ADVANCED ALGEBRA

Exploring Regression

G. BURRILL, J. BURRILL, P. HOPFENSPERGER, J. LANDWEHR

DATA-DRIVEN MATHEMATICS

DALE SEYMOUR PUBLICATIONS®

Exploring Least-Squares Linear Regression

DATA-DRIVEN MATHEMATICS

Gail F. Burrill, Jack C. Burrill, Patrick W. Hopfensperger, and James M. Landwehr

Dale Seymour Publications®
White Plains, New York

This material was produced as a part of the American Statistical Association's Project "A Data-Driven Curriculum Strand for High School" with funding through the National Science Foundation, Grant #MDR-9054648. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Managing Editors: Catherine Anderson, Alan MacDonell

Editorial Manager: John Nelson

Senior Mathematics Editor: Nancy R. Anderson

Project Editor: John Sullivan

Production/Manufacturing Director: Janet Yearian

Production/Manufacturing Manager: Karen Edmonds

Production Coordinator: Roxanne Knoll

Design Manager: Jeff Kelly

Cover and Text Design: Christy Butterfield

Cover Photo: Romilly Lockyer, Image Bank

This book is published by Dale Seymour Publications®, an imprint of Addison Wesley Longman, Inc.

Dale Seymour Publications
10 Bank Street
White Plains, NY 10602
Customer Service: 800-872-1100

Printed in the United States of America.

Order number DS21182

ISBN 1-57232-245-4

1 2 3 4 5 6 7 8 9 10-ML-03 02 01 00 99 98

DALE
SEYMOUR
PUBLICATIONS®

Authors

Gail F. Burrill
Mathematics Science Education Board
Washington, D.C.

Jack C. Burrill
National Center for Mathematics
Sciences Education
University of Wisconsin-Madison
Madison, Wisconsin

Patrick W. Hopfensperger
Homestead High School
Mequon, Wisconsin

James M. Landwehr
Bell Laboratories
Lucent Technologies
Murray Hill, New Jersey

Consultants

Emily Errthum
Homestead High School
Mequon, Wisconsin

Henry Kranendonk
Rufus King High School
Milwaukee, Wisconsin

Maria Mastromatteo
Brown Middle School
Ravenna, Ohio

Vince O'Connor
Milwaukee Public Schools
Milwaukee, Wisconsin

Jeffrey Witmer
Oberlin College
Oberlin, Ohio

Data-Driven Mathematics Leadership Team

Gail F. Burrill
Mathematics Science Education Board
Washington, D.C.

Miriam Clifford
Nicolet High School
Glendale, Wisconsin

James M. Landwehr
Bell Laboratories
Lucent Technologies
Murray Hill, New Jersey

Richard Scheaffer
University of Florida
Gainesville, Florida

Kenneth Sherrick
Berlin High School
Berlin, Connecticut

Acknowledgments

The authors thank the following people for their assistance during the preparation of this module:

- The many teachers who reviewed drafts and participated in the field tests of the manuscripts
- The members of the *Data-Driven Mathematics* leadership team, the consultants, and the writers
- Robert Johnson and Bill Yager for their field testing and evaluation of the original manuscript
- Kathryn Rowe and Wayne Jones for their help in organizing the field-test process and leadership workshops
- Jean Moon for her advice on how to improve the field-test process
- Barbara Shannon for many hours of word processing and secretarial services
- Beth and Bryan Cole for writing the answers for the Teacher's Edition
- The many students at Homestead and Whitnall High Schools who helped shape the ideas as they were being developed

Table of Contents

About *Data-Driven Mathematics*

Historically, the purposes of secondary-school mathematics have been to provide students with opportunities to acquire the mathematical knowledge needed for daily life and effective citizenship, to prepare students for the workforce, and to prepare students for postsecondary education. In order to accomplish these purposes today, students must be able to analyze, interpret, and communicate information from data.

Data-Driven Mathematics is a series of modules meant to complement a mathematics curriculum in the process of reform. The modules offer materials that integrate data analysis with secondary mathematics courses. Using these materials will help teachers motivate, develop, and reinforce concepts taught in current texts. The materials incorporate major concepts from data analysis to provide realistic situations for the development of mathematical knowledge and realistic opportunities for practice. The extensive use of real data provides opportunities for students to engage in meaningful mathematics. The use of real-world examples increases student motivation and provides opportunities to apply the mathematics taught in secondary school.

The project, funded by the National Science Foundation, included writing and field testing the modules, and holding conferences for teachers to introduce them to the materials and to seek their input on the form and direction of the modules. The modules are the result of a collaboration between statisticians and teachers who have agreed on statistical concepts most important for students to know and the relationship of these concepts to the secondary mathematics curriculum.

Using This Module

Why the Content Is Important

Studying mathematics involving data brings with it the notion of fitting a line to a data set. The desire to find a *best* line gives rise to a need to understand least-squares regression and correlation. Most calculators and computer software today create the least-squares regression line and with it often display the correlation coefficient. It is because of this widespread availability and the misconceptions that can accompany these topics that this module came to be written.

In this module, you will explore the development of the least-squares regression line and its application. Why it works, when it is appropriate to use it, and how it should be interpreted are at the heart of the module. While investigating the relationship between data and the line and when the least-squares line is the *best* line, you will become aware of the dependence of the least-squares line upon both residuals and a minimum point determined by plotting the sum of the squared residuals against the slope and intercept of that line. You will also learn to appreciate the effect of outliers upon the line. Knowing how to find and interpret the correlation coefficient and understanding the expression *the strength of a linear relationship between two variables* are two of the desired outcomes of this module. Throughout the module, you will find many real-world applications of these two important topics: least-squares regression line and the correlation coefficient.

Content

Mathematics content: You will be able to:

- Represent linear functions symbolically and graphically.
- Determine and interpret slope and intercepts for linear functions.
- Represent quadratic functions symbolically and graphically.
- Determine the minimum point of a quadratic function.
- Graph the sum of quadratic functions.
- Represent absolute-value functions symbolically and graphically.
- Determine the minimum point of an absolute-value function when possible.
- Graph the sum of absolute-value functions.
- Use summation notation and perform summation arithmetic.
- Use variable notation, including subscripts and superscripts.

Statistics content: You will be able to:

- Calculate residuals.
- Find the sum of squared residuals.
- Find the absolute mean squared error.
- Work with the correlation coefficients r and r^2.
- Describe the linear relationship between two variables.
- Find least-squares regression lines.

INTRODUCTORY LESSON

Why Draw a Line Through Data?

OBJECTIVE

Discover relationships in a scatter plot by drawing lines through the data points.

INVESTIGATE

Estimating Calories

The Food and Drug Administration (FDA) requires nutrition labels on food packages. Below is an example of a label from a box of Lucky Charms breakfast cereal.

Nutrition Facts

Serving Size: 1 cup (30 g)
Servings per Container: about 13

Amount per Serving	Cereal	With $\frac{1}{2}$ cup skim milk
Calories	120	160
Calories from fat	10	15
	% Daily Values	
Total Fat 1 g	2%	2%
Saturated Fat 0 g	0%	0%
Cholesterol 0 mg	0%	1%
Sodium 210 mg	9%	11%
Potassium 55 mg	2%	7%
Total Carbohydrates 25 g	8%	10%
Dietary Fiber 1 g	6%	6%
Sugars 13 g		
Other Carbohydrates 11 g		
Protein 2 g		

Discussion and Practice

Without looking at these labels, how well can you estimate the calories of some selected food items?

1. In the table on page 2 is a list of some food items and their serving sizes. Copy the table. After each item write your estimate for how many calories are in one serving. Use the information above as a guide.

Item	Serving Size	Estimated Calories
Chicken McNuggets	6	_____
French Fries	Regular size	_____
Ben & Jerry's Cookie Dough Ice Cream	$\frac{1}{2}$ cup	_____
Saltine Crackers	5	_____
Beef Ravioli	1 cup	_____
Tomato Soup	$\frac{1}{2}$ cup	_____
Skittles	$1\frac{1}{2}$ oz	_____
Raisins	$\frac{1}{4}$ cup	_____
Parmesan Cheese	1 Tbsp	_____
Rice-a-Roni	$2\frac{1}{2}$ oz	_____
Rice Krispies Cereal	$1\frac{1}{2}$ cup	_____
Cap'n Crunch Cereal	$\frac{3}{4}$ cup	_____

2. How well were you able to estimate the number of calories in one serving of these food items? To help answer this question, use a nutrition book to find the actual number of calories for each item. Then make a scatter plot with your estimate of calories on the horizontal axis and the actual number of calories on the vertical axis.

a. Where will a point lie if your estimate was correct?

b. What line can you draw that represents estimates that are 100% accurate? Write the equation of that line and draw it on the scatter plot.

c. How does this line help you decide if you are a good estimator?

d. If the majority of your points were below the line, would you consider your estimates to be over-estimates or underestimates? Explain your answer.

e. On your scatter plot, draw a vertical line segment from each point to the line that you have drawn. What do these segments represent?

f. Describe how you can find the length of each segment that you drew in part e.

Practice and Applications

The table below shows the number of calories and total fat for hamburgers at various fast-food restaurants.

Basic Burgers	Calories	Total Fat (grams)
McDonald's	255	9
Burger King	260	10
Hardee's	260	10
Jack in the Box	267	11
Wendy's Plain Single	350	15
Burgers with the Works		
McDonald's Big Mac	500	26
Jack in the Box Jumbo Jack	584	34
Burger King Whopper	630	39
Hardee's Frisco Burger	730	47

Source: *Consumer Reports,* August, 1994

3. Do you think there is a relationship between calories and total fat content for hamburgers?

   a. To investigate any possible relationship in these data, construct a scatter plot with calories on the horizontal axis and total fat on the vertical axis.

   b. Is there an association between the variables? Explain.

4. Draw a line on the graph that you think will summarize, or *fit,* the data.

   a. How does this line help describe the relationship between the number of calories and total fat content?

   b. Describe how you could use the line and predict the fat content in a hamburger if it contained 300 calories.

