SECTION 3.2 Least-Squares Regression

LEARNING TARGETS *By the end of the section, you should be able to:*

- Make predictions using regression lines, keeping in mind the dangers of extrapolation.
- Calculate and interpret a residual.
- Interpret the slope and y intercept of a regression line.
- Determine the equation of a least-squares regression line using technology or computer output.
- Construct and interpret residual plots to assess whether a regression model is appropriate.
- Interpret the standard deviation of the residuals and r^2 and use these values to assess how well a least-squares regression line models the relationship between two variables.
- Describe how the least-squares regression line, standard deviation of the residuals, and r^2 are influenced by unusual points.
- 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.

Linear (straight-line) relationships between two quantitative variables are fairly common. In the preceding section, we found linear relationships in settings as varied as Major League Baseball, geysers, and Nobel prizes. Correlation measures the strength and direction of these relationships. When a scatterplot shows a linear relationship, we can summarize the overall pattern by drawing a line on the scatterplot. A **regression line** models the relationship between two variables, but only in a specific setting: when one variable helps explain the other. Regression, unlike correlation, requires that we have an explanatory variable and a response variable.

Sometimes regression lines are referred to as *simple linear regression models.* They are called "simple" because they involve only one explanatory variable.

DEFINITION Regression line

A **regression line** is a line that models how a response variable y changes as an explanatory variable x changes. Regression lines are expressed in the form $\hat{y} = a + bx$ where $\hat{y}$ (pronounced "y-hat") is the predicted value of y for a given value of x.

It is common knowledge that cars and trucks lose value the more they are driven. Can we predict the price of a used Ford F-150 SuperCrew 4×4 truck if we know how many miles it has on the odometer? A random sample of 16 used Ford F-150 SuperCrew 4×4s was selected from among those listed for sale at autotrader.com. The number of miles driven and price (in dollars) were recorded for each of the trucks.[22] Here are the data:

Tim Graham/Alamy

Miles driven	70,583	129,484	29,932	29,953	24,495	75,678	8359	4447
Price ($)	21,994	9500	29,875	41,995	41,995	28,986	31,891	37,991
Miles driven	34,077	58,023	44,447	68,474	144,162	140,776	29,397	131,385
Price ($)	34,995	29,988	22,896	33,961	16,883	20,897	27,495	13,997

Figure 3.6 is a scatterplot of these data. The plot shows a moderately strong, negative linear association between miles driven and price. There are two distinct clusters of trucks: a group of 12 trucks between 0 and 80,000 miles driven and a group of 4 trucks between 120,000 and 160,000 miles driven. The correlation is $r = -0.815$. The line on the plot is a regression line for predicting price from miles driven.

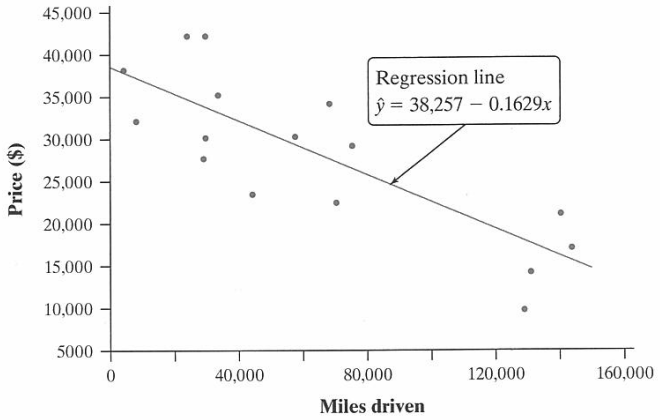


FIGURE 3.6 Scatterplot showing the price and miles driven of used Ford F-150s, along with a regression line.

Prediction

We can use a regression line to predict the value of the response variable for a specific value of the explanatory variable. For the Ford F-150 data, the equation of the regression line is

$$\widehat{\text{price}} = 38{,}257 - 0.1629\ (\text{miles driven})$$

When we want to refer to the predicted value of a variable, we add a hat on top. Here, $\widehat{\text{price}}$ refers to the predicted price of a used Ford F-150.

If a used Ford F-150 has 100,000 miles driven, substitute $x = 100{,}000$ in the equation. The predicted price is

$$\widehat{\text{price}} = 38{,}257 - 0.1629(100{,}000) = \$21{,}967$$

This prediction is illustrated in Figure 3.7.

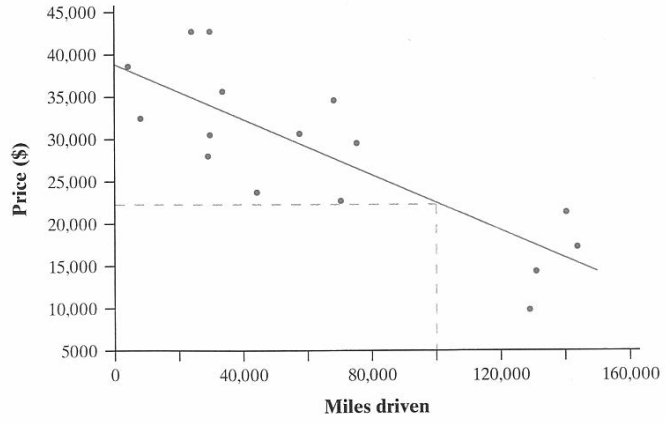


FIGURE 3.7 Using the regression line to predict price for a Ford F-150 with 100,000 miles driven.

Even though the value $\hat{y} = \$21,967$ is unlikely to be the actual price of a truck that has been driven 100,000 miles, it's our best guess based on the linear model using $x =$ miles driven. We can also think of $\hat{y} = \$21,967$ as the average price for a sample of trucks that have each been driven 100,000 miles.

Can we predict the price of a Ford F-150 with 300,000 miles driven? We can certainly substitute 300,000 into the equation of the line. The prediction is

$$\widehat{\text{price}} = 38,257 - 0.1629(300,000) = -\$10,613$$

The model predicts that we would need to *pay* \$10,613 just to have someone take the truck off our hands!

A negative price doesn't make much sense in this context. Look again at Figure 3.7. A truck with 300,000 miles driven is far outside the set of x values for our data. We can't say whether the relationship between miles driven and price remains linear at such extreme values. Predicting the price for a truck with 300,000 miles driven is an **extrapolation** of the relationship beyond what the data show.

DEFINITION Extrapolation

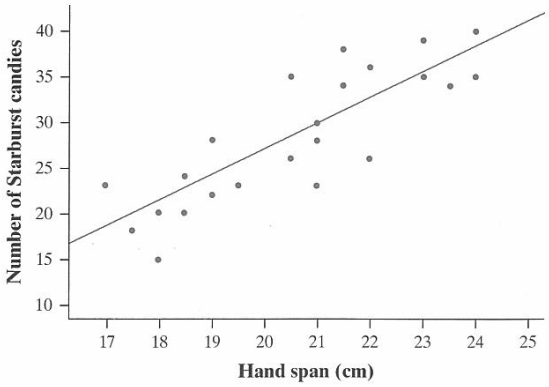
Extrapolation is the use of a regression line for prediction outside the interval of x values used to obtain the line. The further we extrapolate, the less reliable the predictions.

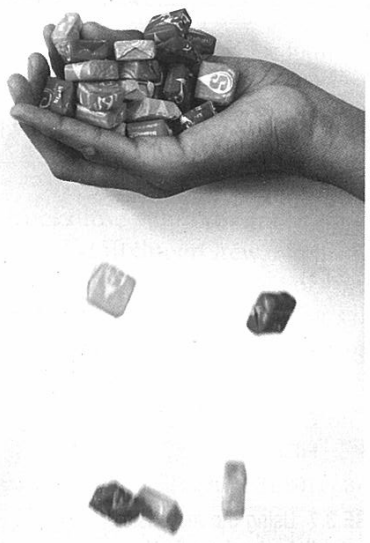
Few relationships are linear for all values of the explanatory variable. **Don't make predictions using values of x that are much larger or much smaller than those that actually appear in your data.**

EXAMPLE How much candy can you grab?

Prediction

PROBLEM: The scatterplot below shows the hand span (in cm) and number of Starburst™ candies grabbed by each student when Mr. Tyson's class did the "Candy grab" activity. The regression line $\hat{y} = -29.8 + 2.83x$ has been added to the scatterplot.




Josh Tabor

(a) Andres has a hand span of 22 cm. Predict the number of Starburst™ candies he can grab.

(b) Mr. Tyson's young daughter McKayla has a hand span of 12 cm. Predict the number of Starburst candies she can grab.

(c) How confident are you in each of these predictions? Explain.

SOLUTION:

(a) $\hat{y} = -29.8 + 2.83(22)$

$\hat{y} = 32.46$ Starburst candies

(b) $\hat{y} = -29.8 + 2.83(12)$

$\hat{y} = 4.16$ Starburst candies

(c) The prediction for Andres is believable because $x = 22$ is within the interval of x-values used to create the model. However, the prediction for McKayla is not trustworthy because $x = 12$ is far outside of the x-values used to create the regression line. The linear form may not extend to hand spans this small.

Don't worry that the predicted number of Starburst candies isn't an integer. Think of 32.46 as the average number of Starburst candies that a group of students, each with a hand span of 22 cm, could grab.

FOR PRACTICE, TRY EXERCISE 37

Residuals

In most cases, no line will pass exactly through all the points in a scatterplot. Because we use the line to predict y from x, the prediction errors we make are errors in y, the vertical direction in the scatterplot.

Figure 3.8 shows a scatterplot of the Ford F-150 data with a regression line added. The prediction errors are marked as bold vertical segments in the graph. These vertical deviations represent "leftover" variation in the response variable after fitting the regression line. For that reason, they are called **residuals**.

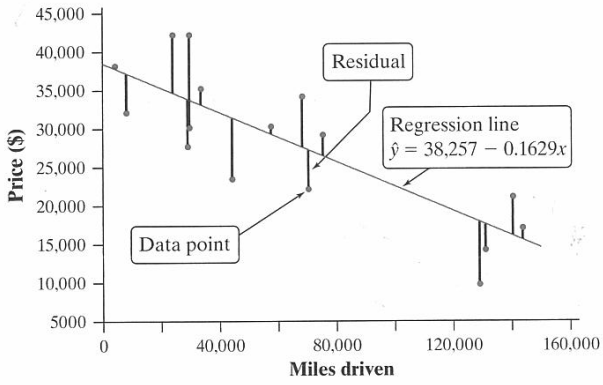


FIGURE 3.8 Scatterplot of the Ford F-150 data with a regression line added. A good regression line should make the residuals (shown as bold vertical segments) as small as possible.

DEFINITION Residual

A **residual** is the difference between the actual value of y and the value of y predicted by the regression line. That is,

$$\text{residual} = \text{actual } y - \text{predicted } y$$
$$= y - \hat{y}$$

In Figure 3.8 above, the highlighted data point represents a Ford F-150 that had 70,583 miles driven and a price of $21,994. The regression line predicts a price of

$$\widehat{\text{price}} = 38{,}257 - 0.1629(70{,}583) = \$26{,}759$$

for this truck, but its actual price was $21,994. This truck's residual is

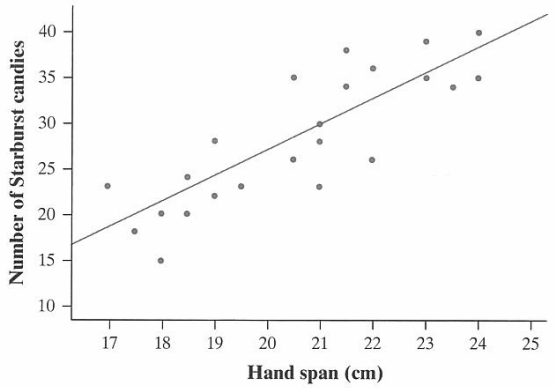
$$\begin{aligned}\text{residual} &= \text{actual } y - \text{predicted } y \\ &= y - \hat{y} \\ &= 21{,}994 - 26{,}759 = -\$4765\end{aligned}$$

The actual price of this truck is $4765 *less* than the cost predicted by the regression line with $x =$ miles driven. Why is the actual price less than predicted? There are many possible reasons. Perhaps the truck needs body work, has mechanical issues, or has been in an accident.

EXAMPLE Can you grab more than expected? Calculating and interpreting a residual

PROBLEM: Here again is the scatterplot showing the hand span (in cm) and number of Starburst™ candies grabbed by each student in Mr. Tyson's class. The regression line is $\hat{y} = -29.8 + 2.83x$.


Josh Tabor



Find and interpret the residual for Andres, who has a hand span of 22 cm and grabbed 36 Starburst candies.

SOLUTION:

Residual = actual y − predicted y
$= y - \hat{y}$

$\hat{y} = -29.8 + 2.83(22) = 32.46$ Starburst candies

Residual $= 36 - 32.46 = 3.54$ Starburst candies

Andres grabbed 3.54 more Starburst candies than the number predicted by the regression line with $x =$ hand span.

FOR PRACTICE, TRY EXERCISE 39

CHECK YOUR UNDERSTANDING

Some data were collected on the weight of a male white laboratory rat for the first 25 weeks after its birth. A scatterplot of y = weight (in grams) and x = time since birth (in weeks) shows a fairly strong, positive linear relationship. The regression equation $\hat{y} = 100 + 40x$ models the data fairly well.

1. Predict the rat's weight at 16 weeks old.
2. Calculate and interpret the residual if the rat weighed 700 grams at 16 weeks old.
3. Should you use this line to predict the rat's weight at 2 years old? Use the equation to make the prediction and discuss your confidence in the result. (There are 454 grams in a pound.)

