

Correlation and Residuals

9



Having pets is not only fun, but it can be good for your health, too. Studies show that there is a correlation between pet ownership and blood pressure, mood, and the ability of your immune system to fight off disease!



9.1 Like a Glove	
Least Squares Regression	523
9.2 Gotta Keep It Correlatin'	
Correlation	533
9.3 The Residual Effect	
Creating Residual Plots	541
9.4 To Fit or Not To Fit? That Is The Question!	
Using Residual Plots.	553
9.5 Who Are You? Who? Who?	
Causation vs. Correlation	563

Like a Glove

Least Squares Regression

LEARNING GOALS

In this lesson, you will:

- Determine and interpret the least squares regression equation for a data set using a formula.
- Use interpolation to make predictions about data.
- Use extrapolation to make predictions about data.

KEY TERMS

- interpolation
- extrapolation
- least squares regression line

How do the nerve cells in your brain communicate with each other? Signals have to be sent all across the brain—from your eyes to your occipital lobe in the back of your brain, from your ears to your temporal lobe, and so on. How does this happen?

In a sense, your nerve cells actually communicate using shapes. When a nerve cell is activated, it releases chemical messengers called neurotransmitters. These messengers have specific shapes, and they fit like keys into the locks on the next cell receiving the message. This message tells the next cell what to do.

And this process happens trillions of times per day!

PROBLEM 1 Music, Anyone?

The table shows the percent of all recorded music sales that came from music stores for the years 1998 through 2004.

Year	1998	1999	2000	2001	2002	2003	2004
Percent of Total Sales from Music Stores	50.8	44.5	42.4	39.7	36.8	33.2	32.5

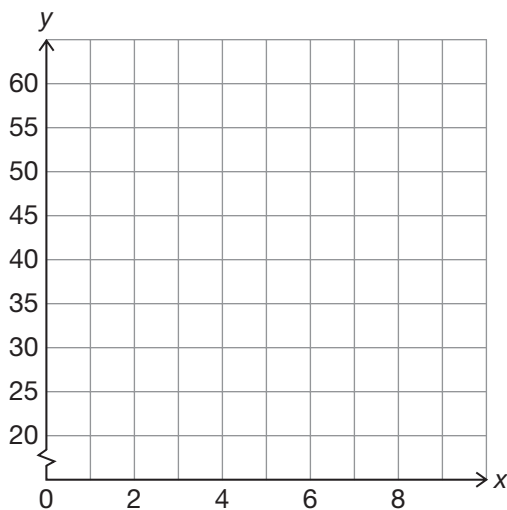
9



1. Represent the data as ordered pairs with the percent of total sales that came from music stores as a function of time. Let x represent the number of years since 1998.

If 1998 is the first year, do I represent it as 1 or 0?

2. Use your calculator to construct a scatter plot of the data. Sketch the scatter plot on the coordinate plane. Label the axes.



3. Describe any patterns you see in the data.

4. Use a graphing calculator to calculate the linear regression equation for the data. Round the values to the nearest hundredth.



5. Interpret the equation of the line in terms of the problem situation.

When you're talking about a line, it's all about the slope and y -intercept.



9

If there is a linear association between the independent and dependent variables of a data set, you can use a linear regression to make predictions within the data set. Using a linear regression to make predictions within the data set is called **interpolation**.



6. Use your equation to predict the percent of total music sales that came from music stores in the year 2000.

7. Compare the predicted percent in 2000 to the actual percent in 2000.

8. Use your equation to predict the percent of total music sales that came from music stores in 2003.

9. Compare the predicted percent in 2003 to the actual percent in 2003.



10. Do you think a prediction made using interpolation will always be close to the actual value? Explain your reasoning.

9

To make predictions for values of x that are outside of the data set is called **extrapolation**.



11. Use the equation to predict the percent of total music sales that would come from music stores in:
- a. 2010.

b. 2020.

c. 1900.

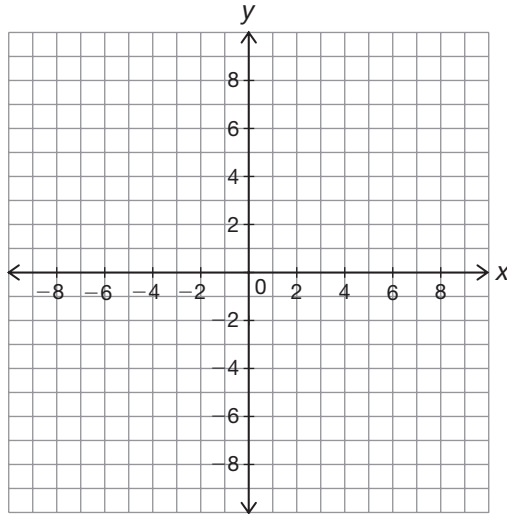


12. Are these predictions reasonable based on the problem situation?

PROBLEM 2 Best vs. Good



1. Suppose a data set is composed of the points $(1, 3)$, $(-3, -7)$, and $(5, 7)$ on a coordinate plane.



Collinear points are points that are located on the same line.



9

2. Are these points collinear? How can you tell?
3. Determine the equation of a line passing through the points at $(1, 3)$ and $(5, 7)$. Graph this line on the coordinate plane.
4. Determine and graph the equation of a line passing through the points at:
- a. $(-3, -7)$ and $(5, 7)$.

There are going to be a number of graphs on your coordinate plane. Be sure to label each as you graph.



- b. $(-3, -7)$ and $(1, 3)$.



5. Would you consider any of the three lines you just graphed to be a line that “best fits” the three points? If yes, explain your reasoning. If no, describe where the line of best fit should be drawn.



One method to determine the line of best fit, or linear regression line, is the method of least squares. A **least squares regression line** is the line of best fit that minimizes the squares of the distances of the points from the line.

For a least squares regression line, ensure the line is written in the form $y = ax + b$. To calculate a and b , use the equations:

$$a = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$b = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n\sum x^2 - (\sum x)^2}$$

where x represents all x -values from the data set, y represents all y -values from the data set, and n represents the number of coordinate pairs in the data set.

Wow! That looks like a pretty complex formula. But I bet if I look at each part separately, it won't be so difficult.



Let's use this formula to determine the least squares regression line using these points:



$(-3, -3), (1, 2),$ and $(3, 4)$



Calculate the values of each part of the equation separately. Then put it all together.



Determine the number of coordinate points in the data set. $n = 3$



Determine the sum of all the x -values in the data set. $\sum x = -3 + 1 + 3 = 1$



Determine the sum of all the y -values in the data set. $\sum y = -3 + 2 + 4 = 3$



Determine the sum of the squares of the x -values.

$$\sum x^2 = (-3)^2 + 1^2 + 3^2 = 19$$

Determine the sum of the products of each coordinate pair.

$$\begin{aligned}\sum xy &= (-3 \cdot -3) + (1 \cdot 2) + (3 \cdot 4) \\ &= 23\end{aligned}$$

Determine the square of the sum of the x -values.

$$(\sum x)^2 = 1^2 = 1$$

Insert each part into the formulas to solve for the values of a and b .

$$\begin{aligned}a &= \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2} \\ &= \frac{(3)(23) - (1)(3)}{(3)(19) - 1} = \frac{66}{56}\end{aligned}$$

$$a \approx 1.18$$

$$\begin{aligned}b &= \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n\sum x^2 - (\sum x)^2} \\ &= \frac{(3)(19) - (1)(23)}{(3)(19) - 1} = \frac{34}{56}\end{aligned}$$

$$b \approx 0.61$$

6. What is the equation of the line of best fit for the points given in the worked example?



7. Margie calculated the least squares linear regression for the worked example.

Margie

$$a = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - \sum x^2}$$

$$a = \frac{(3)(23) - (1)(3)}{(3)(19) - 19} = \frac{66}{38}$$

$$a \approx 1.74$$



Explain to Margie why her solution is incorrect.



8. Calculate the least squares linear regression using the points from Question 1.

a. Calculate the sums and values for each part of the equation.

b. Calculate the values of a and b .

c. Write the equation of the line of best fit.



d. Graph the line on the coordinate plane in Question 1. Does this line “fit” the data better than the others? Explain your reasoning.

PROBLEM 3 One More Time



The table shown displays the median weekly earnings for U.S. workers according to the number of years of schooling they received.