5. Find an equation of the line that you have drawn by finding two points on the line.

6. What is the slope of the line you drew for Problem 4? Explain the slope in terms of the data.

7. Use your equation from Problem 5 to predict the fat content for a hamburger that has 300 calories.

Summary

We draw lines on scatter plots is to assist in the interpretation and analysis of the data. These lines can help identify important data points, summarize relationships between the variables, and predict the value of the variable on the vertical axis from the value of the variable on the horizontal axis.

LESSON 1

What Is a Residual?

How would you like to own a Corvette, or perhaps a Lamborghini?

If the city miles per gallon were known for a certain type of car, could you predict the highway miles per gallon?

The Lamborghini has the lowest city and highway miles per gallon. Does this suggest a general relationship between these two variables?

INVESTIGATE

Cars such as the Corvette and Lamborghini are known for their ability to accelerate but not for their high gas mileage. The following table lists eleven sports cars and the 1997 EPA (Environmental Protection Agency) fuel-economy estimate in miles per gallon (mpg) for city and highway driving for each.

Model	City MPG	Highway MPG
Acura NSX	18	24
Alfa Romeo Spider	22	25
Chevrolet Corvette	17	25
Ferrari 355	10	15
Jaguar XJS	17	24
Lamborghini Diablo	9	14
Lotus Esprit V8	15	23
Mazda Miata MX-5	22	28
Nissan 300ZX	18	23
Porsche 911 Carrera	17	25
Toyota MR2	20	27

Source: *1997 Gas Mileage Guide*, United States Department of Energy

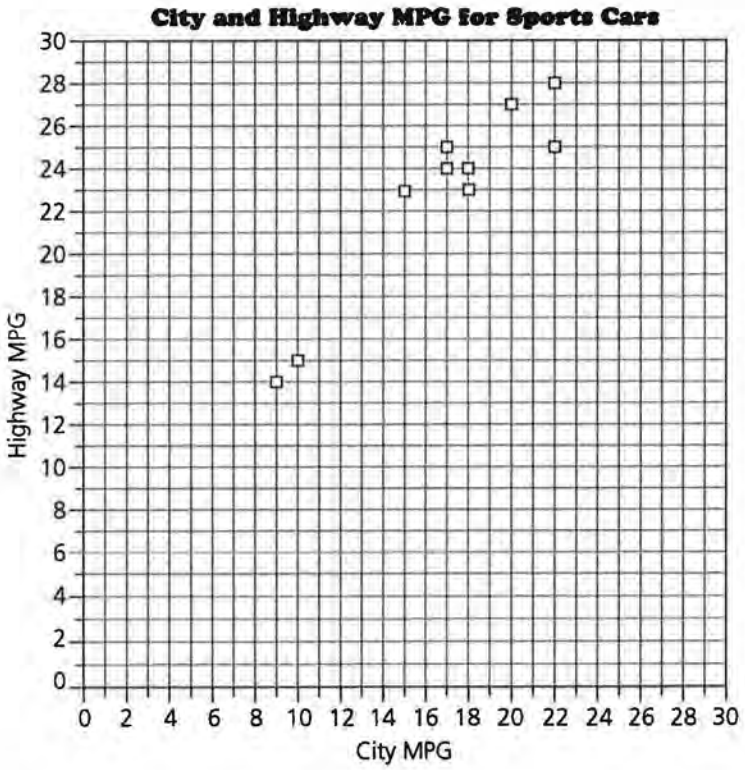
OBJECTIVES

Understand the definition of a residual.

Find the residuals for a line drawn in a scatter plot.

Discussion and Practice

1. Use the table on page 4 to answer the following questions.
   a. Which car seems to be the worst in terms of fuel economy? Which has the best gas-mileage rate?
   b. What kind of relationship would you expect between the miles per gallon for city driving and for highway driving?
2. Below is a scatter plot of the data. Describe any trends or patterns you observe in the plot. Is the trend consistent with your expectations?



3. There seems to be a relationship between miles per gallon for city driving and miles per gallon for highway driving.
   a. A car averages 14 mpg in the city. Use the scatter plot to predict the highway mileage for that car. Explain how you determined your answer.
   b. Compare your prediction with other students' predictions. How did the predictions vary?
4. Suppose you want to buy another sports car not on the list but don't know either its highway mileage or its city mileage. Describe how you might predict the highway mileage.

Predicting from a Graph

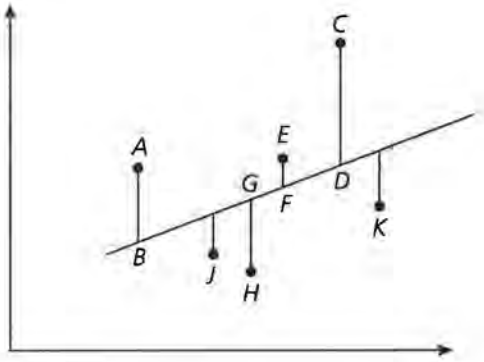
Since there appears to be a linear relationship between city mpg and highway mpg, a line can be drawn on the scatter plot to summarize this relationship. This line can also be used to make predictions.

5. On the scatter plot on *Activity Sheet 1,* draw a line that you think will summarize, or *fit,* the data.

- **a.** The city mileage listed for the Mazda Miata is 22 mpg. How many highway miles per gallon would your line predict for the Miata?
- **b.** The actual number of highway miles per gallon listed for the Mazda Miata was 28. How close was your prediction?
- **c.** Compare your prediction to those made by others in class. Whose was closest?
- **d.** Examine one another's lines. Do you think that the student who was closest also has the line that best summarizes the relationship between city and highway miles per gallon? Justify your answer.

Overall, how well does the line fit the data? The goal is to predict highway miles per gallon when you are given the city miles per gallon. The difference between the actual y-values measured and the predicted y-values determined by the line is called the ***error in prediction.***

6. The error is measured vertically from the point to the line. Why does that make sense?



The scatter plot above uses vertical line segments connecting the data points and the fitted line to show the errors in prediction. These errors are known as ***residuals***. The symbol for the predicted value is $\hat{y}$.

residual = (observed y-value) – (predicted y-value),
symbolically $r = y - \hat{y}$

7. If a data point is above the line, then its y-value, y, is greater than the predicted y-value, $\hat{y}$, and the residual is positive. If the data point is below the line, then its y-value, y, is less than the predicted y-value, $\hat{y}$, and the residual is negative. Use the plot above of points *A*, *E*, *C*, *J*, *H*, and *K* to answer the following questions.

a. $CD > AB$. What does that tell you?

b. Which data point has the greatest residual, in absolute value? How can you tell?

c. List the data points that have negative residuals. What does a negative residual tell you about your prediction?

d. Comment on this sentence: *The residual for A is the same as the residual for H.*

8. Return to the plot of the city and highway miles per gallon.

a. Draw the vertical segments that represent the residuals for your line. For which cars are your predictions too low?

b. What does each vertical unit on the graph represent?

c. Use the graph of your fitted line to find the predicted highway miles per gallon for each city mile per gallon. Then find the residual for each data point. Record your results in a table like that below or in the first table on *Activity Sheet 2.*

(City MPG, Hwy. MPG)	Predicted Hwy. MPG	Residual
(18, 24)	_____	_____
(22, 25)	_____	_____
(17, 25)	_____	_____
(10, 15)	_____	_____
(17, 24)	_____	_____
(9, 14)	_____	_____
(15, 23)	_____	_____
(22, 28)	_____	_____
(18, 23)	_____	_____
(17, 25)	_____	_____
(20, 27)	_____	_____

d. What is the residual for (15, 23)? What does it represent?

e. If all residuals are small, how accurate is the prediction using the line for these cars?

Summary

A residual for a given x-value is the difference between the observed y-value, y, and the predicted y-value, $\hat{y}$, for that x. The observed y is the y-value for the given x. A residual has the same unit as the y-values. Each residual calculated above was in miles per gallon. Residuals also have a direction, either positive or negative, depending upon their relationship to the line drawn through the data.

Predicting from an Equation

Instead of using the graph of the fitted line to find residuals, you can use an equation of that line.

9. Pick two ordered pairs on the fitted line that you drew on the scatter plot of city mpg and highway mpg.

a. Use your ordered pairs to write an equation of the line.

b. Use your equation to predict the highway mileage for a Nissan 300ZX rated by the EPA for city mileage at 18 mpg.

c. The listed highway mileage for the Nissan is 23 mpg. Find the residual. Did your equation predict too low or too high a highway mileage?

10. Use your equation from Problem 9a to find the predicted highway miles per gallon for each given city miles per gallon and the corresponding residual for that data point.

a. Record your results in a table like that below or in the second table on *Activity Sheet 2*.

(City MPG, Hwy. MPG)	Predicted Hwy. MPG	Residual
(18, 24)	_____	_____
(22, 25)	_____	_____
(17, 25)	_____	_____
(10, 15)	_____	_____
(17, 24)	_____	_____
(9, 14)	_____	_____
(15, 23)	_____	_____
(22, 28)	_____	_____
(18, 23)	_____	_____
(17, 25)	_____	_____
(20, 27)	_____	_____

b. What is the residual for a city mileage of 15 mpg?

Summary

A residual can be found both graphically and algebraically. Graphically, a residual is the length of the vertical segment between a data point and the fitted line. Algebraically, a residual is the difference for a given x between the observed y-value, y, and the predicted y-value, $\hat{y}$, using the equation of the fitted line:

$$\text{residual} = \text{observed } y - \text{predicted } y$$
$$r = y - \hat{y}$$

Using Technology to Find Residuals

Calculating residuals can be a long process. A spreadsheet or a graphing calculator allows you to easily calculate, compare, and work with residuals. Throughout this unit, Option A shows you how to use a spreadsheet, Option B a graphing calculator. Suppose the line you have drawn, in slope-intercept form, is $y = 1.2x + 3.5$.

