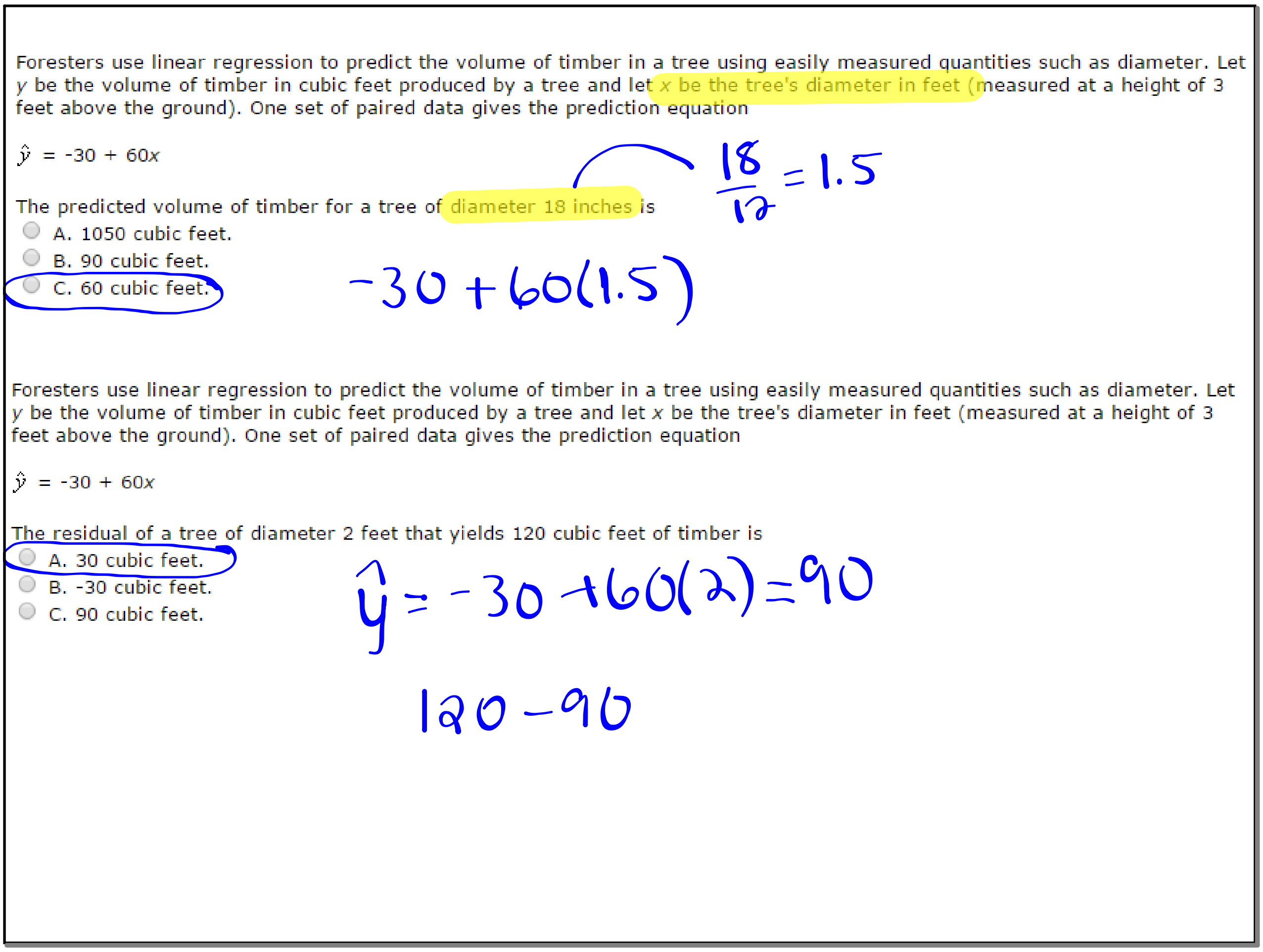


Foresters use linear regression to predict the volume of timber in a tree using easily measured quantities such as diameter. Let y be the volume of timber in cubic feet produced by a tree and let x be the tree's diameter in feet (measured at a height of 3 feet above the ground). One set of paired data gives the prediction equation

$\hat{y} = -30 + 60x$

The predicted volume of timber for a tree of diameter 18 inches is

$\frac{18}{12} = 1.5$

○ A. 1050 cubic feet.