Interpreting a Regression Line

A regression line is a *model* for the data, much like the density curves of Chapter 2. The **y intercept** and **slope** of the regression line describe what this model tells us about the relationship between the response variable y and the explanatory variable x.

The data used to calculate a regression line typically come from a sample. The statistics a and b in the sample regression model estimate the y intercept and slope parameters of the population regression model. You'll learn more about how this works in Chapter 12.

DEFINITION y intercept, Slope

In the regression equation $\hat{y} = a + bx$:

- a is the **y intercept,** the predicted value of y when $x = 0$
- b is the **slope,** the amount by which the predicted value of y changes when x increases by 1 unit

You are probably accustomed to the form $y = mx + b$ for the equation of a line from algebra. Statisticians have adopted a different form for the equation of a regression line. Some use $\hat{y} = b_0 + b_1x$. We prefer the form $\hat{y} = a + bx$ for three reasons: (1) it's simpler, (2) your calculator uses this form, and (3) the formula sheet provided on the AP® exam uses this form. Just remember that the slope is the coefficient of x, no matter what form is used.

Let's return to the Ford F-150 data. The equation of the regression line for these data is $\hat{y} = 38{,}257 - 0.1629x$, where x = miles driven and y = price. The slope $b = -0.1629$ tells us that the *predicted* price of a used Ford F-150 goes down by \$0.1629 (16.29 cents) for each additional mile that the truck has been driven. The y intercept $a = 38{,}257$ is the *predicted* price (in dollars) of a used Ford F-150 that has been driven 0 miles.

AP® EXAM TIP

When asked to interpret the slope or y intercept, it is very important to include the word *predicted* (or equivalent) in your response. Otherwise, it might appear that you believe the regression equation provides actual values of y.

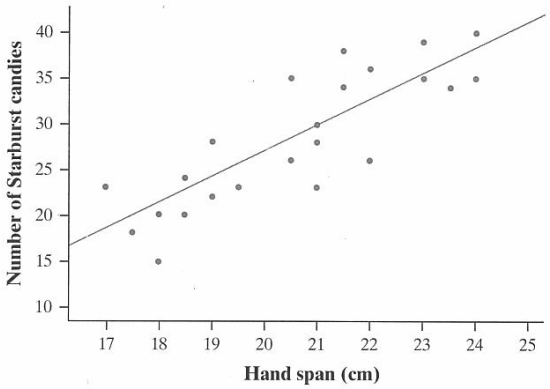
The slope of a regression line is an important numerical description of the relationship between the two variables. Although we need the value of the y intercept to draw the line, it is statistically meaningful only when the explanatory variable can actually take values close to 0, as in the Ford F-150 data. In other cases, using the regression line to make a prediction for $x = 0$ is an extrapolation.

EXAMPLE Grabbing more candy
Interpreting the slope and *y* intercept

PROBLEM: The scatterplot shows the hand span (in cm) and number of Starburst™ candies grabbed by each student in Mr. Tyson's class, along with the regression line $\hat{y} = -29.8 + 2.83x$.

(a) Interpret the slope of the regression line.

(b) Does the value of the y intercept have meaning in this context? If so, interpret the y intercept. If not, explain why.



SOLUTION:

(a) The predicted number of Starburst candies grabbed goes up by 2.83 for each increase of 1 cm in hand span.

Remember that the slope describes how the *predicted* value of y changes, not the actual value of y.

(b) The y intercept does not have meaning in this case, as it is impossible to have a hand span of 0 cm.

Predicting the number of Starburst candies when $x = 0$ is an extrapolation—and results in an unrealistic prediction of −29.8.

FOR PRACTICE, TRY EXERCISE 41

For the Ford F-150 data, the slope $b = -0.1629$ is very close to 0. This does *not* mean that change in miles driven has little effect on price. The size of the slope depends on the units in which we measure the two variables. In this setting, the slope is the predicted change in price (in dollars) when the distance driven increases by 1 mile. There are 100 cents in a dollar. If we measured price in cents instead of dollars, the slope would be 100 times steeper, $b = -16.29$. *You can't say how strong a relationship is by looking at the slope of the regression line.*

CHECK YOUR UNDERSTANDING

Some data were collected on the weight of a male white laboratory rat for the first 25 weeks after its birth. A scatterplot of y = weight (in grams) and x = time since birth (in weeks) shows a fairly strong, positive linear relationship. The regression equation $\hat{y} = 100 + 40x$ models the data fairly well.

1. Interpret the slope of the regression line.
2. Does the value of the y intercept have meaning in this context? If so, interpret the y intercept. If not, explain why.

The Least-Squares Regression Line


Duncan Selby/Alamy

There are many different lines we could use to model the association in a particular scatterplot. A *good* regression line makes the residuals as small as possible.

In the F-150 example, the regression line we used is $\hat{y} = 38{,}257 - 0.1629x$. How does this line make the residuals "as small as possible"? Maybe this line minimizes the *sum* of the residuals. If we add the prediction errors for all 16 trucks, the positive and negative residuals cancel out, as shown in Figure 3.9(a). That's the same issue we faced when we tried to measure deviation around the mean in Chapter 1. We'll solve the current problem in much the same way—by squaring the residuals.

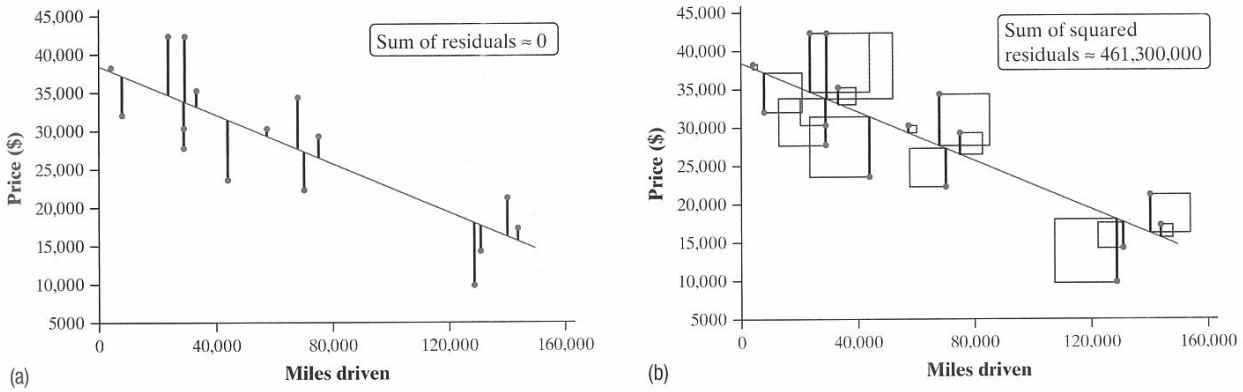


FIGURE 3.9 Scatterplots of the Ford F-150 data with the regression line added. (a) The residuals will add to approximately 0 when using a good regression line. (b) A good regression line should make the sum of squared residuals as small as possible.

A good regression line will have a sum of residuals near 0. But the regression line we prefer is the one that minimizes the sum of the squared residuals. That's what the line shown in Figure 3.9(b) does for the Ford F-150 data, which is why we call it the **least-squares regression line**. No other regression line would give a smaller sum of squared residuals.

In addition to minimizing the sum of squared residuals, the least-squares regression line always goes through the point $(\bar{x}, \bar{y})$.

DEFINITION Least-squares regression line

The **least-squares regression line** is the line that makes the sum of the squared residuals as small as possible.

Your calculator or statistical software will give the equation of the least-squares line from data that you enter. Then you can concentrate on understanding and using the regression line.

AP® EXAM TIP

When displaying the equation of a least-squares regression line, the calculator will report the slope and intercept with much more precision than we need. There is no firm rule for how many decimal places to show for answers on the AP® Statistics exam. Our advice: decide how much to round based on the context of the problem you are working on.

9. Technology Corner CALCULATING LEAST-SQUARES REGRESSION LINES

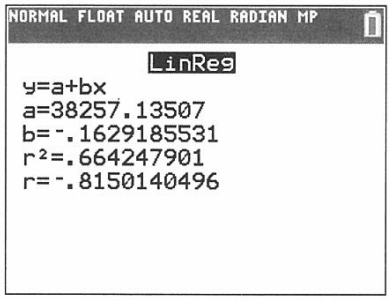
TI-Nspire and other technology instructions are on the book's website at highschool.bfwpub.com/updatedtps6e.

Let's use the Ford F-150 data to show how to find the equation of the least-squares regression line on the TI-83/84. Here are the data again:

Miles driven	70,583	129,484	29,932	29,953	24,495	75,678	8359	4447
Price ($)	21,994	9500	29,875	41,995	41,995	28,986	31,891	37,991
Miles driven	34,077	58,023	44,447	68,474	144,162	140,776	29,397	131,385
Price ($)	34,995	29,988	22,896	33,961	16,883	20,897	27,495	13,997

1. Enter the miles driven data into L1 and the price data into L2.
2. To determine the least-squares regression line, press [STAT]; choose CALC and then LinReg(a+bx).

- **OS 2.55 or later:** In the dialog box, enter the following: Xlist:L1, Ylist:L2, FreqList (leave blank), Store RegEQ (leave blank), and choose Calculate.
- **Older OS:** Finish the command to read LinReg(a+bx) L1,L2 and press [ENTER].

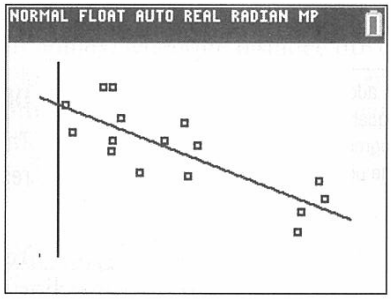


Note: If r^2 and r do not appear on the TI-83/84 screen, do this one-time series of keystrokes:

- **OS 2.55 or later:** Press [MODE] and set STAT DIAGNOSTICS to ON. Then redo Step 2 to calculate the least-squares line. The r^2 and r values should now appear.
- **Older OS:** Press [2nd] [0] (CATALOG), scroll down to DiagnosticOn, and press [ENTER]. Press [ENTER] again to execute the command. The screen should say "Done." Then redo Step 2 to calculate the least-squares line. The r^2 and r values should now appear.

To graph the least-squares regression line on the scatterplot:

1. Set up a scatterplot (see Technology Corner 8 on page 159).
2. Press [Y=] and enter the equation of the least-squares regression line in Y1.
3. Press [ZOOM] and choose ZoomStat to see the scatterplot with the least-squares regression line.



Note: When you calculate the equation of the least-squares regression line, you can have the calculator store the equation to Y1. When setting up the calculation, enter Y1 for the StoreRegEq prompt blank (OS 2.55 or later) or use the following command (older OS): LinReg(a+bx) L1,L2,Y1. Y1 is found by pressing [VARS] and selecting Y-VARS, then Function, then Y1.

Determining if a Linear Model Is Appropriate: Residual Plots

One of the first principles of data analysis is to look for an overall pattern and for striking departures from the pattern. A regression line describes the overall pattern of a linear relationship between an explanatory variable and a response variable. We see departures from this pattern by looking at a **residual plot**.

Some software packages prefer to plot the residuals against the predicted values $\hat{y}$ instead of against the values of the explanatory variable. The basic shape of the two plots is the same because $\hat{y}$ is linearly related to x.

DEFINITION Residual plot

A **residual plot** is a scatterplot that displays the residuals on the vertical axis and the explanatory variable on the horizontal axis.

Residual plots help us assess whether or not a linear model is appropriate. In Figure 3.10(a), the scatterplot shows the relationship between the average income (gross domestic product per person, in dollars) and fertility rate (number of children per woman) in 187 countries, along with the least-squares regression line. The residual plot in Figure 3.10(b) shows the average income for each country and the corresponding residual.

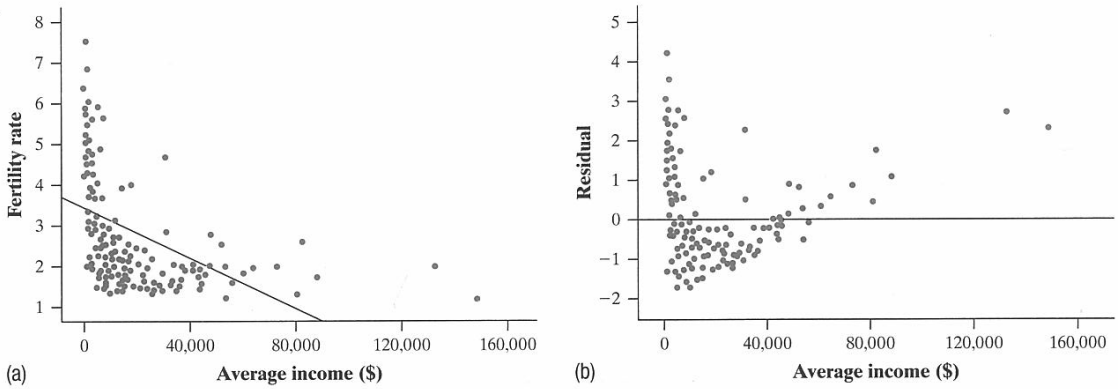


FIGURE 3.10 The (a) scatterplot and (b) residual plot for the linear model relating fertility rate to average income for a sample of countries.

The least-squares regression line clearly doesn't fit this association very well! For most countries with average incomes under $5000, the actual fertility rates are greater than predicted, resulting in positive residuals. For countries with average incomes between $5000 and $60,000, the actual fertility rates tend to be smaller than predicted, resulting in negative residuals. Countries with average incomes above $60,000 all have fertility rates greater than predicted, again resulting in positive residuals. This U-shaped pattern in the residual plot indicates that the linear form of our model doesn't match the form of the association. A curved model might be better in this case.

