,733. The correlation between the number of undergraduates and the population is 0.98367. Use this information to find the least-squares regression line. Show your work.

(d) Add the least-squares line to your scatterplot.

2.76 College students by state without the four largest states. Refer to the previous exercise. Let's eliminate the four largest states, which have populations greater than 15 million. Here are the numerical summaries: for number of undergraduate college students, the mean is 220,134 and the standard deviation is 165,270; for population, the mean is 4,367,448 and the standard deviation is 3,310,957. The correlation between the number of undergraduate college students and the population is 0.97081. Use this information to find the least-squares regression line. Show your work. COL46

2.77 Make predictions and compare. Refer to the two previous exercises. Consider a state with a population of 4 million (this value is approximately the median population for the 50 states). COLLEGE

(a) Using the least-squares regression equation for all 50 states, find the predicted number of undergraduate college students.

(b) Do the same using the least-squares regression equation for the 46 states with populations less than 15 million.

(c) Compare the predictions that you made in parts (a) and (b). Write a short summary of your results and conclusions. Pay particular attention to the effect of including the four states with the largest populations in the prediction equation for a median-sized state.

2.78 College students by state. Refer to Exercise 2.75, where you examined the relationship between the number of undergraduate college students and the populations for the 50 states. Figure 2.21 gives the output from a software package for the regression. Use this output to answer the following questions: COLLEGE

(a) What is the equation of the least-squares regression line?

(b) What is the value of r^2?

(c) Interpret the value of r^2.

(d) Does the software output tell you that the relationship is linear and not, for example, curved? Explain your answer.

IBM SPSS Statistics Viewer

Regression

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.984[a]	.968	.967	65178.746

a. Predictors: (Constant), Population

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-15044.917	12454.662		-1.208	.233
	Population	.053	.001	.984	37.869	.000

a. Dependent Variable: Undergrads

FIGURE 2.21 SPSS output for predicting number of undergraduate college students using population for the 50 states, Exercise 2.78.

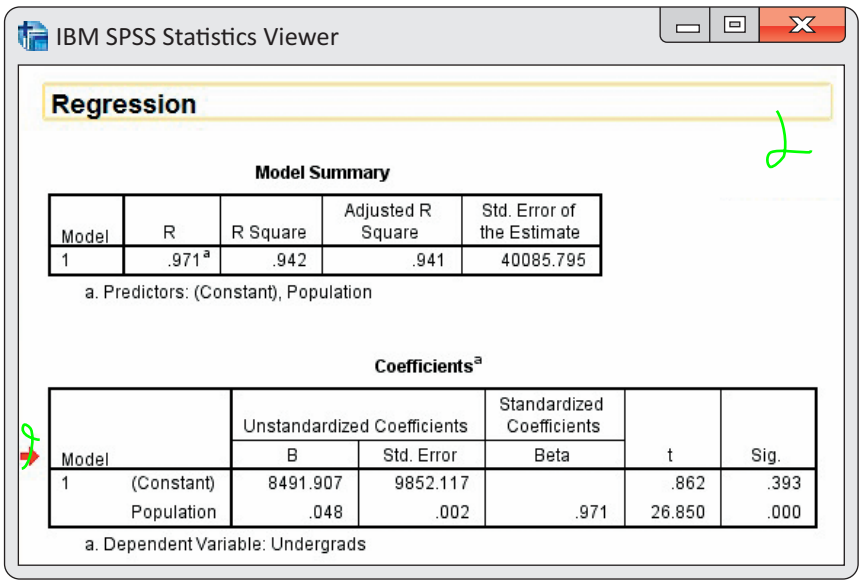


FIGURE 2.22 SPSS output for predicting number of undergraduate college students using population, with the four largest states deleted, Exercise 2.79.

2.79 College students by state without the four largest states. Refer to Exercise 2.76, where you eliminated the four largest states that have populations greater than 15 million. Figure 2.22 gives software output for these data. Answer the questions in the previous exercise for the data set with the 46 states. COL46

2.80 Data generated by software. The following 20 observations on Y and X were generated by a computer program. GENDATA

X	Y	X	Y
22.06	34.38	17.75	27.07
19.88	30.38	19.96	31.17
18.83	26.13	17.87	27.74
22.09	31.85	20.20	30.01
17.19	26.77	20.65	29.61
20.72	29.00	20.32	31.78
18.10	28.92	21.37	32.93
18.01	26.30	17.31	30.29
18.69	29.49	23.50	28.57
18.05	31.36	22.02	29.80

(a) Make a scatterplot and describe the relationship between Y and X.

(b) Find the equation of the least-squares regression line and add the line to your plot.

(c) What percent of the variability in Y is explained by X?

(d) Summarize your analysis of these data in a short paragraph.

2.81 Add an outlier. Refer to Exercise 2.80. Add an additional observation with $y = 25$ and $x = 35$ to the data set. Repeat the analysis that you performed in Exercise 2.80 and summarize your results, paying particular attention to the effect of this outlier. GEN21A

2.82 Add a different outlier. Refer to Exercise 2.80 and the previous exercise. Add an additional observation with $y = 36$ and $x = 30$ to the original data set. GEN21B

(a) Repeat the analysis that you performed in Exercise 2.80 and summarize your results, paying particular attention to the effect of this outlier.

(b) In this exercise and in the previous one, you added an outlier to the original data set and reanalyzed the data. Write a short summary of the changes in correlations that can result from different kinds of outliers.

2.83 Alcohol and calories in beer. Figure 2.12 (page 98) gives a scatterplot of calories versus percent alcohol in 159 brands of domestic beer. BEERD

(a) Find the equation of the least-squares regression line for these data.

(b) Find the value of r^2 and interpret it in the regression context.

AU: OK?

TABLE 2.1 Four Data Sets for Exploring Correlation and Regression

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

(c) Write a short report on the relationship between calories and percent alcohol in beer. Include graphical and numerical summaries for each variable separately as well as graphical and numerical summaries for the relationship in your report.

AU: Okay?

2.84 Alcohol and calories in beer revisited. Refer to the previous exercise. The data that you used includes an outlier. BEERD

(a) Remove the outlier and answer parts (a), (b), and (c) for the new set of data.

(d) Write a short paragraph about the possible effects of outliers on a least-squares regression line and the value of r^2, using this example to illustrate your ideas.

2.85 Always plot your data! Table 2.1 presents four sets of data prepared by the statistician Frank Anscombe to illustrate the dangers of calculating without first plotting the data.[19] ANSC

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

(b) Make a scatterplot for each of the data sets and add the regression line to each plot.

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

2.86 Progress in math scores. Every few years, the National Assessment of Educational Progress asks a national sample of eighth-graders to perform the same math tasks. The goal is to get an honest picture of progress in math. Here are the last few national mean scores, on a scale of 0 to 500:[20] NAEP

Year	1990	1992	1996	2000	2003	2005	2008	2011	2013
Score	263	268	272	273	278	279	281	283	285

(a) Make a time plot of the mean scores, by hand. This is just a scatterplot of score against year. There is a slow linear increasing trend.

(b) Find the regression line of mean score on time step-by-step. First calculate the mean and standard deviation of each variable and their correlation (use a calculator with these functions). Then find the equation of the least-squares line from these. Draw the line on your scatterplot. What percent of the year-to-year variation in scores is explained by the linear trend?

(c) Now use software or the regression function on your calculator to verify your regression line.

2.87 The regression equation. The equation of a least-squares regression line is $y = 15 - 2x$.

(a) What is the value of y for $x = 4$?

(b) If x increases by one unit, what is the corresponding change in y?

(c) What is the intercept for this equation?

2.88 Metabolic rate and lean body mass. Compute the mean and the standard deviation of the metabolic rates

and lean body masses in Exercise 2.37 (page 100) and the correlation between these two variables. Use these values to find the slope of the regression line of metabolic rate on lean body mass. Also find the slope of the regression line of lean body mass on metabolic rate. What are the units for each of the two slopes? BMASS

2.89 Use an applet for progress in math scores. Go to the *Two-Variable Statistical Calculator*. Enter the data for the progress in math scores from Exercise 2.86 using the "User-entered data" option in the "Data" tab. Explore the data by clicking the other tabs in the applet. Using only the only the results provided by the applet, write a short report summarizing the analysis of these data.

2.90 A property of the least-squares regression line. Use the equation for the least-squares regression line to show that this line always passes through the point $(\bar{x}, \bar{y})$.

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

2.5 Cautions about Correlation and Regression

When you complete this section, you will be able to:

- Calculate the residuals for a set of data using the equation of the least-squares regression line and the observed values of the explanatory variable.
- Use a plot of the residuals versus the explanatory variable to assess the fit of a regression line.
- Identify outliers and influential observations by examining scatterplots and residual plots.
- Identify lurking variables that can influence the interpretation of relationships between two variables.
- Explain the difference between association and causality when interpreting the relationship between two variables.

Correlation and regression are among the most common statistical tools. They are used in more elaborate form to study relationships among many variables, a situation in which we cannot see the essentials by studying a single scatterplot. We need a firm grasp of the use and limitations of these tools, both now and as a foundation for more advanced statistics.

Residuals

A regression line describes the overall pattern of a linear relationship between an explanatory variable and a response variable. Deviations from the overall pattern are also important. In the regression setting, we see deviations by looking at the scatter of the data points about the regression line. The vertical distances from the points to the least-squares regression line are as small as possible in the sense that they have the smallest possible sum of squares. Because they represent "leftover" variation in the response after fitting the regression line, these distances are called *residuals*.

RESIDUALS

A **residual** is the difference between an observed value of the response variable and the value predicted by the regression line. That is,

$$\begin{aligned}\text{residual} &= \text{observed } y - \text{predicted } y \\ &= y - \hat{y}\end{aligned}$$

EXAMPLE 2.26

Residuals for fat gain. Example 2.19 (page 107) describes measurements on 16 young people who volunteered to overeat for eight weeks. Those whose nonexercise activity (NEA) spontaneously rose substantially gained less fat than others. Figure 2.23(a) is a scatterplot of these data. The pattern is linear. The least-squares line is

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

One subject's NEA rose by 135 calories. That subject gained 2.7 kilograms of fat. The predicted gain for 135 calories is

$$\hat{y} = 3.505 - (0.00344 \times 135) = 3.04 \text{ kg}$$

The residual for this subject is, therefore,

$$\begin{aligned}\text{residual} &= \text{observed } y - \text{predicted } y \\ &= y - \hat{y} \\ &= 2.7 - 3.04 = -0.34 \text{ kg}\end{aligned}$$

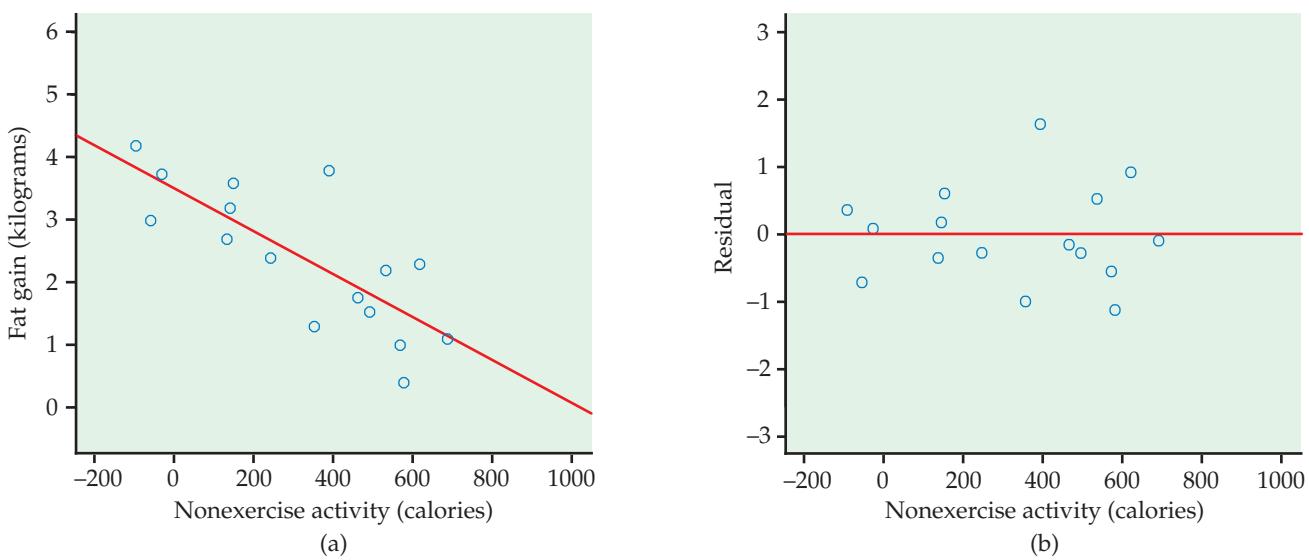


FIGURE 2.23 (a) Scatterplot of fat gain versus increase in nonexercise activity, with the least-squares regression line, Example 2.26. (b) Residual plot for the regression displayed in panel (a); the line at $y = 0$ marks the mean of the residuals.

Most regression software will calculate and store residuals for you.

USE YOUR KNOWLEDGE

2.92 Find the predicted value and the residual. Let's say that we have an individual in the NEA data set who has NEA increase equal to 250 calories and fat gain equal to 2.4 kg. Find the predicted value of fat gain for this individual and then calculate the residual. Explain why this residual is negative.

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

USE YOUR KNOWLEDGE

FIDGET

2.93 Find the sum of the residuals. Here are the 16 residuals for the NEA data rounded to two decimal places:

0.37	−0.70	0.10	−0.34	0.19	0.61	−0.26	−0.98
1.64	−0.18	−0.23	0.54	−0.54	−1.11	0.93	−0.03

Find the sum of these residuals. Note that the sum is not exactly zero because of roundoff error.

You can see the residuals in the scatterplot of Figure 2.23(a) by looking at the vertical deviations of the points from the line. The *residual plot* in Figure 2.23(b) makes it easier to study the residuals by plotting them against the explanatory variable, increase in NEA.

RESIDUAL PLOTS

A **residual plot** is a scatterplot of the regression residuals against the explanatory variable. Residual plots help us assess the fit of a regression line.

Because the mean of the residuals is always zero, the horizontal line at zero in Figure 2.23(b) helps orient us. This line (residual = 0) corresponds to the fitted line in Figure 2.23(a). The residual plot magnifies the deviations from the line to make patterns easier to see. If the regression line catches the overall pattern of the data, there should be *no pattern* in the residuals. That is, the residual plot should show an unstructured horizontal band centered at zero. The residuals in Figure 2.23(b) do have this irregular scatter.