Years of Schooling	Median Weekly Earnings (dollars)
11	444
12	626
13	712
14	767
16	1038
18	1272
22	1510

1. Calculate the equation of the least squares regression line. Define your variables.
2. Interpret the least squares regression equation in terms of this problem situation.

- Predict the weekly earnings of a worker with 12 years of schooling using the least squares regression equation. How does this compare to the actual earnings?



- Predict the weekly earnings of a doctor with 25 years of schooling using the least squares regression equation. How does this compare to the actual earnings?

Talk the Talk



- Why are predictions made by extrapolation more likely to be inaccurate than predictions made by interpolation?



Be prepared to share your solutions and methods.

Gotta Keep It Correlatin'

Correlation

LEARNING GOALS

In this lesson, you will:

- Determine the correlation coefficient using a formula.
- Interpret the correlation coefficient for a set of data.

“**N**ew Study Links Dark Chocolate to Heart Health.” “Video Games Shown to Boost I.Q.” “College Graduates Live Longer, New Study Finds.”

You have probably seen or heard headlines similar to these in magazines, on TV, and online. Each one of these headlines is the result of a correlational study. In a correlational study, researchers compare two variables to see how they are associated. They do this through the use of surveys or even by researching documents such as medical records.

What methods do you think researchers could have used to produce the results mentioned in the headlines above?

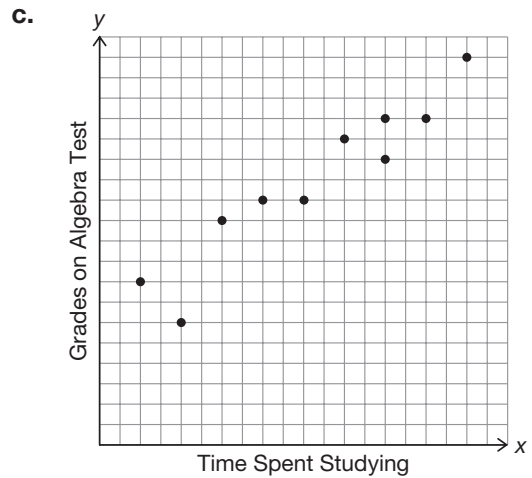
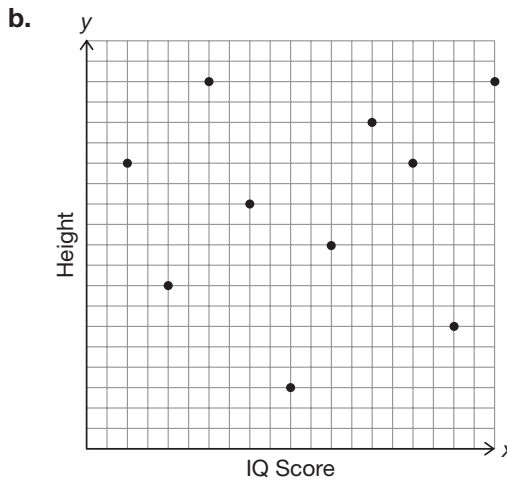
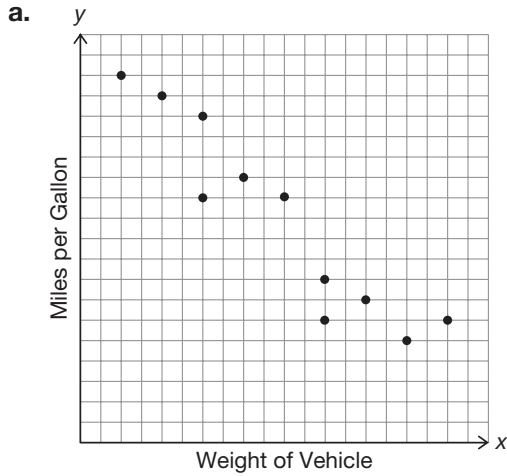
PROBLEM 1 Associate, Formulate, Correlate!



Recall that data comparing two variables can show a positive association, a negative association, or no association.

- Describe the type of association between the independent and dependent variables shown on each scatterplot. Then, draw a line of best fit for each, if possible.

9

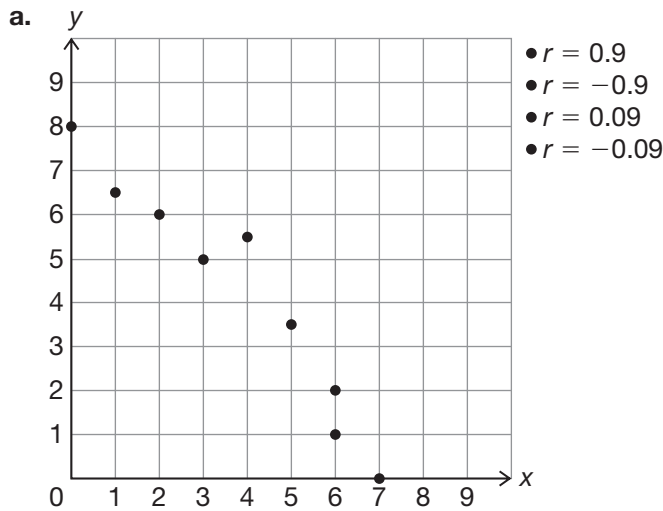


A measure of how well a linear regression line fits a set of data is called correlation. The correlation coefficient is a value between -1 and 1 which indicates how close the data are to forming a straight line. The closer the correlation coefficient is to 1 or -1 , the stronger the linear relationship is between the two variables. The variable r is used to represent the correlation coefficient.

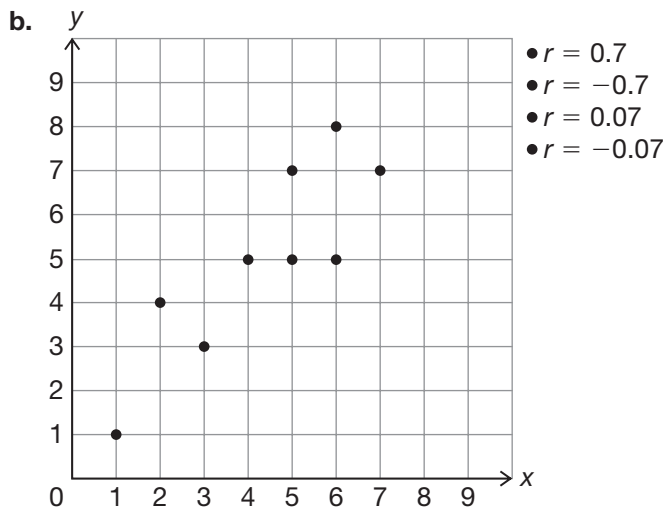
I remember that the correlation coefficient either falls between -1 and 0 if the data show a negative association, or between 0 and 1 if the data show a positive association.

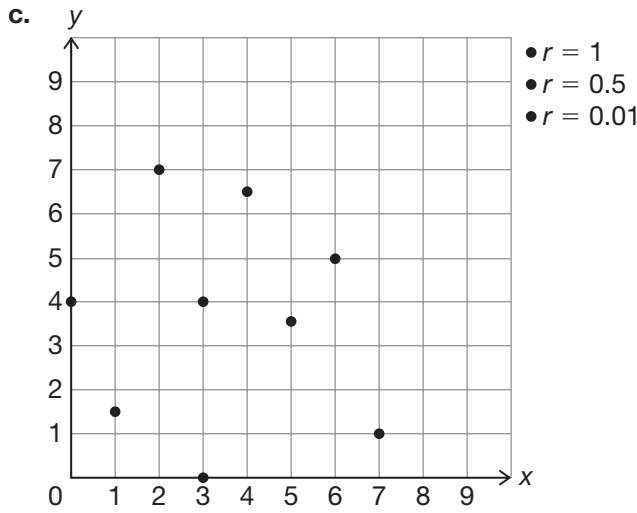


2. Determine whether the points in each scatter plot have a positive correlation, a negative correlation, or no correlation. Four possible r -values are given. Circle the r -value you think is most appropriate. Explain your reasoning for each.



The closer the r -value gets to 0 , the less of a linear relationship there is in the data!





You can calculate the correlation coefficient of a data set using this formula:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

Most of the pieces of this formula look familiar. I think we used them in the formula for standard deviation!



Let's determine the correlation coefficient of this data set using the formula.