○ B. 90 cubic feet.

[✓ ○ C. 60 cubic feet.]

$-30 + 60(1.5)$

Foresters use linear regression to predict the volume of timber in a tree using easily measured quantities such as diameter. Let y be the volume of timber in cubic feet produced by a tree and let x be the tree's diameter in feet (measured at a height of 3 feet above the ground). One set of paired data gives the prediction equation

$\hat{y} = -30 + 60x$

The residual of a tree of diameter 2 feet that yields 120 cubic feet of timber is

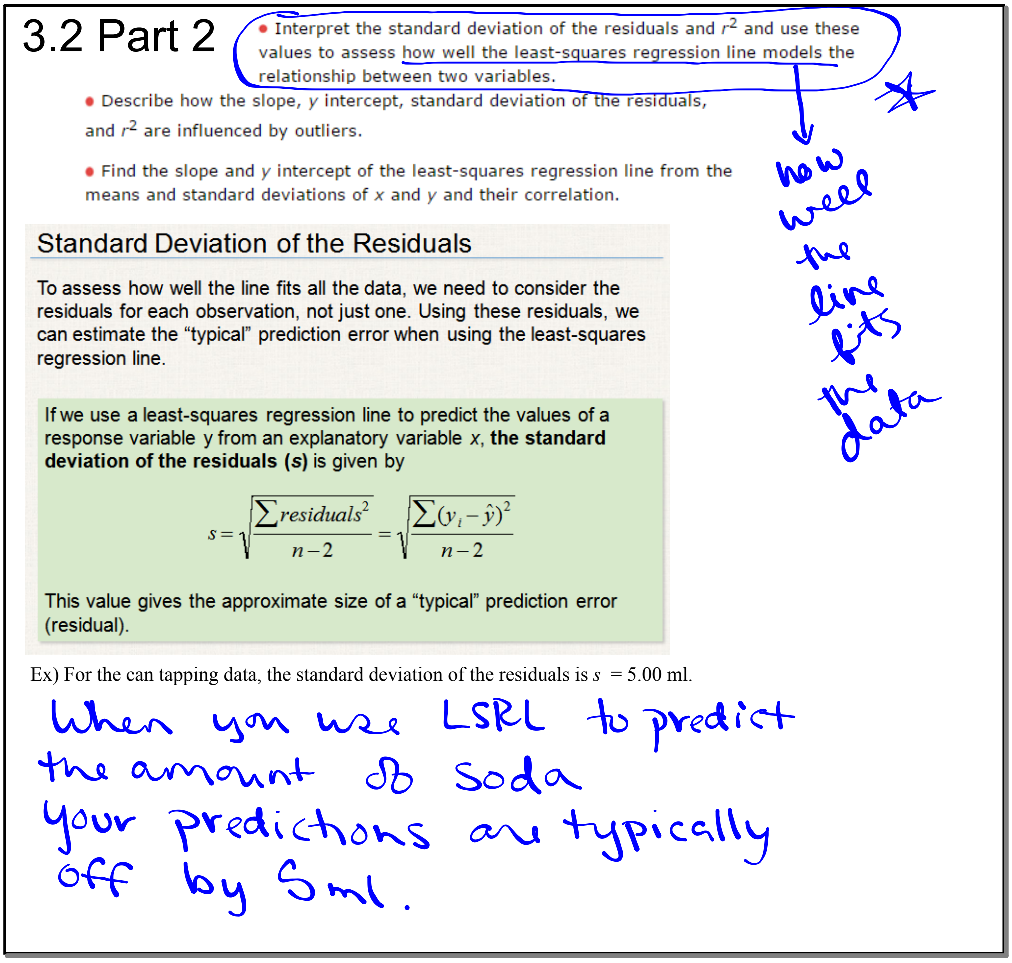
[✓ ○ A. 30 cubic feet.]

○ B. -30 cubic feet.

○ C. 90 cubic feet.

$\hat{y} = -30 + 60(2) = 90$

$120 - 90$



3.2 Part 2

- Interpret the standard deviation of the residuals and r^2 and use these values to assess how well the least-squares regression line models the relationship between two variables. ✗
- Describe how the slope, y intercept, standard deviation of the residuals, and r^2 are influenced by outliers.
- Find the slope and y intercept of the least-squares regression line from the means and standard deviations of x and y and their correlation.

how well the line fits the data

Standard Deviation of the Residuals

To assess how well the line fits all the data, we need to consider the residuals for each observation, not just one. Using these residuals, we can estimate the "typical" prediction error when using the least-squares regression line.