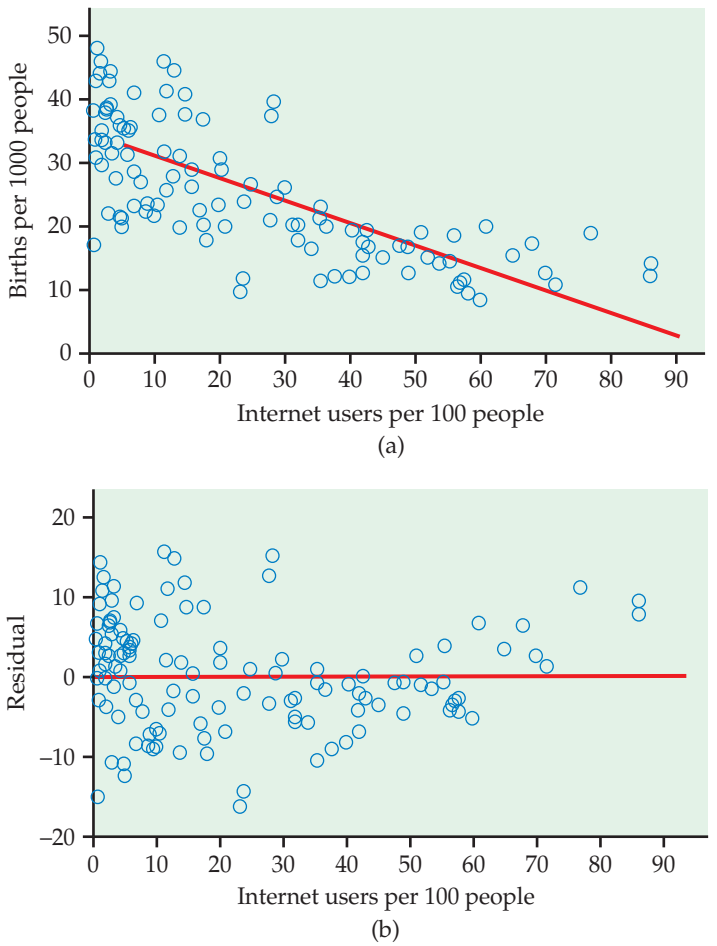
You can see the same thing in the scatterplot of Figure 2.23(a) and the residual plot of Figure 2.23(b). It's just a bit easier in the residual plot. Deviations from an irregular horizontal pattern point out ways in which the regression line fails to catch the overall pattern. Here is an example.

EXAMPLE 2.27

INBIRTH

Patterns in birthrate and Internet user residuals. In Exercise 2.34 (page 99) we used a scatterplot to study the relationship between birthrate and Internet users for 106 countries. In this scatterplot, Figure 2.13, we see that there are many countries with low numbers of Internet users. In addition, the relationship between births and Internet users appears to be curved. For low values of Internet users, there is a clear relationship, while for higher values, the curve becomes relatively flat.

Figure 2.24(a) gives the data with the least-squares regression line, and Figure 2.24(b) plots the residuals. Look at the right part of Figure 2.24(b), where the values of Internet users are high. Here we see that the residuals tend to be positive.



AU: Please check. Okay as adjusted?

FIGURE 2.24 (a) Scatterplot of birthrate versus Internet users, with the least-squares regression line, Example 2.27. (b) Residual plot for the regression displayed in panel (a); the line at $y = 0$ marks the mean of the residuals.

CAUTION

The residual pattern in Figure 2.24(b) is characteristic of a simple curved relationship. *There are many ways in which a relationship can deviate from a linear pattern*. We now have an important tool for examining these deviations. Use it frequently and carefully when you study relationships.

TABLE 2.2 Two Measures of Glucose Level in Diabetics

Subject	HbA1c (%)	FPG (mg/ml)	Subject	HbA1c (%)	FPG (mg/ml)	Subject	HbA1c (%)	FPG (mg/ml)
1	6.1	141	7	7.5	96	13	10.6	103
2	6.3	158	8	7.7	78	14	10.7	172
3	6.4	112	9	7.9	148	15	10.7	359
4	6.8	153	10	8.7	172	16	11.2	145
5	7.0	134	11	9.4	200	17	13.7	147
6	7.1	95	12	10.4	271	18	19.3	255

Outliers and influential observations

When you look at scatterplots and residual plots, look for striking individual points as well as for an overall pattern. Here is an example of data that contain some unusual cases.

EXAMPLE 2.28

HBA1C

Diabetes and blood sugar. People with diabetes must manage their blood sugar levels carefully. They measure their fasting plasma glucose (FPG) several times a day with a glucose meter. Another measurement, made at regular medical checkups, is called HbA1c. This is roughly the percent of red blood cells that have a glucose molecule attached. It measures average exposure to glucose over a period of several months.

This diagnostic test is becoming widely used and is sometimes called A1c by health care professionals. Table 2.2 gives data on both HbA1c and FPG for 18 diabetics five months after they completed a diabetes education class.[21]

Because both FPG and HbA1c measure blood glucose, we expect a positive association. The scatterplot in Figure 2.25(a) shows a

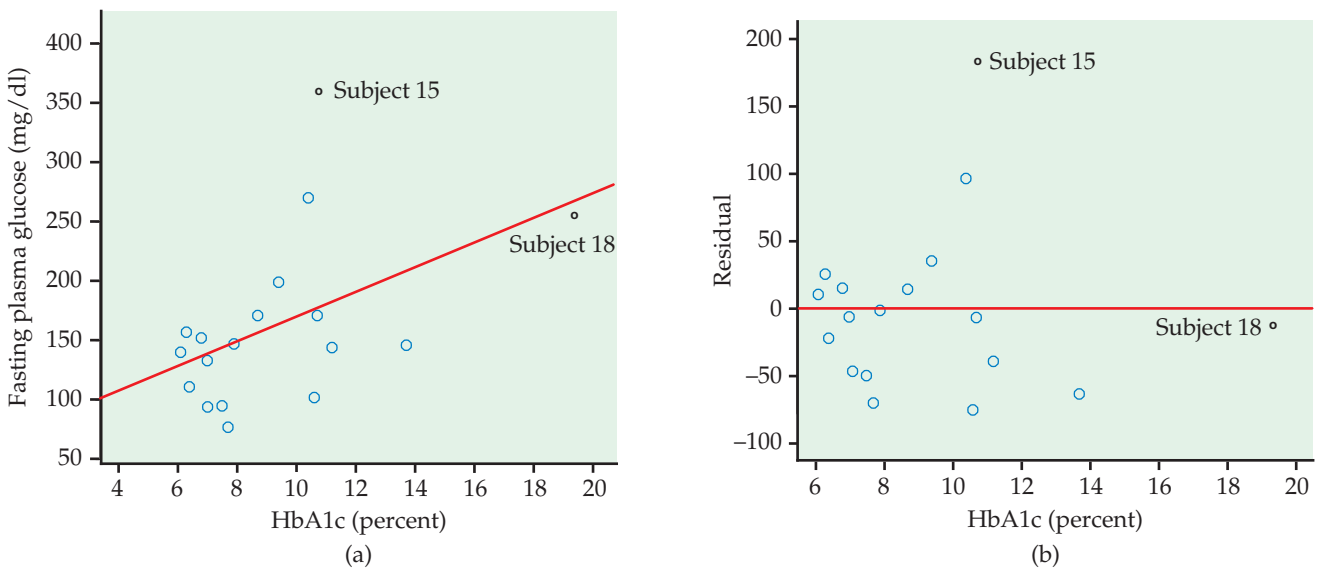


FIGURE 2.25 (a) Scatterplot of fasting plasma glucose against HbA1c (which measures long-term blood glucose), with the least-squares regression line, Example 2.28. (b) Residual plot for the regression of fasting plasma glucose on HbA1c. Subject 15 is an outlier in fasting plasma glucose. Subject 18 is an outlier in HbA1c that may be influential but does not have a large residual.

surprisingly weak relationship, with correlation $r = 0.4819$. The line on the plot is the least-squares regression line for predicting FPG from HbA1c. Its equation is

$$\hat{y} = 66.4 + 10.41x$$

It appears that one-time measurements of FPG can vary quite a bit among people with similar long-term levels, as measured by HbA1c. This is why A1c is an important diagnostic test.

Two unusual cases are marked in Figure 2.25(a). Subjects 15 and 18 are unusual in different ways. Subject 15 has dangerously high FPG and lies far from the regression line in the y direction. Subject 18 is close to the line but far out in the x direction. The residual plot in Figure 2.25(b) confirms that Subject 15 has a large residual and that Subject 18 does not.

Points that are outliers in the x direction, like Subject 18, can have a strong influence on the position of the regression line. Least-squares lines make the sum of squares of the vertical distances to the points as small as possible. A point that is extreme in the x direction with no other points near it pulls the line toward itself.

OUTLIERS AND INFLUENTIAL OBSERVATIONS IN REGRESSION

An **outlier** is an observation that lies outside the overall pattern of the other observations. Points that are outliers in the y direction of a scatterplot have large regression residuals, but other outliers need not have large residuals.

An observation is **influential** for a statistical calculation if removing it would markedly change the result of the calculation. Points that are outliers in the x direction of a scatterplot are often influential for the least-squares regression line.

Influence is a matter of degree—how much does a calculation change when we remove an observation? It is difficult to assess influence on a regression line without actually doing the regression both with and without the suspicious observation. A point that is an outlier in x is often influential. But if the point happens to lie close to the regression line calculated from the other observations, then its presence will move the line only a little and the point will not be influential.

The influence of a point that is an outlier in y depends on whether there are many other points with similar values of x that hold the line in place. Figures 2.25(a) and (b) identify two unusual observations. How influential are they?

EXAMPLE 2.29

HBA1C

Influential observations. Subjects 15 and 18 both influence the correlation between FPG and HbA1c, in opposite directions. Subject 15 weakens the linear pattern; if we drop this point, the correlation increases from $r = 0.4819$ to $r = 0.5684$. Subject 18 extends the linear pattern; if we omit this subject, the correlation drops from $r = 0.4819$ to $r = 0.3837$.

To assess influence on the least-squares line, we recalculate the line leaving out a suspicious point. Figure 2.26 shows three least-squares lines. The solid line is the regression line of FPG on HbA1c based on all 18 subjects. This is the same line that appears in Figure 2.25(a). The dotted line is calculated from all subjects except Subject 18. You see that point 18 does pull the line down toward itself. But the influence of Subject 18 is not very large—the dotted and solid lines are close together for HbA1c values between 6 and 14, the range of all except Subject 18.

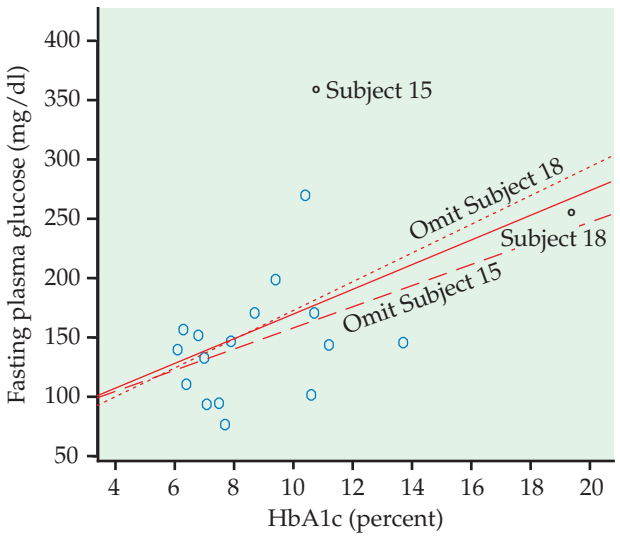


FIGURE 2.26 Three regression lines for predicting fasting plasma glucose from HbA1c, Example 2.29. The solid line uses all 18 subjects. The dotted line leaves out Subject 18. The dashed line leaves out Subject 15. "Leaving one out" calculations are the surest way to assess influence.

The dashed line omits Subject 15, the outlier in y. Comparing the solid and dashed lines, we see that Subject 15 pulls the regression line up. The influence is again not large, but it exceeds the influence of Subject 18.

We did not need the distinction between outliers and influential observations in Chapter 1. A single large salary that pulls up the mean salary $\bar{x}$ for a group of workers is an outlier because it lies far above the other salaries. It is also influential because the mean changes when it is removed. In the regression setting, however, not all outliers are influential. Because influential observations draw the regression line toward themselves, we may not be able to spot them by looking for large residuals.

Beware of the lurking variable

Correlation and regression are powerful tools for measuring the association between two variables and for expressing the dependence of one variable on

the other. These tools must be used with an awareness of their limitations. We have seen that

- Correlation measures *only linear association*, and fitting a straight line makes sense only when the overall pattern of the relationship is linear. Always plot your data before calculating.
- *Extrapolation* (using a fitted model far outside the range of the data that we used to fit it) often produces unreliable predictions.
- Correlation and least-squares regression are *not resistant*. Always plot your data and look for potentially influential points.

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

LURKING VARIABLE

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

EXAMPLE 2.30

Discrimination in medical treatment? Studies show that men who complain of chest pain are more likely to get detailed tests and aggressive treatment such as bypass surgery than are women with similar complaints. Is this association between sex and treatment due to discrimination?

Perhaps not. Men and women develop heart problems at different ages—women are, on the average, between 10 and 15 years older than men. Aggressive treatments are more risky for older patients, so doctors may hesitate to recommend them. Lurking variables—the patient's age and condition—may explain the relationship between sex and doctors' decisions.

Here is an example of a different type of lurking variable.

EXAMPLE 2.31

Gas and electricity bills. A single-family household receives bills for gas and electricity each month. The 12 observations for a recent year are plotted with the least-squares regression line in Figure 2.27. We have arbitrarily chosen to put the electricity bill on the x axis and the gas bill on the y axis. There is a clear negative association. Does this mean that a high electricity bill causes the gas bill to be low and vice versa?

To understand the association in this example, we need to know a little more about the two variables. In this household, heating is done by gas and cooling is done by electricity. Therefore, in the winter months the gas bill will be relatively high and the electricity bill will be relatively low. The pattern is reversed in the summer months. The association that we see in this example is due to a lurking variable: time of year.

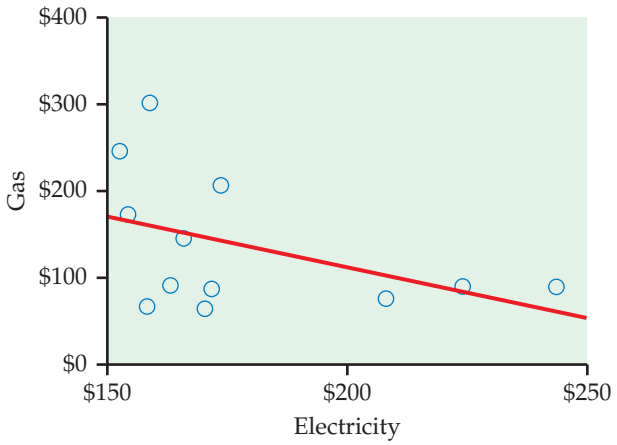


FIGURE 2.27 Scatterplot with least-squares regression line for predicting a household's monthly charges for gas using its monthly charges for electricity, Example 2.31.

