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.

RESPONSE VARIABLE, EXPLANATORY VARIABLE

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

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.

independent variable / dependent variable

Some of the statistical techniques in this chapter require us to distinguish explanatory from response variables; others make no use of this distinction. You will often see explanatory variables called independent variables variable and response variables called dependent variables. These terms express 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.

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.

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 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.

DIRECTION: 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.

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 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.

CATEGORICAL VARIABLES IN SCATTERPLOTS

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

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.

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.

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

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

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

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.

- 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 illustrate how values of r closer to 1 or −1 correspond to stronger linear relationships. To make the essential meaning of r clear, the standard deviations of both variables in these plots are equal, and the horizontal and vertical scales are the same. In general, it is not so easy to guess the value of r from the appearance of a scatterplot. Remember that changing the plotting scales in a scatterplot may mislead our eyes, but it does not change the standardized values of the variables and, therefore, cannot change the correlation.

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.

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 $-1 \leq r \leq 1$ and indicates the strength of a relationship by how close it is to −1 or 1. Perfect correlation, $r = \pm 1$, occurs only when the points lie exactly on a straight line.
- Correlation ignores the distinction between explanatory and response variables. The value of r is not affected by changes in the unit of measurement of either variable. Correlation is not resistant, so outliers can greatly change the value of r.

2.4 Least-Squares Regression

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

REGRESSION LINE

A regression line is a straight line that describes how a response variable y changes as an explanatory variable x changes. We often use a regression line to predict the value of y for a given value of x. Regression, unlike correlation, requires that we have an explanatory variable and a response variable.

Fitting a line to data

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 $x = 0$.

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.

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.

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.

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 $\overline{x}$ and s_x for x and $\overline{y}$ and s_y for y, and the correlation between x and y is r. The **equation of the least-squares regression line** of y on x is

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

with **slope**

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

and **intercept**

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

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 $(\overline{x}, \overline{y})$ 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 $(\overline{x}, \overline{y})$. We can describe regression entirely in terms of the basic descriptive measures $\overline{x}$, s_x, $\overline{y}$, 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.

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 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 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.

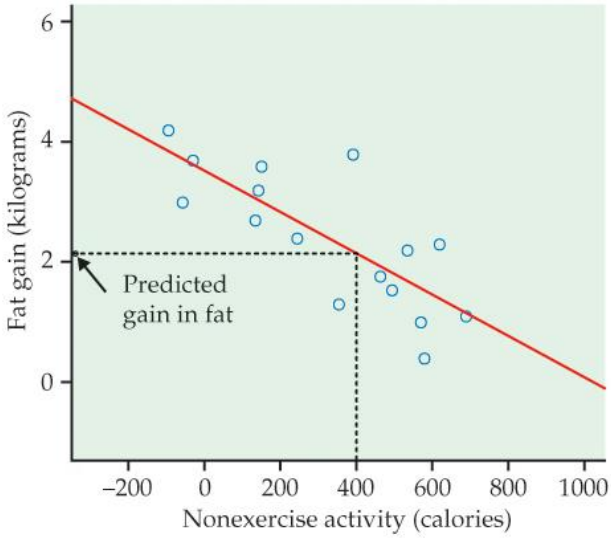

FIGURE 2.17 A regression line fitted to the nonexercise activity data and used to predict fat gain for an NEA increase of 400 calories, Examples 2.20 and 2.21.

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 $\hat{y} = b_0 + b_1x$ 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** b_1 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, b_1 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** b_0 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.

2.5 Cautions about Correlation and Regression

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,

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

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.

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.

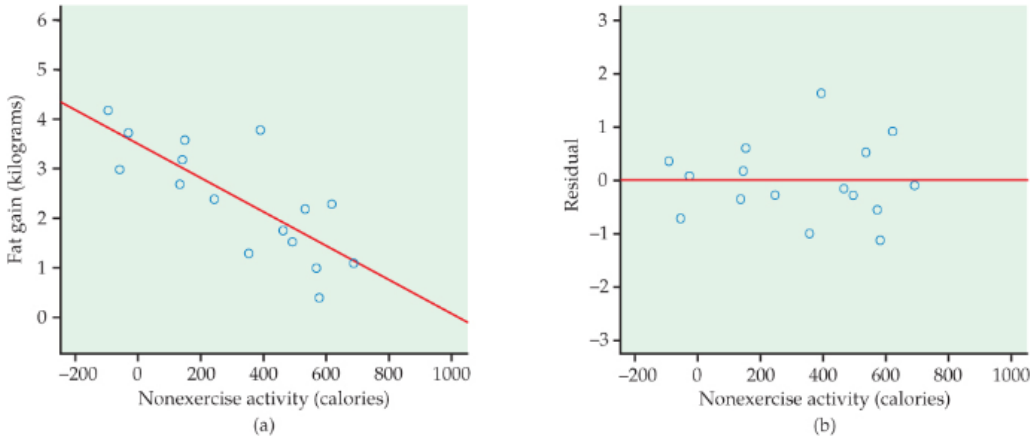

FIGURE 2.23 (a) Scatterplot of fat gain versus increase in nonexercise activity, with the least-squares regression line, Example 2.26. (b) Residual plot for the regression displayed in panel (a); the line at $y = 0$ marks the mean of the residuals.