Option A: Spreadsheet

The following shows how a spreadsheet can be set up to find the residuals. Enter the slope of 1.2 in Cell B1 and the intercept 3.5 in Cell B2. After you have typed the equation of the fitted line in C4 and the rule for the residuals in D4, use the fill down command in both columns to calculate the values. The formula in C4 uses the cell locations B1, A4, and B2.

	A	B	C	D
1	Slope =	1.2		
2	Intercept =	3.5		
3	City MPG	Hwy. MPG	Predicted Hwy. MPG	Actual – Predicted
4	18	24	=$B$1*A4+$B$2	=B4–C4
5	22	25		
6	17	25		
7	10	15		
8	17	24		
9	9	14		
10	15	23		
11	22	28		
12	18	23		
13	17	25		
14	20	27		

11. Create a spreadsheet using the format above.

a. Why are some of the values in column D of your spreadsheet negative? What do these values represent?

b. Change the value of the slope in the spreadsheet. What effect did this change have on the individual residuals?

Option B: Calculator

Another method to find residuals is to use a graphing calculator. The steps below describe how to calculate the residuals on a TI-83. Enter the equation of the line $y = 1.2x + 3.5$ into Y1 = . Then select **STAT EDIT**.

Type the city miles per gallon in List 1, L1.

Type the highway miles per gallon in List 2, L2.

Define L3 = Y1(L1) by moving the cursor to the top of L3.

Type the following: " Y1(L1) ".

Quotation marks are part of the typing.

L1	L2	L3
18	24	
22	25	
17	25	
10	15	
17	24	
9	14	
15	23	
" Y1 (L1) "		

(Note: Only the first seven entries appear in L1 and L2; remember there are eleven entries in each List.)

Place the cursor on L4. Define L4 by typing " L2 – L3 " with the cursor above L4 as pictured below.

L2	L3	L4
24	25.1	
25	29.9	
25	23.9	
15	15.5	
24	23.9	
14	14.3	
15	21.5	
L4 = " L2–L3 "		

12. Answer the following questions.

a. What does the formula you used to define L3 calculate?

b. What does the entry in L3(3) represent?

c. What will the entries in L4 represent?

13. Enter the data and follow the steps above to find L3 and L4.

a. Why are some of the values in column L4 negative?

b. Change the equation that you have entered in Y1 = . Did the residuals increase or decrease?

14. Write a paragraph summarizing what a residual represents and how to find the residuals for a data set and a line.

Practice and Applications

15. In BMX dirt-bike racing, jumping, or *getting air,* depends on many factors: the rider's skill, the angle of the jump, and the weight of the bike. Here are data about the maximum heights for various bike weights for the same rider.

Weight (pounds)	Height (inches)
19.0	10.35
19.5	10.30
20.0	10.25
20.5	10.20
21.0	10.10
22.0	9.85
22.5	9.80
23.0	9.79
23.5	9.70
24.0	9.60

Source: *Statistics Across the Curriculum*

a. Make a scatter plot of (weight, height). Draw a line on the graph that you think will fit the data.

b. Predict the height a 21.5-pound bike could clear. Explain how you made your prediction.

c. Predict the height a 35-pound bike could clear. Do you think there is a 35-pound bike?

d. Use either a spreadsheet or graphing calculator to find the residuals and explain what they represent.

16. Compare your line and residuals with those of another student.

a. What conclusions can you make about the residuals?

b. What did you do to come up with your conclusion? What evidence can you give to support your conclusion?

LESSON 2

Finding a Measure of Fit

Did you ever notice that around the time that votes are cast for the Academy Awards, Hollywood releases a number of new films?

Many times the major film companies will show these films in a great number of theaters throughout the country. Why do you think they do this?

Do you think there is a relationship between the number of screens on which a movie is shown and the amount of money taken in at the box office?

INVESTIGATE

Movies often have unusual titles. Did you ever hear of *Fried Green Tomatoes?* What do you think *Stop or My Mom Will Shoot* was all about?

The table on page 14 contains the information on the top ten films for the weekend of February 28 to March 1, 1992, about one month before the presentation of the Academy Awards. The box-office revenue column is the amount of money that the movie *grossed,* or took in, in units of $10,000.

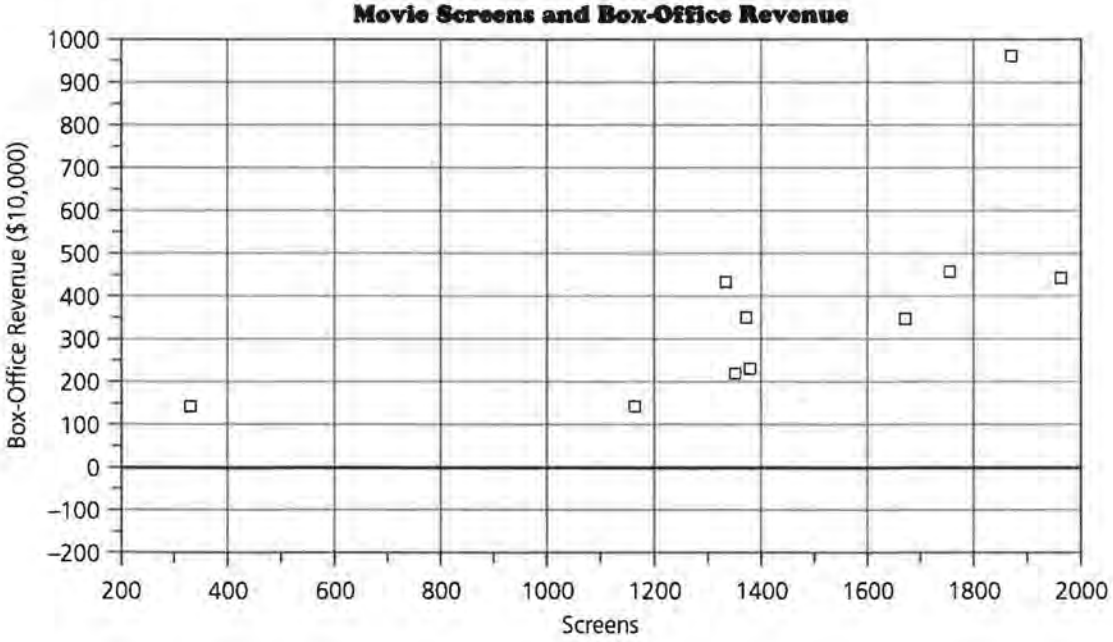
OBJECTIVE

Investigate different ways to combine residuals to determine the *best* line using the sum of the absolute values of the residuals and the sum of the squared residuals.

Film	Number of Screens	Box-Office Revenue ($10,000s)
Wayne's World	1878	964
Memoirs of an Invisible Man	1753	460
Stop or My Mom Will Shoot	1963	448
Fried Green Tomatoes	1329	436
Medicine Man	1363	353
The Hand That Rocks the Cradle	1679	352
Final Analysis	1383	230
Beauty and the Beast	1346	212
Mississippi Burning	325	150
The Prince of Tides	1163	146

Source: *Entertainment Data Inc.* and *Variety*, 1992

1. Use the data in the table to answer the following questions.
   a. How much money did *Medicine Man* gross during that weekend?
   b. On the average, how much money did *Beauty and the Beast* gross per screen?
   c. In 1992, the average ticket price for a movie was about $6. Use that average price for a ticket to estimate the number of people who saw *Beauty and the Beast* over the 3-day period.
2. On the first plot on *Activity Sheet 3*, which contains the scatter plot below, draw a line that you think could be used to predict box-office revenue from the number of screens showing a movie.



a. Compare your line to that of a classmate. Which line do you think is better?

b. What criteria did you use to make the decision?

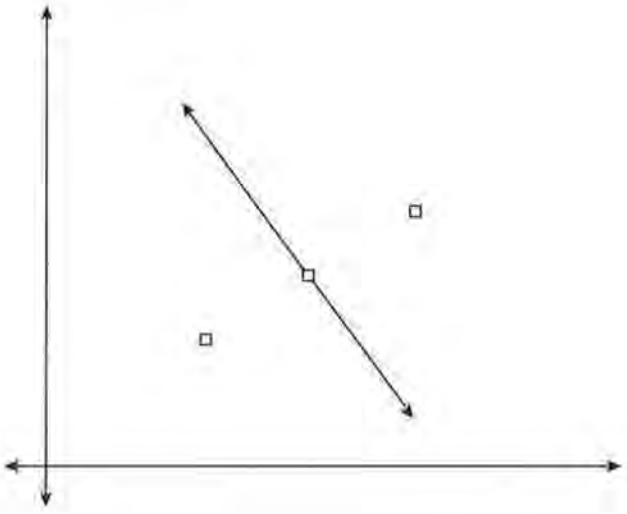
The investigations in this lesson will show you how residuals can be used to determine how well a line summarizes a relationship between two variables. In Lesson 1, you found that looking at residuals is one way to determine how accurately your line predicted outcomes. If the residuals were small, you may have felt that your line summarized the data very well. But the residuals change for each new line you draw. Because it is not possible to decide whether the line fits overall on the basis of one or two data points and their residuals, it is reasonable to look at some methods that combine residuals of a given fitted line into a single measure. A single measure can then be helpful in comparing different lines.

Sum of Residuals

3. One method of combining the residuals is to find the sum.

a. Draw in the residuals for the following plot and line.

b. Estimate the sum of the residuals for the following plot and line.



c. Comment on this statement: *If the sum of the residuals is zero, the line is a good fit for the data.*

The sign of the residual indicates whether a data point is above or below the fitted line but seems to cause a problem with the sum of the residuals. There are two simple mathematical operations that eliminate negative signs: take the absolute value or take the square.

Consider the data for the box-office revenue and the movie screens. Suppose you drew a line and found the residuals. You could consider the sum of the absolute value of the residuals, written as $\Sigma|\text{residuals}|$, or you could consider the sum of the squares of the residuals, written $\Sigma(\text{residuals})^2$.