Correlations that are due to lurking variables are sometimes called "nonsense correlations." The correlation is real. What is nonsense is the suggestion that the variables are directly related so that changing one of the variables *causes* changes in the other. The question of causation is important enough to merit separate treatment in Section 2.7. For now, just *remember that an association between two variables x and y can reflect many types of relationships among x, y, and one or more lurking variables*.

ASSOCIATION DOES NOT IMPLY CAUSATION

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

Lurking variables sometimes create a correlation between x and y, as in Examples 2.30 and 2.31. *When you observe an association between two variables, always ask yourself if the relationship that you see might be due to a lurking variable*. As in Example 2.31, time is often a likely candidate.

Beware of correlations based on averaged data

Regression or correlation studies sometimes work with averages or other measures that combine information from many individuals. For example, if we plot the average height of young children against their age in months, we will see a very strong positive association with correlation near 1. But individual children of the same age vary a great deal in height. A plot of height against age for individual children will show much more scatter and lower correlation than the plot of average height against age.

A correlation based on averages over many individuals is usually higher than the correlation between the same variables based on data for individuals. This fact reminds us again of the importance of noting exactly what variables a statistical study involves.

Beware of restricted ranges

restricted-range problem

The range of values for the explanatory variable in a regression can have a large impact on the strength of the relationship. For example, if we use age as a predictor of reading ability for a sample of students in the third grade, we will probably see little or no relationship. However, if our sample includes students from grades 1 through 8, we would expect to see a relatively strong relationship. We call this phenomenon the **restricted-range problem**.

EXAMPLE 2.32

A test for job applicants. Your company gives a test of cognitive ability to job applicants before deciding whom to hire. Your boss has asked you to use company records to see if this test really helps predict the performance ratings of employees. The restricted-range problem may make it difficult to see a strong relationship between test scores and performance ratings. The current employees were selected by a mechanism that is likely to result in scores that tend to be higher than those of the entire pool of applicants.

BEYOND THE BASICS

Data Mining

Chapters 1 and 2 of this text are devoted to the important aspect of statistics called *exploratory data analysis* (EDA). We use graphs and numerical summaries to examine data, searching for patterns and paying attention to striking deviations from the patterns we find. In discussing regression, we advanced to using the pattern we find (in this case, a linear pattern) for prediction.

Suppose now that we have a truly enormous database, such as all purchases recorded by the cash register scanners of a national retail chain during the past week. Surely this treasure chest of data contains patterns that might guide business decisions. If we could see clearly the types of activewear preferred in large California cities and compare the preferences of small Midwest cities—right now, not at the end of the season—we might improve profits in both parts of the country by matching stock with demand. This sounds much like EDA, and indeed it is. Exploring really large databases in the hope of finding useful patterns is called **data mining**. Here are some distinctive features of data mining:

data mining

- When you have terrabytes of data, even straightforward calculations and graphics become very time-consuming. So efficient algorithms are very important.
- The structure of the database and the process of storing the data (the fashionable term is *data warehousing*), perhaps by unifying data scattered across many departments of a large corporation, require careful consideration.
- Data mining requires automated tools that work based on only vague queries by the user. The process is too complex to do step-by-step as we have done in EDA.

All these features point to the need for sophisticated computer science as a basis for data mining. Indeed, data mining is often viewed as a part of computer science. Yet many statistical ideas and tools—mostly tools for dealing with multidimensional data, not the sort of thing that appears in a first statistics course—are very helpful. Like many other modern developments, data mining crosses the boundaries of traditional fields of study.

Do remember that the perils we associate with blind use of correlation and regression are yet more perilous in data mining, where the fog of an immense database can prevent clear vision. Extrapolation, ignoring lurking variables, and confusing association with causation are traps for the unwary data miner.

SECTION 2.5 SUMMARY

- You can examine the fit of a regression line by plotting the **residuals**, which are the differences between the observed and predicted values of y. Be on the lookout for points with unusually large residuals and also for nonlinear patterns and uneven variation about the line.
- Also look for **influential observations**, individual points that substantially change the regression line. Influential observations are often outliers in the x direction, but they need not have large residuals.
- Correlation and regression must be **interpreted with caution**. Plot the data to be sure that the relationship is roughly linear and to detect outliers and influential observations.
- **Lurking variables** may explain the relationship between the explanatory and response variables. Correlation and regression can be misleading if you ignore important lurking variables.
- We cannot conclude that there is a cause-and-effect relationship between two variables just because they are strongly associated. **High correlation does not imply causation**.
- **A correlation based on averages** is usually higher than if we used data for individuals.

SECTION 2.5 EXERCISES

For Exercise 2.92, see page 125; and for Exercise 2.93 see page 125.

2.94 Bone strength. Exercise 2.24 (page 97) gives the bone strengths of the dominant and the nondominant arms for 15 men who were controls in a study. The least-squares regression line for these data is

$$\text{dominant} = 2.74 + (0.936 \times \text{nondominant})$$

Here are the data for four cases:

ID	Nondominant	Dominant	ID	Nondominant	Dominant
5	12.0	14.8	7	12.3	13.1
6	20.0	19.8	8	14.4	17.5

Calculate the residuals for these four cases. ARMSTR

2.95 Bone strength for baseball players. Refer to the previous exercise. Similar data for baseball players is given in Exercise 2.25 (page 98). The equation of the least-squares line for the baseball players is

$$\text{dominant} = 0.886 + (1.373 \times \text{nondominant})$$

Here are the data for the first four cases:

ID	Nondominant	Dominant	ID	Nondominant	Dominant
20	21.0	40.3	22	31.5	36.9
21	14.6	20.8	23	14.9	21.2

Calculate the residuals for these four cases. ARMSTR

2.96 Least-squares regression for radioactive decay. Refer to Exercise 2.32 (page 99) for the data on radioactive decay of barium-137m. Here are the data: DECAY

Time	1	3	5	7
Count	578	317	203	118

(a) Using the least-squares regression equation

$$\text{count} = 602.8 - (74.7 \times \text{time})$$

and the observed data, find the residuals for the counts.

(b) Plot the residuals versus time.

(c) Write a short paragraph assessing the fit of the least-squares regression line to these data based on your interpretation of the residual plot.

2.97 Least-squares regression for the log counts. Refer to Exercise 2.33 (page 99), where you analyzed the radioactive decay of barium-137m data using log counts. Here are the data: DECAY

Time	1	3	5	7
Log count	6.35957	5.75890	5.31321	4.77068

(a) Using the least-squares regression equation

$$\text{log count} = 6.593 - (0.2606 \times \text{time})$$

and the observed data, find the residuals for the counts.

(b) Plot the residuals versus time.

(c) Write a short paragraph assessing the fit of the least-squares regression line to these data based on your interpretation of the residual plot.

2.98 College students by state. Refer to Exercise 2.75 (page 119), where you examined the relationship between the number of undergraduate college students and the populations for the 50 states. COLLEGE

(a) Make a scatterplot of the data with the least-squares regression line.

(b) Plot the residuals versus population.

(c) Focus on California, the state with the largest population. Is this state an outlier when you consider only the distribution of population? Explain your answer and describe what graphical and numerical summaries you used as the basis for your conclusion.

(d) Is California an outlier in the distribution of undergraduate college students? Explain your answer and describe what graphical and numerical summaries you used as the basis for your conclusion.

(e) Is California an outlier when viewed in terms of the relationship between number of undergraduate college students and population? Explain your answer and describe what graphical and numerical summaries you used as the basis for your conclusion.

(f) Is California influential in terms of the relationship between number of undergraduate college students and population? Explain your answer and describe what graphical and numerical summaries you used as the basis for your conclusion.

2.99 College students by state using logs. Refer to the previous exercise. Answer parts (a) through (f) for that exercise using the logs of both variables. Write a short paragraph summarizing your findings and comparing them with those from the previous exercise. COLLEGE

2.100 Make some scatterplots. For each of the following scenarios, make a scatterplot with 10 observations that show a moderate positive association, plus one that illustrates the unusual case. Explain each of your answers.

(a) An outlier in x that is not influential for the regression.

(b) An outlier in x that is influential for the regression.

(c) An influential observation that is not an outlier in x.

(d) An observation that is influential for the intercept but not for the slope.

2.101 What's wrong? Each of the following statements contains an error. Describe each error and explain why the statement is wrong.

(a) An influential observation will always have a large residual.

(b) High correlation is never present when there is causation.

(c) If we have data at values of x equal to 1, 2, 3, 4, and 5, and we try to predict the value of y for $x = 2.5$ using a least-squares regression equation, we are doing an extrapolation.

2.102 What's wrong? Each of the following statements contains an error. Describe each error and explain why the statement is wrong.

(a) If the residuals are all negative, this implies that there is a negative relationship between the response variable and the explanatory variable.

(b) A strong negative relationship does not imply that there is an association between the explanatory variable and the response variable.

(c) A lurking variable is always something that can be measured.

2.103 Internet use and babies. Exercise 2.34 (page 99) explores the relationship between Internet use and birthrate for 106 countries. Figure 2.13 (page 99) is a scatterplot of the data. It shows a negative association between these two variables. Do you think that this plot indicates that Internet use causes people to have fewer babies? Give another possible explanation for why these two variables are negatively associated. INBIRTH

2.104 A lurking variable. The effect of a lurking variable can be surprising when individuals are divided into groups. In recent years, the mean SAT score of all high school seniors has increased. But the mean SAT score has decreased for students at each level of high school grades (A, B, C, and so on). Explain how grade inflation in high school (the lurking variable) can account for this pattern. *A relationship that holds for each group within a population need not hold for the population as a whole. In fact, the relationship can even change direction.*

2.105 How's your self-esteem? People who do well tend to feel good about themselves. Perhaps helping people feel good about themselves will help them do better in their jobs and in life. For a time, raising self-esteem became a goal in many schools and companies. Can you think of explanations for the association between high self-esteem and good performance other than "Self-esteem causes better work"?

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

2.107 Does herbal tea help nursing-home residents? A group of college students believes that herbal tea has remarkable powers. To test this belief, they make weekly visits to a local nursing home, where they visit with the residents and serve them herbal tea. The nursing-home staff reports that after several months many of the residents are healthier and more cheerful. We should commend the students for their good deeds but doubt that herbal tea helped the residents. Identify the explanatory and response variables in this informal study. Then explain what lurking variables account for the observed association.

2.108 Price and ounces. In Example 2.2 (page 80) and Exercise 2.3 (page 82), we examined the relationship between the price and the size of a Mocha Frappuccino. The 12-ounce Tall drink costs $3.95, the 16-ounce Grande is $4.45, and the 24-ounce Venti is $4.95.

(a) Plot the data and describe the relationship. (Explain why you should plot size in ounces on the x axis.)

(b) Find the least-squares regression line for predicting the price using size. Add the line to your plot.

(c) Draw a vertical line from the least-squares line to each data point. This gives a graphical picture of the residuals.

(d) Find the residuals and verify that they sum to zero.

(e) Plot the residuals versus size. Interpret this plot.

2.109 Use the applet. It isn't easy to guess the position of the least-squares line by eye. Use the *Correlation and Regression* applet to compare a line you draw with the least-squares line. Click on the scatterplot to create a group of 15 points from lower left to upper right with a clear, positive straight-line pattern (correlation around 0.6). Click the "Draw line" button and use the mouse to draw a line through the middle of the cloud of points from lower left to upper right. Note the "thermometer" that appears above the plot. The red portion is the sum of the squared vertical distances from the points in the plot to the least-squares line. The green portion is the "extra" sum of squares for your line—it shows by how much your line misses the smallest possible sum of squares.

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

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

2.110 Use the applet. Go to the *Correlation and Regression* applet. Click on the scatterplot to create a group of 12 points in the lower-right corner of the scatterplot with a strong straight-line pattern (correlation about −0.8). Now click the "Show least-squares line" box to display the regression line.

(a) Add one point at the upper left that is far from the other 12 points but exactly on the regression line. Why does this outlier have no effect on the line even though it changes the correlation?

(b) Now drag this last point down until it is opposite the group of 12 points. You see that one end of the least-squares line chases this single point, while the other end remains near the middle of the original group of 12. What makes the last point so influential?

2.111 Education and income. There is a strong positive correlation between years of education and

income for economists employed by business firms. (In particular, economists with doctorates earn more than economists with only a bachelor's degree.) There is also a strong positive correlation between years of education and income for economists employed by colleges and universities. But when all economists are considered, there is a *negative* correlation between education and income. The explanation for this is that business pays high salaries and employs mostly economists with bachelor's degrees, while colleges pay lower salaries and employ mostly economists with doctorates. Sketch a scatterplot with two groups of cases (business and academic) that illustrates how a strong positive correlation within each group and a negative overall correlation can occur together.

2.112 Dangers of not looking at a plot. Table 2.1 (page 122) presents four sets of data prepared by the statistician Frank Anscombe to illustrate the dangers of calculating without first plotting the data.[22] ANSC

(a) Use x to predict y for each of the four data sets. Find the predicted values and residuals for each of the four regression equations.

(b) Plot the residuals versus x for each of the four data sets.

(c) Write a summary of what the residuals tell you for each data set, and explain how the residuals help you to understand these data.

2.6 Data Analysis for Two-Way Tables

When you complete this section, you will be able to:

- Identify the row variable, the column variable, and the cells in a two-way table.
- Find and interpret the joint distribution in a two-way table.
- Find and interpret the marginal distributions in a two-way table.
- Use the conditional distributions to describe the relationship displayed in a two-way table.
- Determine the joint distribution, the marginal distributions, and the conditional distributions in a two-way table from software output.
- Interpret examples of Simpson's paradox.

LOOK BACK
quantitative and categorical variables, p. 3

When we study relationships between two variables, one of the first questions we ask is whether each variable is quantitative or categorical. For two quantitative variables, we use a scatterplot to examine the relationship, and we fit a line to the data if the relationship is approximately linear. If one of the variables is quantitative and the other is categorical, we can use the methods in Chapter 1 to describe the distribution of the quantitative variable for each value of the categorical variable. This leaves us with the situation where both variables are categorical. In this section, we discuss methods for studying these relationships.