Patterns in birthrate and Internet user residuals. 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.

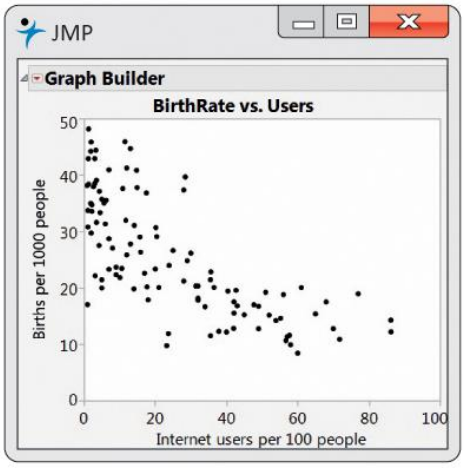

FIGURE 2.13 Scatterplot of births (per 1000 people) versus Internet users (per 100 people) for 106 countries, Exercise 2.34.

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.

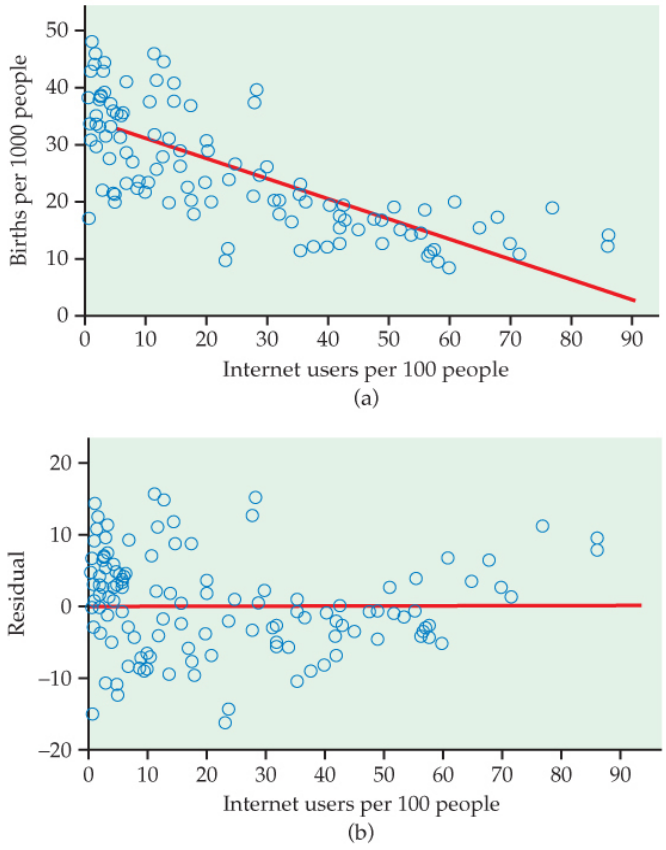

FIGURE 2.24 (a) Scatterplot of birthrate versus Internet users, with the least-squares regression line, Example 2.27. (b) Residual plot for the regression displayed in panel (a); the line at $y = 0$ marks the mean of the residuals.

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.

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.

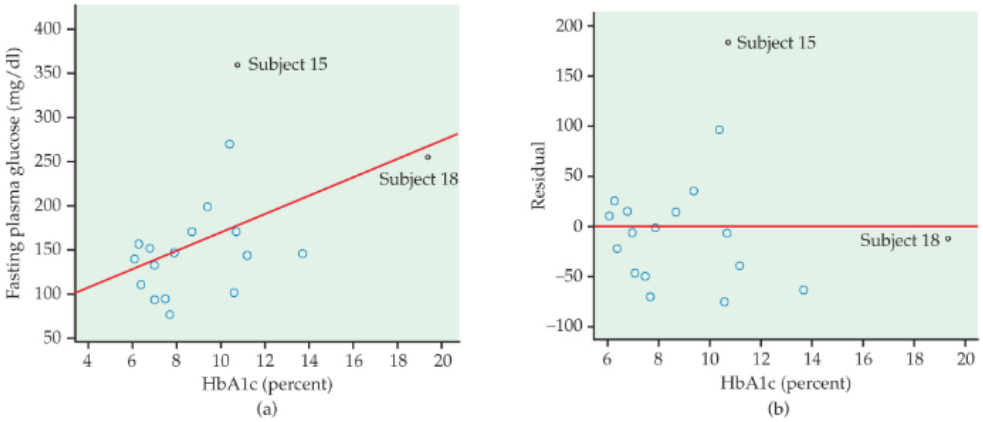

FIGURE 2.25 (a) Scatterplot of fasting plasma glucose against HbA1c (which measures long-term blood glucose), with the least-squares regression line, Example 2.28. (b) Residual plot for the regression of fasting plasma glucose on HbA1c. Subject 15 is an outlier in fasting plasma glucose. Subject 18 is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

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.

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.

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.

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.

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 2.6 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 2.7 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.