4. Suppose the $\Sigma|\text{residuals}| = 3{,}900$ for the box-office data, and $\Sigma(\text{residuals})^2$ is 25,000.

   a. What is the average of the absolute values of the residual? This value is called the ***average absolute residual.*** What does the average absolute residual tell you about predicting the revenue given the number of screens?

   b. What is the average of the squares of the residuals? This value is called ***average squared residual.***

   c. The square root of the average squared residual is called the ***root mean squared error.*** When predicting the revenue given the number of screens, the root mean squared error gives an indication of the amount of error in the predictions. Find the root mean squared error for the residuals in part b.

What are the sums of the absolute values and squared residuals for your line? How do they compare to those from the lines drawn by others in class? As in Lesson 1, either work with a spreadsheet similar to the one shown on page 17 or use a graphing calculator with a list function.

Option A: Spreadsheet

The spreadsheet below is set up to give the results when the line $y = 0.4x - 93$ is used to predict the box-office revenue, y, from the number of screens, x.

	A	B	C	D	E	F
1	**Slope =**	**0.4**				
2	**Intercept =**	**–93**				
3	**Screens**	**Box-Office Revenue**	**Predicted Revenue**	**Residual**	**Absolute Value of Residual**	**Square of Residual**
4	1878	964	=$B$1*A4+$B$2	=B4–C4	=abs(D4)	=(D4)^2
5	1753	460				
6	1963	448				
7	1329	436				
8	1363	353				
9	1679	352				
10	1383	230				
11	1346	212				
12	325	150				
13	1163	146				
14	**Sum of absolute residuals =**				=Sum (E4:E13)	
15	**Sum of squared residuals =**					=Sum (F4:F13)

The results of this spreadsheet are shown on page 18.

	A	B	C	D	E	F
1	**Slope =**	**0.4**				
2	**Intercept =**	**–93**				
3	**Screens**	**Box-Office Revenue**	**Predicted Revenue**	**Residual**	**Absolute Value of Residual**	**Square of Residuals**
4	1878	964	658.2	305.8	305.8	93513.64
5	1753	460	608.2	–148.2	148.2	21963.24
6	1963	448	692.2	–244.2	244.2	59633.64
7	1329	436	438.6	–2.6	2.6	6.76
8	1363	353	452.2	–99.2	99.2	9840.64
9	1679	352	578.6	–226.6	226.6	51347.56
10	1383	230	460.2	–230.2	230.2	52992.04
11	1346	212	445.4	–233.4	233.4	54475.56
12	325	150	37	113	113	12769
13	1163	146	372.2	–226.2	226.2	51166.44
14	**Sum of absolute residuals =**				1829.4	
15	**Sum of squared residuals =**					407708.52

5. Refer to the results in the spreadsheet to answer the following.

a. Explain how the spreadsheet calculated the value of 658.2 in cell C4. What does this value represent?

b. Explain how the spreadsheet calculated the value of 305.8 in cell E4. What does this value represent?

c. Explain how the spreadsheet calculated the value of 93513.64 in cell F4. What does this value represent?

d. What does the value found in cell E14 represent?

e. The residuals are almost all negatives. What does this tell you about the line?

f. Find the value of the average absolute residual and the root mean squared error. What do these values represent?

6. Find an equation of the line you drew on the movie screens and revenue plot.

a. Change the slope and the intercept on the spreadsheet to match your line's slope and intercept, then record the values $\Sigma|\text{residuals}|$ and $\Sigma(\text{residuals})^2$ for your line.

b. What is the root mean squared error for your line? What does this tell you about the typical error in predicting the revenue from the number of screens?

c. How does your line seem to summarize the relationship between the number of movie screens and the box office revenue compared to the lines drawn by others in class?

Option B: Calculator

Enter the equation in Y1: Y1 = $0.4x - 93$.

Define: L3 as " Y1(L1) ",

L4 as " L2 – L3 ",

L5 as " abs (L4) ", and

L6 as " L4^2 ".

Screens	Box Office	Predicted	Residual	Absolute	Square
L1	L2	L3	L4	L5	L6
1878	964	658.2	305.8	305.8	93514
1753	460	608.2	–148.2	148.2	21963
1963	448	692.2	–244.2	244.2	59634
1329	436	438.6	–2.6	2.6	6.76
1363	353	452.2	–99.2	99.2	9840.6
1679	352	578.6	–226.6	226.6	51348
1383	230	460.2	–230.2	230.2	52992
L3 = " Y1(L1) "					

7. Use the above results to answer the following.

a. Explain how the calculator found the value of 658.2 in L3(1).

b. Explain how the calculator found the value of 305.8 in L4(1).

c. Explain how the calculator found the value of 93,514 in L6(1).

d. The residuals are almost all negatives. What does this tell you about the line?

e. To find the sum of the absolute or squared residuals, use STAT/CALC/1-Var Stats and select the appropriate list. What are $\Sigma|\text{residuals}|$ and $\Sigma(\text{residuals})^2$?

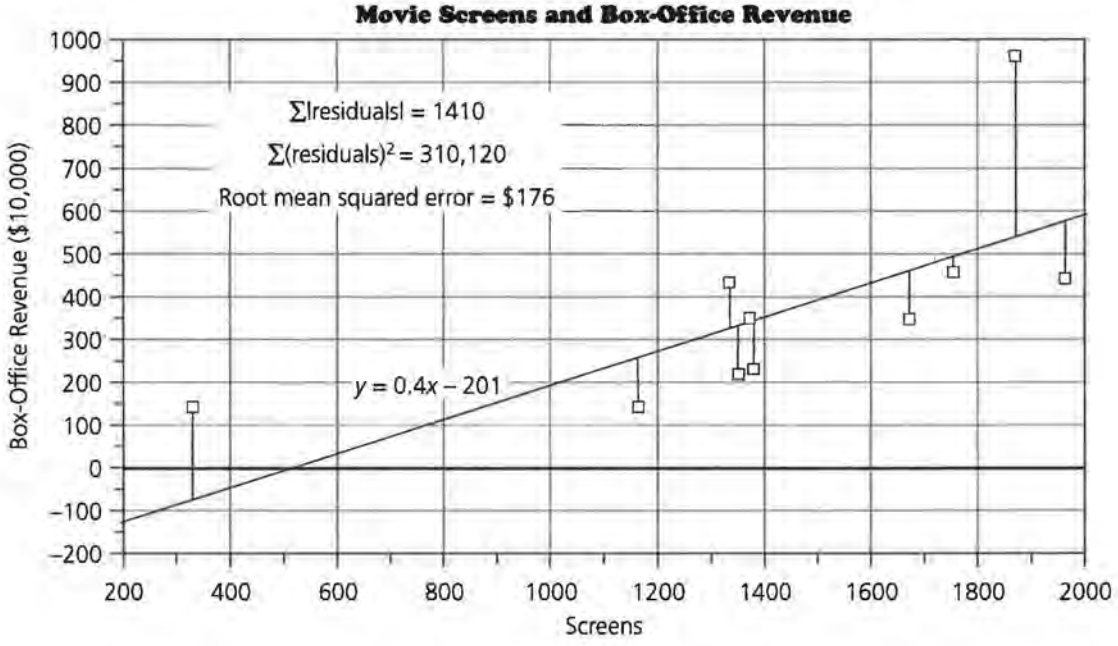
8. Find an equation of the line you drew on the movie-screens-and-revenue plot.

   a. Change the equation in Y1 to your equation. Then record the values $\Sigma|\text{residuals}|$ and $\Sigma(\text{residuals})^2$ for your line.

   b. Find the value of the root mean squared error. What does this tell you about the typical error in predicting the revenue from the number of screens?

   c. How does your line seem to summarize the relationship between the number of movie screens and the box-office revenue compared to the lines drawn by others in class?

Study the plot below. Remember that a residual is the difference between the observed y-value and the predicted y-value for a given value of x. It can be represented by the vertical distance between the line and the data point. The sum of the absolute value of the residuals or the sum of the squares of the residuals will be smallest in the line that fits a data set best.



So far, the slope and intercept for a line, the sum of the absolute values of the residuals, and the sum of the squared residuals for that line have been determined. To find the *best* line, find a line that will minimize the sum of the absolute residuals and a line that will minimize the sum of the squared residuals. Which line is better? Will the same line minimize both sums?

9. Was $\Sigma|\text{residuals}|$ for your line less than the value listed in the figure above? Was the $\Sigma(\text{residuals})^2$ smaller? Do you think your line is a *better* line than the one shown above? Explain.

10. From the lines of other groups, collect the slope, y-intercept, $\Sigma|\text{residuals}|$ and $\Sigma(\text{residuals})^2$.

- **a.** If the definition of the *best* line is the one that minimizes the $\Sigma|\text{residuals}|$, what is the equation of the line from your group that has the least $\Sigma|\text{residuals}|$? Graph this line on the second plot on *Activity Sheet 3*.
- **b.** If the definition of the best line is the one that minimizes the $\Sigma(\text{residuals})^2$, what is the equation of the line that has the least $\Sigma(\text{residuals})^2$? Graph this line on the second plot on *Activity Sheet 3*.
- **c.** Are the two lines identical? If not, explain why there are two lines that could be the *best* line to summarize data presented on a scatter plot.

11. Describe the two methods that can be used to decide which line better summarizes data presented on a scatter plot.

Summary

When finding an equation of the *best* line, it is possible to arrive at different answers depending on the definition that is used. In this section, two definitions were used to define *best* line. One definition defined *best* as the line that minimized the sum of the absolute values of the residuals. Another definition minimized the sum of the squares of the residuals. Lesson 3 will compare these two definitions and discuss which definition is better.

Practice and Applications

12. Recall that in BMX dirt-bike racing, jumping depends on many factors: the rider's skill, the angle of the jump, and the weight of the bike. Here again are the data presented in Lesson 1 about the maximum height for various bike weights.