In Figure 3.11(a), the scatterplot shows the Ford F-150 data, along with the least-squares regression line. The corresponding residual plot is shown in Figure 3.11(b).

Looking at the scatterplot, the line seems to be a good fit for this relationship. You can "see" that the line is appropriate by the lack of a leftover curved pattern in the residual plot. In fact, the residuals look randomly scattered around the residual = 0 line.

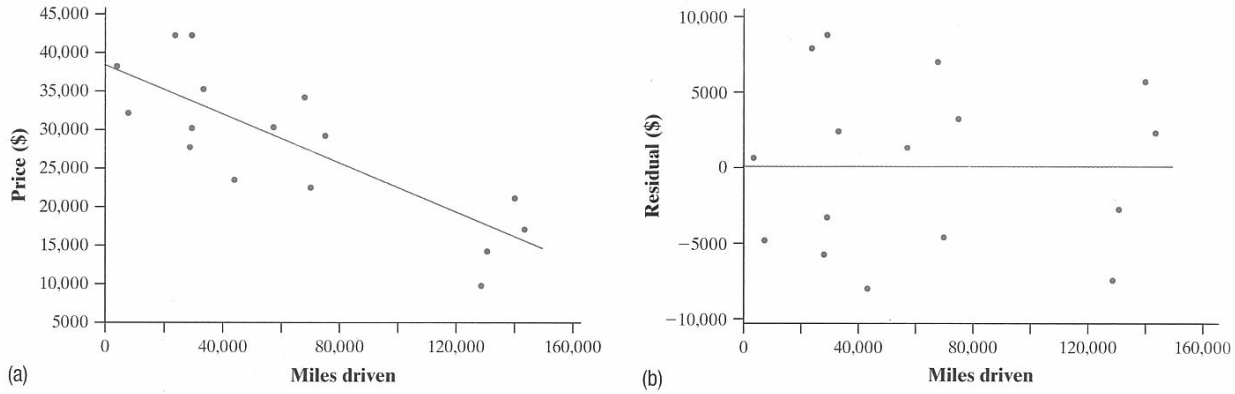


FIGURE 3.11 The (a) scatterplot and (b) residual plot for the linear model relating price to miles driven for Ford F-150s.

HOW TO INTERPRET A RESIDUAL PLOT

To determine whether the regression model is appropriate, look at the residual plot.

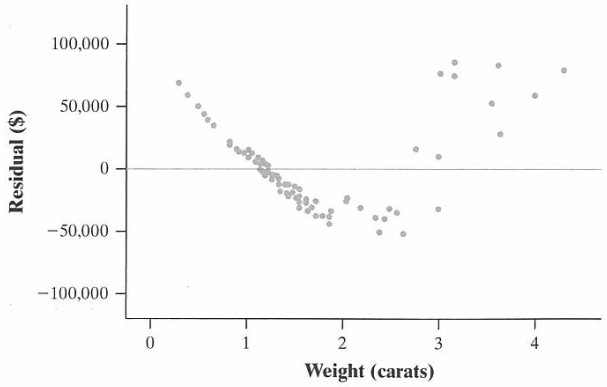
- If there is no leftover curved pattern in the residual plot, the regression model is appropriate.
- If there is a leftover curved pattern in the residual plot, consider using a regression model with a different form.

EXAMPLE

Pricing diamonds
Interpreting a residual plot


JGI/Getty Images

PROBLEM: Is a linear model appropriate to describe the relationship between the weight (in carats) and price (in dollars) of round, clear, internally flawless diamonds with excellent cuts? We calculated a least-squares regression line using x = weight and y = price and made the corresponding residual plot shown.[23] Use the residual plot to determine if the linear model is appropriate.



SOLUTION:

The linear model relating price to carat weight is not appropriate because there is a U-shaped pattern left over in the residual plot.

FOR PRACTICE, TRY EXERCISE 47

Think About It

WHY DO WE LOOK FOR PATTERNS IN RESIDUAL PLOTS? The word *residual* comes from the Latin word *residuum*, meaning "left over." When we calculate a residual, we are calculating what is left over after subtracting the predicted value from the actual value:

$$\text{residual} = \text{actual } y - \text{predicted } y$$

Likewise, when we look at the form of a residual plot, we are looking at the form that is left over after subtracting the form of the model from the form of the association:

$$\text{form of residual plot} = \text{form of association} - \text{form of model}$$

When there is a leftover form in the residual plot, the form of the association and form of the model are not the same. However, if the form of the association and form of the model are the *same*, the residual plot should have no form, other than random scatter.

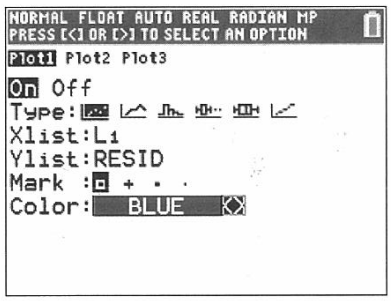
10. Technology Corner MAKING RESIDUAL PLOTS

TI-Nspire and other technology instructions are on the book's website at highschool.bfwpub.com/updatedtps6e.

Let's continue the analysis of the Ford F-150 miles driven and price data from Technology Corner 9 (page 184). You should have already made a scatterplot, calculated the equation of the least-squares regression line, and graphed the line on the scatterplot. Now, we want to calculate residuals and make a residual plot. Fortunately, your calculator has already done most of the work. Each time the calculator computes a regression line, it computes the residuals and stores them in a list named RESID.

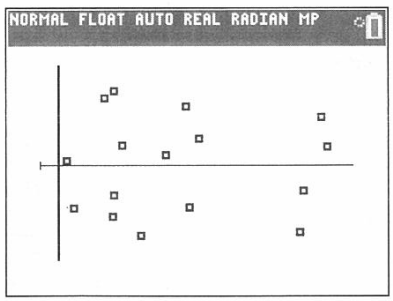
1. Set up a scatterplot in the statistics plots menu.

   - Press 2nd Y= (STAT PLOT).
   - Press ENTER or 1 to go into Plot1.
   - Adjust the settings as shown. The RESID list is found in the List menu by pressing 2nd STAT. *Note:* You have to calculate the equation of the least-squares regression line using the calculator *before* making a residual plot. Otherwise, the RESID list will include the residuals from a different least-squares regression line.



2. Use ZoomStat to let the calculator choose an appropriate window.

   - Press ZOOM and choose 9: ZoomStat.



Note: If you want to see the values of the residuals, you can have the calculator put them in L3 (or any list). In the list editor, highlight the heading of L3, choose the RESID list from the LIST menu, and press ENTER.

CHECK YOUR UNDERSTANDING

In Exercises 3 and 7, we asked you to make and describe a scatterplot for the hiker data shown in the table.

Body weight (lb)	120	187	109	103	131	165	158	116
Backpack weight (lb)	26	30	26	24	29	35	31	28

1. Calculate the equation of the least-squares regression line.
2. Make a residual plot for the linear model in Question 1.
3. What does the residual plot indicate about the appropriateness of the linear model? Explain your answer.

How Well the Line Fits the Data: The Role of s and r^2 in Regression

We use a residual plot to determine if a least-squares regression line is an appropriate model for the relationship between two variables. Once we determine that a least-squares regression line is appropriate, it makes sense to ask a follow-up question: How well does the line work? That is, if we use the least-squares regression line to make predictions, how good will these predictions be?

THE STANDARD DEVIATION OF THE RESIDUALS We already know that a residual measures how far an actual y value is from its corresponding predicted value $\hat{y}$. Earlier in this section, we calculated the residual for the Ford F-150 with 70,583 miles driven and price \$21,994. As shown in Figure 3.12, the residual was −\$4765, meaning that the actual price was \$4765 *less* than we predicted.

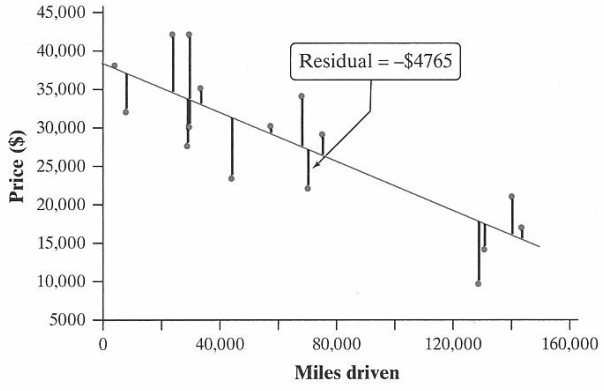


FIGURE 3.12 Scatterplot of the Ford F-150 data with a regression line added. Residuals for each truck are shown with vertical line segments.

To assess how well the line fits *all* the data, we need to consider the residuals for each of the trucks, not just one. Here are the residuals for all 16 trucks:

−4765	−7664	−3506	8617	7728	3057	−5004	458
2289	1183	−8121	6858	2110	5572	−5973	−2857

Using these residuals, we can estimate the "typical" prediction error when using the least-squares regression line. To do this, we calculate the **standard deviation of the residuals** s.

DEFINITION Standard deviation of residuals s

The **standard deviation of the residuals s** measures the size of a typical residual. That is, s measures the typical distance between the actual y values and the predicted y values.

To calculate s, use the following formula:

$$s = \sqrt{\frac{\text{sum of squared residuals}}{n-2}} = \sqrt{\frac{\Sigma(y_i - \hat{y}_i)^2}{n-2}}$$

For the Ford F-150 data, the standard deviation of the residuals is

$$s = \sqrt{\frac{(-4765)^2 + (-7664)^2 + \cdots + (-2857)^2}{16-2}} = \sqrt{\frac{461,264,136}{14}} = \$5740$$

Interpretation: The actual price of a Ford F-150 is typically about \$5740 away from the price predicted by the least-squares regression line with x = miles driven. If we look at the residual plot in Figure 3.11, this seems like a reasonable value. Although some of the residuals are close to 0, others are close to \$10,000 or −\$10,000.

Think About It

DOES THE FORMULA FOR s LOOK SLIGHTLY FAMILIAR? It should. In Chapter 1, we defined the standard deviation of a set of quantitative data as

$$s_x = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}}$$

We interpreted the resulting value as the "typical" distance of the data points from the mean. In the case of two-variable data, we're interested in the typical (vertical) distance of the data points from the regression line. We find this value in much the same way: first add up the squared deviations, then average them (again, in a funny way), and take the square root to get back to the original units of measurement.

THE COEFFICIENT OF DETERMINATION r^2 There is another numerical quantity that tells us how well the least-squares line predicts values of the response variable y. It is r^2, the **coefficient of determination**. Some computer packages call it "R-sq." You may have noticed this value in some of the output that we showed earlier. Although it's true that r^2 is equal to the square of the correlation r, there is much more to this story.

Some people interpret r^2 as the proportion of variation in the response variable that is explained by the explanatory variable in the model.

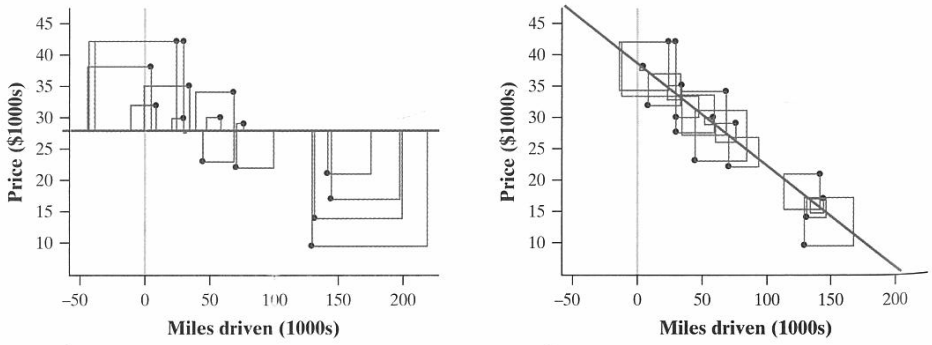
DEFINITION The coefficient of determination r^2

The **coefficient of determination r^2** measures the percent reduction in the sum of squared residuals when using the least-squares regression line to make predictions, rather than the mean value of y. In other words, r^2 measures the percent of the variability in the response variable that is accounted for by the least-squares regression line.

Holly Albrecht

Suppose we wanted to predict the price of a particular used Ford F-150, but we didn't know how many miles it had been driven. Our best guess would be the average cost of a used Ford F-150, $\bar{y} = \$27,834$. Of course, this prediction is unlikely to be very good, as the prices vary quite a bit from the mean ($s_y = \$9570$). If we knew how many miles the truck had been driven, we could use the least-squares regression line to make a better prediction. How much better are predictions that use the least-squares regression line with $x =$ miles driven, rather than predictions that use only the average price? The answer is r^2.

The scatterplot in Figure 3.13(a) shows the squared residuals along with the sum of squared residuals (approximately 1,374,000,000) when using the average price as the predicted value. The scatterplot in Figure 3.13(b) shows the squared residuals along with the sum of squared residuals (approximately 461,300,000) when using the least-squares regression line with $x =$ miles driven to predict the price. Notice that the squares in part (b) are quite a bit smaller.



$\widehat{\text{Price}} = 27{,}834$
(a) Sum of squares ≈ 1,374,000,000

$\widehat{\text{Price}} = 38{,}257 - 0.1629$ Miles driven; $r^2 = 0.66$
(b) Sum of squares ≈ 461,300,000

FIGURE 3.13 (a) The sum of squared residuals is about 1,374,000,000 if we use the mean price as our prediction for all 16 trucks. (b) The sum of squared residuals from the least-squares regression line is about 461,300,000.