If we use a least-squares regression line to predict the values of a response variable y from an explanatory variable x, **the standard deviation of the residuals (s)** is given by

$$s = \sqrt{\frac{\sum residuals^2}{n-2}} = \sqrt{\frac{\sum (y_i - \hat{y})^2}{n-2}}$$

This value gives the approximate size of a "typical" prediction error (residual).

Ex) For the can tapping data, the standard deviation of the residuals is s = 5.00 ml.

When you use LSRL to predict the amount of soda your predictions are typically off by 5ml.

The Coefficient of Determination r^2

The standard deviation of the residuals gives us a numerical estimate of the average size of our prediction errors. There is another numerical quantity that tells us how well the least-squares regression line predicts values of the response *y*.

The **coefficient of determination** r^2 is the fraction of the variation in the values of *y* that is accounted for by the least-squares regression line of *y* on *x*. We can calculate r^2 using the following formula:

$$r^2 = 1 - \frac{\sum \text{residuals}^2}{\sum (y_i - \bar{y})^2}$$

$r = \sqrt{.854} = .924$

for sign look at sign of slope

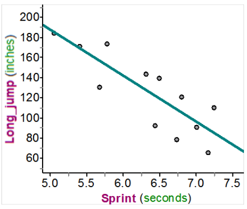
r^2 tells us how much better the LSRL does at predicting values of *y* than simply guessing the mean *y* for each value in the dataset.

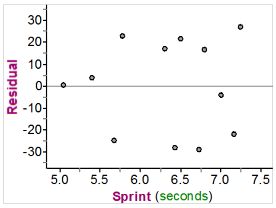
Ex) Tapping Cans r^2 = .854

85.4% of the variability in amount of soda is accounted for by LSRL relating predicted amount of soda to tapping time

AP® EXAM TIP Students often have a hard time interpreting the value of r^2 on AP® exam questions. They frequently leave out key words in the definition. Our advice: Treat this as a fill-in-the-blank exercise. Write "____% of the variation in [response variable name] is accounted for by the linear model relating [response variable name] to [explanatory variable name]."

Back to the track!
In Section 3.1, we looked at the relationship between the 40-yard sprint time (in seconds) and the long-jump distance (in inches) for a small statistics class with 12 students. A scatterplot with the least-squares regression line $\hat{y}$ = 414.79- 45.74*x*) and a residual plot are shown below. Also, *s* = 22.38 and r^2 = 0.702.



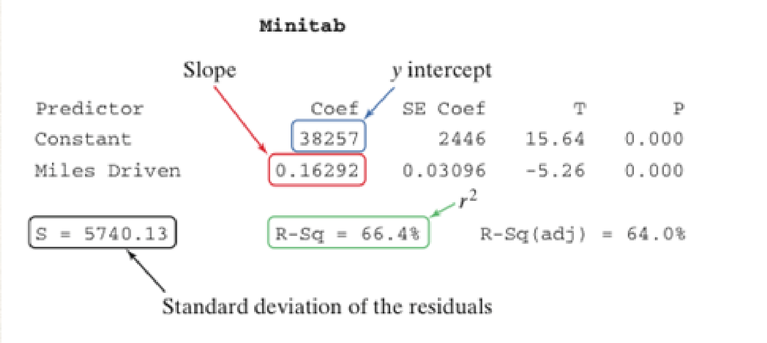


(a) Calculate and interpret the residual for Christian, who had a sprint time of 7.25 seconds and a long jump of 110 inches.
(b) Is a linear model appropriate for these data? Explain.
(c) Interpret the value of *s*.
(d) Interpret the value of r^2 .