Weight (pounds)	Height (inches)
19.0	10.35
19.5	10.30
20.0	10.25
20.5	10.20
21.0	10.10
22.0	9.85
22.5	9.80
23.0	9.79
23.5	9.70
24.0	9.60

Source: *Statistics Across the Curriculum*

a. Find the slopes and the intercepts of three different lines that you think might summarize the data. Then find an equation of each line.

b. Create a spreadsheet similar to the one shown in this lesson or use a graphing calculator to find the sum of the absolute values of the residuals and the sum of squared residuals for each line.

c. Which equation minimizes the sum of the absolute values? Find the average absolute residual.

d. Which equation of the line minimizes the sum of squared residuals? Find the root mean squared error.

e. Use the equations from parts c and d to predict the height for a 20.5-pound bike.

f. Which of the two equations do you think is better at predicting maximum height for a bike? Explain your answer.

LESSON 3

Squaring or Absolute Value?

If it is important to combine residuals in some manner that will eliminate the negative sign, which method should be used?

Is one method of eliminating the negative sign of residuals easier to work with for all sets of data?

Lesson 2 involved finding the residuals for a given data set and its fitted line and investigated both the absolute value and the square of those residuals as a method of eliminating any negative signs created.

OBJECTIVES

Recognize and describe the graph of quadratic and absolute-value functions.

Recognize what happens when you combine two or more absolute-value functions or two or more quadratic functions.

INVESTIGATE

Absolute-Value Functions

An answer to the question in the title can be found by investigating how functions such as $y = (x - 2)^2$ and $y = |x - 2|$ behave and by comparing the results from adding quadratics to the results from adding absolute values.

Discussion and Practice

1. Graph the function $y = |x|$. What is the minimum point of the graph?
2. Graph the function $y = |x - 2|$. What is the minimum point of the graph?
3. Write a comparison of the graphs of $y = |x|$ and $y = |x - 2|$.
4. Graph each absolute-value function.
   a. $y = |x - 3|$ b. $y = |3 - x|$ c. $y = |x + 4|$
   d. $y = |x - 5|$ e. $y = |5 - x|$

5. Use the examples on page 23 for the following.

   a. In general, what does the graph of $y = |x - a|$ look like for $a > 0$?

   b. In general, what does the graph of $y = |a - x|$ look like for $a > 0$?

   c. Write a comparison of the graphs of $y = |x - a|$ and of $y = |a - x|$.

Summary

The graph of the ***absolute-value function,*** of the form $y = |x - a|$, is shaped like a V. The sides of the V are rays, and each has a constant slope, although the slope of one ray is positive and the other is negative. The graph is symmetric around the line $x = a$, which is parallel to the y-axis. This line is called the ***axis of symmetry.***

Quadratic Functions

6. Graph the function $y = x^2$. What is the minimum point of the graph?

7. Graph the function $y = (x - 2)^2$. What is the minimum point of the graph?

8. Write a comparison of the graphs of $y = x^2$ and $y = (x - 2)^2$.

9. Graph each quadratic function.

   a. $y = (x - 3)^2$ b. $y = (3 - x)^2$ c. $y = (x + 4)^2$

   d. $y = (x - 5)^2$ e. $y = (5 - x)^2$

10. Use the examples above for the following.

    a. In general, what does the graph of $y = (x - a)^2$ look like for $a > 0$?

    b. In general, what does the graph of $y = (a - x)^2$ look like for $a > 0$?

    c. Write a comparison of the graphs of $y = (x - a)^2$ and of $y = (a - x)^2$.

Summary

In general, if a is any real number, the equation of the form $y = (x - a)^2$ describes a ***quadratic function.*** The graph of a quadratic equation is called a ***parabola.*** The parabola is generally shaped like a U, and the vertex of this U is called a ***turning***

point. If the graph opens up, this point is a ***minimum point.*** If the graph opens down, the point is a ***maximum point.*** The graph is symmetric around the line $x = a$, which is parallel to the y-axis. This line is called the ***axis of symmetry.***

Combining Absolute-Value Functions

In Lesson 2, the sum of the absolute values of the residuals and the sum of the squares of the residuals were found.

To help decide which of these two methods might be better or easier to work with, we will investigate functions that are sums of absolute values and functions that are sums of squares.

11. Consider $y = |5 - x| + |2 - x|$.

- **a.** Describe the graph you think this sum of absolute-value expressions will have.
- **b.** Graph the function. How successfully did you describe the graph?
- **c.** What is the minimum y-value? Describe the x-values that give this minimum.

12. Now, introduce a third absolute-value difference: $y = |5 - x| + |2 - x| + |8 - x|$.

- **a.** Describe the graph you think this sum of absolute-value expressions will have.
- **b.** Graph the function. How successfully did you describe the graph?
- **c.** What is the minimum y-value? Describe the x-values that give this minimum.
- **d.** Comment on this statement: *The graph of a function made up of the sum of a set of absolute values is shaped like the absolute-value function,* $y = |x|$.

Combining Quadratic Functions

13. Now consider $y = (5 - x)^2 + (2 - x)^2$.

- **a.** Describe the graph you think this sum of squares has.
- **b.** Graph the function. How successfully did you describe the graph?
- **c.** What is the minimum y-value? Describe the x-values that give this minimum.

14. What will the graph of the sum of three squared differences look like?

a. Make a conjecture about the shape of the graph $y = (5 - x)^2 + (2 - x)^2 + (8 - x)^2$.

b. Graph the function.

c. What is the minimum y-value? Describe the x-values that give this minimum.

d. Comment on this statement: *The graph of a function made up of the sum of a set of quadratic expressions is shaped like the graph of a quadratic equation, $y = x^2$.*

Summary

When investigating the absolute-value function, it became difficult to predict what the graph would look like as the number of absolute-value terms increased. The graph of $y = |x|$ is V-shaped, but as more absolute-value terms are added, the shape of the graph changes and is difficult to analyze. The graph of $y = |x|$ has one ordered pair that is a minimum, but as terms are added there may be many ordered pairs with the minimum y-value. The graph of a set of squared differences, however, does not cause the same confusion. No matter how many quadratic terms are summed, the graph remains a parabola. Each graph has one and only one x-value that corresponds to the minimum y-value. The sum of the squares will be a familiar curve, a parabola, which is easy to graph and analyze, and always has one minimum point. Statisticians normally work with a line of best fit that uses squared differences rather than absolute differences. That line is called the ***least-squares regression line.***

Practice and Applications

Graph each equation and determine its minimum point(s).

15. **a.** $y = |x|$ **b.** $y = |3 - x|$

c. $y = |3 - x| + |5 - x|$

d. $y = |3 - x| + |5 - x| + |4 + x|$

16. **a.** $y = x^2$ **b.** $y = (3 - x)^2$

c. $y = (3 - x)^2 + (5 - x)^2$

d. $y = (3 - x)^2 + (5 - x)^2 + (4 + x)^2$

LESSON 4

Finding the *Best* Slope

How can you know you really do have the *best* line?

How can you be sure that different people investigating a problem will come up with exactly the same results?

How can the *best* line be found without trying all the possibilities?

Can the slope be found that minimizes the sum of squared residuals?

In the last lesson, the method of minimizing the sum of squared residuals was accepted as a means of determining a *best* line. Trying a variety of equations eventually gives you a line that has a small sum of squared residuals.

OBJECTIVE

Find the slope of a line that minimizes the sum of the squared residuals.

INVESTIGATE

This lesson investigates the relationship between the slope of a line and the sum of the squared residuals of the line; that is, you will find the slope that minimizes the sum of squared residuals.

There are very sophisticated methods that can be employed to determine this line, but the method introduced here produces the same result. The equation of a line is determined by a point on the line and its slope, and these values vary from line to line. One way to determine the slope of a line that minimizes the sum of squared residuals is to fix a point and vary the slope. The procedure will be to guess that the best line will most likely contain the center point of the data. Based on that assumption, identify the center point and consider lines with different slopes that would pass through that point. From the sum of the squared residuals for each of these lines, it will be possible to find the slope that yields the minimum squared residuals.

In Lesson 5, you will use that slope and vary the y-intercept. From the sum of the squared residuals for each of these lines, it will be possible to determine the y-intercept of the line that has the minimum sum of squared residuals.

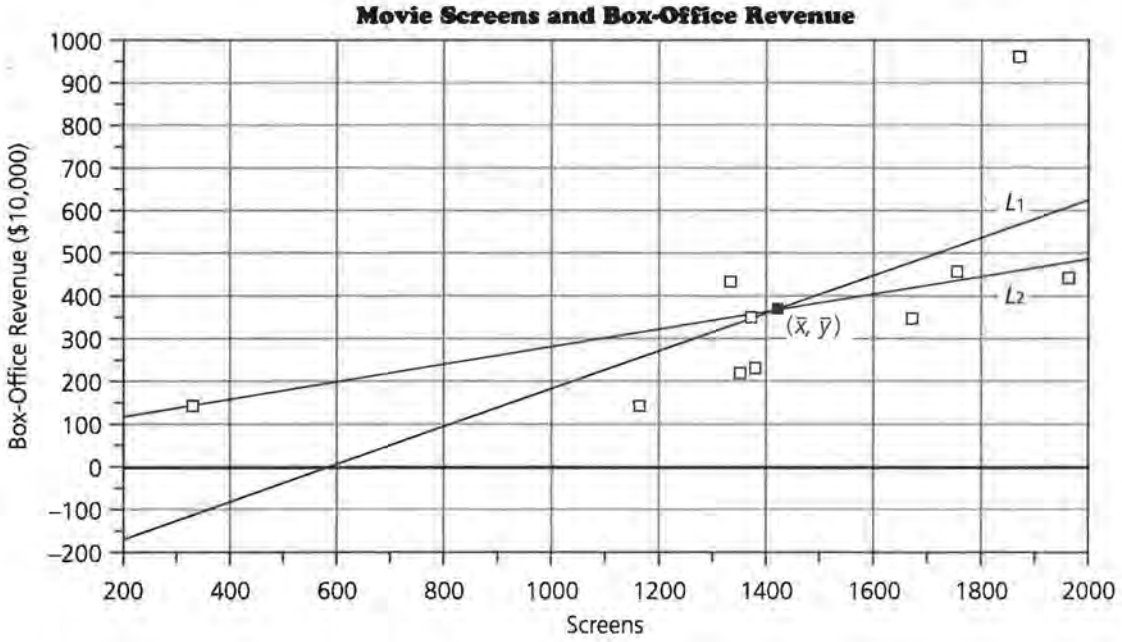
Discussion and Practice

The first step is to decide which point should be fixed. The box-office revenue for any movie could be estimated without knowing the number of screens by using $\bar{y}$, the mean box-office revenue of the movies on the list. Likewise, the mean number for all the screens, that is, $\bar{x}$, can serve as an estimate for the number of screens for any movie being studied.