To find r^2, calculate the percent reduction in the sum of squared residuals:

$$r^2 = \frac{1{,}374{,}000{,}000 - 461{,}300{,}000}{1{,}374{,}000{,}000} = \frac{912{,}700{,}000}{1{,}374{,}000{,}000} = 0.66$$

The sum of squared residuals has been reduced by 66%.

Interpretation: About 66% of the variability in the price of a Ford F-150 is accounted for by the least-squares regression line with x = miles driven. The remaining 34% is due to other factors, including age, color, and condition.

If all the points fall directly on the least-squares line, the sum of squared residuals is 0 and $r^2 = 1$. Then all the variation in y is accounted for by the linear relationship with x. In the worst-case scenario, the least-squares line does no better at predicting y than $y = \bar{y}$ does. Then the two sums of squared residuals are the same and $r^2 = 0$.

It's fairly remarkable that the coefficient of determination r^2 is actually the square of the correlation. This fact provides an important connection between correlation and regression. When you see a linear association, square the correlation to get a better feel for how well the least-squares line fits the data.

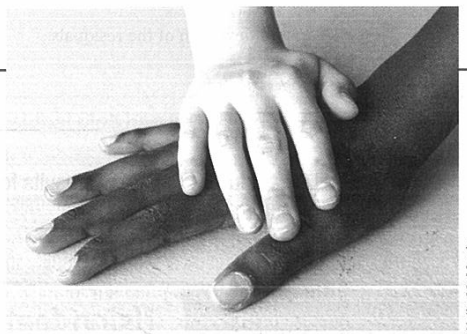
Think About It

WHAT'S THE RELATIONSHIP BETWEEN s AND r^2? Both s and r^2 are calculated from the sum of squared residuals. They also both measure how well the line fits the data. The standard deviation of the residuals reports the size of a typical prediction error, in the same units as the response variable. In the truck example, $s = \$5740$. The value of r^2, however, does not have units and is usually expressed as a percentage between 0% and 100%, such as $r^2 = 66\%$. Because these values assess how well the line fits the data in different ways, we recommend you follow the example of most statistical software and report both.

Knowing how to interpret s and r^2 is much more important than knowing how to calculate them. Consequently, we typically let technology do the calculations.

EXAMPLE

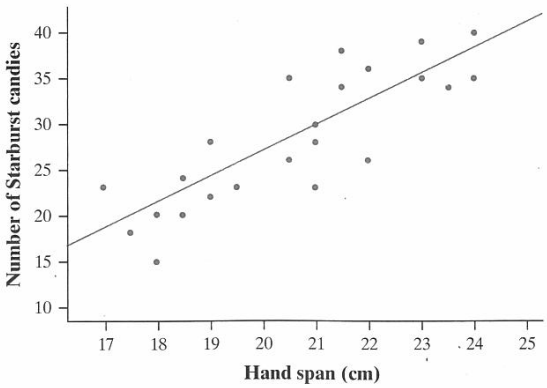
Grabbing candy, again
Interpreting s and r^2


Kylie McManis

PROBLEM: The scatterplot shows the hand span (in centimeters) and number of Starburst™ candies grabbed by each student in Mr. Tyson's class, along with the regression line $\hat{y} = -29.8 + 2.83x$. For this model, technology gives $s = 4.03$ and $r^2 = 0.697$.

(a) Interpret the value of s.

(b) Interpret the value of r^2.



SOLUTION:

(a) The actual number of Starburst™ candies grabbed is typically about 4.03 away from the number predicted by the least-squares regression line with x = hand span.

(b) About 69.7% of the variability in number of Starburst candies grabbed is accounted for by the least-squares regression line with x = hand span.

FOR PRACTICE, TRY EXERCISE 55

Interpreting Computer Regression Output

Figure 3.14 displays the basic regression output for the Ford F-150 data from two statistical software packages: Minitab and JMP. Other software produces very similar output. Each output records the slope and y intercept of the least-squares line. The software also provides information that we don't yet need, although we will use much of it later. Be sure that you can locate the slope, the y intercept, and the values of s (called *root mean square error* in JMP) and r^2 on both computer outputs. *Once you understand the statistical ideas, you can read and work with almost any software output.*

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

r^2

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

Standard deviation of the residuals

JMP

r^2

Summary of Fit	
RSquare	0.664248
RSquare Adj	0.640266
Root Mean Square Error	5740.131
Mean of Response	27833.69
Observations (or Sum Wgts)	16

Standard deviation of the residuals

Parameter Estimates				
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	38257.135	2445.813	15.64	<.0001
Miles Driven	-0.162919	0.030956	-5.26	0.0001

y intercept

Slope

FIGURE 3.14 Least-squares regression results for the Ford F-150 data from Minitab and JMP statistical software. Other software produces similar output.

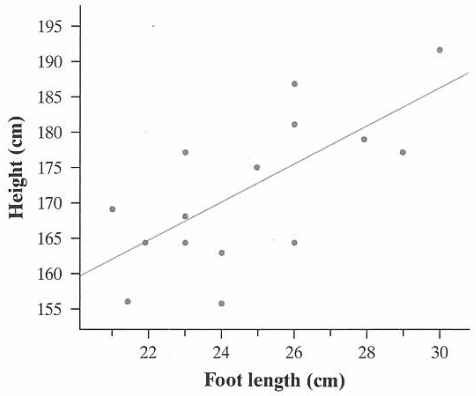
EXAMPLE

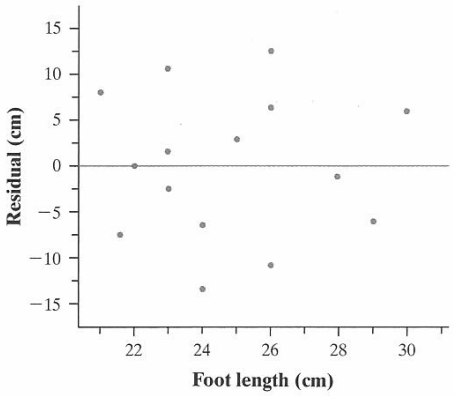
Using feet to predict height

Interpreting regression output

PROBLEM: A random sample of 15 high school students was selected from the U.S. Census At School database. The foot length (in centimeters) and height (in centimeters) of each student in the sample were recorded. Here are a scatterplot with the least-squares regression line added, a residual plot, and some computer output:

© Fancy/Alamy





Predictor	Coef	SE Coef	T	P
Constant	103.4100	19.5000	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%

(a) Is a line an appropriate model to use for these data? Explain how you know.

(b) Find the correlation.

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

(d) By about how much do the actual heights typically vary from the values predicted by the least-squares regression line with x = foot length?

SOLUTION:

(a) *Because the scatterplot shows a linear association and the residual plot has no obvious leftover curved patterns, a line is an appropriate model to use for these data.*

(b) $r = \sqrt{0.486} = 0.697$

The correlation r is the square root of r^2, where r^2 is a value between 0 and 1. Because the square root function on your calculator will always give a positive result, make sure to consider whether the correlation is positive or negative. If the slope is negative, so is the correlation.

(c) $\widehat{\text{height}} = 103.41 + 2.7469\ (\text{foot length})$

(d) $s = 7.95$, *so the actual heights typically vary by about 7.95 cm from the values predicted by the regression line with* x = *foot length.*

We could also write the equation as $\hat{y} = 103.41 + 2.7469x$, where $\hat{y}$ = predicted height (cm) and x = foot length (cm).

FOR PRACTICE, TRY EXERCISE 59

CHECK YOUR UNDERSTANDING

In Section 3.1, you read about the Old Faithful geyser in Yellowstone National Park. The computer output shows the results of a regression of y = interval of time until the next eruption (in minutes) and x = duration of the most recent eruption (in minutes) for each eruption of Old Faithful in a particular month.

Summary of Fit

RSquare	0.853725
RSquare Adj	0.853165
Root Mean Square Error	6.493357
Mean of Response	77.543730
Observations (or Sum Wgts)	263.000000

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	33.347442	1.201081	27.76	<.0001*
Duration	13.285406	0.340393	39.03	<.0001*

1. What is the equation of the least-squares regression line that models the relationship between interval and duration? Define any variables that you use.
2. Interpret the slope of the least-squares regression line.
3. Identify and interpret the standard deviation of the residuals.
4. What percent of the variability in interval is accounted for by the least-squares regression line with x = duration?

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. Exploring this method will highlight an important relationship between the correlation and the slope of a least-squares regression line—and reveal why we include the word *regression* in the expression *least-squares regression line.*

HOW TO CALCULATE THE LEAST-SQUARES REGRESSION LINE USING SUMMARY STATISTICS

We have data on an explanatory variable x and a response variable y for n individuals and want to calculate the least-squares regression line $\hat{y} = a + bx$. From the data, calculate the means $\bar{x}$ and $\bar{y}$ and the standard deviations s_x and s_y of the two variables and their correlation r. The **slope** is:

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

Because the least-squares regression line passes through the point $(\bar{x}, \bar{y})$, the **y intercept** is:

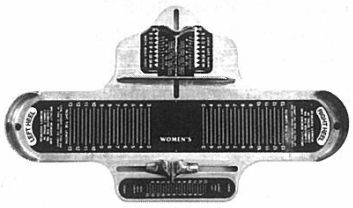
$$a = \bar{y} - b\bar{x}$$

The formula for the slope reminds us that the distinction between explanatory and response variables is important in regression. Least-squares regression makes the distances of the data points from the line small only in the y direction. If we reverse the roles of the two variables, the values of s_x and s_y will reverse in the slope formula, resulting in a different least-squares regression line. This is *not* true for correlation: switching x and y does *not* affect the value of r.

The formula for the y intercept comes from the fact that the least-squares regression line always passes through the point $(\bar{x}, \bar{y})$. Once we know the slope (b) and that the line goes through the point $(\bar{x}, \bar{y})$, we can use algebra to solve for the y intercept. Substituting $(\bar{x}, \bar{y})$ into the equation $\hat{y} = a + bx$ produces the equation $\bar{y} = a + b\bar{x}$. Solving this equation for a gives the equation shown in the definition box, $a = \bar{y} - b\bar{x}$. To see how these formulas work in practice, let's look at an example.

EXAMPLE More about feet and height
Calculating the least-squares regression line

PROBLEM: In the preceding 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 $\bar{x} = 24.76$ and $s_x = 2.71$. The mean and standard deviation of the heights are $\bar{y} = 171.43$ and $s_y = 10.69$. The correlation between foot length and height is $r = 0.697$. Find the equation of the least-squares regression line for predicting height from foot length.



panpote/Getty Images

SOLUTION:

$$b = r\frac{s_y}{s_x}$$
$$a = \bar{y} - b\bar{x}$$

$$b = 0.697\frac{10.69}{2.71} = 2.75$$

$$a = 171.43 - 2.75(24.76) = 103.34$$

The equation of the least-squares regression line is $\hat{y} = 103.34 + 2.75x$.

FOR PRACTICE, TRY EXERCISE 63

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

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

This equation says that along the regression line, a change of 1 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. For example, if $r = 1$ and x is 2 standard deviations above $\bar{x}$, then the corresponding value of $\hat{y}$ will be 2 standard deviations above $\bar{y}$.

However, if the variables are not perfectly correlated ($-1 < r < 1$), the change in $\hat{y}$ is *less than* the change in x, when measured in standard deviation units. To illustrate this property, let's return to the foot length and height data from the preceding example.

Figure 3.15 shows the scatterplot of height versus foot length and the regression line $\hat{y} = 103.34 + 2.75x$. We have added four more lines to the graph: a vertical line at the mean foot length $\bar{x}$, a vertical line at $\bar{x} + s_x$ (1 standard deviation above the mean foot length), a horizontal line at the mean height $\bar{y}$, and a horizontal line at $\bar{y} + s_y$ (1 standard deviation above the mean height).

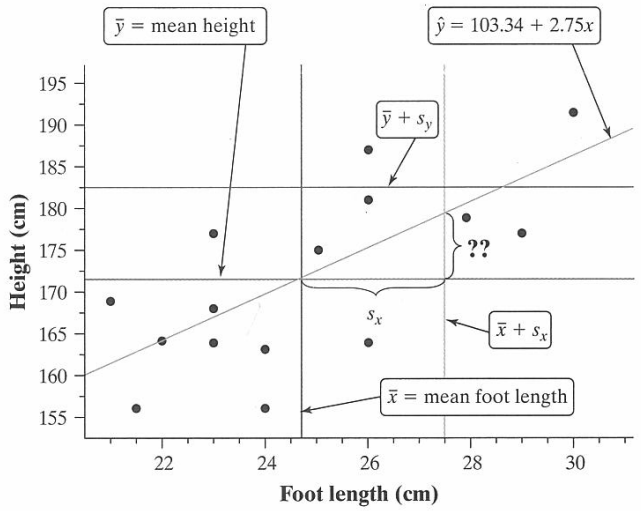


FIGURE 3.15 Scatterplot showing the relationship between foot length and height for a sample of students, along with lines showing the means of x and y and the values 1 standard deviation above each mean.

When a student's foot length is 1 standard deviation above the mean foot length $\bar{x}$, the predicted height $\hat{y}$ is above the mean height $\bar{y}$—but not an entire standard deviation above the mean. How far above the mean is the value of $\hat{y}$?

From the graph, we can see that

$$b = \text{slope} = \frac{\text{change in } y}{\text{change in } x} = \frac{??}{s_x}$$

From earlier, we know that

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

Setting these two equations equal to each other, we have

$$\frac{??}{s_x} = \frac{r \cdot s_y}{s_x}$$

Thus, $\hat{y}$ must be $r \cdot s_y$ above the mean $\bar{y}$.