Some variables—such as sex, race, and occupation—are inherently categorical. Other categorical variables are created by grouping values of a quantitative variable into classes. Published data are often reported in grouped form to save space. To describe categorical data, we use the *counts* (frequencies) or *percents* (relative frequencies) of individuals that fall into various categories.

The two-way table

two-way table

A key idea in studying relationships between two variables is that both variables must be measured on the same individuals or cases. When both variables are categorical, the raw data are summarized in a **two-way table** that gives counts of observations for each combination of values of the two categorical variables. Here is an example.

EXAMPLE 2.33


IOM

Is the calcium intake adequate? Young children need calcium in their diet to support the growth of their bones. The Institute of Medicine provides guidelines for how much calcium should be consumed for people of different ages.[23] One study examined whether or not a sample of children consumed an adequate amount of calcium, based on these guidelines. Because there are different requirements for children aged 5 to 10 years and for children aged 11 to 13 years of age, the children were classified into these two age groups. For each student, his or her calcium intake was classified as meeting or not meeting the requirement. There were 2029 children in the study. Here are the data:[24]

Two-way table for "met requirement" and age

	Age (years)	
Met requirement	5 to 10	11 to 13
No	194	557
Yes	861	417

We see that 194 children aged 5 to 10 did not meet the calcium requirement, and 861 children aged 5 to 10 years met the calcium requirement.

USE YOUR KNOWLEDGE

2.113 Read the table. Refer to the table in Example 2.33. How many children aged 11 to 13 met the requirement? How many did not?


IOM

row variable

column variable

cell

For the calcium requirement example, we could view age as an explanatory variable and "met requirement" as a response variable. This is why we put age in the columns (like the x axis in a scatterplot) and "met requirement" in the rows (like the y axis in a scatterplot). We call "met requirement" the **row variable** because each horizontal row in the table describes whether or not the requirement was met. Age is the **column variable** because each vertical column describes one age group. Each combination of values for these two variables is called a **cell**. For example, the cell corresponding to children who are 5 to 10 years old and who have not met the requirement contains the number 194. This table is called a 2×2 table because there are two rows and two columns.

To describe relationships between two categorical variables, we compute different types of percents. Our job is easier if we expand the basic two-way table by adding various totals. We illustrate the idea with our calcium requirement example.

EXAMPLE 2.34


IOM

Add the margins to the table. We expand the table in Example 2.33 by adding the totals for each row, for each column, and the total number of all the observations. Here is the result:

Two-way table for "met requirement" and age

	Age (years)		
Met requirement	5 to 10	11 to 13	Total
No	194	557	751
Yes	861	417	1278
Total	1055	974	2029

In this study there were 1055 children aged 5 to 10. The total number of children who did not meet the calcium requirement is 751, and the total number of children in the study is 2029.

USE YOUR KNOWLEDGE

2.114 Read the margins of the table. How many children aged 11 to 13 were subjects in the calcium requirement study? What is the total number of children who met the calcium requirement?

In this example, be sure that you understand how the table is obtained from the raw data. Think about a data file with one line per subject. There would be 2029 lines or records in this data set. In the two-way table, each individual is counted once and only once. As a result, the sum of the counts in the table is the total number of individuals in the data set. *Most errors in the use of categorical-data methods come from a misunderstanding of how these tables are constructed.*

Joint distribution

We are now ready to compute some proportions that help us understand the data in a two-way table. Suppose that we are interested in the children aged 5 to 10 years who do not meet the calcium requirement. The proportion of these is simply 194 divided by 2029, or 0.0956. We would estimate that 9.56% of children in the population from which this sample was drawn are 5- to 10-year-olds who do not meet the calcium requirement. For each cell, we can compute a proportion by dividing the cell entry by the total sample size. The collection of these proportions is the **joint distribution** of the two categorical variables.

joint distribution

EXAMPLE 2.35

IOM

The joint distribution. For the calcium requirement example, the joint distribution of age "met requirement" and age is

Joint distribution of "met requirement" and age		
	Age (years)	
Met requirement	5 to 10	11 to 13
No	0.0956	0.2745
Yes	0.4243	0.2055

Because this is a distribution, the sum of the proportions should be 1. For this example the sum is 0.9999. The difference is due to roundoff error.

USE YOUR KNOWLEDGE

2.115 Explain the computation. Explain how the entry for the children aged 5 to 10 who met the calcium requirement in Example 2.35 is computed from the table in Example 2.34.

IOM

How might we use the information in the joint distribution for this example? Suppose that we were to develop an outreach unit to increase the consumption of calcium. The distribution suggests that the older students should be targeted if we have to make a choice because of limited funds. Children who are 11 to 13 years old and do not meet the calcium requirement are 27.45 of the total;

AU: %?

AU: %?

however, children who are 5 to 10 years old and do not meet the requirement are only 9.56 of the total. For other uses of these data, we may need to calculate different numerical summaries. Let's look at the distribution of age.

Marginal distributions

marginal distribution

When we examine the distribution of a single variable in a two-way table, we are looking at a **marginal distribution**. There are two marginal distributions, one for each categorical variable in the two-way table. They are very easy to compute.

EXAMPLE 2.36

IOM

The marginal distribution of age. Look at the table in Example 2.34. The total numbers of children aged 5 to 10 and children aged 11 to 13 are given in the bottom row, labeled "Total." Our sample has 1055 children aged 5 to 10 and 974 children aged 11 to 13. To find the marginal distribution of age, we simply divide these numbers by the total sample size, 2029. The marginal distribution of age is

Marginal distribution of age		
	5 to 10	11 to 13
Proportion	0.52	0.48

Note that the proportions sum to 1; there is no roundoff error.

Often, we prefer to use percents rather than proportions. Here is the marginal distribution of age described with percents:

Marginal distribution of age		
	5 to 10	11 to 13
Percent	52%	48%

Which form do you prefer?

The percent of children in each age group is approximately the same. This is interesting because the first category includes six ages (5, 6, 7, 8, 9, and 10); whereas the second includes only three ages (11, 12, and 13). Recall that the age categories were chosen in this way because the Institute of Medicine defined the calcium requirement differently for these age groups. In this study, the children were selected from grades 4, 5, and 6. The distribution of ages within these grades explains the marginal distribution of age for our sample.

The other marginal distribution for this example is the distribution of "met requirement."

EXAMPLE 2.37

IOM

The marginal distribution of "met requirement." Here is the marginal distribution of "met requirement," in percents:

Marginal distribution of "met requirement"		
	No	Yes
Percent	37.01%	62.99%

USE YOUR KNOWLEDGE

2.116 Explain the marginal distribution. Explain how the marginal distribution of "met requirement" given in Example 2.37 is computed from the entries in the table given in Example 2.34.

IOM

Each marginal distribution from a two-way table is a distribution for a single categorical variable. We can use a bar graph or a pie chart to display such a distribution. For our two-way table, we will be content with numerical summaries: for example, 52% of the children are aged 5 to 10, and 37% of the children are not meeting their calcium requirement. When we have more rows or columns, the graphical displays are particularly useful.

LOOK BACK
bar graphs and pie charts, p. 9

Describing relations in two-way tables

The table in Example 2.34 contains much more information than the two marginal distributions of age alone and "met requirement" alone. We need to do a little more work to examine the relationship. *Relationships among categorical variables are described by calculating appropriate percents from the counts given.* What percents do you think we should use to describe the relationship between age and meeting the calcium requirement?

EXAMPLE 2.38

Meeting the calcium requirement for children aged 5 to 10. What percent of the children aged 5 to 10 in our sample met the calcium requirement? This is the count of the children who are 5 to 10 years old and who met the calcium requirement as a percent of the number of children who are 5 to 10 years old:

IOM

$$\frac{861}{1055} = 0.8161 = 82\%$$

USE YOUR KNOWLEDGE

2.117 Find the percent. Refer to the table in Example 2.34 (page 137). Show that the percent of children 11 to 13 years old who met the calcium requirement is 43%.

IOM

Conditional distributions

In Example 2.38, we looked at the children aged 5 to 10 alone and examined the distribution of the other categorical variable, "met requirement." Another way to say this is that we *conditioned* on the value of age, 5 to 10 years old. Similarly, we can condition on the value of age being 11 to 13 years old. When we condition on the value of one variable and calculate the distribution of the other variable, we obtain a **conditional distribution**. Note that in Example 2.38, we calculated only the percent for children aged 5 to 10 years. The complete conditional distribution gives the proportions or percents for all possible values of the conditioning variable.

conditional distribution

EXAMPLE 2.39

Conditional distribution of "met requirement" for children aged 5 to 10. For children aged 5 to 10 years, the conditional distribution of the "met requirement" variable in terms of percents is

IOM

Conditional distribution of "met requirement" for children aged 5 to 10		
	No	Yes
Percent	18.39%	81.61%

Note that we have included the percents for both of the possible values, Yes and No, of the "met requirement" variable. These percents sum to 100%.

USE YOUR KNOWLEDGE

IOM

2.118 A conditional distribution. Perform the calculations to show that the conditional distribution of "met requirement" for children aged 11 to 13 years is

Conditional distribution of "met requirement" for children aged 11 to 13		
	No	Yes
Percent	57.19%	42.81%

Comparing the conditional distributions (Example 2.39 and Exercise 2.118) reveals the nature of the association between age and meeting the calcium requirement. In this set of data, the older children are more likely to fail to meet the calcium requirement.

Bar graphs can help us to see relationships between two categorical variables. No single graph (such as a scatterplot) portrays the form of the relationship between categorical variables, and no single numerical measure (such as the correlation) summarizes the strength of an association. Bar graphs are flexible enough to be helpful, but you must think about what comparisons you want to display. For numerical measures, we must rely on well-chosen percents or on more advanced statistical methods.[25]

CAUTION

A two-way table contains a great deal of information in compact form. Making that information clear almost always requires finding percents. You must decide which percents you need. Of course, we prefer to use software to compute the joint, marginal, and conditional distributions.

EXAMPLE 2.40

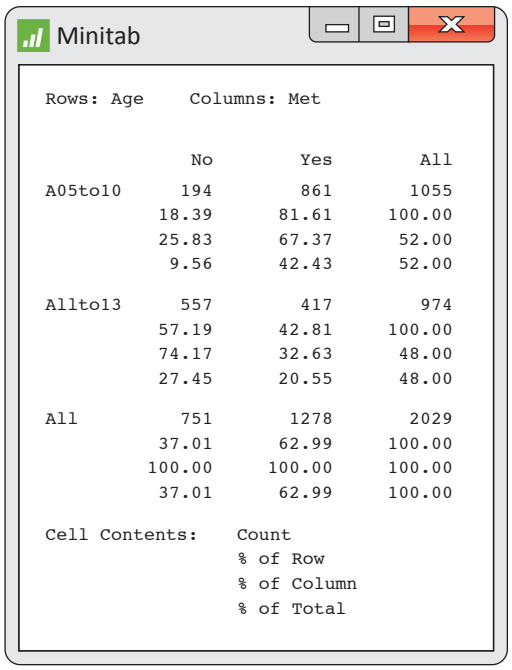
IOM

Software output. Figure 2.28 gives computer output for the data in Example 2.33 using Minitab, SPSS, and JMP. There are minor variations among software packages, but these outputs are typical of what is usually produced. Each cell in the 2×2 table has four entries. These are the count (the number of observations in the cell), the conditional distributions for rows and columns, and the joint distribution. Note that all of these are expressed as percents rather than proportions. Marginal totals and distributions are given in the rightmost column and the bottom row.

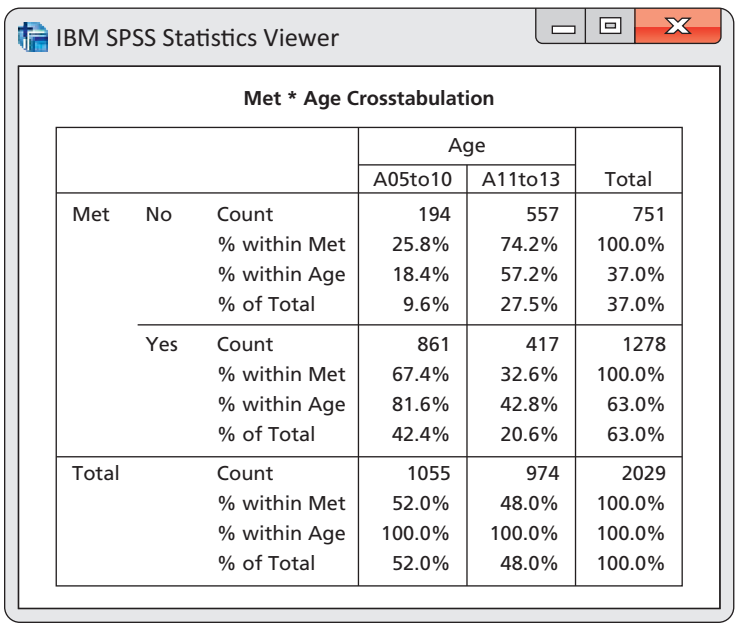
Most software packages order the row and column labels numerically or alphabetically. In general, it is better to use words rather than numbers for the column labels. This sometimes involves some additional work, but it avoids the kind of confusion that can result when you forget the real

values associated with each numerical value. You should verify that the entries in Figure 2.28 correspond to the calculations that we performed in Examples 2.34 through 2.39. In addition, verify the calculations for the conditional distributions of age for each value of "met requirement."

FIGURE 2.28 Computer output for the calcium requirement study, Example 2.40: (a) Minitab; (b) SPSS; (c) JMP.