$(-3, -3), (1, 2)$ and $(3, 4)$



Look at the numerator of the formula first.



$$\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$



Determine the mean of the x -values and the mean of the y -values. $\bar{x} = \frac{1}{3}$ $\bar{y} = 1$



Keep in mind that the notation $\sum_{i=1}^n$ just tells you that you will be determining the sum of all the data values.



Notice these differences are used throughout the formula.



Determine the difference between each data value and the mean for both the x-coordinates and the y-coordinates.

$(x_i - \bar{x})$	$(y_i - \bar{y})$
$(-3 - \frac{1}{3}) = -\frac{10}{3}$	$(-3 - 1) = -4$
$(1 - \frac{1}{3}) = \frac{2}{3}$	$(2 - 1) = 1$
$(3 - \frac{1}{3}) = \frac{8}{3}$	$(4 - 1) = 3$

Determine the product of the differences in each pair. Then, determine the sum of those products. This is your numerator.

$\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$	
$(-\frac{10}{3} \cdot -4) = \frac{40}{3}$	} $\frac{40}{3} + \frac{2}{3} + 8 = 22$
$(\frac{2}{3} \cdot 1) = \frac{2}{3}$	
$(\frac{8}{3} \cdot 3) = 8$	

Now let's analyze the denominator of the formula.

$$\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \quad \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}$$

Determine the sum of the squares of the differences between each value and its mean.

$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
$(-\frac{10}{3})^2 = \frac{100}{9}$	$(-4)^2 = 16$
$(\frac{2}{3})^2 = \frac{4}{9}$	$(1)^2 = 1$
$(\frac{8}{3})^2 = \frac{64}{9}$	$(3)^2 = 9$
} = $\frac{56}{3}$	

Determine the square root of each sum.

$\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}$	$\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}$
$\sqrt{\frac{56}{3}} \approx 4.32$	$\sqrt{26} \approx 5.099$

Determine the product of these two values. This is your denominator.

$(4.32)(5.099) = 22.02768$



3. Put the pieces together. Determine the correlation coefficient of the data set.



4. Interpret the correlation coefficient of the data set.

PROBLEM 2 The Doctor Will See You Now



The Center for Disease Control collected data on the percent of children, aged 12 to 19, that were considered obese between the years 1971 and 2007. The data are given in the table.

Year	Percent of Obese Children
1971	6.4
1976	5.0
1988	10.5
1999	14.8
2001	16.7
2003	17.4
2005	17.8
2007	18.1

What do you notice as you read through the data?

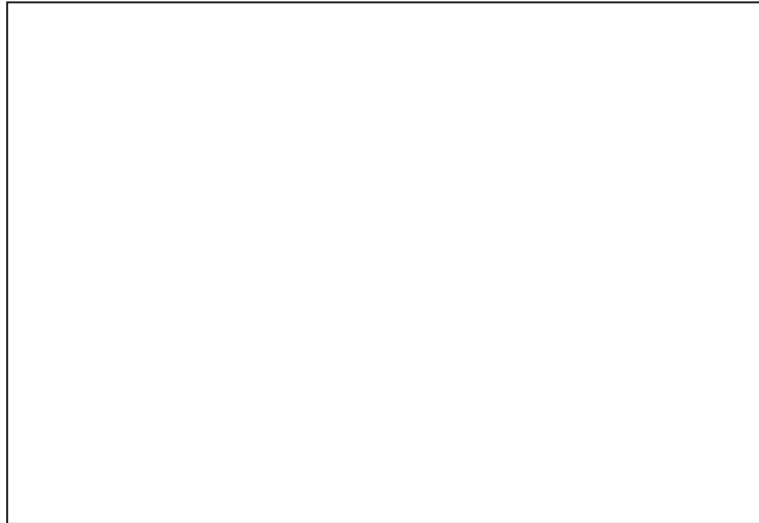




1. Identify the independent and dependent quantities in this problem situation.

2. Construct a scatter plot of the data using your graphing calculator.

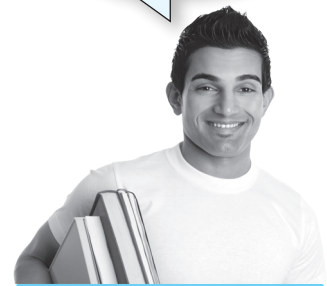
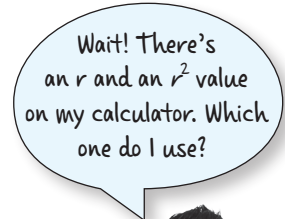
a. Sketch the scatter plot. Label the axes.



b. Do you think a linear regression equation would best describe this situation? Explain your reasoning.

3. Use a graphing calculator to determine whether a line of best fit is appropriate for these data.

a. Determine and interpret the linear regression equation.



b. Determine the correlation coefficient.



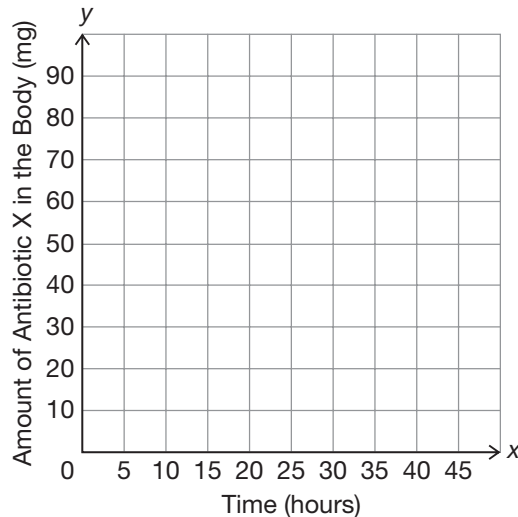
c. Would a line of best fit be appropriate for this data set? Explain your reasoning.



4. The amount of antibiotic that remains in your body over a period of time varies from one drug to the next. The table given shows the amount of Antibiotic X that remains in your body over a period of two days.

Time (hours)	0	6	12	18	24	30	36	42	48
Amount of Antibiotic X in Body (mg)	60	36	22	13	7.8	4.7	2.8	1.7	1

- Determine and interpret a linear regression equation for this data set.
- Determine and interpret the correlation coefficient of this data set.
- Does it seem appropriate to use a line of best fit? If no, explain your reasoning. If yes, determine and interpret the least squares regression equation.
- Sketch a scatter plot of the data.



- Look at the graph of the data. Do you still agree with your answer to part (c)? Explain your reasoning.



Be prepared to share your solutions and methods.

The Residual Effect

Creating Residual Plots

LEARNING GOALS

In this lesson, you will:

- Create residual plots.
- Analyze the shapes of residual plots.

KEY TERMS

- residual
- residual plot

Maybe you once made a lot of spelling mistakes in an essay that you wrote. The next time you wrote an essay, you made sure to do a spell check (or use a dictionary). Maybe you noticed that you missed a lot of free throws in basketball games. You decided to practice your free throw shooting to improve. Maybe you told a joke that hurt your friend's feelings. You remembered to be more sensitive around him or her in the future.

We all learn from our mistakes. In mathematics, too, you can learn a lot about data by looking at error. That's what this lesson is all about!

PROBLEM 1 Hit the Brakes!



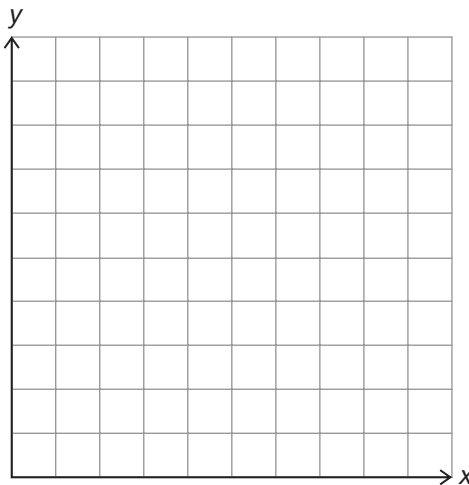
You have used the shape of data in a scatter plot and the correlation coefficient to help you determine whether a linear model is an appropriate model for a data set. For some data sets, these measures may not provide enough information to determine if a linear model is most appropriate.

In order to be a safe driver, there are a lot of things to consider. For example, you have to leave enough distance between your car and the car in front of you in case you need to stop suddenly. The table shows the braking distance for a particular car when traveling at different speeds.