In other words, for an increase of 1 standard deviation in the value of the explanatory variable x, the least-squares regression line predicts an increase of *only* r standard deviations in the response variable y. When the correlation isn't $r = 1$ or $r = -1$, the predicted value of y is closer to its mean $\bar{y}$ than the value of x is to its mean $\bar{x}$. *This is called regression to the mean, because the values of y "regress" to their mean.*

Sir Francis Galton (1822–1911) is often credited with discovering the idea of regression to the mean. He looked at data on the heights of children versus the heights of their parents. He found that taller-than-average parents tended to have

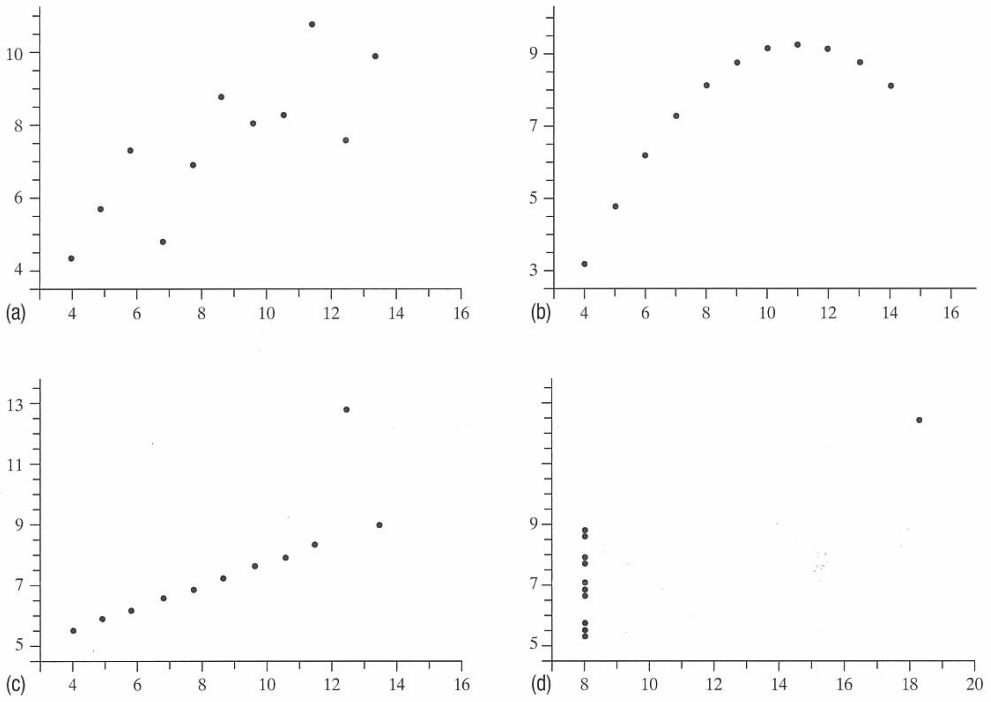
children who were taller than average but not quite as tall as their parents. Likewise, shorter-than-average parents tended to have children who were shorter than average but not quite as short as their parents. Galton used the symbol r for the correlation because of its important relationship to regression.

Correlation and Regression Wisdom

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

CORRELATION AND REGRESSION LINES DESCRIBE ONLY LINEAR RELATIONSHIPS You can calculate the correlation and the least-squares line for any relationship between two quantitative variables, but the results are useful only if the scatterplot shows a linear pattern. *Always plot your data first!*

The following four scatterplots show very different relationships. Which one do you think shows the greatest correlation?

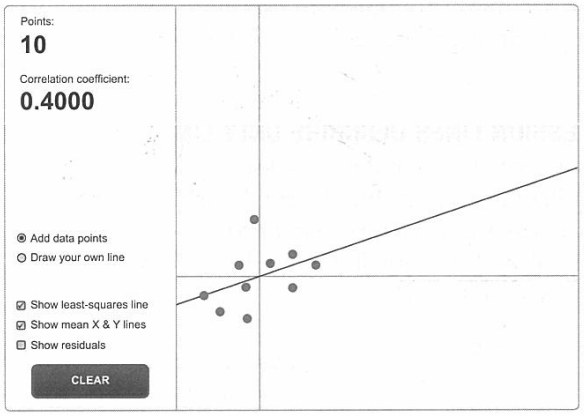


Answer: All four have the same correlation, $r = 0.816$. Furthermore, the least-squares regression line for each relationship is exactly the same, $\hat{y} = 3 + 0.5x$. These four data sets, developed by statistician Frank Anscombe, illustrate the importance of graphing data before doing calculations.[24]

CORRELATION AND LEAST-SQUARES REGRESSION LINES ARE NOT RESISTANT You already know that the correlation r is not resistant. One unusual point in a scatterplot can greatly change the value of r. Is the least-squares line resistant? The following activity will help you answer this question.

ACTIVITY Investigating properties of the least-squares regression line

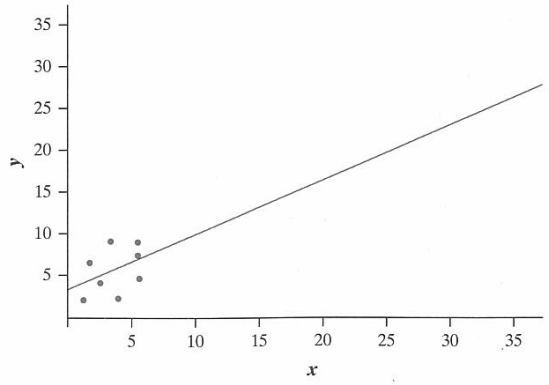
In this activity, you will use the *Correlation and Regression* applet to explore some properties of the least-squares regression line.



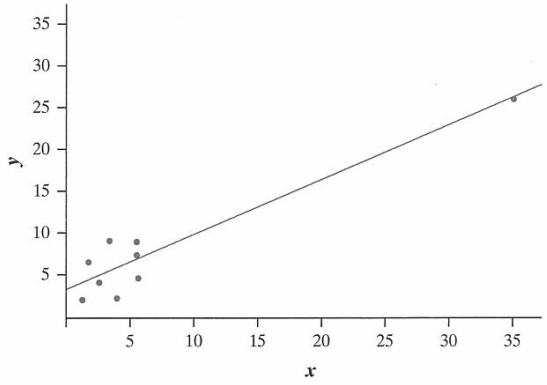
1. Launch the applet at highschool.bfwpub.com/updatedtps6e.
2. Click on the graphing area to add 10 points in the lower-left corner so that the correlation is about $r = 0.40$. Also, check the boxes to show the "Least-Squares Line" and the "Mean X & Y" lines as in the screen shot. Notice that the least-squares regression line goes though the point $(\bar{x}, \bar{y})$.
3. If you were to add a point on the least-squares regression line at the right edge of the graphing area, what do you think would happen to the least-squares regression line? To the value of r^2? (Remember that r^2 is the square of the correlation coefficient, which is provided by the applet.) Add the point to see if you were correct.
4. Click on the point you just added, and drag it up and down along the right edge of the graphing area. What happens to the least-squares regression line? To the value of r^2?
5. Now, move this point so that it is on the vertical $\bar{x}$ line. Drag the point up and down on the $\bar{x}$ line. What happens to the least-squares regression line? To the value of r^2?
6. Briefly summarize how unusual points influence the least-squares regression line.

As you learned in the activity, unusual points may or may not have an influence on the least-squares regression line and the coefficient of determination r^2. The same is true for the correlation r and the standard deviation of the residuals s. Here are four scatterplots that summarize the possibilities. In all four scatterplots, the 8 points in the lower left are the same.

Case 1: No unusual points



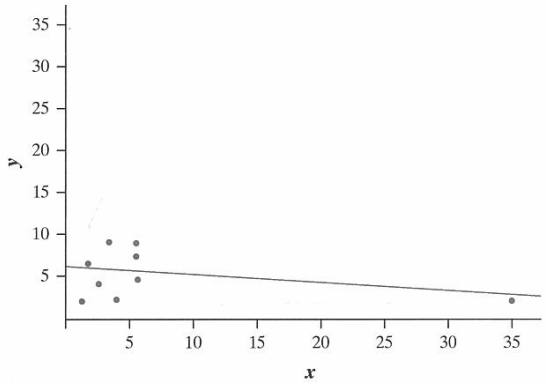
Case 2: A point that is far from the other points in the x direction, but in the same pattern.



$\hat{y} = 3.04 + 0.65x$
$r = 0.94$
$r^2 = 0.89$
$s = 2.56$

Compared to Case 1, the equation of the least-squares regression line remained the same, but the values of r and r^2 greatly increased. The standard deviation of the residuals got a bit smaller because the additional point has a very small residual.

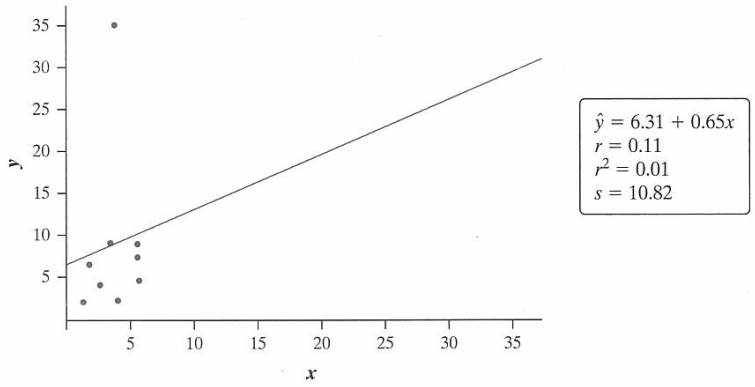
Case 3: A point that is far from the other points in the x direction, and not in the same pattern.



$\hat{y} = 5.80 - 0.09x$
$r = -0.34$
$r^2 = 0.18$
$s = 2.89$

Compared to Case 1, the equation of the least-squares regression line is much different, with the slope going from positive to negative and the y intercept increasing. The value of r is now negative while the value of r^2 stayed almost the same. Even though the new point has a relatively small residual, the standard deviation of the residuals got a bit larger because the line doesn't fit the remaining points nearly as well.

Case 4: A point that is far from the other points in the y direction, and not in the same pattern.



Compared to Case 1, the slope of the least-squares regression line is the same, but the y intercept is a little larger as the line appears to have shifted up slightly. The values of r and r^2 are much smaller than before. Because the new point has such a large residual, the standard deviation of the residuals is much larger.

In Cases 2 and 3, the unusual point had a much bigger x value than the other points. Points whose x values are much smaller or much larger than the other points in a scatterplot have **high leverage**. In Case 4, the unusual point had a very large residual. Points with large residuals are called **outliers**. All three of these unusual points are considered **influential points** because adding them to the scatterplot substantially changed either the equation of the least-squares regression line or one or more of the other summary statistics (r, r^2, s).

DEFINITION High leverage, Outlier, Influential point

Points with **high leverage** in regression have much larger or much smaller x values than the other points in the data set.

An **outlier** in regression is a point that does not follow the pattern of the data and has a large residual.

An **influential point** in regression is any point that, if removed, substantially changes the slope, y intercept, correlation, coefficient of determination, or standard deviation of the residuals.

Outliers and high-leverage points are often influential in regression calculations! The best way to investigate the influence of such points is to do regression calculations with and without them to see how much the results differ. Here is an example that shows what we mean.

Does the age at which a child begins to talk predict a later score on a test of mental ability? A study of the development of young children recorded the age in months at which each of 21 children spoke their first word and their Gesell Adaptive Score, the result of an aptitude test taken much later.[25] A scatterplot of the data appears in Figure 3.16, along with a residual plot, and computer output. Two points, child 18 and child 19, are labeled on each plot.

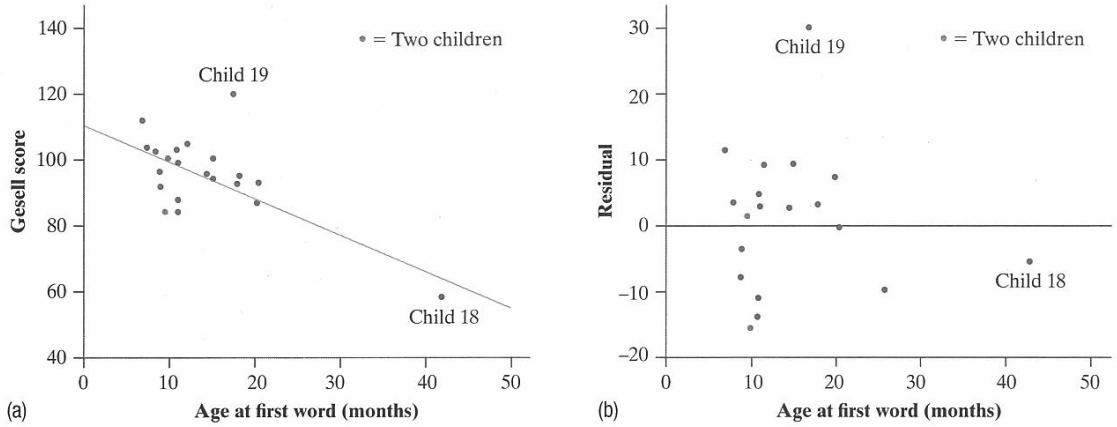


FIGURE 3.16 (a) Scatterplot of Gesell Adaptive Scores versus the age at first word for 21 children, along with the least-squares regression line. (b) Residual plot for the linear model. The point for Child 18 has high leverage and the point for Child 19 is an outlier. Each purple point in the graphs stands for two individuals.

Clover No.7 Photography/Getty Images

The point for Child 18 has high leverage because its x value is much larger than the x values of other points. The point for Child 19 is an outlier because it falls outside the pattern of the other points and has a very large residual. How do these two points affect the regression? Figure 3.17 shows the results of removing each of these points on the equation of the least-squares regression line, the standard deviation of the residuals, and r^2.