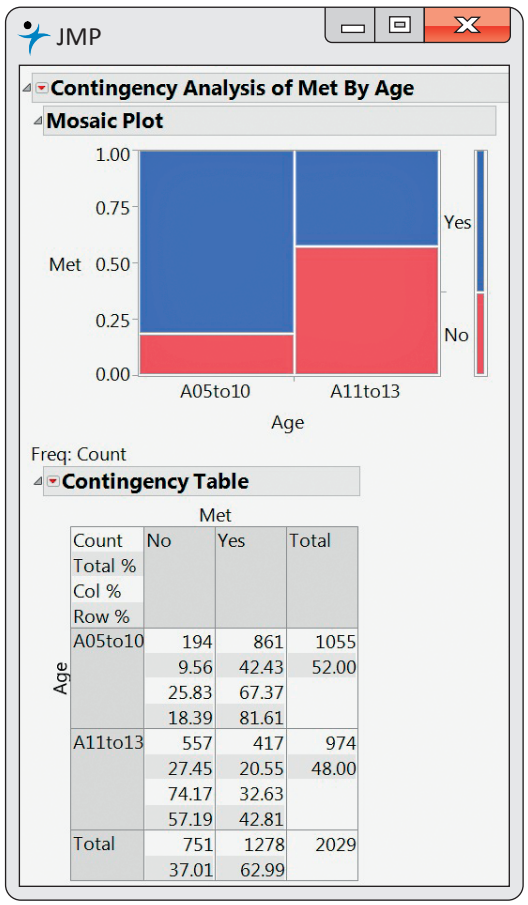
(a) Minitab



			Age		
			A05to10	A11to13	Total
Met	No	Count	194	557	751
		% within Met	25.8%	74.2%	100.0%
		% within Age	18.4%	57.2%	37.0%
		% of Total	9.6%	27.5%	37.0%
	Yes	Count	861	417	1278
		% within Met	67.4%	32.6%	100.0%
		% within Age	81.6%	42.8%	63.0%
		% of Total	42.4%	20.6%	63.0%
Total		Count	1055	974	2029
		% within Met	52.0%	48.0%	100.0%
		% within Age	100.0%	100.0%	100.0%
		% of Total	52.0%	48.0%	100.0%

(b) SPSS

FIGURE 2.28 *Continued*



(c) JMP

mosaic plot

The JMP output in Figure 2.28 includes a graphical display of the data called a **mosaic plot**. The sizes of the four boxes display the joint distribution. The narrow bar to the right shows the marginal distribution of Met and the widths of the vertical bars show the marginal distribution of age. The conditional distribution of Met for each Age is represented in each of these vertical bars by the heights of the blue and red sections. Notice that they always add to one.

Simpson's paradox

As is the case with quantitative variables, the effects of lurking variables can strongly influence relationships between two categorical variables. Here is an example that demonstrates the surprises that can await the unsuspecting consumer of data.

EXAMPLE 2.41

Which customer service representative is better? A customer service center has a goal of resolving customer questions in 10 minutes or less. Here are the records for two representatives:

	Representative	
Goal met	Alexis	Peyton
Yes	172	118
No	28	82
Total	200	200

Alexis has met the goal 172 times out of 200, a success rate of 86%. For Peyton, the success rate is 118 out of 200, or 59%. Alexis clearly has the better success rate.

Let's look at the data in a little more detail. The data summarized come from two different weeks in the year.

EXAMPLE 2.42

Look at the data more carefully. Here are the counts broken down by week:

	Week 1		Week 2	
Goal met	Alexis	Peyton	Alexis	Peyton
Yes	162	19	10	99
No	18	1	10	81
Total	180	20	20	180

For Week 1, Alexis met the goal 90% of the time (162/180), while Peyton met the goal 95% of the time (19/20). Peyton had the better performance in Week 1. What about Week 2? Here, Alexis met the goal 50% of the time (10/20), while the success rate for Peyton was 55% (99/180). Peyton again had the better performance. How does this analysis compare with the analysis that combined the counts for the two weeks? That analysis clearly showed that Alexis had the better performance, 86% versus 59%.

These results can be explained by a lurking variable, Week. The first week was during a period when the product had been in use for several months. Most of the calls to the customer service center concerned problems that had been encountered before. The representatives were trained to answer these questions and usually had no trouble in meeting the goal of resolving the problems quickly. On the other hand, the second week occurred shortly after the release of a new version of the product. Most of the calls during this week concerned new problems that the representatives had not yet encountered. Many more of these questions took longer than the 10-minute goal to resolve.

Look at the totals in the bottom row of the detailed table. During the first week, when calls were easy to resolve, Alexis handled 180 calls and Peyton handled 20. The situation was exactly the opposite during the second week, when the calls were difficult to resolve. There were 20 calls for Alexis and 180 for Peyton.

The original two-way table, which did not take account of week, was misleading. This example illustrates *Simpson's paradox*.

SIMPSON'S PARADOX

An association or comparison that holds for all of several groups can reverse direction when the data are combined to form a single group. This reversal is called **Simpson's paradox**.

The lurking variables in our Simpson's paradox example, Week and problem difficulty, are categorical. That is, they break the observations into groups by workweek. *Simpson's paradox is an extreme form of the fact that observed associations can be misleading when there are lurking variables.*

three-way table

The data in Example 2.42 are given in a **three-way table** that reports counts for each combination of three categorical variables: week, representative, and whether or not the goal was met. In our example, we constructed the three-way table by constructing two two-way tables for representative by goal, one for each week. The original table in Example 2.41 can be obtained by adding the corresponding counts for these two tables. This process is called **aggregating** the data. When we aggregated data in Example 2.41, we ignored the variable week, which then became a lurking variable. *Conclusions that seem obvious when we look only at aggregated data can become quite different when the data are examined in more detail.*

aggregation

SECTION 2.6 SUMMARY

- A **two-way table** of counts organizes data about two categorical variables. Values of the **row variable** label the rows that run across the table, and values of the **column variable** label the columns that run down the table. Two-way tables are often used to summarize large amounts of data by grouping outcomes into categories.

- The **joint distribution** of the row and column variables is found by dividing the count in each cell by the total number of observations.

- The **row totals** and **column totals** in a two-way table give the **marginal distributions** of the two variables separately. It is clearer to present these distributions as percents of the table total. Marginal distributions do not give any information about the relationship between the variables.

- To find the **conditional distribution** of the row variable for one specific value of the column variable, look only at that one column in the table. Find each entry in the column as a percent of the column total.

- There is a conditional distribution of the row variable for each column in the table. Comparing these conditional distributions is one way to describe the association between the row and the column variables. It is particularly useful when the column variable is the explanatory variable. When the row variable is explanatory, find the conditional distribution of the column variable for each row and compare these distributions.

- **Bar graphs** are a flexible means of presenting categorical data. There is no single best way to describe an association between two categorical variables.

- We present data on three categorical variables in a **three-way table**, printed as separate two-way tables for each level of the third variable. A comparison between two variables that holds for each level of a third

variable can be changed or even reversed when the data are **aggregated** by summing over all levels of the third variable. **Simpson's paradox** refers to the reversal of a comparison by aggregation. It is an example of the potential effect of lurking variables on an observed association.

SECTION 2.6 EXERCISES

For Exercise 2.113, see page 137; for 2.114, see page 138; for 2.115, see page 138; for 2.116, see page 140; for 2.117, see page 140; and for 2.118, see page 141.

2.119 Does driver's ed help? A study is planned to look at the effect of driver education programs on accidents. The driving records of all drivers under 18 in a given year will classify each driver as having taken a driver's education course or not. The drivers will also classified with respect to the number of accidents that they had in the year after they received their license. The categories are zero, one, and two or more accidents.

(a) There are two variables in this study. Do you think one is an explanatory variable and the other is a response variable? Explain your answer.

(b) Sketch a two-way table that could be used to organize the data. Which variable is the row variable? Which variable is the column variable?

(c) How many cells are in the table? Describe in words what each of the cells will contain when the data are collected.

2.120 Music and video games. You are planning a study of undergraduates in which you will examine the relationship between listening to music and playing video games. The study subjects will be asked how much time they spend in each of these activities during a typical day. The choices for both activities will be a half hour or less, more than a half hour but less than an hour, and more than an hour.

(a) There are two variables in this study. Do you think one is an explanatory variable and the other is a response variable? Explain your answer.

(b) Sketch a two-way table that could be used to organize the data. Which variable is the row variable? Which variable is the column variable?

(c) How many cells are in the table? Describe in words what each of the cells will contain when the data are collected.

2.121 Eight is enough. A healthy body needs good food, and healthy teeth are needed to chew our food so that it can nourish our bodies. The U.S. Army has recognized this fact and requires recruits to pass a dental examination. If you wanted to be a soldier in the Spanish American War, which took place in 1898, you needed to have at least eight teeth. Here is the statement of the requirement: TEETH

> *Unless an applicant has at least four sound double teeth, one above and one below on each side of the mouth, and so opposed as to serve the purpose of mastication, he should be rejected.*

A study reported the rejection data for enlistment candidates classified by age. Here are the data:[26]

	Age					
Rejected	Under 20	20 to 25	25 to 30	30 to 35	35 to 40	Over 40
Yes	68	647	1114	1783	2887	3801
No	58,884	77,992	55,597	43,994	47,569	39,985

(a) Which variable is the explanatory variable? Which variable is the response variable? Give reasons for your answer.

(b) Find the joint distribution. Write a brief summary explaining the major features of this distribution.

(c) Find the two marginal distributions. Write a brief summary explaining the major features of these distributions.

(d) Which conditional distribution would you choose to explain the relationship between these two variables? Explain your answer.

(e) Find the conditional distribution that you chose in part (d), and write a summary that includes your interpretation of the relationship based on this conditional distribution.

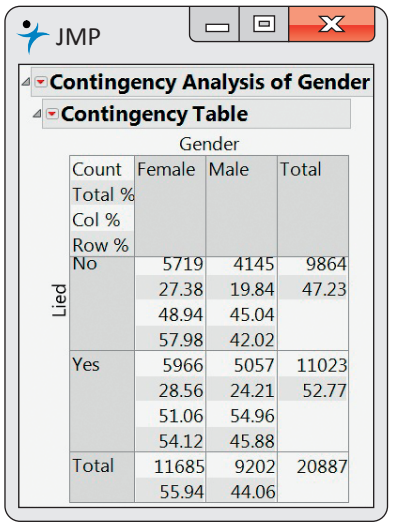
2.122 Survival and class on the *Titanic*. In Exercise 1.27 (page 24), you created a graphical summary of the number of passengers who survived classified by the accommodations that they had on the ship: first, second, or third class. Let's look at these data with a two-way table. TITANIC

(a) Create a two-way table that you could use to explore the relationship between survival and class.

(b) Which variable is the explanatory variable and which is the response variable? Give reasons for your answers.

(c) Find the two marginal distributions. Write a brief summary explaining the major features of these distributions.

(d) Which conditional distribution would you choose to explain the relationship between these two variables? Explain your answer.


JMP

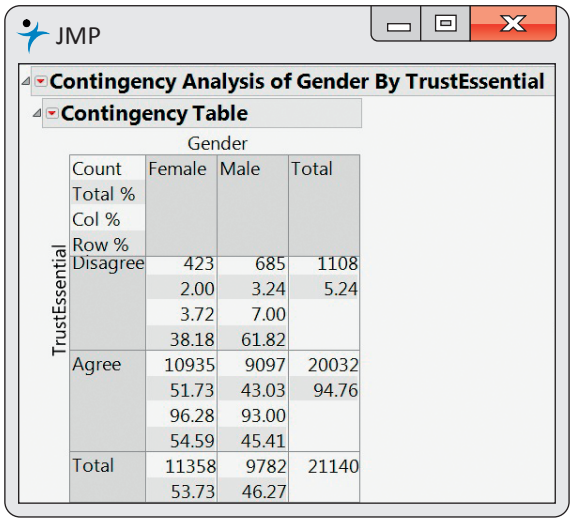
Contingency Analysis of Gender

Contingency Table

Gender

Lied	Count / Total % / Col % / Row %	Female	Male	Total
No		5719	4145	9864
		27.38	19.84	47.23
		48.94	45.04	
		57.98	42.02	
Yes		5966	5057	11023
		28.56	24.21	52.77
		51.06	54.96	
		54.12	45.88	
Total		11685	9202	20887
		55.94	44.06	

FIGURE 2.29 Computer output for the lying to a teacher data, Exercise 2.123.


JMP

Contingency Analysis of Gender By TrustEssential

Contingency Table

Gender

TrustEssential	Count / Total % / Col % / Row %	Female	Male	Total
Disagree		423	685	1108
		2.00	3.24	5.24
		3.72	7.00	
		38.18	61.82	
Agree		10935	9097	20032
		51.73	43.03	94.76
		96.28	93.00	
		54.59	45.41	
Total		11358	9782	21140
		53.73	46.27	

FIGURE 2.30 Computer output for the trust and honesty in the workplace data, Exercise 2.124.

(e) Find the conditional distribution that you chose in part (d) and, write a summary that includes your interpretation of the relationship based on this conditional distribution.

2.123 Lying to a teacher. One of the questions in a survey of high school students asked about lying to teachers.[27] The data set LYING gives the numbers of students who said that they lied to a teacher about something significant at least once during the past year, classified by sex. Figure 2.29 gives software output for these data. Use this output to analyze these data and write a report summarizing your work. Be sure to include a discussion of whether or not you consider this relationship to involve an explanatory variable and a response variable. LYING

2.124 Trust and honesty in the workplace. The students surveyed in the study described in the previous exercise were also asked whether they thought trust and honesty were essential in business and the workplace. Figure 2.30 gives software output for these data. Use this output to analyze these data and write a report summarizing your work. Be sure to include a discussion of whether or not you consider this relationship to involve an explanatory variable and a response variable. TRUST

2.125 Exercise and adequate sleep. A survey of 656 boys and girls, who were 13 to 18 years old, asked about adequate sleep and other health-related behaviors. The recommended amount of sleep is six to eight hours per night.[28] In the survey, 59% of the respondents reported that they got less than this amount of sleep on school nights. An exercise scale was developed and used to classify the students as above or below the median in this domain. Here is the 2×2 table of counts with students classified as getting or not getting adequate sleep and by the exercise variable: SLEEP

	Exercise	
Enough sleep	High	Low
Yes	151	115
No	148	242

(a) Find the distribution of adequate sleep for the high exercisers.

(b) Do the same for the low exercisers.

(c) Summarize the relationship between adequate sleep and exercise using the results of parts (a) and (b).

2.126 Adequate sleep and exercise. Refer to the previous exercise. SLEEP

(a) Find the distribution of exercise for those who get adequate sleep.

(b) Do the same for those who do not get adequate sleep.

(c) Write a short summary of the relationship between adequate sleep and exercise using the results of parts (a) and (b).

(d) Compare this summary with your summary from part (c) of the previous exercise. Which do you prefer? Give a reason for your answer.

2.127 Which hospital is safer? Insurance companies and consumers are interested in the performance of

hospitals. The government releases data about patient outcomes in hospitals that can be useful in making informed health care decisions. Here is a two-way table of data on the survival of patients after surgery in two hospitals. All patients undergoing surgery in a recent time period are included. "Survived" means that the patient lived at least six weeks following surgery.

HOSP

	Hospital A	Hospital B
Died	63	16
Survived	2037	784
Total	2100	800

What percent of Hospital A patients died? What percent of Hospital B patients died? These are the numbers one might see reported in the media.

2.128 Patients in "poor" or "good" condition. Refer to the previous exercise. Not all surgery cases are equally serious, however. Patients are classified as being in either "poor" or "good" condition before surgery. Here are the data broken down by patient condition. The entries in the original two-way table are just the sums of the "poor" and "good" entries in this pair of tables.