Speed (mph)	Braking Distance (feet)
30	48
40	80
50	120
60	180
70	240
80	320



1. Construct a scatter plot of the data.



2. Based on the shape of the scatter plot, do you think a linear model is appropriate? Explain your reasoning.

3. Calculate the line of best fit for the data. Write a function $d(s)$ to represent the line of best fit.

4. Interpret the function in terms of the problem situation.



5. Determine and interpret the correlation coefficient.



In addition to the shape of the scatter plot and the correlation coefficient, one additional method to determine if a linear model is appropriate for the data is to analyze the *residuals*. A **residual** is the distance between an observed data value and its predicted value using the regression equation.

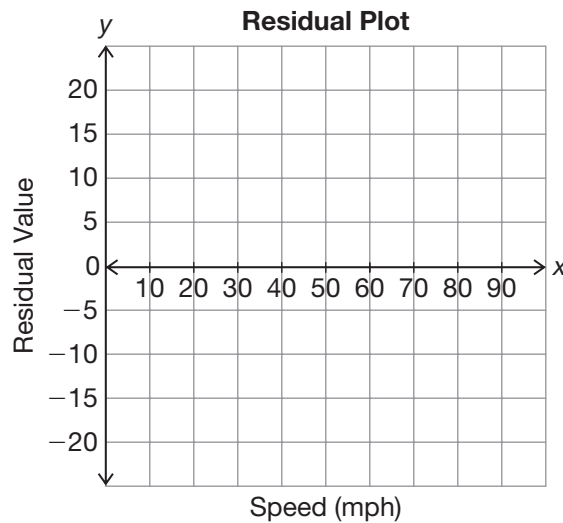
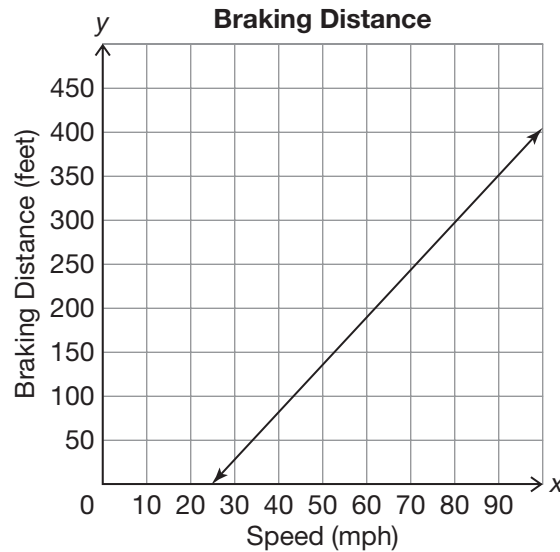


6. Complete the table to determine the residuals for the braking distance data.

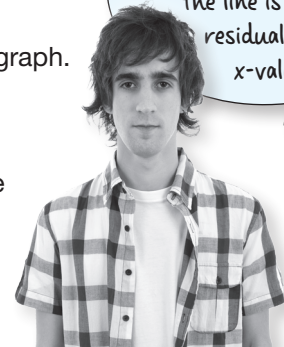
Speed (mph)	Observed Braking Distance (feet)	Predicted Braking Distance (feet)	Residual Value Observed Value – Predicted
30	48		
40	80		
50	120		
60	180		
70	240		
80	320		

Now, let's analyze the relationship between the observed braking distances and the predicted braking distances using graphs. The graph of the line of best fit for the observed braking distances is shown.

Use the graph to answer Questions 7–9 and then construct a residual plot.



7. For each data point, there is a residual equal to the difference between the observed measured braking distance and the value predicted by the line of best fit.
- Plot each observed value on the Braking Distance graph.
 - Connect each observed value to its predicted value using a vertical line.



The vertical distance from each observed data point to the line is called the residual for that x -value.

8. Examine the scatter plot and the residual values.
- When does a residual have a positive value?
 - When does a residual have a negative value?

The residual data can now be used to create a *residual plot*. A **residual plot** is a scatter plot of the independent variable on the x -axis and the residuals on the y -axis.

The residual plot displays the residual values you calculated in the table.

9. Construct a residual plot of the speed and braking distance data.



10. Interpret each residual in the context of the problem situation.
- At 30 mph, the braking distance is 20 feet greater than predicted.
 - At 40 mph, the braking distance is _____.
 - At 50 mph, the braking distance is _____.
 - At 60 mph, the braking distance is _____.
 - At 70 mph, the braking distance is _____.
 - At 80 mph, the braking distance is _____.



11. What pattern, if any, do you notice in the residuals?



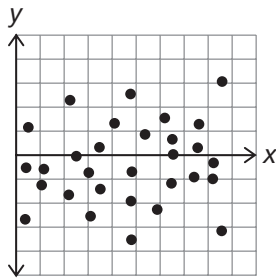
The shape of the residual plot can be useful to determine whether there may be a more appropriate model other than a linear model for a data set.

If a residual plot results in no identifiable pattern or a flat pattern, then the data may be linearly related. If there is a pattern in the residual plot, the data may not be linearly related. Even if the data are not linearly related, the data may still have some other type of non-linear relationship.

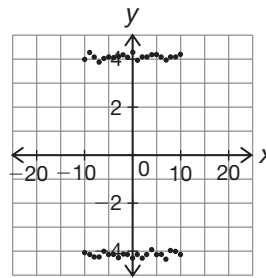
A residual plot can't tell you whether a linear model is appropriate. It can only tell you that there may be a model other than linear that is more appropriate.



Residual Plots Indicating a Possible Linear Relationship

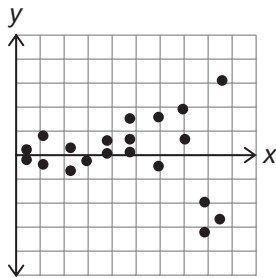


There is no pattern in the residual plot. The data may be linearly related.

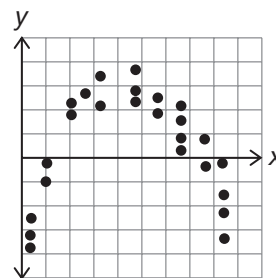


There is a flat pattern in the residual plot. The data may be linearly related.

Residual Plots Indicating a Non-Linear Relationship



There is a pattern in the residual plot. As the x -value increases, the residuals become more spread out. The data may not be linearly related.



There is a pattern in the residual plot. The residuals form a curved pattern. The data may not be linearly related.

12. Interpret the residual plot for the braking distance data.



13. Anita thinks the residual plot looks like it forms a curve. She says that this means the data must be more quadratic than linear. Is Anita correct? Why or why not?



Keep in mind that this only represents a portion of the entire data set.



14. Is the least squares regression line you determined in Question 3 a good fit for this data set? Explain your reasoning.

PROBLEM 2 Attendance Matters

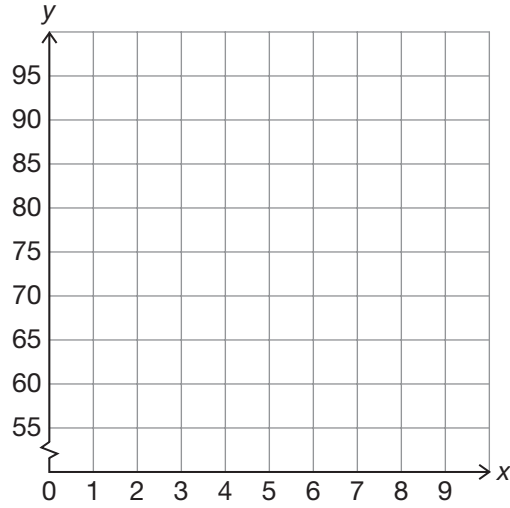


Over the last semester, Mr. Finch kept track of the number of student absences. Now that the semester is over, he wants to see if there is a linear relationship between the number of absences and a student's grade for the semester. The data he collected are given in the table.