Since the mean is a value that can be used to summarize the center of a distribution, it seems reasonable to use $(\bar{x}, \bar{y})$ as the fixed point. This point is called the ***centroid.*** With this point fixed, you can then draw lines with different slopes through this point to find the slope that minimizes the sum of the squared residuals.

1. Use the data on the top ten films from Lesson 2.
   a. Determine the mean number of screens and the mean reported box-office revenue. Call this point $(\bar{x}, \bar{y})$.
   b. Locate the point $(\bar{x}, \bar{y})$ on the scatter plot of the box-office revenue on a clean copy of *Activity Sheet 3.*

There are many lines that pass through the point $(\bar{x}, \bar{y})$. Investigating some of these will help determine the relationship between the slope and the sum of the squared residuals.



c. A useful way to write an equation given a point (x_1, y_1) and the slope, m, is $y = y_1 + m(x - x_1)$. Use the centroid and another point on each line from the graph above to calculate the slope. Then write an equation for each line in the form $y = y_1 + m(x - x_1)$.

d. Draw a line through the point $(\bar{x}, \bar{y})$ that you think summarizes the data fairly well. Find the slope of the line and then write an equation of your line in the form $y = \bar{y} + m(x - \bar{x})$. Compare your equation with those of your classmates. How are the equations similar? How are the equations different?

The next step is to find the sum of squared residuals for each line.

2. Use a spreadsheet or calculator to find the sum of squared residuals. Refer to the example that follows. Record the slope of your line and the sum.

Option A: Spreadsheet

Enter the slope you are using in B1. Type the equation in C3, the rule for the squared difference in D3, and fill down both columns.

	A	B	C	D
1	**Slope=**			
2	**Screens**	**Box-Office Revenue**	**Predicted Revenue**	**Square of Residual**
3	1878	964	=375.1+$B$1*(A3–1418.2)	=(B3–C3)^2
4	1753	460		
5	1963	448		
6	1329	436		
7	1363	353		
8	1679	352		
9	1383	230		
10	1346	212		
11	325	150		
12	1163	146		
13			**Sum =**	=Sum(D3:D12)

3. Use the spreadsheet above to answer the following.

a. The formula in cell C3 uses cell locations B1 and A3. What data are stored in each of these cells?

b. What do the values 1418.2 and 375.1 represent in the formula in C3?

c. What does the formula in C3 calculate?

d. What does the formula in D3 calculate?

Option B: Calculator

Define L3 using quotation marks and your equation.

Define L4 as " (L2 – L3)^2 ".

Screens	Box Office	L3	L4
1878	964		
1753	460		
1963	448		
1329	436		
1363	353		
1679	352		
1383	230		
1346	212		
325	150		
1163	146		
L3 =			

4. Use the slope 0.4 and the point (1418.2, 375.1).
   a. What equation do you enter in Y1?
   b. What does the value in L3(3) represent?
   c. What does the value in L4(2) represent?
   d. Find the sum of squared residuals.

At this point you have the equation for one line through the point $(\bar{x}, \bar{y})$. The goal is to find a slope that minimizes the sum of squared residuals. Try several more lines, each line with a different slope, but all passing through $(\bar{x}, \bar{y})$.

5. Record the slopes and sum of squared residuals found by the rest of the class.

6. Data collected by three students in a mathematics class are at the right.

Slope	$\Sigma(\text{residuals})^2$
0.22	323,975
0.45	327,886
0.26	309,714

   a. Describe how the value of 323,975 was obtained.
   b. Write an equation of the line used to obtain 300,000 for $\Sigma(\text{residuals})^2$.
   c. Use the value 300,000 to find the value of the root mean squared error. What does this value tell you about predicting the movie revenue from the number of screens?

7. Plot the three ordered pairs (slope, $\Sigma(\text{residuals})^2$) from Problem 6. A grid is provided on *Activity Sheet 4*.

   a. Plot the ordered pairs (slope, $\Sigma(\text{residuals})^2$) you found from the lines you and other students in class used. What pattern do you see in the scatter plot?

   b. Draw a smooth curve through the ordered pairs on the graph. What kind of equation might be used to describe this graph?

   c. Find the x-coordinate of the point that has the least y-coordinate. Write the coordinates of this point.

   d. Describe what the x-coordinate and y-coordinate of this point represent.

Summary

The overall goal has been to find the equation of the line that *best* fits the data.

In this lesson, you found a slope that minimized the sum of squared residuals starting with the fixed point $(\bar{x}, \bar{y})$. The slope from any line through this point generated its own $\Sigma(\text{residuals})^2$. The ordered pairs (slope, $\Sigma(\text{residuals})^2$) formed a parabola where the x-coordinate of the minimum point represented the slope of the line that had the least $\Sigma(\text{residuals})^2$.

Practice and Applications

8. Write a paragraph discussing how to find an estimate for the value of the slope that minimizes the sum of squared residuals.

9. Use the data on BMX dirt-bike racing found at the end of Lesson 1 to find the value of the slope that minimizes the sum of squared residuals, starting with the point $(\bar{x}, \bar{y})$.

LESSON 5

Finding the *Best* Intercept

What do you think will happen if you keep the slope the same but change the point that the line passes through?

INVESTIGATE

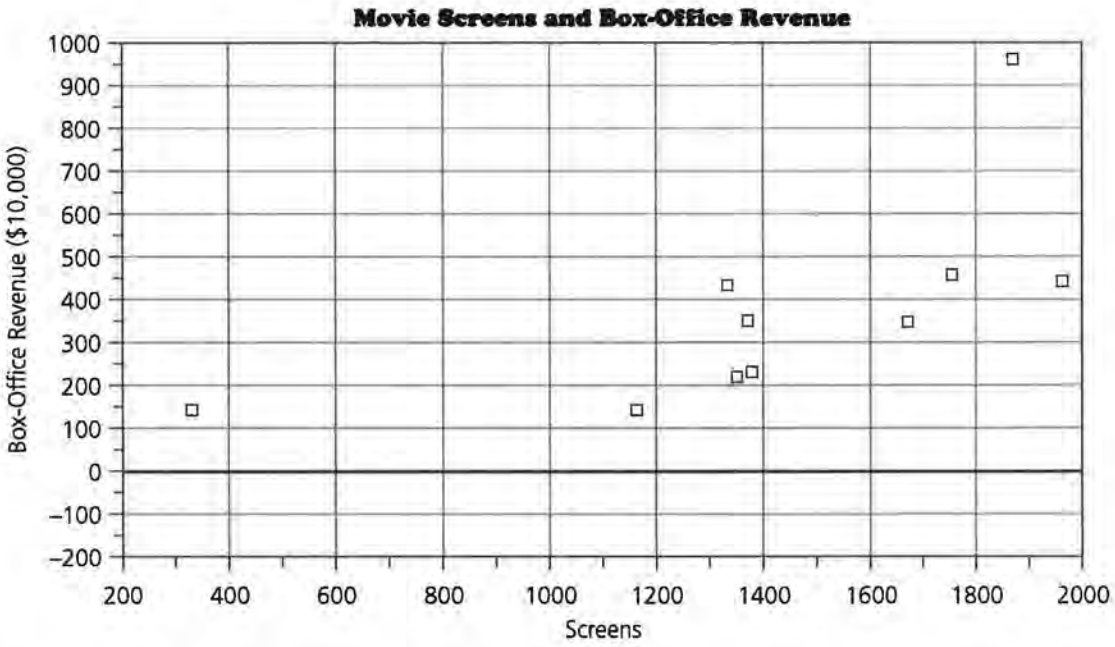
The goals in Lesson 4 and in this lesson are to find the slope and the y-intercept of the line that minimizes the sum of squared residuals. By drawing several lines through the point (1418.2, 375.1) in Lesson 4, you found that a slope of approximately 0.33 gave the least sum of squared residuals. But remember, we made the assumption that the line had to pass through the point with coordinates $(\bar{x}, \bar{y})$, the mean of the number of screens and the mean of the box-office revenue. In this lesson, the slope will be held constant and the point will be varied. Remember, the goal is to find the point that minimizes the sum of squared residuals.

OBJECTIVE

Investigate the relationship between the intercept and the sum of squared residuals.

Discussion and Practice

1. The scatter plot on page 34 is reproduced on *Activity Sheet 3*. Use a clean copy of the activity sheet, and draw the line p that passes through the point (1418.2, 375.1) and has a slope of 0.33.



a. Write an equation of line p in the form $y = y_1 + m(x - x_1)$, where m is the slope and (x_1, y_1) is any point on the line.

b. Write an equation of line p in the form $y = mx + b$, where b is the y-coordinate of the y-intercept and m is the slope. What is the y-intercept of the line with slope equal to 0.33 that passes through $(\bar{x}, \bar{y})$?

2. On your scatter plot from Problem 1, draw a line q that is parallel to line p and passes through the point (1600, 370).

a. What is the slope of line q? How did you find your answer?

b. Write an equation of line q in the form $y = y_1 + m(x - x_1)$.