HOSP

	Good condition	
	Hospital A	Hospital B
Died	6	8
Survived	594	592
Total	600	600

	Poor condition	
	Hospital A	Hospital B
Died	57	8
Survived	1443	192
Total	1500	200

(a) Find the death rate for Hospital A patients who were classified as "poor" before surgery. Do the same for Hospital B. In which hospital do "poor" patients fare better?

(b) Repeat part (a) for patients classified as "good" before surgery.

(c) What is your recommendation to someone facing surgery and choosing between these two hospitals?

(d) How can Hospital A do better in both groups, yet do worse overall? Look at the data and carefully explain how this can happen.

2.129 Complete the table. Here are the row and column totals for a two-way table with two rows and two columns:

a	b	300
c	d	200
300	200	500

Find *two different* sets of counts a, b, c, and d for the body of the table that give these same totals. This shows that the relationship between two variables cannot be obtained from the two individual distributions of the variables.

2.130 Construct a table with no association. Construct a 3×4 table of counts where there is no apparent association between the row and column variables.

2.7 The Question of Causation*

When you complete this section, you will be able to:

- Identify the differences among causation, common response, and confounding in explaining an association.
- Apply the five criteria for establishing causation.

In many studies of the relationship between two variables, the goal is to establish that changes in the explanatory variable *cause* changes in the response variable. Even when a strong association is present, however, the conclusion that this association is due to a causal link between the variables is often hard to justify. What ties between two variables (and others lurking in the background) can explain an observed association? What constitutes good evidence for causation? We begin our consideration of these questions with a set of observed associations. In each

*This section is optional.

case, there is a clear association between variable x and variable y. Moreover, the association is positive whenever the direction makes sense.

Explaining association

EXAMPLE 2.43

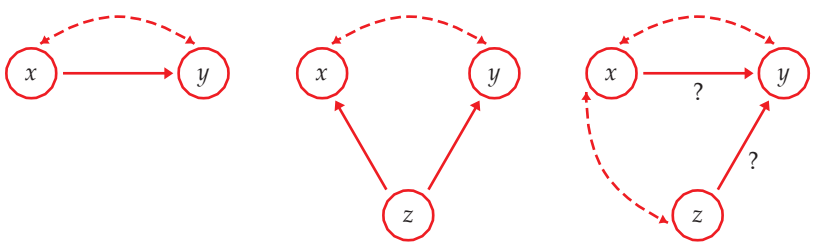
AU: data icon removed here. Okay?

Observed associations. Here are some examples of observed association between x and y:

1. x = mother's body mass index
   y = daughter's body mass index
2. x = amount of the artificial sweetener saccharin in a rat's diet
   y = count of tumors in the rat's bladder
3. x = a student's SAT score as a high school senior
   y = a student's first-year college grade point average
4. x = monthly flow of money into stock mutual funds
   y = monthly rate of return for the stock market
5. x = whether a person regularly attends religious services
   y = how long the person lives
6. x = the number of years of education a worker has
   y = the worker's income

Explaining association: Causation Figure 2.31 shows in outline form how a variety of underlying links between variables can explain association. The dashed double-arrow line represents an observed association between the variables x and y. Some associations are explained by a direct cause-and-effect link between these variables. The first diagram in Figure 2.31 shows "x causes y" by a solid arrow running from x to y.

Items 1 and 2 in Example 2.43 are examples of direct causation. *Even when direct causation is present, very often it is not a complete explanation of an association between two variables.* The best evidence for causation comes from experiments that actually change x while holding all other factors fixed. If y changes, we have good reason to think that x caused the change in y.



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

FIGURE 2.31 Possible explanations for an observed association. The dashed double-arrow lines show an association. The solid arrows show a cause-and-effect link. The variable x is explanatory, y is a response variable, and z is a lurking variable.

common response

Explaining association: Common response "Beware of the lurking variable" is good advice when thinking about an association between two variables. The second diagram in Figure 2.31 illustrates **common response**. The observed association between the variables x and y is explained by a lurking variable z. Both x and y change in response to changes in z. This common response creates an association even though there may be no direct causal link between x and y.

The third and fourth items in Example 2.43 illustrate how common response can create an association. What would be a good candidate for the variable z in these two examples?

Explaining association: Confounding For the first item in Example 2.43, we expect that inheritance explains part of the association between the body mass indexes (BMIs) of daughters and their mothers. Can we use r or r^2 to say how much inheritance contributes to the daughters' BMIs? No. It may well be that mothers who are overweight also set an example of little exercise, poor eating habits, and lots of television. Their daughters pick up these habits to some extent, so the influence of heredity is mixed up with influences from the girls' environment. We call this mixing of influences *confounding*.

CONFOUNDING

Two variables are **confounded** when their effects on a response variable cannot be distinguished from each other. The confounded variables may be either explanatory variables or lurking variables or both.

When many uncontrolled variables are related to a response variable, you should always ask whether or not confounding of several variables prevents you from drawing conclusions about causation. The third diagram in Figure 2.31 illustrates confounding. Both the explanatory variable x and the lurking variable z may influence the response variable y. Because x is confounded with z, we cannot distinguish the influence of x from the influence of z. We cannot say how strong the direct effect of x on y is. In fact, it can be hard to say if x influences y at all.

The last two associations in Example 2.43 (Items 5 and 6) are explained in part by confounding. What would be a good candidate for the confounding variable z in these two examples?

Many observed associations are at least partly explained by lurking variables. Both common response and confounding involve the influence of a lurking variable (or variables) z on the response variable y. The distinction between these two types of relationship is less important than the common element, the influence of lurking variables. The most important lesson of these examples is one we have already emphasized: **even a very strong association between two variables is not by itself good evidence that there is a cause-and-effect link between the variables**.

Establishing causation

How can a direct causal link between x and y be established? The best method—indeed, the only fully compelling method—of establishing causation is to conduct a carefully designed experiment in which the effects of possible lurking variables are controlled. Chapter 3 explains how to design convincing experiments.

Many of the sharpest disputes in which statistics plays a role involve questions of causation that cannot be settled by experiment. Does gun control reduce violent crime? Does living near power lines cause cancer? Has "outsourcing" work to overseas locations reduced overall employment in the United States? All these questions have become public issues. All concern associations among variables. And all have this in common: they try to pinpoint cause and effect in a setting involving complex relations among many interacting variables. Common response and confounding, along with the number of potential lurking variables, make observed associations misleading. Experiments are not possible for ethical or practical reasons. We can't assign some people to live near power lines or compare the same nation with and without strong gun controls.

EXAMPLE 2.44

Power lines and leukemia. Electric currents generate magnetic fields. So living with electricity exposes people to magnetic fields. Living near power lines increases exposure to these fields. Really strong fields can disturb living cells in laboratory studies. Some people claim that the weaker fields we experience if we live near power lines cause leukemia in children.

It isn't ethical to do experiments that expose children to magnetic fields. It's hard to compare cancer rates among children who happen to live in more and less exposed locations because leukemia is rare and locations vary in many ways other than magnetic fields. We must rely on studies that compare children who have leukemia with children who don't.

A careful study of the effect of magnetic fields on children took five years and cost $5 million. The researchers compared 638 children who had leukemia and 620 who did not. They went into the homes and actually measured the magnetic fields in the children's bedrooms, in other rooms, and at the front door. They recorded facts about nearby power lines for the family home and also for the mother's residence when she was pregnant. Result: no evidence of more than a chance connection between magnetic fields and childhood leukemia.[29]

"No evidence" that magnetic fields are connected with childhood leukemia doesn't prove that there is no risk. It says only that a careful study could not find any risk that stands out from the play of chance that distributes leukemia cases across the landscape. Critics continue to argue that the study failed to measure some lurking variables or that the children studied don't fairly represent all children. Nonetheless, a carefully designed study comparing children with and without leukemia is a great advance over haphazard and sometimes emotional counting of cancer cases.

EXAMPLE 2.45

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

Doctors had long observed that most lung cancer patients were smokers. Comparison of smokers and similar nonsmokers showed a very strong association between smoking and death from lung cancer. Could the association be due to common response? Might there be, for example,

a genetic factor that predisposes people both to nicotine addiction and to lung cancer? Smoking and lung cancer would then be positively associated even if smoking had no direct effect on the lungs. Or perhaps confounding is to blame. It might be that smokers live unhealthy lives in other ways (diet, alcohol, lack of exercise) and that some other habit confounded with smoking is a cause of lung cancer. How were these objections overcome?

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

- *The association is strong.* The association between smoking and lung cancer is very strong.
- *The association is consistent.* Many studies of different kinds of people in many countries link smoking to lung cancer. That reduces the chance that a lurking variable specific to one group or one study explains the association.
- *Higher doses are associated with stronger responses.* People who smoke more cigarettes per day or who smoke over a longer period get lung cancer more often. People who stop smoking reduce their risk.
- *The alleged cause precedes the effect in time.* Lung cancer develops after years of smoking.
- *The alleged cause is plausible.* Experiments show that tars from cigarette smoke cause cancer when applied to the backs of mice.

Medical authorities do not hesitate to say that smoking causes lung cancer. The U.S. Surgeon General states that cigarette smoking is "the largest avoidable cause of death and disability in the United States."[30] The evidence for causation is strong—but it is not as strong as the evidence provided by well-designed experiments.

SECTION 2.7 SUMMARY

- Some observed associations between two variables are due to a **cause-and-effect** relationship between these variables, but others are explained by **lurking variables**.
- The effect of lurking variables can operate through **common response** if changes in both the explanatory and the response variables are caused by changes in lurking variables. **Confounding** of two variables (either explanatory or lurking variables or both) means that we cannot distinguish their effects on the response variable.
- Establishing that an association is due to causation is best accomplished by conducting an **experiment** that changes the explanatory variable while controlling other influences on the response.
- In the absence of experimental evidence, be cautious in accepting claims of causation. Good evidence of causation requires (1) a strong association, (2) that appears consistently in many studies, (3) that has higher doses associated with stronger responses, (4) with the alleged cause preceding the effect in time, and (5) that is plausible.

SECTION 2.7 EXERCISES

2.131 Examples of association. Give three examples of association: one due to causation, one due to common response, and one due to confounding. Use your examples to write a short paragraph explaining the differences among these three explanations for an observed association.

2.132 The five criteria for establishing causation. Consider the five criteria for establishing causation. Explain how each of these, if not established seriously, weakens the case that an association is due to causation.

2.133 Iron and anemia. A lack of adequate iron in the diet is associated with anemia, a condition in which the body does not have enough red blood cells. However, anemia is also associated with malaria and infections with worms called helminths. Discuss these observed associations using the framework of Figure 2.31. (page 149)

2.134 Stress and lack of sleep in college students. Studies of college students have shown that stress and lack of sleep are associated. Do you think that lack of sleep causes stress or that stress causes lack of sleep? Write a short paragraph summarizing your opinions.

2.135 Online courses. Many colleges offer online versions of some courses that are also taught in the classroom. It often happens that the students who enroll in the online version do better than the classroom students on the course exams. This does not show that online instruction is more effective than classroom teaching because the people who sign up for online courses are often quite different from the classroom students. Suggest some student characteristics that you think could be confounded with online versus classroom. Use a diagram like Figure 2.31(c) to illustrate your ideas.

2.136 Marriage and income. Data show that men who are married, and also divorced or widowed men, earn quite a bit more than men who have never been married. This does not mean that a man can raise his income by getting married. Suggest several lurking variables that you think are confounded with marital status and that help explain the association between marital status and income. Use a diagram like Figure 2.31(c) to illustrate your ideas.

2.137 Exercise and self-confidence. A college fitness center offers an exercise program for staff members who choose to participate. The program assesses each participant's fitness, using a treadmill test, and also administers a personality questionnaire. There is a moderately strong positive correlation between fitness score and score for self-confidence. Is this good evidence that improving fitness increases self-confidence? Explain why or why not.

2.138 Computer chip manufacturing and miscarriages. A study showed that women who work in the production of computer chips have abnormally high numbers of miscarriages. The union claimed that exposure to chemicals used in production caused the miscarriages. Another possible explanation is that these workers spend most of their work time standing up. Illustrate these relationships in a diagram like one of those in Figure 2.31.

2.139 Hospital size and length of stay. A study shows that there is a positive correlation between the size of a hospital (measured by its number of beds x) and the median number of days y that patients remain in the hospital. Does this mean that you can shorten a hospital stay by choosing a small hospital? Use a diagram like one of those in Figure 2.31 to explain the association.

2.140 Watching TV and low grades. Children who watch many hours of television get lower grades in school, on average, than those who watch less TV. Explain clearly why this fact does not show that watching TV *causes* poor grades. In particular, suggest some other variables that may be confounded with heavy TV viewing and may contribute to poor grades.

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

2.142 Exercise and mortality. A sign in a fitness center says, "Mortality is halved for men over 65 who walk at least 2 miles a day."

(a) Mortality is eventually 100% for everyone. What do you think "mortality is halved" means?

(b) Assuming that the claim is true, explain why this fact does not show that exercise *causes* lower mortality.

2.143 Effect of a math skills refresher initiative. Students enrolling in an elementary statistics course take a pretest that assesses their math skills. Those who receive low scores are given the opportunity to take three one-hour refresher sessions designed to review the basic math skills needed for the statistics course. Those who took the refresher sessions performed worse than those who did not on the final exam in the statistics course. Can you conclude that the refresher course has a negative impact on performance in the statistics course? Explain your answer.

CHAPTER 2 EXERCISES

2.144 Dwelling permits and sales for 23 countries. The Organisation for Economic Co-operation and Development collects data on main economic indicators (MEIs) for many countries. Each variable is recorded as an index with the year 2000 serving as a base year. This means that the variable for each year is reported as a ratio of the value for the year divided by the value for 2000. Use of indices in this way makes it easier to compare values for different countries. Table 2.3 gives the values of three MEIs for 23 countries.[31] MEIS

(a) Make a scatterplot with sales as the response variable and permits issued for new dwellings as the explanatory variable. Describe the relationship. Are there any outliers or influential observations?

(b) Find the least-squares regression line and add it to your plot.

(c) Interpret the slope of the line in the context of this exercise.

(d) Interpret the intercept of the line in the context of this exercise. Explain whether or not this interpretation is useful in explaining the relationship between these two variables.

(e) What is the predicted value of sales for a country that has an index of 224 for dwelling permits?

AU: "European" deleted (since list is not of all European countries). Okay?