Student	Number of Absences	Grade (percent)
James	0	95
Tiona	5	73
Mikala	3	84
Paul	1	92
Danasia	2	92
Erik	3	80
Rachael	10	65
Cheyanne	0	90
Chen	6	70
Javier	1	88



1. Construct a scatter plot of the data.



2. Describe the association shown in the scatter plot.

3. Determine the equation of the least squares regression line. Interpret the equation for this problem situation.

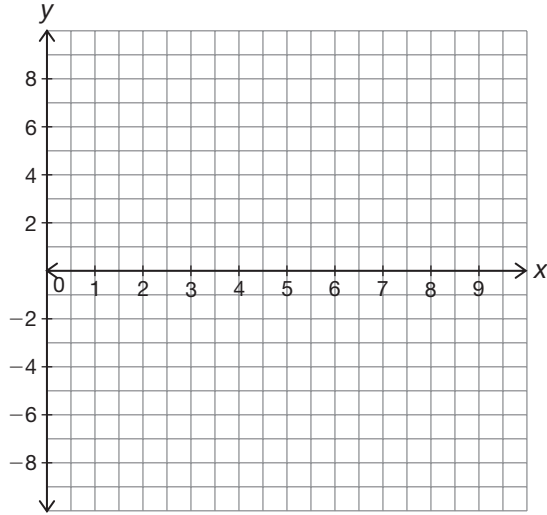
4. Determine and interpret the correlation coefficient.

5. Determine the residuals for the data. Interpret each residual.

Student	Number of Absences	Algebra Grade (percent)	Predicted Value	Residual	Interpretation
James	0	95	92.6	2.4	For 0 absences the actual grade is 2.4% greater than predicted.
Tiona	5	73			
Mikala	3	84			
Paul	1	92			
Danasia	2	92			
Erik	3	80			
Rachael	10	65			
Cheyenne	0	90			
Chen	6	70			
Javier	1	88			



6. Construct and interpret a residual plot of the data.

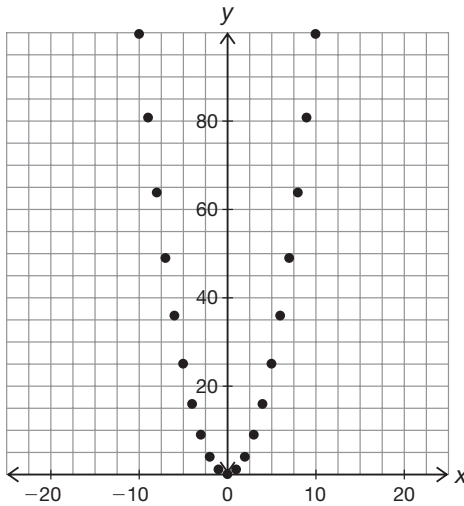


Talk the Talk

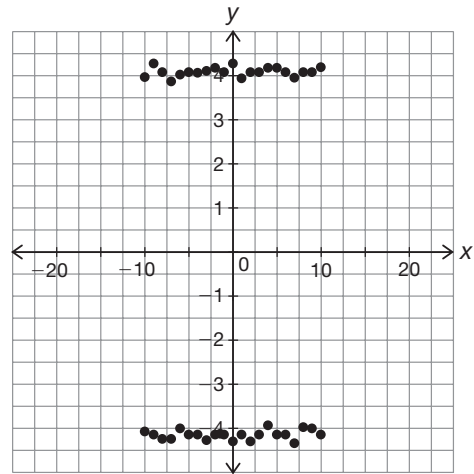


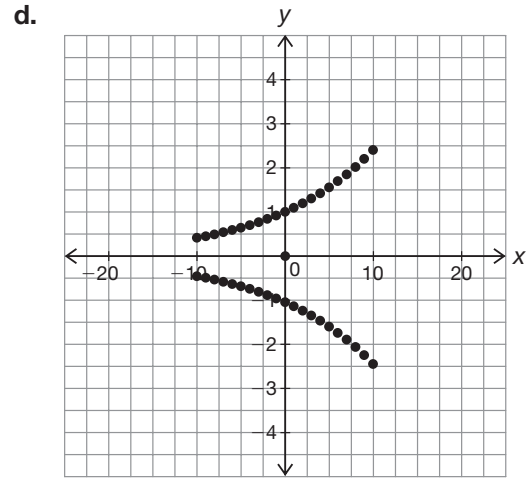
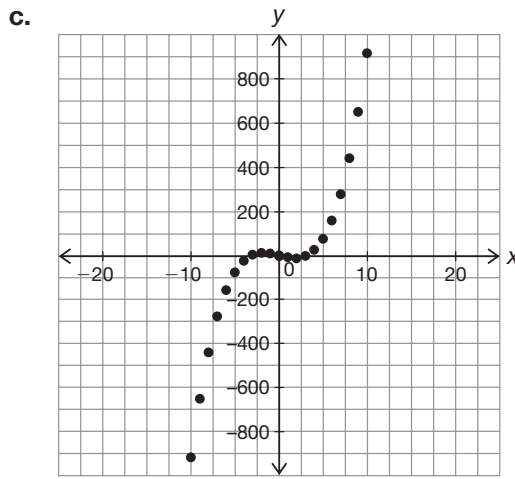
1. Explain what you can conclude from each residual plot about whether a linear model is appropriate.

a.



b.





2. How would you describe the difference between “line of best fit” and “most appropriate model”?



Be prepared to share your solutions and methods.

To Fit or Not To Fit? That Is The Question!

Using Residual Plots

LEARNING GOALS

In this lesson, you will:

- Use scatter plots and correlation coefficients to determine whether a linear regression is a good fit for data.
- Use residual plots to help determine whether a linear regression is the best fit for data.

Have you ever had to make a big decision? One characteristic of a “big” decision is that you often need to use many different sources of information to tackle it.

What kind of car should you drive? To make this decision, you have to think about finances, safety, how you will use the car, gas mileage, and so on. What college should I attend? For this big decision, you might think about the reputation of the school, its distance from home, cost, and so on.

In this lesson, you will learn that even in mathematics we often need multiple sources of information to help us make the best decisions.

PROBLEM 1 Wanna Buy a Car?



The table shows the number of franchised car dealerships in the United States since 1990. Sandy wants to know if the relationship between the time since 1990 and the number of car dealerships can be best modeled with a linear function.

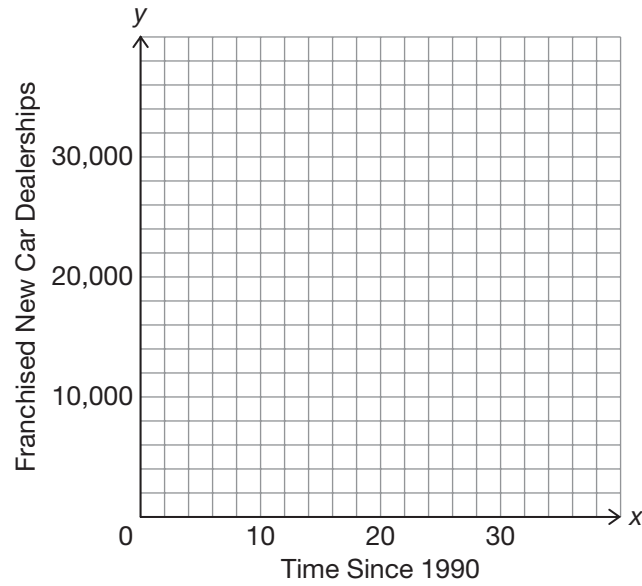
Time Since 1990	Number of Franchised New Car Dealerships
0	24,825
10	22,250
13	21,650
14	21,640
15	21,495
16	21,200
17	20,770
18	20,010
19	18,460
20	17,700

What does this table tell you?





1. Construct a scatter plot of the data on the coordinate plane shown.



2. Based on the shape of the scatter plot, do you think a linear model is a good fit for the data? Why or why not?
3. Calculate the line of best fit for the data. Write a function $c(t)$ to represent the line of best fit. Interpret the line of best fit in terms of this problem situation. Then, graph the line of best fit on the same coordinate plane as the scatter plot.

Don't always trust what you see. A little more analysis is in order!



4. Determine and interpret the correlation coefficient.

Does the correlation coefficient change your opinion on whether a linear model is good?



5. Based on the correlation coefficient, do you think a linear model is a good fit for the data? Why or why not?

6. Use the line of best fit to predict the number of car dealerships in each year.

a. 1995

b. 2015

c. 2025

7. Calculate and interpret the residuals for the data.

Time Since 1990	Number of Franchised New Car Dealerships	Predicted Value	Residual	Interpretation
0	24,825			
10	22,250			
13	21,650			
14	21,640			
15	21,495			
16	21,200			
17	20,770			
18	20,010			
19	18,460			
20	17,700			

You can use a graphing calculator to plot residuals.