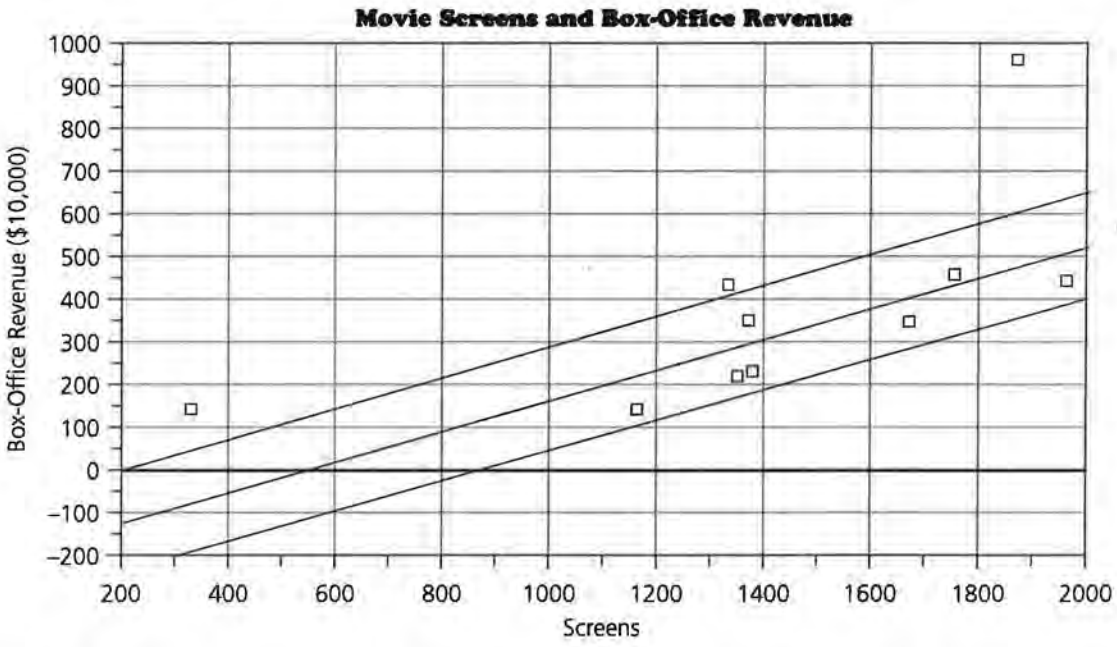
c. Write an equation of line q in the form $y = mx + b$. What is the y-intercept of this line?

3. Suppose you choose another point and draw a line parallel to line p.

a. What effect will this have on the slope and the y-intercept?

b. What effect does changing the point have on the equation of line p when the equation is written in the $y = mx + b$ form?

4. The lines below can be considered a *family* of lines. Describe the similarities and differences.



The goal of this lesson is to draw a family of lines that all have the same slope, 0.33, and to find the *y*-intercept of the line that gives the least value for the sum of squared residuals. Just as before, use a spreadsheet or graphing calculator to investigate the change in the sum of squared residuals as different lines are drawn on the plot of (number of screens, box-office revenue). Remember this important point: Holding the slope constant and changing from the centroid to other points for the line to pass through also changes the *y*-intercept.

5. Find the sum of squared residuals for line *q* from Problem 2. If you use a spreadsheet, you have to enter the *y*-intercept of the line, so you must write the equation of the line in slope-intercept ($y = mx + b$) form. Record your results in a table similar to one of those on page 36.

Option A: Spreadsheet

Enter the intercept you are using in B1, the equation in C3, and the rule for the squared difference in D3. Fill down columns C and D.

	A	B	C	D
1	*y*-Intercept=			
2	Screens	Box-Office Revenue	Predicted Revenue	Square of Residual
3	1878	964	=0.33*A3+$B$1	=(B3–C3)^2
4	1753	460		
5	1963	448		
6	1329	436		
7	1363	353		
8	1679	352		
9	1383	230		
10	1346	212		
11	325	150		
12	1163	146		
13			Sum =	=Sum(D3:D12)

Option B: Calculator

Type the equation you are using in Y1.

Define L3 as Y1(L1) and L4 as $(L2 - L3)^2$.

L1	L2	L3
1878	964	
1753	460	
1963	448	
1329	436	
1363	353	
1679	352	
1383	230	
1346	212	
325	150	
1163	146	
L3 = Y1(L1)		

6. Draw at least two more lines parallel to line p on your scatter plot from Problems 1 and 2. Write an equation of each line in the form $y = 0.33x + b$ and find the sum of the squared residuals. Record your results in a table like the following, or use *Activity Sheet 4*.

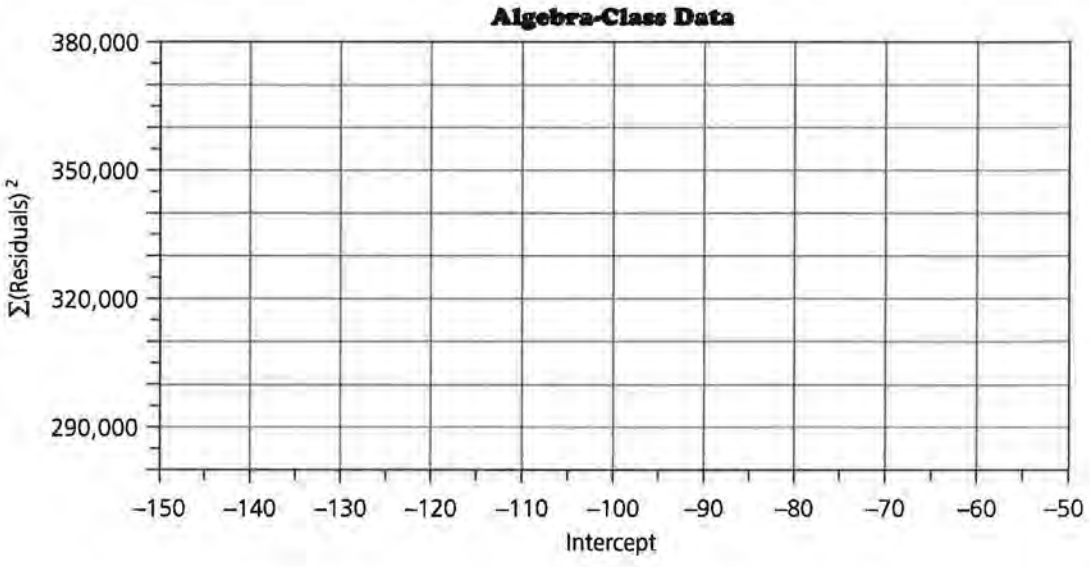
Slope	Point	Intercept	Sum of Squared Residuals
0.33	(1679, 352)	_____	_____
0.33	_____	_____	_____
0.33	_____	_____	_____
0.33	_____	_____	_____
0.33	_____	_____	_____

7. Data collected by several students in an algebra class are below.

Slope	Intercept	Σ(Residuals)^2
0.33	–100	300,398
0.33	–96	299,990
0.33	–95	299,938

a. Write an equation of the line that gave $\Sigma(\text{residuals})^2 = 299{,}990$. What does this equation tell you about the number of screens and box-office revenue?

b. Use a grid like the one below, reproduced on *Activity Sheet 4*, or your graphing calculator and plot the three ordered pairs (intercept, $\Sigma(\text{residuals})^2$) from the table shown in this problem.



c. On the plot, add the ordered pairs (intercept, $\Sigma(\text{residuals})^2$) you found from the lines you drew. Add ordered pairs from your classmates. Describe any patterns you observe in the scatter plot.

d. Draw a smooth curve through the ordered pairs on the graph. What kind of equation might be used to describe this curve?

e. From the scatter plot, determine the x-coordinate of the point that has the least y-coordinate. Write the coordinates of this point.

f. Describe what the x-coordinate and the y-coordinate of this point represent. Then compare this intercept with the intercept from Problem 1b.

8. In this lesson, you fixed the slope at 0.33 and then drew lines with this slope. Write a paragraph discussing how to find the value of the intercept that minimizes the sum of squared residuals for this fixed slope.

Summary

In this lesson the slope was fixed and points (intercept, $\Sigma(\text{residuals})^2$) were generated. The plot of these points was a parabola with the point that gave the smallest sum of squared residuals at the minimum point, $(\overline{x}, \overline{y})$, the centroid.

Practice and Applications

9. Use the data on BMX dirt-bike racing found at the end of Lesson 1 to find the value of the y-intercept that minimizes the sum of squared residuals. Use the value of the slope found in Problem 9 of Lesson 4.

Attention: Even though varying the slope while passing the line through the centroid led to the discovery of a slope that created the least sum of squared residuals and varying the point the line was passing through while holding the slope constant created the least sum of squared residuals, care must be taken **to not assume** that the *best* line will be the one in which we do those changes simultaneously. Lesson 6 will consider the effect of making those changes simultaneously and how that might be done.

LESSON 6

The *Best* Slope and Intercept

What happens to the sum of squared residuals if *both* the slope and intercept are varied?