TABLE 2.3 Dwelling Permits, Sales, and Production for 21 Countries

Country	Dwelling permits	Sales	Production
Australia	116	137	109
Belgium	125	105	112
Canada	224	122	101
Czech Republic	178	134	162
Denmark	121	126	109
Finland	105	136	125
France	145	121	104
Germany	54	100	119
Greece	117	136	102
Hungary	109	140	155
Ireland	92	123	144
Japan	86	99	109
Korea	158	110	156
Luxembourg	145	161	118
Netherlands	160	107	109
New Zealand	127	139	112
Norway	125	136	94
Poland	163	139	159
Portugal	53	112	105
Spain	122	123	108
Sweden	180	142	116

(f) Canada has an index of 224 for dwelling permits. Find the residual for this country.

(g) What percent of the variation in sales is explained by dwelling permits?

2.145 Dwelling permits and production. Refer to the previous exercise. MEIS

(a) Make a scatterplot with production as the response variable and permits issued for new dwellings as the explanatory variable. Describe the relationship. Are there any outliers or influential observations?

(b) Find the least-squares regression line and add it to your plot.

(c) Interpret the slope of the line in the context of this exercise.

(d) Interpret the intercept of the line in the context of this exercise. Explain whether or not this interpretation is useful in explaining the relationship between these two variables.

(e) What is the predicted value of production for a country that has an index of 224 for dwelling permits?

(f) Canada has an index of 224 for dwelling permits. Find the residual for this country.

(g) What percent of the variation in production is explained by dwelling permits? How does this value compare with the value that you found in the previous exercise for the percent of variation in sales that is explained by building permits?

2.146 Sales and production. Refer to the previous two exercises. MEIS

(a) Make a scatterplot with sales as the response variable and production as the explanatory variable. Describe the relationship. Are there any outliers or influential observations?

(b) Find the least-squares regression line and add it to your plot.

(c) Interpret the slope of the line in the context of this exercise.

(d) Interpret the intercept of the line in the context of this exercise. Explain whether or not this interpretation is useful in explaining the relationship between these two variables.

(e) What is the predicted value of sales for a country that has an index of 109 for production?

(f) The Netherlands has an index of 109 for production. Find the residual for this country.

(g) What percent of the variation in sales is explained by production? How does this value compare with the percents of variation that you calculated in the two previous exercises?

Excel

	A	B	C	D
1	ProvinceOrTerritory	Population	Pct15&Under	Pct65&Over
2	Alberta	4121.7	18.3	11.4
3	British Columbia	4631.3	14.6	17.0
4	Manitoba	1282.0	18.7	14.6
5	New Brunswick	753.9	14.6	18.3
6	Newfoundland & Labrador	527.0	14.4	17.7
7	Northwest Territories	43.6	21.4	6.6
8	Nova Scotia	942.7	14.1	18.3
9	Nunavut	36.6	31.1	3.7
10	Ontario	13678.7	16.0	15.6
11	Prince Edward Island	146.3	15.9	17.9
12	Quebec	8214.7	15.4	17.1
13	Saskatchewan	1125.4	18.9	14.5
14	Yukon	36.5	16.6	10.5

FIGURE 2.32 Percent of the population over 65 years and percent of the population under 15 years in the 13 Canadian provinces and territories, Exercise 2.147.

2.147 Population in Canadian provinces and territories. Statistics Canada provides a great deal of demographic data organized in different ways.[32] Figure 2.32 gives the percent of the population aged 65 years and older and the percent aged 15 years and younger for each of the 13 Canadian provinces and territories. Figure 2.33 is a scatterplot of the percent of the population over 65 versus the percent under 15. CANADAP

(a) Write a short paragraph explaining what the plot tells you about these two demographic groups in the 13 Canadian provinces and territories.

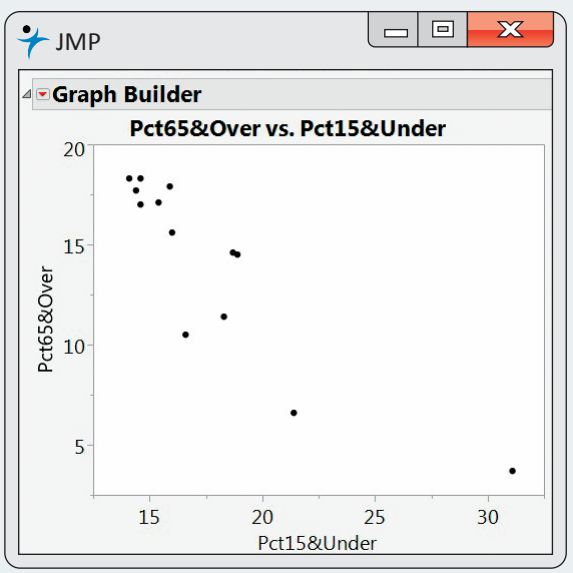


FIGURE 2.33 Scatterplot of percent of the population over 65 years versus percent of the population under 15 years for the 13 Canadian provinces and territories, Exercise 2.147.

(b) Find the correlation between the percent of the population over 65 and the percent under 15. Does the correlation give a good numerical summary of the strength of this relationship? Explain your answer.

2.148 Nunavut. Refer to the previous exercise and Figures 2.32 and 2.33. CANADAP

(a) Do you think that Nunavut is an outlier?

(b) Make a residual plot for these data. Comment on the size of the residual for Nunavut. Use this information to expand on your answer to part (a).

(c) Find the value of the correlation without Nunavut. How does this compare with the value you computed in part (b) of the previous exercise?

(d) Write a short paragraph about Nunavut based on what you have found in this exercise and the previous one.

2.149 Compare the provinces with the territories. Refer to the previous exercise. The three Canadian territories are the Northwest Territories, Nunavut, and the Yukon Territories. All the other entries in Figure 2.32 are provinces. CANADAP

(a) Generate a scatterplot of the Canadian demographic data similar to Figure 2.33 but with the points labeled "P" for provinces and "T" for territories.

(b) Use your new scatterplot to write a new summary of the demographics for the 13 Canadian provinces and territories.

2.150 Records for men and women in the 10K. Table 2.4 shows the progress of world record times (in seconds) for the 10,000-meter run for both men and women.[33] TENK

TABLE 2.4 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	1963	1695.6	1967	2286.4
1921	1840.2	1965	1659.3	1970	2130.5
1924	1835.4	1972	1658.4	1975	2100.4
1924	1823.2	1973	1650.8	1975	2041.4
1924	1806.2	1977	1650.5	1977	1995.1
1937	1805.6	1978	1642.4	1979	1972.5
1938	1802.0	1984	1633.8	1981	1950.8
1939	1792.6	1989	1628.2	1981	1937.2
1944	1775.4	1993	1627.9	1982	1895.3
1949	1768.2	1993	1618.4	1983	1895.0
1949	1767.2	1994	1612.2	1983	1887.6
1949	1761.2	1995	1603.5	1984	1873.8
1950	1742.6	1996	1598.1	1985	1859.4
1953	1741.6	1997	1591.3	1986	1813.7
1954	1734.2	1997	1587.8	1993	1771.8
1956	1722.8	1998	1582.7		
1956	1710.4	2004	1580.3		
1960	1698.8	2005	1577.5		
1962	1698.2				

(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?

2.151 Remote deposit capture. The Federal Reserve has called remote deposit capture (RDC) "the most important development the [U.S.] banking industry has seen in years." This service allows users to scan checks and to transmit the scanned images to a bank for posting.[34] In its annual survey of community banks, the American Bankers Association asked banks whether or not they offered this service.[35] Here are the results classified by the asset size (in millions of dollars) of the bank: RDCSIZE

	Offer RDC	
Asset size	Yes	No
Under $100	63	309
$101 to $200	59	132
$201 or more	112	85

Summarize the results of this survey question numerically and graphically. Write a short paragraph explaining the relationship between the size of a bank, measured by assets, and whether or not RDC is offered.

2.152 How does RDC vary across the country? The survey described in the previous exercise also classified community banks by region. Here is the 6 × 2 table of counts:[36] RDCREG

	Offer RDC	
Region	Yes	No
Northeast	28	38
Southeast	57	61
Central	53	84
Midwest	63	181
Southwest	27	51
West	61	76

Summarize the results of this survey question numerically and graphically. Write a short paragraph explaining the relationship between the location of a bank and whether or not RDC is offered.

2.153 Fields of study for college students. The following table gives the number of students (in thousands) graduating from college with degrees in several fields of study for seven countries:[37] FOS

Field of study	Canada	France	Germany	Italy	Japan	U.K.	U.S.
Social sciences, business, law	64	153	66	125	250	152	878
Science, mathematics, engineering	35	111	66	80	136	128	355
Arts and humanities	27	74	33	42	123	105	397
Education	20	45	18	16	39	14	167
Other	30	289	35	58	97	76	272

(a) Calculate the marginal totals and add them to the table.

(b) Find the marginal distribution of country and give a graphical display of the distribution.

(c) Do the same for the marginal distribution of field of study.

2.154 Fields of study by country for college students. In the previous exercise you examined data on fields of study for graduating college students from seven countries. FOS

(a) Find the seven conditional distributions giving the distribution of graduates in the different fields of study for each country.

(b) Display the conditional distributions graphically.

(c) Write a paragraph summarizing the relationship between field of study and country.

2.155 Graduation rates. One of the factors used to evaluate undergraduate programs is the proportion of incoming students who graduate. This quantity, called the graduation rate, can be predicted by other variables such as the SAT or ACT scores and the high school records of the incoming students. One of the components that *U.S. News & World Report* uses when evaluating colleges is the difference between the actual graduation rate and the rate predicted by a regression equation.[38] In this chapter, we call this quantity the residual. Explain why the residual is a better measure to evaluate college graduation rates than the raw graduation rate.

2.156 Salaries and raises. For this exercise, we consider a hypothetical employee who starts working in Year 1 with a salary of $50,000. Each year her salary increases by approximately 5%. By Year 20, she is earning $126,000. The following table gives her salary for each year (in thousands of dollars): RAISES

Year	Salary	Year	Salary	Year	Salary	Year	Salary
1	50	6	63	11	81	16	104
2	53	7	67	12	85	17	109
3	56	8	70	13	90	18	114
4	58	9	74	14	93	19	120
5	61	10	78	15	99	20	126

FIGURE 2.34 Plot of salary versus year for an individual who receives approximately a 5% raise each year for 20 years, with the least-squares regression line, Exercise 2.156.

(a) Figure 2.34 is a scatterplot of salary versus year, with the least-squares regression line. Describe the relationship between salary and year for this person.

(b) The value of r^2 for these data is 0.9832. What percent of the variation in salary is explained by year? Would you say that this is an indication of a strong linear relationship? Explain your answer.

2.157 Look at the residuals. Refer to the previous exercise. Figure 2.35 is a plot of the residuals versus year. RAISES

(a) Interpret the residual plot.

(b) Explain how this plot highlights the deviations from the least-squares regression line that you can see in Figure 2.34.

2.158 Try logs. Refer to the previous two exercises. Figure 2.36 is a scatterplot with the least-squares regression line for log salary versus year. For this model, $r^2 = 0.9995$. RAISES

(a) Compare this plot with Figure 2.34. Write a short summary of the similarities and the differences.

(b) Figure 2.37 is a plot of the residuals for the model using year to predict log salary. Compare this plot with Figure 2.35 and summarize your findings.

FIGURE 2.35 Plot of residuals versus year for an individual who receives approximately a 5% raise each year for 20 years, Exercise 2.157.

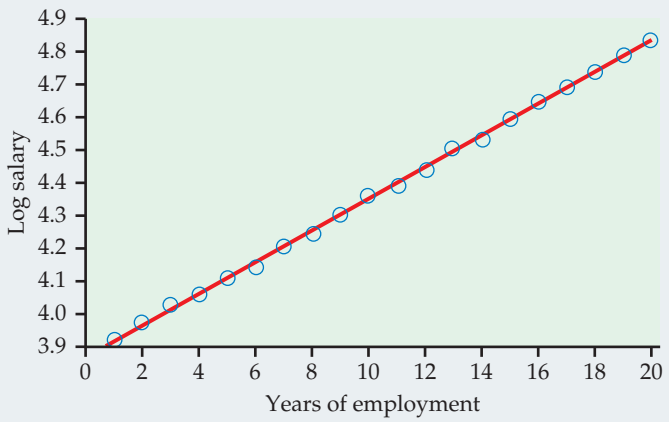


FIGURE 2.36 Plot of log salary versus year for an individual who receives approximately a 5% raise each year for 20 years, with the least-squares regression line, Exercise 2.158.

2.159 Make some predictions. The individual whose salary we have been studying wants to do some financial planning. Specifically, she would like to predict her salary five years into the future, that is, for Year 25. She is willing to assume that her employment situation will be stable for the next five years and that it will be similar to the last 20 years. RAISES

(a) Predict her salary for Year 25 using the least-squares regression equation constructed to predict salary from year.

(b) Predict her salary for Year 25 using the least-squares regression equation constructed to predict log salary from year. Note that you will need to take the predicted log salary and convert this value back to the predicted salary. Many calculators have a function that will perform this operation.

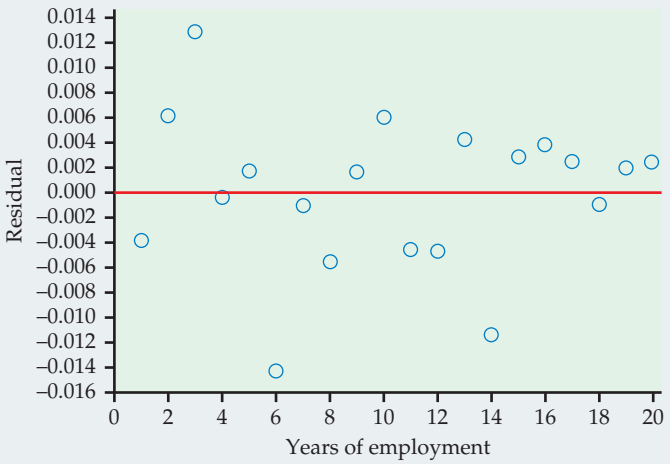


FIGURE 2.37 Plot of residuals, based on log salary, versus year for an individual who receives approximately a 5% raise each year for 20 years, Exercise 2.158.

(c) Which prediction do you prefer? Explain your answer.

(d) Someone looking at the numerical summaries and not the plots for these analyses says that because both models have very high values of r^2, they should perform equally well in doing this prediction. Write a response to this comment.

(e) Discuss the value of graphical summaries and the problems of extrapolation using what you have learned in studying these salary data.

2.160 Faculty salaries. Here are the salaries for a sample of professors in a mathematics department at a large midwestern university for the academic years 2014–2015 and 2015–2016. FACULTY

2014–2015 salary ($)	2015–2016 salary ($)	2014–2015 salary ($)	2015–2016 salary ($)
145,700	147,700	136,650	138,650
112,700	114,660	132,160	134,150
109,200	111,400	74,290	76,590
98,800	101,900	74,500	77,000
112,000	113,000	83,000	85,400
111,790	113,800	141,850	143,830
103,500	105,700	122,500	124,510
149,000	150,900	115,100	117,100

(a) Construct a scatterplot with the 2015–2016 salaries on the vertical axis and the 2014–2015 salaries on the horizontal axis.