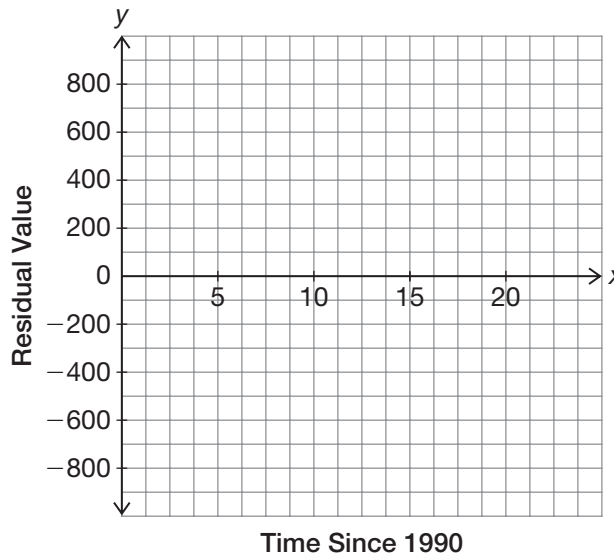
You can use a graphing calculator to show how the actual values of a data set differ from the values predicted by a linear regression.

- Step 1:** Enter the data values, press **STAT**, select **CALC**, and then select **4:LinReg(ax+b)**. Scroll down to **Store RegEQ:** Press **VARS**, select **Y-VARS** at the top, and then press **1** two times. Then select **Calculate**.
- Step 2:** Press **STAT** and then **1**. Then press the right arrow key until you get to **L6**. Press the up arrow key and then the right arrow key.
- Step 3:** If the list of residuals is not already displayed, press **2ND** and then **LIST**. Select **7↓RESID**. Press **ENTER**.
- Step 4:** Press **2ND**, **STAT PLOT**, **1** to turn on the plot and choose the type of display for the graph. Press **ZOOM** and then **9** to show the data and the line of best fit.

You can also use a graphing calculator to graph a residual plot.

- Step 5:** Press **STAT** and then **1**. Copy the data from the residuals list to **L6**. You can round the data values if you wish.
- Step 6:** Press **2ND**, **STAT PLOT**, and then **1**. Make sure **L1** is entered next to **Xlist** and **L6** is entered next to **Ylist**.
- Step 7:** Press **STAT**, select **CALC**, and then select **2:2-Var Stats**. Make sure **L1** is entered next to **Xlist** and **L6** is entered next to **Ylist**. Select **Calculate** and then press **ZOOM**, **9** to see the residual plot.

8. Create a residual plot of the data using a graphing calculator on the coordinate plane shown.



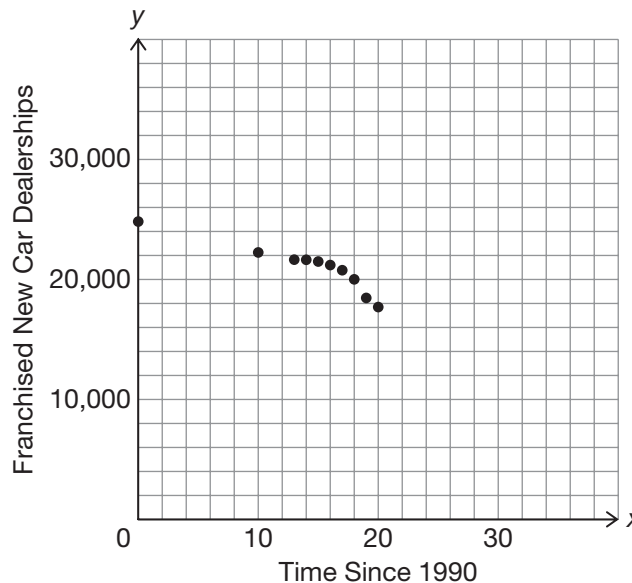
9. Based on the residual plot, do you think a linear model is a good fit for the data? Why or why not?


Remember that a residual plot can't tell you whether a linear model is appropriate. It can only tell you that there may be something better.

You used the shape of the scatter plot, the correlation coefficient, and the residual plot to determine whether a linear model was a good fit for the data. Let's consider a different function family.



10. Graph the function $q(t) = -15.657t^2 - 1.2709t + 24,650$ on the same coordinate plane as the scatter plot.



11. Do you think the function $q(t)$ is a better fit for the data than the line of best fit? Explain your reasoning.
12. Use the function $q(t)$ to predict the number of car dealerships in each year.
- a. 1995
 - b. 2015
 - c. 2025
-  13. Compare the predictions using the line of best fit $c(t)$ and the predictions using the function $q(t)$. What do you notice?

Talk the Talk



1. Explain how you can use each to help determine if a linear model is an appropriate fit for a data set.
 - a. Shape of scatter plot
 - b. Correlation coefficient
 - c. Residual plot
2. Why is it important to use more than one measure to determine if a linear model is a good fit for a data set?
3. Do you think determining the best fit for a data set is more important for interpolation or extrapolation? Explain your reasoning.

9



Be prepared to share your solutions and methods.

Who Are You? Who? Who?

Causation vs. Correlation

LEARNING GOALS

In this lesson, you will:

- Understand the difference between correlation and causation.
- Understand necessary conditions.
- Understand sufficient conditions.

KEY TERMS

- causation
- necessary condition
- sufficient condition
- common response
- confounding variable

Contrary to what you might see on TV, forensic scientists don't always catch the criminals. It is a complex science, and often a forensic team is not able to gather enough evidence to prove to a court that a criminal should be charged with a crime. In many cases, the criminal or criminals aren't found at all.

Some investigations get shelved for long periods of time until new evidence or information arrives. These are often referred to as "cold cases." DNA evidence has made it possible to solve many cold cases that were shelved before DNA testing was used. In 2011, DNA evidence was used to convict a man for a crime he committed 43 years earlier!

PROBLEM 1 Experiments and Conclusions



Students in an Atlanta classroom were asked to design an experiment, gather data, determine the correlation between the quantities, and draw conclusions about their results. For each experiment, decide whether the students' conclusions are supported by their results or are in error. Explain your reasoning.

9

1. One group of students found that the number of people that carried umbrellas is highly correlated to the days that it rained. Their conclusion was that people carrying umbrellas caused it to rain.
2. Another group found that the number of snow cones sold by a sidewalk vendor is highly correlated to the temperature. They concluded that the number of snow cones sold causes higher temperatures.
3. A third group found that high rates of school absenteeism are correlated to lower grades. They concluded that high rates of school absenteeism caused students to have lower grades.



PROBLEM 2 Proving Causation



The experiments in Problem 1, *Experiments and Conclusions*, showed us that even though two quantities are correlated, this does not mean that one quantity caused the other. This is one of the most misunderstood and misapplied uses of statistics.

Causation is when one event causes a second event. A correlation is a **necessary condition** for causation, but a correlation is not a **sufficient condition** for causation. While determining a correlation is straightforward, using statistics to establish causation is very difficult.



1. Many medical studies have tried to prove that smoking causes lung cancer.
 - a. Is smoking a necessary condition for lung cancer? Why or why not?
 - b. Is smoking a sufficient condition for lung cancer? Why or why not?
 - c. Is there a correlation between people who smoke and people who get lung cancer? Explain your reasoning.
 - d. Is it true that smoking causes lung cancer? If so, how was it proven?

2. It is often said that teenage drivers cause automobile accidents.
 - a. Is being a teenage driver a necessary condition to have an automobile accident? Why or why not?
 - b. Is being a teenage driver a sufficient condition to have an automobile accident? Why or why not?
 - c. Is there a correlation between teenage drivers and automobile accidents? Explain your reasoning.



- d. Is it true that teenage drivers cause automobile accidents? Explain your reasoning.



3. Let's revisit the example of school absenteeism causing poor performance in school. A correlation between the independent variable of days absent to the dependent variable of grades makes sense. However, this alone does not prove causation. In order to prove that the number of days that a student is absent causes the student to get poor grades, we would need to conduct more controlled experiments.

a. List several ways that you could design additional experiments to attempt to prove this assertion.

b. Will any of these experiments prove the assertion? Explain your reasoning.

4. There are two relationships that are often mistaken for causation. A **common response** is when some other reason may cause the same result. A **confounding variable** is when there are other variables that are unknown or unobserved.

a. In North Carolina, the number of shark attacks increases when the temperature increases. Therefore, a temperature increase appears to cause sharks to attack. List two or more common responses that could also cause this result.

b. A company claims that their weight loss pill caused people to lose 20 pounds when following the accompanying exercise program. List two or more confounding variables that could have had an effect on this claim.