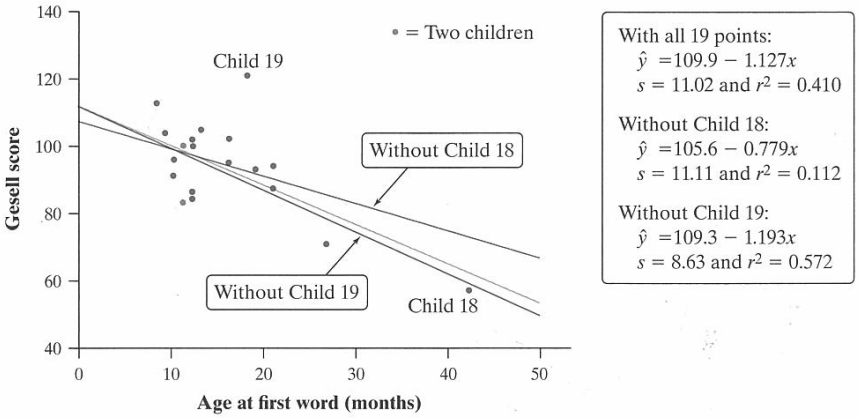


FIGURE 3.17 Three least-squares regression lines of Gesell score on age at first word. The green line is calculated from all the data. The dark blue line is calculated leaving out only Child 18. The red line is calculated leaving out only Child 19.

You can see that removing the point for Child 18 moves the line quite a bit. Because of Child 18's extreme position on the age (x) scale, removing this high-leverage point makes the slope closer to 0 and the y intercept smaller. Removing Child 18 also increases the standard deviation of the residuals because its small residual was making the typical distance from the regression line smaller. Finally, removing Child 18 also decreases r^2 (and makes the correlation closer to 0) because the linear association is weaker without this point.

Child 19's Gesell score was far above the least-squares regression line, but this child's age (17 months) is very close to $\bar{x} = 14.4$ months, making this point an outlier with low leverage. Thus, removing Child 19 has very little effect on the least-squares regression line. The line shifts down slightly from the original regression line, but not by much. Child 19 has a bigger influence on the standard deviation of the residuals: without Child 19's big residual, the size of the typical residual goes from $s = 11.02$ to $s = 8.63$. Likewise, without Child 19, the strength of the linear association increases and r^2 goes from 0.410 to 0.572.

Think About It

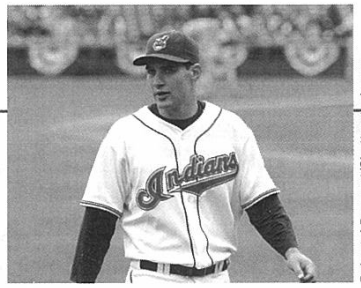
WHAT SHOULD WE DO WITH UNUSUAL POINTS? The strong influence of Child 18 makes the original regression of Gesell score on age at first word misleading. The original data have $r^2 = 0.41$. That is, the least-squares line with $x =$ age at which a child begins to talk accounts for 41% of the variability in Gesell score. This relationship is strong enough to be interesting to parents. If we leave out Child 18, r^2 drops to only 11%. The apparent strength of the association was largely due to a single influential observation.

What should the child development researcher do? She must decide whether Child 18 is so slow to speak that this individual should not be allowed to influence the analysis. If she excludes Child 18, much of the evidence for a connection between the age at which a child begins to talk and later ability score vanishes. If she keeps Child 18, she needs data on other children who were also slow to begin talking, so the analysis no longer depends as heavily on just one child.

EXAMPLE

Dodging the pattern?

Outliers and high-leverage points

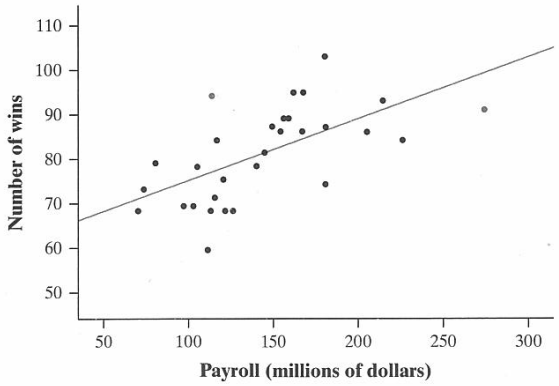


Robert J Daveant/Shutterstock.com

PROBLEM: The scatterplot shows the payroll (in millions of dollars) and number of wins for Major League Baseball teams in 2016, along with the least-squares regression line. The points highlighted in red represent the Los Angeles Dodgers (far right) and the Cleveland Indians (upper left).

(a) Describe what influence the point representing the Los Angeles Dodgers has on the equation of the least-squares regression line. Explain your reasoning.

(b) Describe what influence the point representing the Cleveland Indians has on the standard deviation of the residuals and r^2. Explain your reasoning.



SOLUTION:

(a) Because the point for the Los Angeles Dodgers is on the right and below the least-squares regression line, it is making the slope of the line closer to 0 and the y intercept greater. If the Dodgers' point was removed, the line would be steeper.

The point for the Dodgers has high leverage because its x value is much larger than the others.

(b) Because the point for the Cleveland Indians has a large residual, it is making the standard deviation of the residuals greater and the value of r^2 smaller.

The point for the Indians is an outlier because it has a large residual.

FOR PRACTICE, TRY EXERCISE 67

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.

Remember: It only makes sense to talk about the *correlation* between two *quantitative* variables. If one or both variables are categorical, you should refer to the *association* between the two variables. To be safe, use the more general term *association* when describing the relationship between any two variables.

A study once found that people with two cars live longer than people who own only one car.[26] 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. Can 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.

Section 3.2 Summary

- A **regression line** models how a response variable y changes as an explanatory variable x changes. You can use a regression line to **predict** the value of y for any value of x by substituting this x value into the equation of the line.
- The **slope** b of a regression line $\hat{y} = a + bx$ describes how the predicted value of y changes for each increase of 1 unit in x.
- The y **intercept** a of a regression line $\hat{y} = a + bx$ is the predicted value of y when the explanatory variable x equals 0. This prediction does not have a logical interpretation unless x can actually take values near 0.
- Avoid **extrapolation**, using a regression line to make predictions using values of the explanatory variable outside the values of the data from which the line was calculated.
- The most common method of fitting a line to a scatterplot is least squares. The **least-squares regression line** is the line that minimizes the sum of the squares of the vertical distances of the observed points from the line.

- You can examine the fit of a regression line by studying the **residuals,** which are the differences between the actual values of y and predicted values of y: Residual $= y - \hat{y}$. Be on the lookout for curved patterns in the **residual plot,** which indicate that a linear model may not be appropriate.
- The **standard deviation of the residuals** s measures the typical size of a residual when using the regression line.
- The **coefficient of determination** r^2 is the percent of the variation in the response variable that is accounted for by the least-squares regression line using a particular explanatory variable.
- The least-squares regression line of y on x is the line with slope $b = r\frac{s_y}{s_x}$ and intercept $a = \bar{y} - b\bar{x}$. This line always passes through the point $(\bar{x}, \bar{y})$.
- **Influential points** can greatly affect correlation and regression calculations. Points with x values far from $\bar{x}$ have **high leverage** and can be very influential. Points with large residuals are called **outliers** and can also affect correlation and regression calculations.
- Most of all, be careful not to conclude that there is a cause-and-effect relationship between two variables just because they are strongly associated.

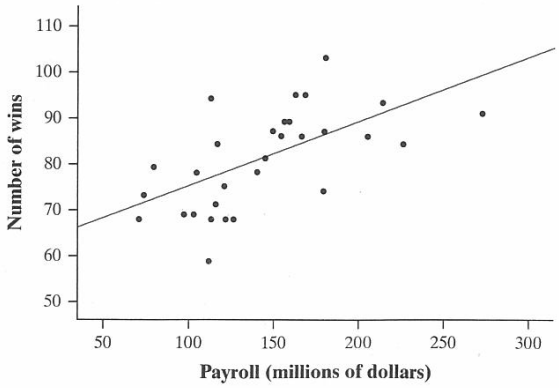
3.2 Technology Corners

TI-Nspire and other technology instructions are on the book's website at highschool.bfwpub.com/updatedtps6e.

Section 3.2 Exercises

37. **Predicting wins** Earlier we investigated the relationship between x = payroll (in millions of dollars) and y = number of wins for Major League Baseball teams in 2016. Here is a scatterplot of the data, along with the regression line $\hat{y} = 60.7 + 0.139x$:

pg 178



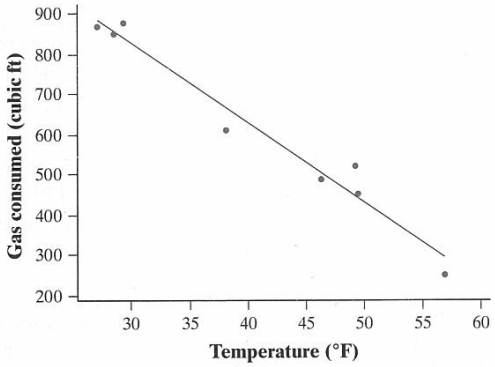
(a) Predict the number of wins for a team that spends $200 million on payroll.

(b) Predict the number of wins for a team that spends $400 million on payroll.

(c) How confident are you in each of these predictions? Explain your reasoning.

38. **How much gas?** Joan is concerned about the amount of energy she uses to heat her home. The scatterplot (on page 205) shows the relationship between x = mean temperature in a particular month and y = mean amount of natural gas used per day (in cubic feet) in that month, along with the regression line $\hat{y} = 1425 - 19.87x$.

(a) Predict the mean amount of natural gas Joan will use per day in a month with a mean temperature of 30°F.



(b) Predict the mean amount of natural gas Joan will use per day in a month with a mean temperature of 65°F.

(c) How confident are you in each of these predictions? Explain your reasoning.

39. **Residual wins** Refer to Exercise 37. The Chicago Cubs won the World Series in 2016. They had 103 wins and spent $182 million on payroll. Calculate and interpret the residual for the Cubs.

pg 180

40. **Residual gas** Refer to Exercise 38. During March, the average temperature was 46.4°F and Joan used an average of 490 cubic feet of gas per day. Calculate and interpret the residual for this month.

41. **More wins?** Refer to Exercise 37.

pg 182

(a) Interpret the slope of the regression line.

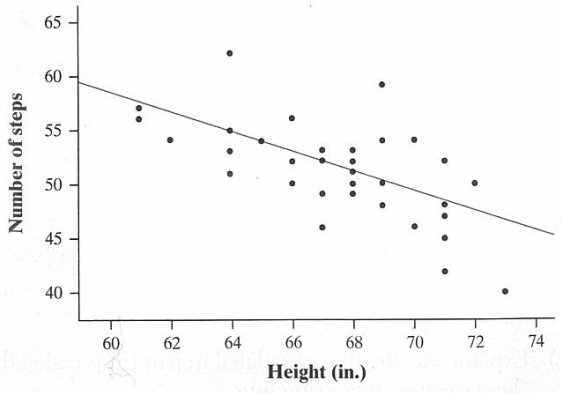
(b) Does the value of the y intercept have meaning in this context? If so, interpret the y intercept. If not, explain why.

42. **Less gas?** Refer to Exercise 38.

(a) Interpret the slope of the regression line.

(b) Does the value of the y intercept have meaning in this context? If so, interpret the y intercept. If not, explain why.

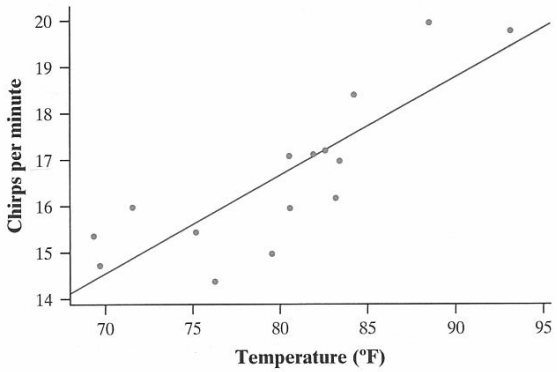
43. **Long strides** The scatterplot shows the relationship between x = height of a student (in inches) and y = number of steps required to walk the length of a school hallway, along with the regression line $\hat{y} = 113.6 - 0.921x$.



(a) Calculate and interpret the residual for Kiana, who is 67 inches tall and took 49 steps to walk the hallway.

(b) Matthew is 10 inches taller than Samantha. About how many fewer steps do you expect Matthew to take compared to Samantha?

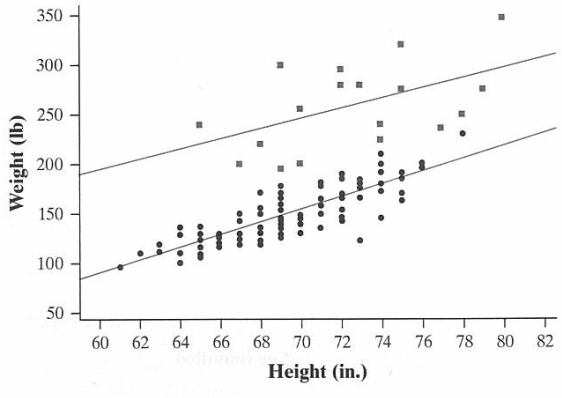
44. **Crickets chirping** The scatterplot shows the relationship between x = temperature in degrees Fahrenheit and y = chirps per minute for the striped ground cricket, along with the regression line $\hat{y} = -0.31 + 0.212x$.[27]



(a) Calculate and interpret the residual for the cricket who chirped 20 times per minute when the temperature was 88.6°F.