(b) Comment on the form, direction, and strength of the relationship in your scatterplot.

(c) What proportion of the variation in 2015–2016 salaries is explained by 2014–2015 salaries?

2.161 Find the line and examine the residuals. Refer to the previous exercise. FACULTY

(a) Find the least-squares regression line for predicting 2015–2016 salaries from 2014–2015 salaries.

(b) Analyze the residuals, paying attention to any outliers or influential observations. Write a summary of your findings.

2.162 Bigger raises for those earning less. Refer to the previous two exercises. The 2014–2015 salaries do an excellent job of predicting the 2015–2016 salaries. Is there anything more that we can learn from these data? In this department, there is a tradition of giving higher-than-average percent raises to those whose salaries are lower. Let's see if we can find evidence to support this idea in the data. FACULTY

(a) Compute the percent raise for each faculty member. Take the difference between the 2015–2016 salary and the 2014–2015 salary, divide by the 2014–2015 salary, and then multiply by 100. Make a scatterplot with raise as the response variable and the 2014–2015 salary as the explanatory variable. Describe the relationship that you see in your plot.

(b) Find the least-squares regression line and add it to your plot.

(c) Analyze the residuals. Are there any outliers or influential cases? Make a graphical display and include this in a short summary of your conclusions.

(d) Is there evidence in the data to support the idea that greater percent raises are given to those with lower salaries? Include numerical and graphical summaries to support your conclusion.

2.163 Firefighters and fire damage. Someone says, "There is a strong positive correlation between the number of firefighters at a fire and the amount of damage the fire does. So sending lots of firefighters just causes more damage." Explain why this reasoning is wrong.

2.164 Predicting text pages. The editor of a statistics text would like to plan for the next edition. A key variable is the number of pages that will be in the final version. Text files are prepared by the authors using a word processor called LaTeX, and separate files contain figures and tables. For the previous edition of the text, the number of pages in the LaTeX files can easily be determined, as well as the number of pages in the final version of the text. Here are the data: TEXTP

Chapter	1	2	3	4	5	6	7	8	9	10	11	12	13
LaTeX pages	77	73	59	80	45	66	81	45	47	43	31	46	26
Text pages	99	89	61	82	47	68	87	45	53	50	36	52	19

(a) Plot the data and describe the overall pattern.

(b) Find the equation of the least-squares regression line and add the line to your plot.

(c) Find the predicted number of pages for the next edition if the number of LaTeX pages is 62.

(d) Write a short report for the editor explaining to her how you constructed the regression equation and how she could use it to estimate the number of pages in the next edition of the text.

2.165 Plywood strength. How strong is a building material such as plywood? To be specific, support a 24-inch by 2-inch strip of plywood at both ends and apply force in the middle until the strip breaks. The modulus of rupture (MOR) is the force needed to break the strip. We would like to be able to predict MOR without actually breaking the wood. The modulus of elasticity (MOE) is found by bending the wood without breaking it. Both MOE and MOR are measured in pounds per square inch. Here are data for 32 specimens of the same type of plywood:[39] MOEMOR

MOE	MOR	MOE	MOR	MOE	MOR	MOE	MOR
2,005,400	11,591	2,181,910	12,702	1,774,850	10,541	1,747,010	11,794
1,166,360	8,542	1,559,700	11,209	1,457,020	10,314	1,791,150	11,413
1,842,180	12,750	2,372,660	12,799	1,959,590	11,983	2,535,170	13,920
2,088,370	14,512	1,580,930	12,062	1,720,930	10,232	1,355,720	9,286
1,615,070	9,244	1,879,900	11,357	1,355,960	8,395	1,646,010	8,814
1,938,440	11,904	1,594,750	8,889	1,411,210	10,654	1,472,310	6,326
2,047,700	11,208	1,558,770	11,565	1,842,630	10,223	1,488,440	9,214
2,037,520	12,004	2,212,310	15,317	1,984,690	13,499	2,349,090	13,645

Can we use MOE to predict MOR accurately? Use the data to write a discussion of this question.

2.166 Distribution of the residuals. Some statistical methods require that the residuals from a regression line have a Normal distribution. The residuals for the nonexercise activity example are given in Exercise 2.93 (page 125). Is their distribution close to Normal? Make a Normal quantile plot to find out. FIDGET

2.167 An example of Simpson's paradox. Mountain View University has professional schools in business and law. Here is a three-way table of applicants to these

professional schools, categorized by sex, school, and admission decision:[40] ADMITS

Business			Law		
	Admit			Admit	
Sex	Yes	No	Sex	Yes	No
Male	400	200	Male	90	110
Female	200	100	Female	200	200

(a) Make a two-way table of sex by admission decision for the combined professional schools by summing entries in the three-way table.

(b) From your two-way table, compute separately the percents of male and female applicants admitted. Male applicants are admitted to Mountain View's professional schools at a higher rate than female applicants.

(c) Now compute separately the percents of male and female applicants admitted by the business school and by the law school.

(d) Explain carefully, as if speaking to a skeptical reporter, how it can happen that Mountain View appears to favor males when this is not true within each of the professional schools.

2.168 Simpson's paradox and regression. Simpson's paradox occurs when a relationship between variables within groups of observations reverses when all of the data are combined. The phenomenon is usually discussed in terms of categorical variables, but it also occurs in other settings. Here is an example: SIMREG

y	x	Group	y	x	Group
10.1	1	1	18.3	6	2
8.9	2	1	17.1	7	2
8.0	3	1	16.2	8	2
6.9	4	1	15.1	9	2
6.1	5	1	14.3	10	2

(a) Make a scatterplot of the data for Group 1. Find the least-squares regression line and add it to your plot. Describe the relationship between y and x for Group 1.

(b) Do the same for Group 2.

(c) Make a scatterplot using all 10 observations. Find the least-squares line and add it to your plot.

(d) Make a plot with all of the data using different symbols for the two groups. Include the three regression lines on the plot. Write a paragraph about Simpson's paradox for regression using this graphical display to illustrate your description.

2.169 Class size and class level. A university classifies its classes as either "small" (fewer than 40 students) or "large." A dean sees that 62% of Department A's classes are small, while Department B has only 40% small classes. She wonders if she should cut Department A's budget and insist on larger classes. Department A responds to the dean by pointing out that classes for third- and fourth-year students tend to be smaller than classes for first- and second-year students. The following three-way table gives the counts of classes by department, size, and student audience. Write a short report for the dean that summarizes these data. Start by computing the percents of small classes in the two departments and include other numerical and graphical comparisons as needed. Here are the numbers of classes to be analyzed: CSIZE

	Department A			Department B		
Year	Large	Small	Total	Large	Small	Total
First	2	0	2	18	2	20
Second	9	1	10	40	10	50
Third	5	15	20	4	16	20
Fourth	4	16	20	2	14	16

2.170 More smokers live at least 20 more years! You can see the headlines: "More smokers than nonsmokers live at least 20 more years after being contacted for study!" A medical study contacted randomly chosen people in a district in England. Here are data on the 1314 women contacted who were either current smokers or who had never smoked. The tables classify these women by their smoking status and age at the time of the survey and whether they were still alive 20 years later.[41] SMOKERS

	Age 18 to 44		Age 45 to 64		Age 65+	
	Smoker	Not	Smoker	Not	Smoker	Not
Dead	19	13	78	52	42	165
Alive	269	327	167	147	7	28

(a) From these data, make a two-way table of smoking (yes or no) by dead or alive. What percent of the smokers stayed alive for 20 years? What percent of the nonsmokers survived? It seems surprising that a higher percent of smokers stayed alive.

(b) The age of the women at the time of the study is a lurking variable. Show that within each of the three age groups in the data, a higher percent of nonsmokers remained alive 20 years later. This is another example of Simpson's paradox.

(c) The study authors give this explanation: "Few of the older women (over 65 at the original survey) were smokers, but many of them had died by the time of follow-up." Compare the percent of smokers in the three age groups to verify the explanation.

2.171 Recycled product quality. Recycling is supposed to save resources. Some people think recycled products are lower in quality than other products, a fact that makes recycling less practical. People who actually use a recycled product may have different opinions from those who don't use it. Here are data on attitudes toward coffee filters made of recycled paper among people who do and don't buy these filters:[42] RECYCLE

	Think the quality of the recycled product is:		
	Higher	The same	Lower
Buyers	20	7	9
Nonbuyers	29	25	43

(a) Find the marginal distribution of opinion about quality. Assuming that these people represent all users of coffee filters, what does this distribution tell us?

(b) How do the opinions of buyers and nonbuyers differ? Use conditional distributions as a basis for your answer. Include a mosaic plot if you have access to the needed software. Can you conclude that using recycled filters causes more favorable opinions? If so, giving away samples might increase sales.

2.172 Survival and sex on the *Titanic*. In Exercise 2.122, you examined the relationship between survival and class on the *Titanic*. The data file TITANIC contains data on the sex of the *Titanic* passengers. Examine the relationship between survival and sex and write a short summary of your findings. TITANIC

2.173 Survival, class, and sex on the *Titanic*. Refer to the previous exercise and Exercise 2.122 (page 146). When we looked at survival and class, we ignored sex. When we looked at survival and sex, we ignored class. Are we missing something interesting about these data when we choose this approach to the analysis? Here is one way to answer this question. TITANIC

(a) Create two separate two-way tables. One for survival and class for the women and another for survival and class for the men.

(b) Perform an analysis of the relationship between survival and class for the women. Summarize your findings.

(c) Perform an analysis of the relationship between survival and class for the men. Summarize your findings.

(d) Compare the analyses that you performed in parts (b) and (c). Write a short report on the relationship between survival and the two explanatory variables, class and sex.

2.174 Blueberries and anthocyanins. Refer to Exercises 1.165 and 1.166 (page 77). Figure 2.38 gives JMP output for examining the relationship between Antho2 and Antho1. Use this output to write a summary of this relationship using the methods and ideas that you learned in this chapter. BERRIES

2.175 Averaged date for blueberries and anthocyanins. Refer to the previous exercise where you examined the relationship between Antho2 and Antho1. The variables Antho2M and Antho1M were computed by averaging Antho2 and Antho1 for values of Antho1 in the intervals (0, 0.5), [0.5, 1.0), [1.0, 1.5), [1.5, 2.0), [2.0, 2.5), [2.5, 3.0), and [3.0, 3.5). Analyze the relationship between Antho2M and Antho1M, and compare your results that you found in the previous exercise using Antho2 and Antho1. Summarize what the comparison tells you about relationships with averaged data. BERRIEM

2.176 Restricting the range for blueberries and anthocyanins. Refer to Exercise 2.174 where you examined the relationship between Antho2 and Antho1. The data file BERRIER was created from the data file BERRIES by excluding cases with values of Antho1 that are less than 1.5 and cases with values of Antho1 that are greater than 3. Analyze the relationship between Antho2 and Antho1 for this restricted range data set, and compare your results that you found in Exercise 2.174 for the complete data set. Summarize what the comparison tells you about relationships with a restricted range. BERRIER

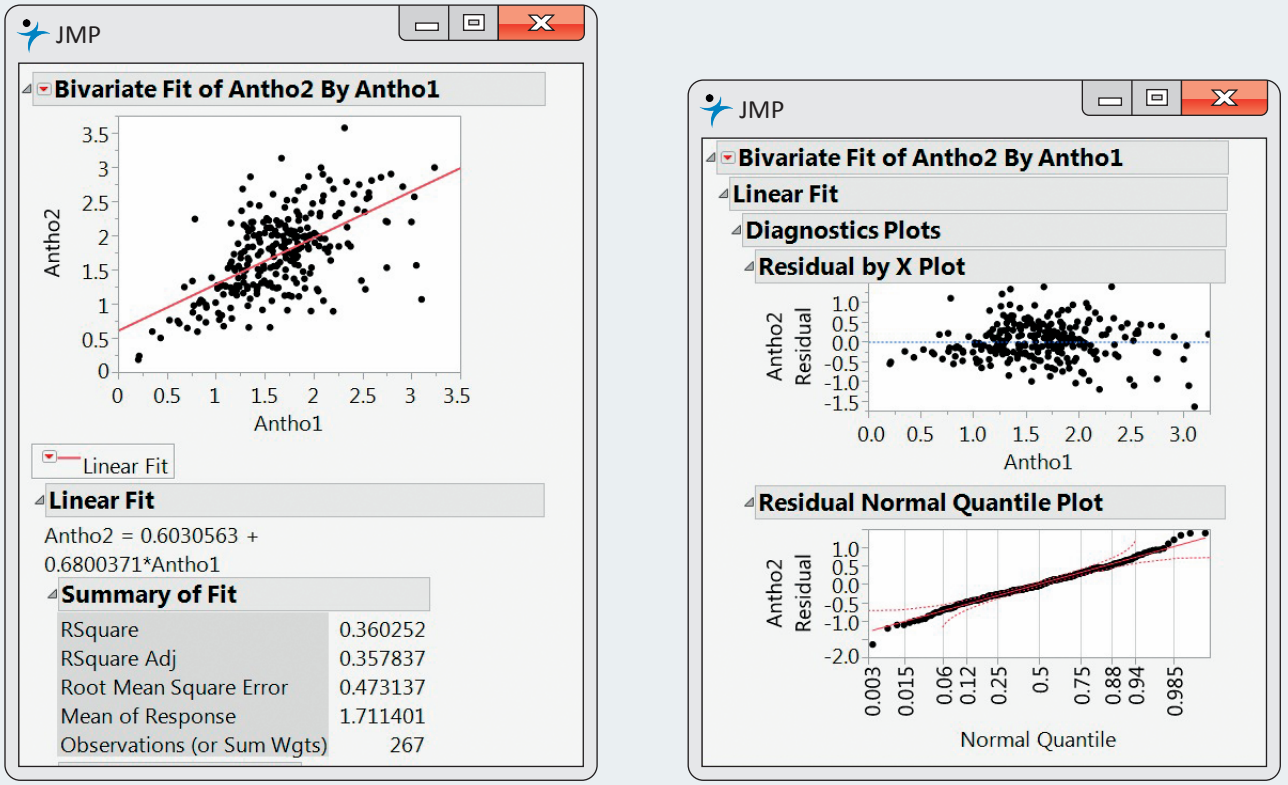


FIGURE 2.38 Selected JMP output for examining the relationship between Antho2 and Antho1, Exercise 2.174.