5. For each, decide whether the correlation implies causation. List reasons why or why not.
- The number of cavities in the teeth of elementary school children is highly negatively correlated to the students' reading vocabulary.



- The number of homeless people who sleep in shelters is negatively correlated to the number of ice cream cones sold.

Talk the Talk



- Look in magazines or online for stories that report on correlational studies. Identify the variables being compared, the type of association, and the method used (if mentioned) to gather the data.
- For each of your stories, identify possible confounding variables or common responses.



Be prepared to share your solutions and methods.

Chapter 9 Summary

KEY TERMS

- interpolation (9.1)
- extrapolation (9.1)
- least squares regression line (9.1)
- residual (9.3)
- residual plot (9.3)
- causation (9.5)
- necessary condition (9.5)
- sufficient condition (9.5)
- common response (9.5)
- confounding variable (9.5)

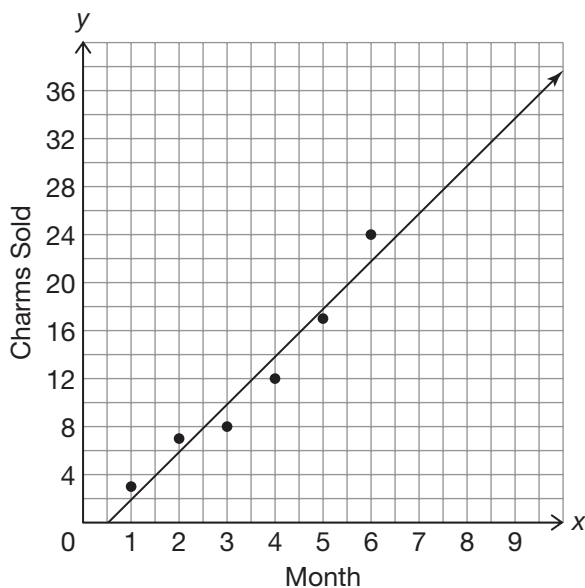
9.1 Interpreting a Linear Regression Equation

If there is a linear association between the independent and dependent variables, a linear regression can be used to make predictions within the data set. Using a linear regression to make predictions within the data set is called interpolation. To make predictions outside the data set is called extrapolation.

Example

Nina makes keychain charms that she sells to her classmates. She tracked the sales of her charms over the months since she began selling them.

Month	1	2	3	4	5	6
Charms Sold	3	7	8	12	17	24



The linear regression equation is:

$$y = 3.97x - 2.07.$$

Using the equation to interpolate, Nina should sell about 14 charms in the fourth month.

$$\begin{aligned} y &= 3.97x - 2.07 \\ &= 3.97(4) - 2.07 \\ &= 13.81 \end{aligned}$$

Using the equation to extrapolate, Nina should sell about 30 charms in the eighth month.

$$\begin{aligned} y &= 3.97x - 2.07 \\ &= 3.97(8) - 2.07 \\ &= 29.69 \end{aligned}$$

9.1 Determining a Least Squares Regression Equation

A least squares regression line is the line of best fit that minimizes the squares of the distances of the points from the line. A least squares regression line is written in the form $y = ax + b$. To calculate a and b , use these formulas:

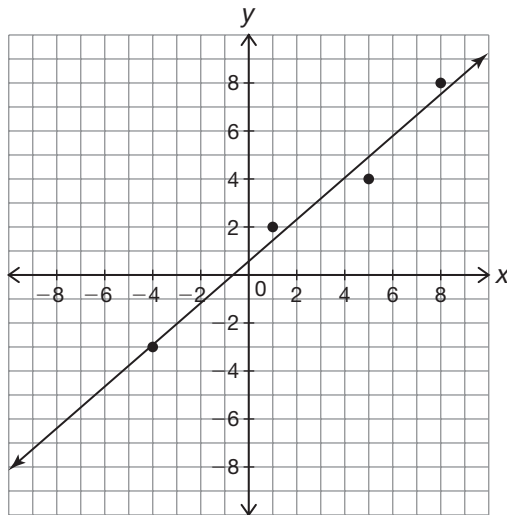
$$a = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2} \qquad b = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n\sum x^2 - (\sum x)^2}$$

where x represents all x -values from the data set, y represents all y -values from the data set, and n represents the number of coordinate pairs in the data set. A graphing calculator can also be used to determine a least squares regression equation.

Example

Data set: $(-4, -3)$, $(1, 2)$, $(5, 4)$, $(8, 8)$

The equation of the line of best fit is $y = 0.87x + 0.57$.



9.2 Analyzing Correlation Using the Correlation Coefficient

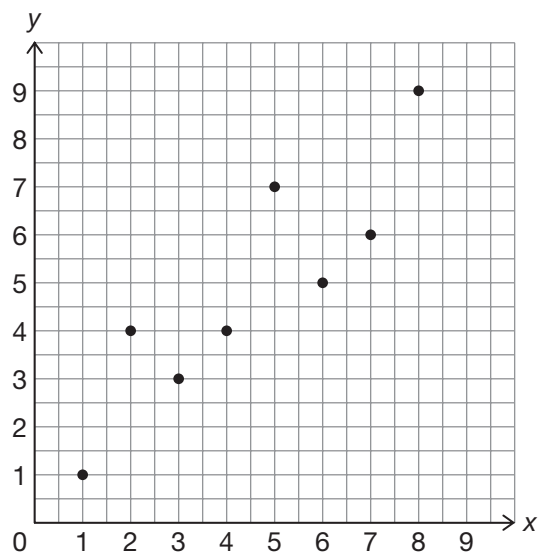
A measure of how well a linear regression line fits a set of data is called correlation. When dealing with regression equations, the variable r is used to represent a value called the correlation coefficient. The correlation coefficient indicates how close the data are to forming a straight line. The correlation coefficient either falls between -1 and 0 if the data show a negative association or between 0 and 1 if the data show a positive association. The closer the r -value gets to 0 , the less of a linear relationship there is in the data.

Example

Possible choices for r :

- $r = -0.88$
- $r = -0.11$
- $r = 0.88$
- $r = 0.11$

The data has a positive correlation. Because of this the value of r must be positive. Also, the data are fairly close to forming a straight line so of the choices, $r = 0.88$ would be the most accurate.



9.2 Determining and Interpreting the Correlation Coefficient

The correlation coefficient of a data set can be determined using this formula:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

A graphing calculator can also be used to determine the correlation coefficient.

Example

Hours of Video Games Played per Day	3	1	2	4	0
Hours of Sleep per Night	5	9	8	7	11

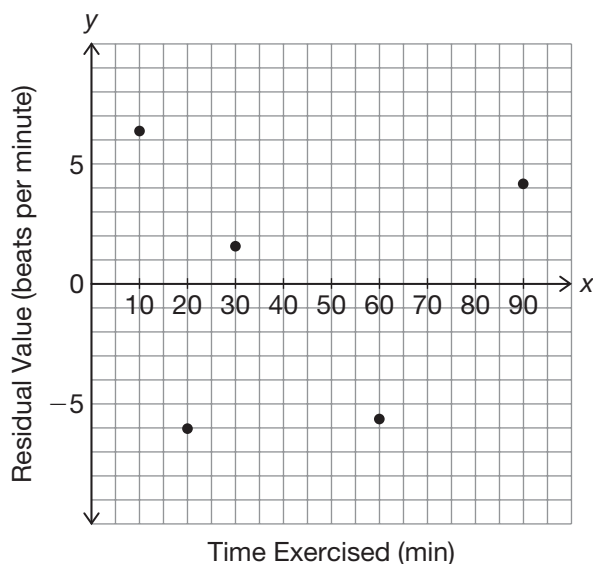
The correlation coefficient of this data set is -0.85 . The correlation coefficient indicates that the data set has a negative association and is closer to being linear than not.

9.3 Creating Residual Plots

An additional method used to determine if a linear model is appropriate for a data set is to analyze the residuals. A residual is the distance between an observed data value and its predicted value using the regression equation. Once residuals are determined, this residual data can be used to create a residual plot. A residual plot is a scatter plot of the independent variable on the x -axis and the residuals on the y -axis.