Interpreting Computer Regression Output

A number of statistical software packages produce similar regression output. Be sure you can locate

- the slope *b*
- the *y* intercept *a*
- the values of *s* and r^2



Minitab

Slope

y intercept

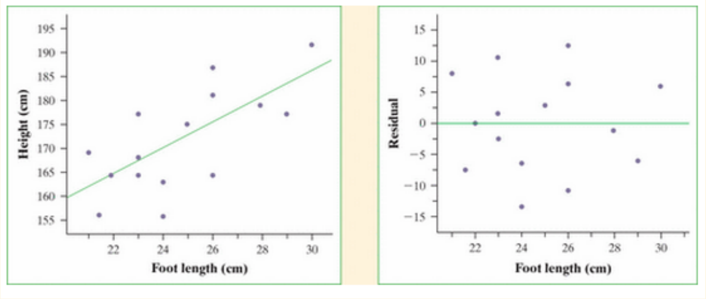
```
Predictor         Coef    SE Coef      T       P
Constant         38257       2446  15.64   0.000
Miles Driven   0.16292    0.03096  -5.26   0.000

S = 5740.13       R-Sq = 66.4%    R-Sq(adj) = 64.0%
```

r^2

Standard deviation of the residuals

A random sample of 15 high school students was selected from the U.S. CensusAtSchool database. The foot length (in centimeters) and height (in centimeters) of each student in the sample were recorded. Least-squares regression was performed on the data. A scatterplot with the regression line added, a residual plot, and some computer output from the regression are shown below.

```
Predictor          Coef     SE Coef     T       P
Constant           103.41   19.50       5.30    0.000
Foot length        2.7469   0.7833      3.51    0.004
S = 7.95126     R-Sq = 48.6%     R-Sq(adj) = 44.7%
```



PROBLEM:

(a) What is the equation of the least-squares regression line that describes the relationship between foot length and height? Define any variables that you use.

(b) Interpret the slope of the regression line in context.

(c) Find the correlation.

d) Interpret r^2

e) Interpret s

f) Is the linear model appropriate? Explain.

Regression to the Mean

Using technology is often the most convenient way to find the equation of a least-squares regression line. It is also possible to calculate the equation of the least- squares regression line using only the means and standard deviations of the two variables and their correlation.

How to Calculate the Least-Squares Regression Line

We have data on an explanatory variable *x* and a response variable *y* for *n* individuals. From the data, calculate the means and the standard deviations of the two variables and their correlation *r*.

The least-squares regression line is the line $\hat{y} = a + bx$ with **slope**

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

And ***y* intercept** $a = \bar{y} - b\bar{x}$

AP® EXAM TIP The formula sheet for the AP® exam uses different notation for these equations: $b_1 = r\frac{s_y}{s_x}$ and $b_0 = \bar{y} - b_1\bar{x}$. That's because the least-squares line is written as $\hat{y} = b_0 + b_1 x$. We prefer our simpler versions without the subscripts!

Example 15 Using Feet to Predict Height

Worked Example Videos

Calculating the least-squares regression line

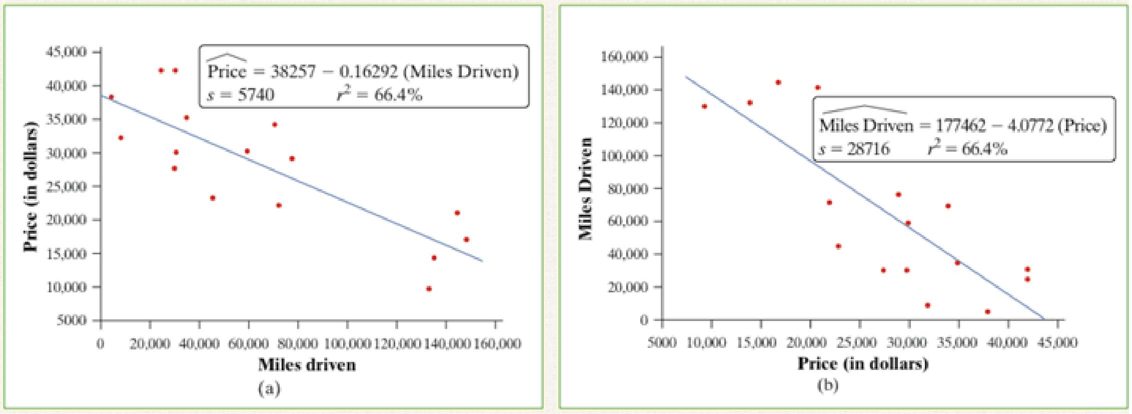
In the previous example, we used data from a random sample of 15 high school students to investigate the relationship between foot length (in centimeters) and height (in centimeters). The mean and standard deviation of the foot lengths are $\bar{x}$ = 24.76 cm and s_x = 2.71 cm. The mean and standard deviation of the heights are $\bar{y}$ = 171.43 cm and s_y = 10.69 cm. The correlation between foot length and height is r = 0.697.