(b) About how many additional chirps per minute do you expect a cricket to make if the temperature increases by 10°F?

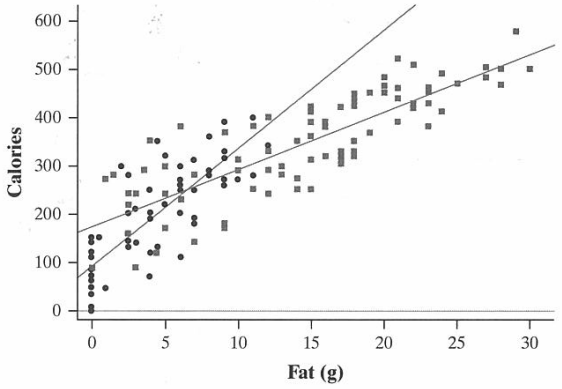
45. **More Olympic athletes** In Exercises 5 and 11, you described the relationship between height (in inches) and weight (in pounds) for Olympic track and field athletes. The scatterplot shows this relationship, along with two regression lines. The regression line for the shotput, hammer throw, and discus throw athletes (blue squares) is $\hat{y} = -115 + 5.13x$. The regression line for the remaining athletes (black dots) is $\hat{y} = -297 + 6.41x$.



(a) How do the regression lines compare?

(b) How much more do you expect a 72-inch discus thrower to weigh than a 72-inch sprinter?

46. **More Starbucks** In Exercises 6 and 12, you described the relationship between fat (in grams) and the number of calories in products sold at Starbucks. The scatterplot shows this relationship, along with two regression lines. The regression line for the food products (blue squares) is $\hat{y} = 170 + 11.8x$. The regression line for the drink products (black dots) is $\hat{y} = 88 + 24.5x$.

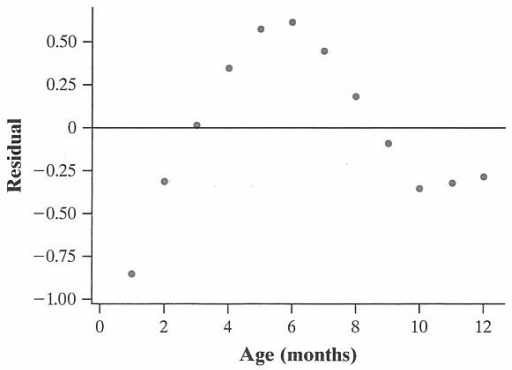


(a) How do the regression lines compare?

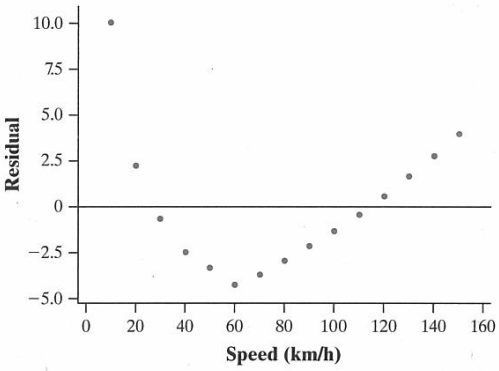
(b) How many more calories do you expect to find in a food item with 5 grams of fat compared to a drink item with 5 grams of fat?

47. **Infant weights in Nahya** A study of nutrition in developing countries collected data from the Egyptian village of Nahya. Researchers recorded the mean weight (in kilograms) for 170 infants in Nahya each month during their first year of life. A hasty user of statistics enters the data into software and computes the least-squares line without looking at the scatterplot first. The result is $\widehat{\text{weight}} = 4.88 + 0.267\,(\text{age})$. Use the residual plot to determine if this linear model is appropriate.

pg 186



48. **Driving speed and fuel consumption** Exercise 9 (page 171) gives data on the fuel consumption y of a car at various speeds x. Fuel consumption is measured in liters of gasoline per 100 kilometers driven, and speed is measured in kilometers per hour. A statistical software package gives the least-squares regression line $\hat{y} = 11.058 - 0.01466x$. Use the residual plot to determine if this linear model is appropriate.



49. **Actual weight** Refer to Exercise 47. Use the equation of the least-squares regression line and the residual plot to estimate the *actual* mean weight of the infants when they were 1 month old.

50. **Actual consumption** Refer to Exercise 48. Use the equation of the least-squares regression line and the residual plot to estimate the *actual* fuel consumption of the car when driving 20 kilometers per hour.

51. **Movie candy** Is there a relationship between the amount of sugar (in grams) and the number of calories in movie-theater candy? Here are the data from a sample of 12 types of candy:

Name	Sugar (g)	Calories	Name	Sugar (g)	Calories
Butterfinger Minis	45	450	Reese's Pieces	61	580
Junior Mints	107	570	Skittles	87	450
M&M'S®	62	480	Sour Patch Kids	92	490
Milk Duds	44	370	SweeTarts	136	680
Peanut M&M'S®	79	790	Twizzlers	59	460
Raisinets	60	420	Whoppers	48	350

(a) Sketch a scatterplot of the data using sugar as the explanatory variable.

(b) Use technology to calculate the equation of the least-squares regression line for predicting the number of calories based on the amount of sugar. Add the line to the scatterplot from part (a).

(c) Explain why the line calculated in part (b) is called the "least-squares" regression line.

52. **Long jumps** Here are the 40-yard-dash times (in seconds) and long-jump distances (in inches) for a small class of 12 students:

Dash time (sec)	5.41	5.05	7.01	7.17	6.73	5.68
Long-jump distance (in.)	171	184	90	65	78	130
Dash time (sec)	5.78	6.31	6.44	6.50	6.80	7.25
Long-jump distance (in.)	173	143	92	139	120	110

(a) Sketch a scatterplot of the data using dash time as the explanatory variable.

(b) Use technology to calculate the equation of the least-squares regression line for predicting the long-jump distance based on the dash time. Add the line to the scatterplot from part (a).

(c) Explain why the line calculated in part (b) is called the "least-squares" regression line.

53. **More candy** Refer to Exercise 51. Use technology to create a residual plot. Sketch the residual plot and explain what information it provides.

54. **More long jumps** Refer to Exercise 52. Use technology to create a residual plot. Sketch the residual plot and explain what information it provides.

55. **Longer strides** In Exercise 43, we modeled the relationship between x = height of a student (in inches) and y = number of steps required to walk the length of a school hallway, with the regression line $\hat{y} = 113.6 - 0.921x$. For this model, technology gives $s = 3.50$ and $r^2 = 0.399$.

pg 191

(a) Interpret the value of s.

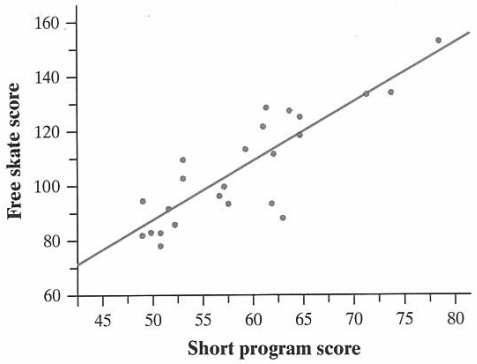
(b) Interpret the value of r^2.

56. **Crickets keep chirping** In Exercise 44, we modeled the relationship between x = temperature in degrees Fahrenheit and y = chirps per minute for the striped ground cricket, with the regression line $\hat{y} = -0.31 + 0.212x$. For this model, technology gives $s = 0.97$ and $r^2 = 0.697$.

(a) Interpret the value of s.

(b) Interpret the value of r^2.

57. **Olympic figure skating** For many people, the women's figure skating competition is the highlight of the Olympic Winter Games. Scores in the short program x and scores in the free skate y were recorded for each of the 24 skaters who competed in both rounds during the 2010 Winter Olympics in Vancouver, Canada.[28] Here is a scatterplot with least-squares regression line $\hat{y} = -16.2 + 2.07x$. For this model, $s = 10.2$ and $r^2 = 0.736$.



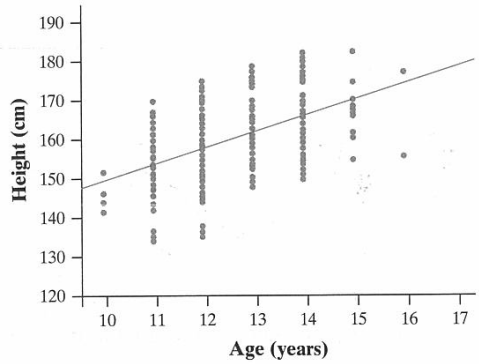
(a) Calculate and interpret the residual for the 2010 gold medal winner Yu-Na Kim, who scored 78.50 in the short program and 150.06 in the free skate.

(b) Interpret the slope of the least-squares regression line.

(c) Interpret the standard deviation of the residuals.

(d) Interpret the coefficient of determination.

58. **Age and height** A random sample of 195 students was selected from the United Kingdom using the Census At School data selector. The age x (in years) and height y (in centimeters) were recorded for each student. Here is a scatterplot with the least-squares regression line $\hat{y} = 106.1 + 4.21x$. For this model, $s = 8.61$ and $r^2 = 0.274$.



(a) Calculate and interpret the residual for the student who was 141 cm tall at age 10.

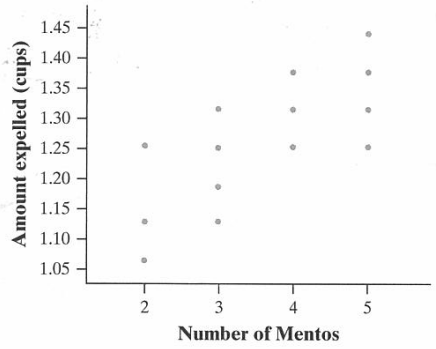
(b) Interpret the slope of the least-squares regression line.

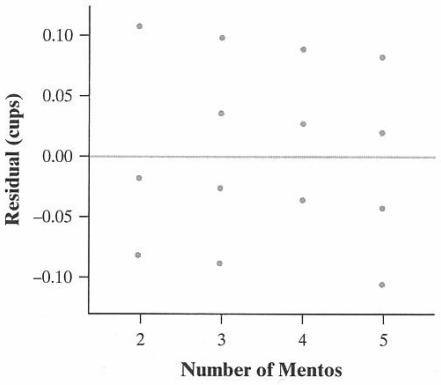
(c) Interpret the standard deviation of the residuals.

(d) Interpret the coefficient of determination.

59. **More mess?** When Mentos are dropped into a newly opened bottle of Diet Coke, carbon dioxide is released from the Diet Coke very rapidly, causing the Diet Coke to be expelled from the bottle. To see if using more Mentos causes more Diet Coke to be expelled, Brittany and Allie used twenty-four 2-cup bottles of Diet Coke and randomly assigned each bottle to receive either 2, 3, 4, or 5 Mentos. After waiting for the fizzing to stop, they measured the amount expelled (in cups) by subtracting the amount remaining from the original amount in the bottle.[29] Here is computer output from a linear regression of y = amount expelled on x = number of Mentos:

pg 192





Term	Coef	SE Coef	T-Value	P-Value
Constant	1.0021	0.0451	22.21	0.000
Mentos	0.0708	0.0123	5.77	0.000

S = 0.06724 R-Sq = 60.21% R-Sq(adj) = 58.40%

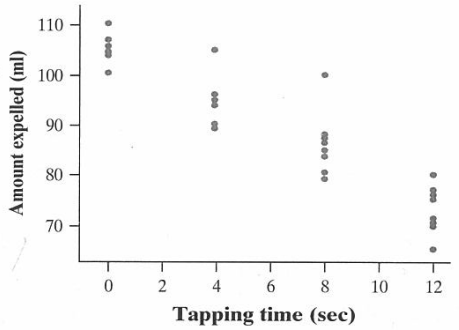
(a) Is a line an appropriate model to use for these data? Explain how you know.

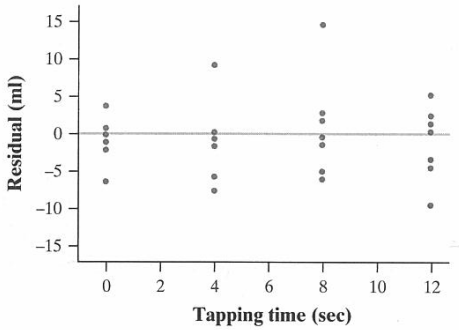
(b) Find the correlation.

(c) What is the equation of the least-squares regression line? Define any variables that you use.

(d) Interpret the values of s and r^2.

60. **Less mess?** Kerry and Danielle wanted to investigate whether tapping on a can of soda would reduce the amount of soda expelled after the can has been shaken. For their experiment, they vigorously shook 40 cans of soda and randomly assigned each can to be tapped for 0 seconds, 4 seconds, 8 seconds, or 12 seconds. After waiting for the fizzing to stop, they measured the amount expelled (in milliliters) by subtracting the amount remaining from the original amount in the can.[30] Here is computer output from a linear regression of y = amount expelled on x = tapping time:





Term	Coef	SE Coef	T-Value	P-Value
Constant	106.360	1.320	80.34	0.000
Tapping_time	-2.635	0.177	-14.90	0.000

S = 5.00347 R-Sq = 85.38% R-Sq(adj) = 84.99%

(a) Is a line an appropriate model to use for these data? Explain how you know.

(b) Find the correlation.

(c) What is the equation of the least-squares regression line? Define any variables that you use.

(d) Interpret the values of s and r^2.