Example

Time Exercised per Day (minutes)	Resting Heart Rate (beats per minute)	Predicted Resting Heart Rate (beats per minute)	Residual Value Actual Value – Predicted
10	90	$y = -0.26(10) + 86.23 = 83.63$	$90 - 83.63 = 6.37$
20	75	$y = -0.26(20) + 86.23 = 81.03$	$75 - 81.03 = -6.03$
30	80	$y = -0.26(30) + 86.23 = 78.43$	$80 - 78.43 = 1.57$
60	65	$y = -0.26(60) + 86.23 = 70.63$	$65 - 70.63 = -5.63$
90	67	$y = -0.26(90) + 86.23 = 62.83$	$67 - 62.83 = 4.17$



9.3

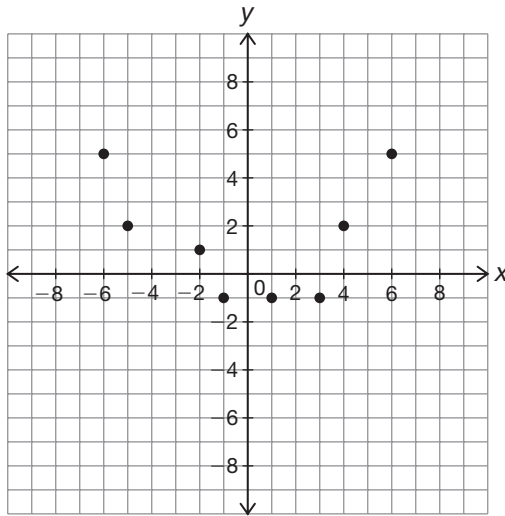
Analyzing the Shapes of Residual Plots

The shape of a residual plot can be useful when determining the most appropriate model for a data set. When a linear model is a good fit for the data, the shape of the residual plot is flat. When a linear model may not be the best fit for the data, the shape of the residual plot is a curve.

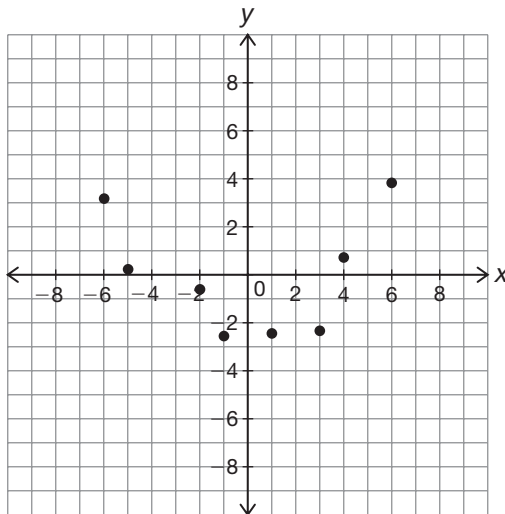
Examples

Data Set A

Scatter plot for Data A:
The scatter plot does not look like a linear model.

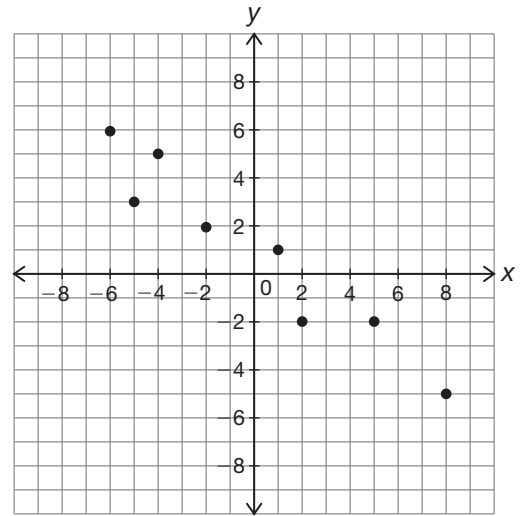


Residual plot for Data A:
The residual plot is curved, indicating a linear model may not be the best fit.

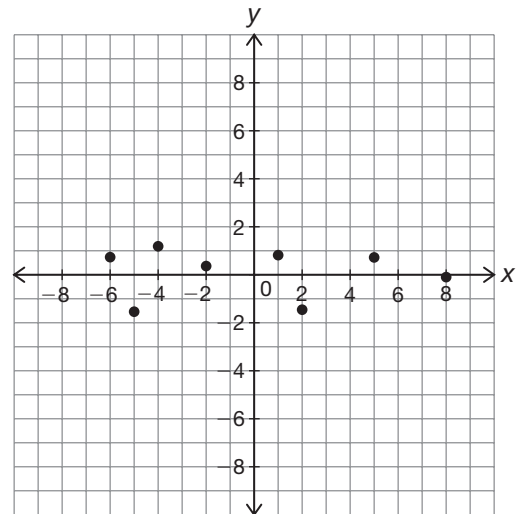


Data Set B

Scatter plot for Data B:
The scatter plot looks like a linear model.



Residual plot for Data B:
The residual plot is flat, indicating a linear model may be a good fit.



9.4

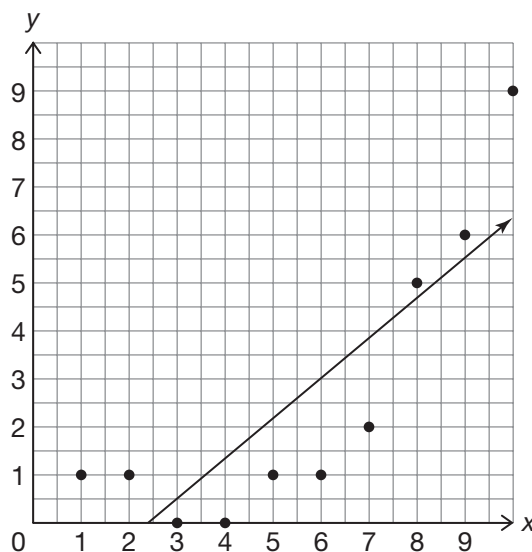
Determining Whether a Linear Regression Is a Good Fit for Data

To determine if a linear model is an appropriate fit for a data set, consider the shape of the scatter plot, the correlation coefficient, or the residual plot. It is always a good idea to look at the data in multiple ways because one measure may show you something that isn't obvious with another measure. If the points on a scatter plot appear to lie along a line, then a linear model may be appropriate. A correlation coefficient close to -1 or 1 indicates that a linear model may be appropriate. If the residual plot is curved, then a linear model may not be the most appropriate model for the data.

9

Example

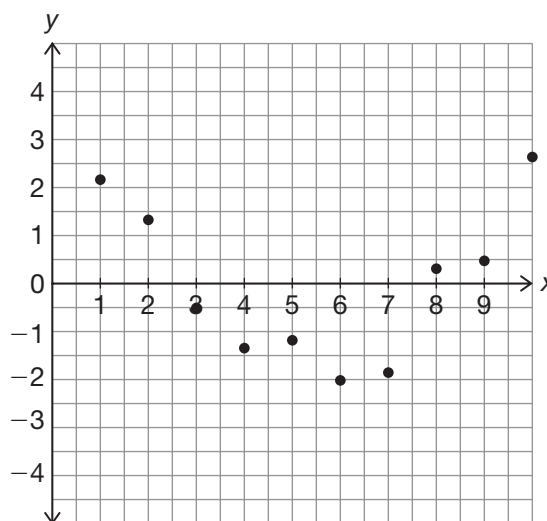
x	y	Residual Value
1	1	2.164
2	1	1.327
3	0	-0.509
4	0	-1.345
5	1	-1.182
6	1	-2.018
7	2	-1.855
8	5	0.309
9	6	0.473
10	9	2.636



The regression equation is $y = 0.836x - 2$ and the r -value is 0.837 .

From the scatter plot and the r -value, it seems like the regression equation is a good fit for the data.

The residual plot indicates that a linear model may not be the best fit for the data because the residual plot is not flat.



Examining Correlation Vs. Causation

When interpreting the correlation between two variables, you are looking at the association between the variables. While an association may exist, that does not mean there is causation between the variables. Causation is when one event causes a second event. A correlation is a necessary condition for causation, but a correlation is not a sufficient condition for causation.

Example

A group of college students conducted an experiment and found that more class absences correlated to rainy days. Therefore they concluded that rain causes students to be sick.

This correlation does not imply causation. Rain is neither a necessary condition (because students can get sick on days that do not rain) nor a sufficient condition (because not every student who is absent is necessarily sick) for students being sick.