PROBLEM: Find the equation of the least-squares regression line for predicting height from foot length. Show your work.

Correlation and Regression Wisdom

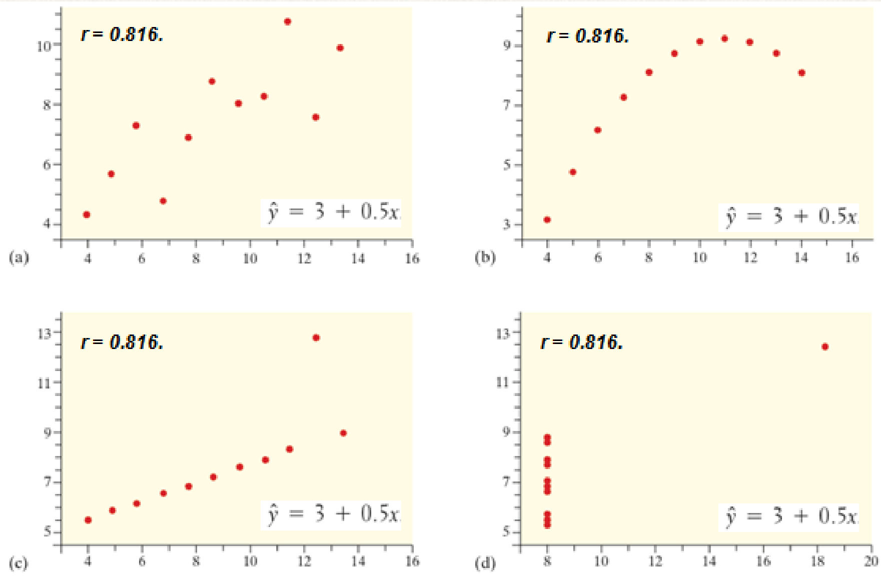
Correlation and regression are powerful tools for describing the relationship between two variables. When you use these tools, be aware of their limitations.