61. **Temperature and wind** The average temperature (in degrees Fahrenheit) and average wind speed (in miles per hour) were recorded for 365 consecutive days at Chicago's O'Hare International Airport. Here is computer output for a regression of y = average wind speed on x = average temperature:

Summary of Fit

RSquare	0.047874
RSquare Adj	0.045251
Root Mean Square Error	3.655950
Mean of Response	9.826027
Observations (or Sum Wgts)	365

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	11.897762	0.521320	22.82	<.0001*
Avg temp	-0.041077	0.009615	-4.27	<.0001*

(a) Calculate and interpret the residual for the day where the average temperature was 42°F and the average wind speed was 2.2 mph.

(b) Interpret the slope.

(c) By about how much do the actual average wind speeds typically vary from the values predicted by the least-squares regression line with x = average temperature?

(d) What percent of the variability in average wind speed is accounted for by the least-squares regression line with x = average temperature?

62. **Beetles and beavers** Do beavers benefit beetles? Researchers laid out 23 circular plots, each 4 meters in diameter, in an area where beavers were cutting down cottonwood trees. In each plot, they counted the number of stumps from trees cut by beavers and the number of clusters of beetle larvae. Ecologists believe that the new sprouts from stumps are more tender than other cottonwood growth, so beetles prefer them. If so, more stumps should produce more beetle larvae.[31] Here is computer output for a regression of y = number of beetle larvae on x = number of stumps:

Summary of Fit

RSquare	0.839144
RSquare Adj	0.831484
Root Mean Square Error	6.419386
Mean of Response	25.086960
Observations (or Sum Wgts)	23

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-1.286104	2.853182	-0.45	0.6568
Number of stumps	11.893733	1.136343	10.47	<.0001*

(a) Calculate and interpret the residual for the plot that had 2 stumps and 30 beetle larvae.

(b) Interpret the slope.

(c) By about how much do the actual number of larvae typically vary from the values predicted by the least-squares regression line with x = number of stumps?

(d) What percent of the variability in number of larvae is accounted for by the least-squares regression line with x = number of stumps?

63. **Husbands and wives** The mean height of married pg 195 American women in their early 20s is 64.5 inches and the standard deviation is 2.5 inches. The mean height of married men the same age is 68.5 inches with standard deviation 2.7 inches. The correlation between the heights of husbands and wives is about $r = 0.5$.

(a) Find the equation of the least-squares regression line for predicting a husband's height from his wife's height for married couples in their early 20s.

(b) Suppose that the height of a randomly selected wife was 1 standard deviation below average. Predict the height of her husband *without* using the least-squares line.

64. **The stock market** Some people think that the behavior of the stock market in January predicts its behavior for the rest of the year. Take the explanatory variable x to be the percent change in a stock market index in January and the response variable y to be the change in the index for the entire year. We expect a positive correlation between x and y because the change during January contributes to the full year's change. Calculation from data for an 18-year period gives

$$\bar{x} = 1.75\% \quad s_x = 5.36\% \quad \bar{y} = 9.07\%$$
$$s_y = 15.35\% \quad r = 0.596$$

(a) Find the equation of the least-squares line for predicting full-year change from January change.

(b) Suppose that the percent change in a particular January was 2 standard deviations above average. Predict the percent change for the entire year *without* using the least-squares line.

65. **Will I bomb the final?** We expect that students who do well on the midterm exam in a course will usually also do well on the final exam. Gary Smith of Pomona College looked at the exam scores of all 346 students who took his statistics class over a 10-year period.[32] Assume that both the midterm and final exam were scored out of 100 points.

(a) State the equation of the least-squares regression line if each student scored the same on the midterm and the final.

(b) The actual least-squares line for predicting final-exam score y from midterm-exam score x was $\hat{y} = 46.6 + 0.41x$. Predict the score of a student who scored 50 on the midterm and a student who scored 100 on the midterm.

(c) Explain how your answers to part (b) illustrate regression to the mean.

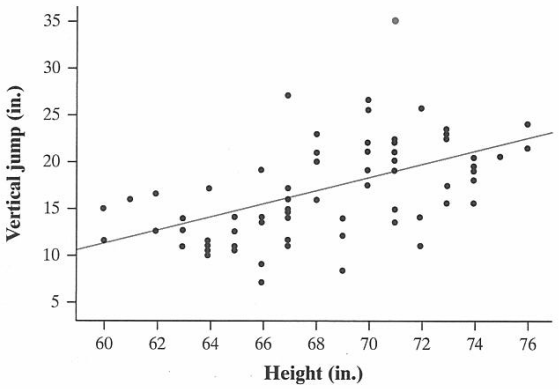
66. **It's still early** We expect that a baseball player who has a high batting average in the first month of the season will also have a high batting average the rest of the season. Using 66 Major League Baseball players from a recent season,[33] a least-squares regression line was calculated to predict rest-of-season batting average y from first-month batting average x. *Note:* A player's batting average is the proportion of times at bat that he gets a hit. A batting average over 0.300 is considered very good in Major League Baseball.

(a) State the equation of the least-squares regression line if each player had the same batting average the rest of the season as he did in the first month of the season.

(b) The actual equation of the least-squares regression line is $\hat{y} = 0.245 + 0.109x$. Predict the rest-of-season batting average for a player who had a 0.200 batting average the first month of the season and for a player who had a 0.400 batting average the first month of the season.

(c) Explain how your answers to part (b) illustrate regression to the mean.

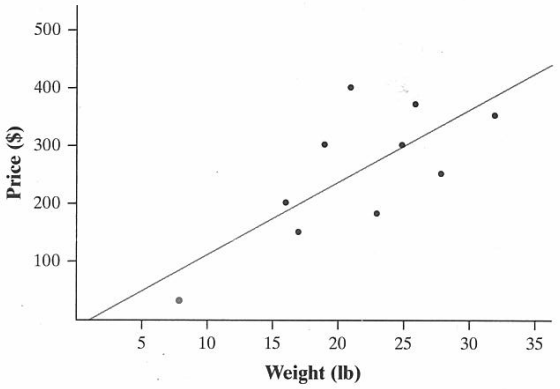
67. **Who's got hops?** Haley, Jeff, and Nathan measured
pg 202
the height (in inches) and vertical jump (in inches) of 74 students at their school.[34] Here is a scatterplot of the data, along with the least-squares regression line. Jacob (highlighted in red) had a vertical jump of nearly 3 feet!



(a) Describe the influence that Jacob's point has on the equation of the least-squares regression line.

(b) Describe the influence that Jacob's point has on the standard deviation of the residuals and r^2.

68. **Stand mixers** The scatterplot shows the weight (in pounds) and cost (in dollars) of 11 stand mixers.[35] The mixer from Walmart (highlighted in red) was much lighter—and cheaper—than the other mixers.



(a) Describe what influence the highlighted point has on the equation of the least-squares regression line.

(b) Describe what influence the highlighted point has on the standard deviation of the residuals and r^2.

69. **Managing diabetes** People with diabetes measure their fasting plasma glucose (FPG, measured in milligrams per milliliter) after fasting for at least 8 hours. Another measurement, made at regular medical checkups, is called HbA. 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. The table gives data on both HbA and FPG for 18 diabetics five months after they had completed a diabetes education class.[36]

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

(a) Make a scatterplot with HbA as the explanatory variable. Describe what you see.

(b) Subject 18 has an unusually large x value. What effect do you think this subject has on the correlation? What effect do you think this subject has on the equation of the least-squares regression line? Calculate the correlation and equation of the least-squares regression line with and without this subject to confirm your answer.

(c) Subject 15 has an unusually large y value. What effect do you think this subject has on the correlation? What effect do you think this subject has on the equation of the least-squares regression line? Calculate the correlation and equation of the least-squares regression line with and without this subject to confirm your answer.

70. **Rushing for points** What is the relationship between rushing yards and points scored in the National Football League? The table gives the number of rushing yards and the number of points scored for each of the 16 games played by the Jacksonville Jaguars in a recent season.[37]

Game	Rushing yards	Points scored	Game	Rushing yards	Points scored
1	163	16	9	141	17
2	112	3	10	108	10
3	128	10	11	105	13
4	104	10	12	129	14
5	96	20	13	116	41
6	133	13	14	116	14
7	132	12	15	113	17
8	84	14	16	190	19

(a) Make a scatterplot with rushing yards as the explanatory variable. Describe what you see.

(b) Game 16 has an unusually large x value. What effect do you think this game has on the correlation? On the equation of the least-squares regression line? Calculate the correlation and equation of the least-squares regression line with and without this game to confirm your answers.

(c) Game 13 has an unusually large y value. What effect do you think this game has on the correlation? On the equation of the least-squares regression line? Calculate the correlation and equation of the least-squares regression line with and without this game to confirm your answers.

Multiple Choice: *Select the best answer for Exercises 71–78.*

71. Which of the following is *not* a characteristic of the least-squares regression line?

(a) The slope of the least-squares regression line is always between –1 and 1.

(b) The least-squares regression line always goes through the point $(\bar{x}, \bar{y})$.

(c) The least-squares regression line minimizes the sum of squared residuals.

(d) The slope of the least-squares regression line will always have the same sign as the correlation.

(e) The least-squares regression line is not resistant to outliers.

72. Each year, students in an elementary school take a standardized math test at the end of the school year. For a class of fourth-graders, the average score was 55.1 with a standard deviation of 12.3. In the third grade, these same students had an average score of 61.7 with a standard deviation of 14.0. The correlation between the two sets of scores is $r = 0.95$. Calculate the equation of the least-squares regression line for predicting a fourth-grade score from a third-grade score.

(a) $\hat{y} = 3.58 + 0.835x$

(b) $\hat{y} = 15.69 + 0.835x$

(c) $\hat{y} = 2.19 + 1.08x$

(d) $\hat{y} = -11.54 + 1.08x$

(e) Cannot be calculated without the data.

73. Using data from the LPGA tour, a regression analysis was performed using x = average driving distance and y = scoring average. Using the output from the regression analysis shown below, determine the equation of the least-squares regression line.

```
Predictor            Coef     SE Coef     T       P
Constant          87.974000  2.391000  36.78   0.000
Driving Distance  -0.060934  0.009536  -6.39   0.000

S = 1.01216     R-Sq = 22.1%     R-Sq(adj) = 21.6%
```

(a) $\hat{y} = 87.974 + 2.391x$

(b) $\hat{y} = 87.974 + 1.01216x$

(c) $\hat{y} = 87.974 - 0.060934x$

(d) $\hat{y} = -0.060934 + 1.01216x$

(e) $\hat{y} = -0.060934 + 87.947x$

Exercises 74 to 78 refer to the following setting. Measurements on young children in Mumbai, India, found this least-squares line for predicting y = height (in cm) from x = arm span (in cm):[38]

$$\hat{y} = 6.4 + 0.93x$$

74. By looking at the equation of the least-squares regression line, you can see that the correlation between height and arm span is

(a) greater than zero.

(b) less than zero.

(c) 0.93.

(d) 6.4.

(e) Can't tell without seeing the data.

75. In addition to the regression line, the report on the Mumbai measurements says that $r^2 = 0.95$. This suggests that

(a) although arm span and height are correlated, arm span does not predict height very accurately.

(b) height increases by $\sqrt{0.95} = 0.97$ cm for each additional centimeter of arm span.

(c) 95% of the relationship between height and arm span is accounted for by the regression line.

(d) 95% of the variation in height is accounted for by the regression line with x = arm span.

(e) 95% of the height measurements are accounted for by the regression line with x = arm span.

76. One child in the Mumbai study had height 59 cm and arm span 60 cm. This child's residual is

(a) −3.2 cm. (b) −2.2 cm. (c) −1.3 cm.

(d) 3.2 cm. (e) 62.2 cm.

77. Suppose that a tall child with arm span 120 cm and height 118 cm was added to the sample used in this study. What effect will this addition have on the correlation and the slope of the least-squares regression line?

(a) Correlation will increase, slope will increase.

(b) Correlation will increase, slope will stay the same.

(c) Correlation will increase, slope will decrease.

(d) Correlation will stay the same, slope will stay the same.

(e) Correlation will stay the same, slope will increase.

78. Suppose that the measurements of arm span and height were converted from centimeters to meters by dividing each measurement by 100. How will this conversion affect the values of r^2 and s?

(a) r^2 will increase, s will increase.

(b) r^2 will increase, s will stay the same.

(c) r^2 will increase, s will decrease.

(d) r^2 will stay the same, s will stay the same.

(e) r^2 will stay the same, s will decrease.

Recycle and Review

79. **Fuel economy** (2.2) In its recent *Fuel Economy Guide*, the Environmental Protection Agency (EPA) gives data on 1152 vehicles. There are a number of outliers, mainly vehicles with very poor gas mileage or hybrids with very good gas mileage. If we ignore the outliers, however, the combined city and highway gas mileage of the other 1120 or so vehicles is approximately Normal with mean 18.7 miles per gallon (mpg) and standard deviation 4.3 mpg.

(a) The Chevrolet Malibu with a four-cylinder engine has a combined gas mileage of 25 mpg. What percent of the 1120 vehicles have worse gas mileage than the Malibu?

(b) How high must a vehicle's gas mileage be in order to fall in the top 10% of the 1120 vehicles?

80. **Marijuana and traffic accidents** (1.1) Researchers in New Zealand interviewed 907 drivers at age 21. They had data on traffic accidents and they asked the drivers about marijuana use. Here are data on the numbers of accidents caused by these drivers at age 19, broken down by marijuana use at the same age:[39]

	Marijuana use per year			
	Never	1–10 times	11–50 times	51+ times
Number of drivers	452	229	70	156
Accidents caused	59	36	15	50

(a) Make a graph that displays the accident rate for each category of marijuana use. Is there evidence of an association between marijuana use and traffic accidents? Justify your answer.

(b) Explain why we can't conclude that marijuana use *causes* accidents based on this study.