1. The distinction between explanatory and response variables is important in regression.



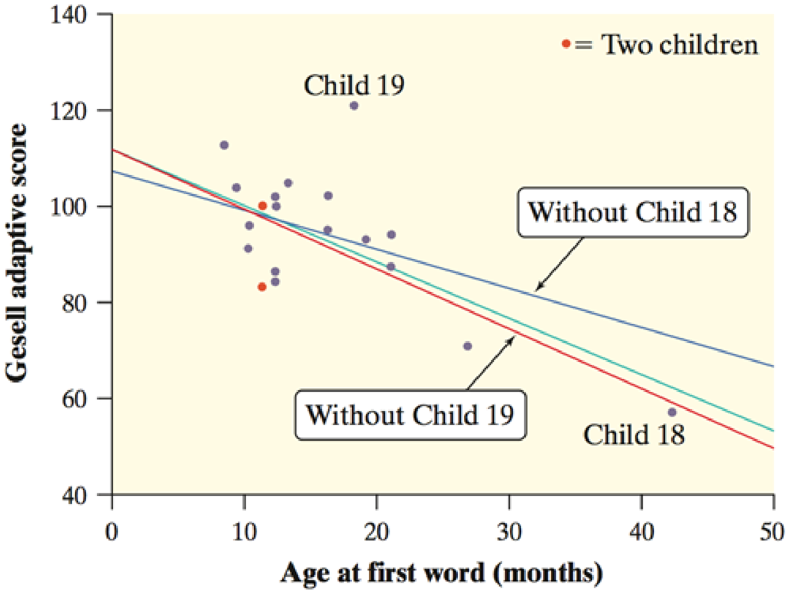
Correlation and Regression Wisdom

2. Correlation and regression lines describe only linear relationships.



Correlation and Regression Wisdom

3. Correlation and least-squares regression lines are not resistant.



With all 19 children:
$r = -0.64$
$\hat{y} = 109.874 - 1.127x$

Without Child 19:
$r = -0.76$
$\hat{y} = 109.305 - 1.193x$

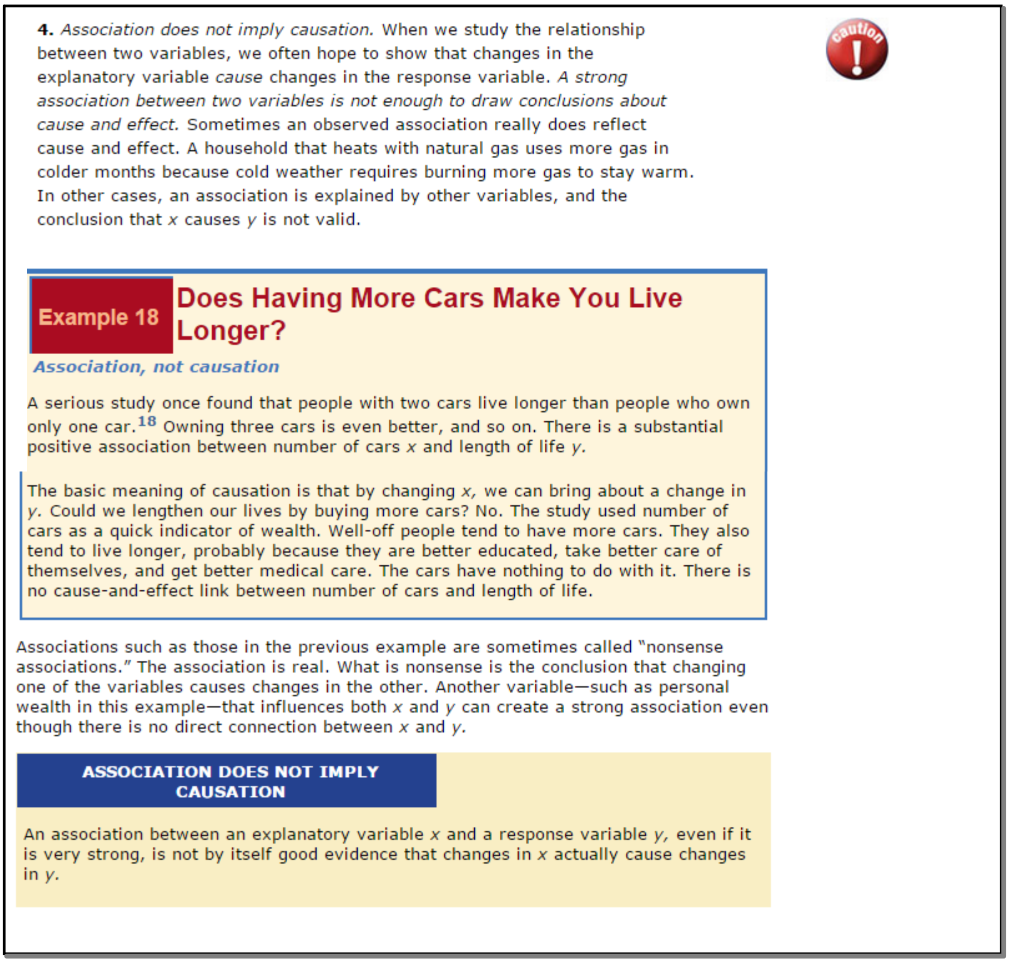
Without Child 18:
$r = -0.33$
$\hat{y} = 105.630 - 0.779x$

Outliers and Influential Observations in Regression

Least-squares lines make the sum of the 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. We call such points **influential**.

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



4. *Association does not imply causation.* When we study the relationship between two variables, we often hope to show that changes in the explanatory variable *cause* changes in the response variable. *A strong association between two variables is not enough to draw conclusions about cause and effect.* Sometimes an observed association really does reflect cause and effect. A household that heats with natural gas uses more gas in colder months because cold weather requires burning more gas to stay warm. In other cases, an association is explained by other variables, and the conclusion that *x* causes *y* is not valid.

Example 18 Does Having More Cars Make You Live Longer?

Association, not causation

A serious study once found that people with two cars live longer than people who own only one car.[18] Owning three cars is even better, and so on. There is a substantial positive association between number of cars *x* and length of life *y*.

The basic meaning of causation is that by changing *x*, we can bring about a change in *y*. Could we lengthen our lives by buying more cars? No. The study used number of cars as a quick indicator of wealth. Well-off people tend to have more cars. They also tend to live longer, probably because they are better educated, take better care of themselves, and get better medical care. The cars have nothing to do with it. There is no cause-and-effect link between number of cars and length of life.

Associations such as those in the previous example are sometimes called "nonsense associations." The association is real. What is nonsense is the conclusion that changing one of the variables causes changes in the other. Another variable—such as personal wealth in this example—that influences both *x* and *y* can create a strong association even though there is no direct connection between *x* and *y*.

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*.