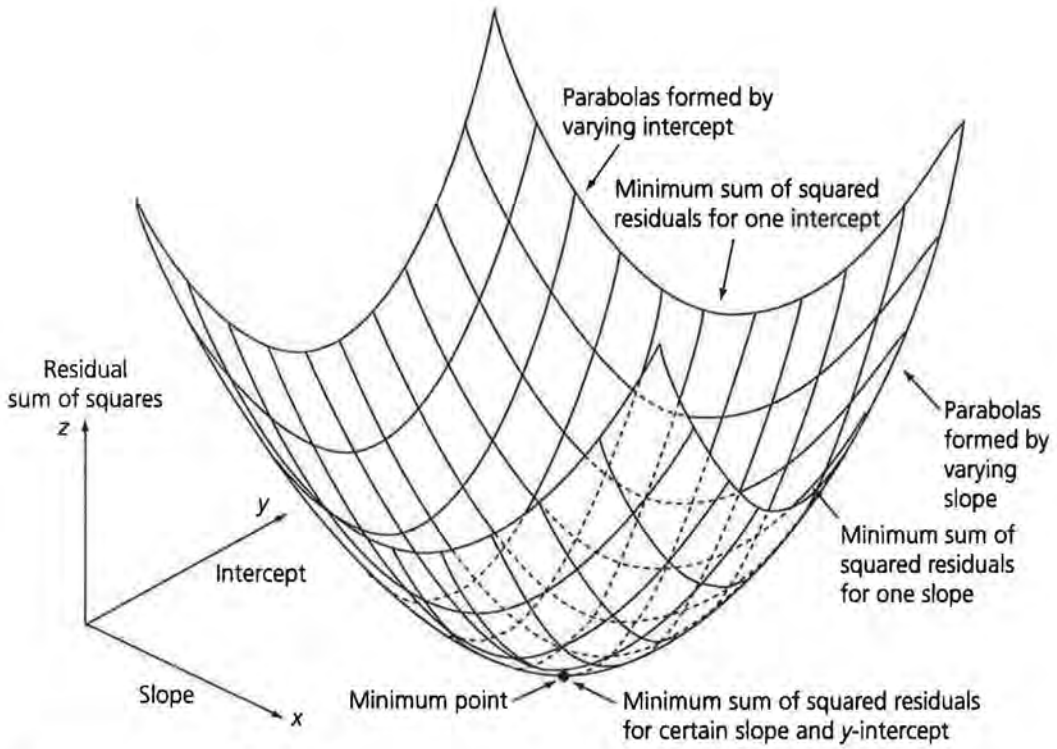
OBJECTIVE

Investigate how the sum of squared residuals depends jointly on the slope and the *y*-intercept.

INVESTIGATE

Recall that the goal was to find the slope and the intercept of the line that best summarizes the data and that can best be used to make predictions. This line was defined to be the line minimizing the sum of squared residuals. To accomplish this goal, the centroid, $(\bar{x}, \bar{y})$ having the value (1418.2, 375.1) was fixed. It was found that a slope of 0.33 minimized the sum of squared residuals. Next, the slope was fixed at 0.33 and the point was varied. In this case, it was found that a *y*-intercept of –93 minimized the sum of squared residuals. The equation of the line was $y = 0.33x - 93$, and it contained the point (1418.2, 375.1). Is this the best line?

The picture that follows represents the families of curves that occur if both the slope and intercept are varied. It is a *paraboloid;* the lowest point of the paraboloid is the point that has the minimum sum of squared residuals.

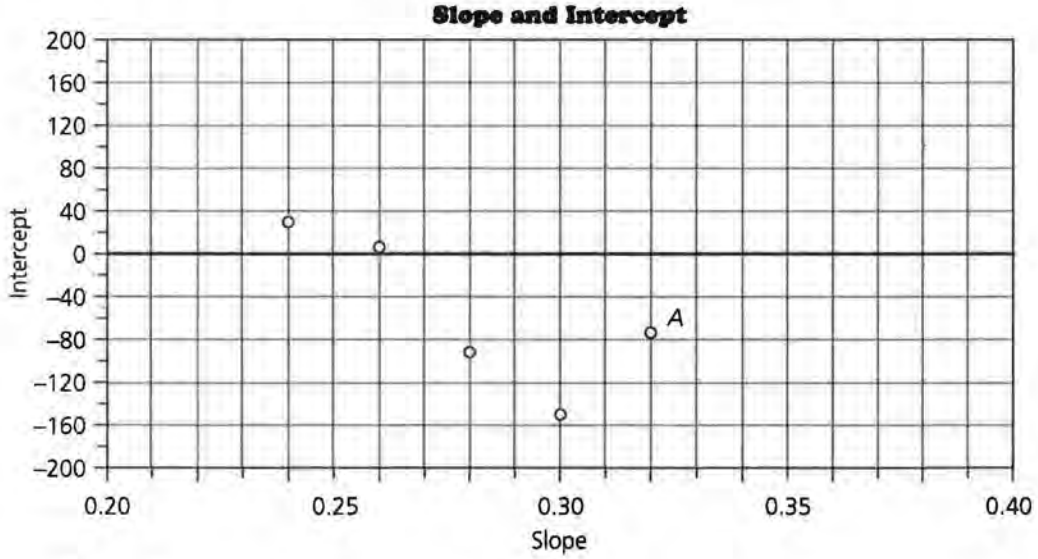


Discussion and Practice

In Lessons 2, 4, and 5, you collected data relating the sum of squared residuals to the slopes and intercepts of several lines. A sample of these data is listed at the right.

Slope	Intercept	Σ(Residuals)^2
0.30	−150	401,025
0.28	−93	355,358
0.26	6	309,713
0.32	−78	300,117
0.24	34	316,058

1. The plot below shows the ordered pairs (slope, intercept) from these data. On *Activity Sheet 5*, plot at least eight more points from the data collected in the last lessons. Include the point from Lesson 5, Problem 7e.



a. What does point *A* represent?

b. What is an equation of the line that generated point *A*?

c. Can you find a point where $\Sigma(\text{residuals})^2$ is less than the $\Sigma(\text{residuals})^2$ for point *A*? If so, what is an equation of the line that generated the point?

d. For each point, write the $\Sigma(\text{residuals})^2$ that you collected for the line with the given slope and *y*-intercept.

e. Find the point that has the least $\Sigma(\text{residuals})^2$. What is an equation of the line that generated this point?

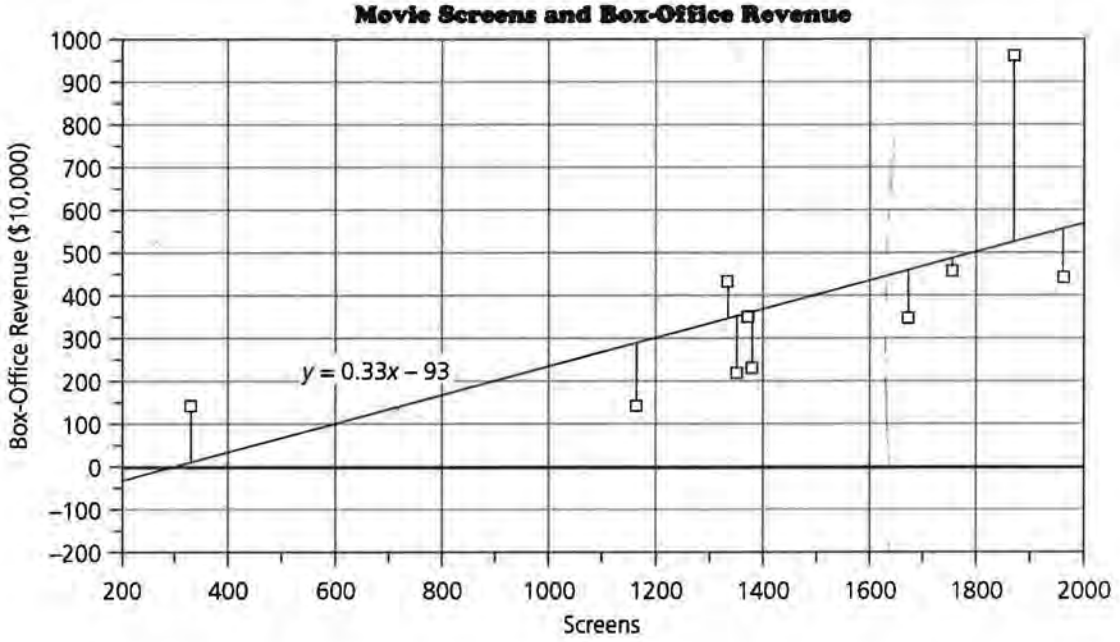
2. Use a clean copy of *Activity Sheet 3* to plot the number of screens and box-office revenue and graph the line whose equation you found in Problem 1e.

a. What does this line represent?

b. How does the line help you to summarize the data?

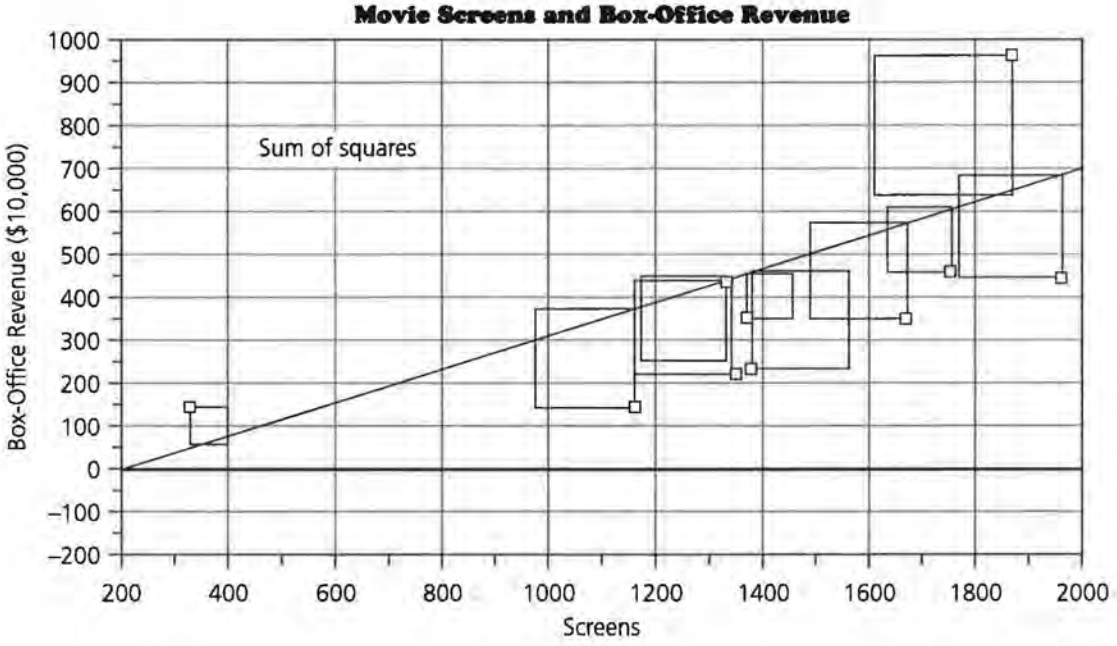
Summary

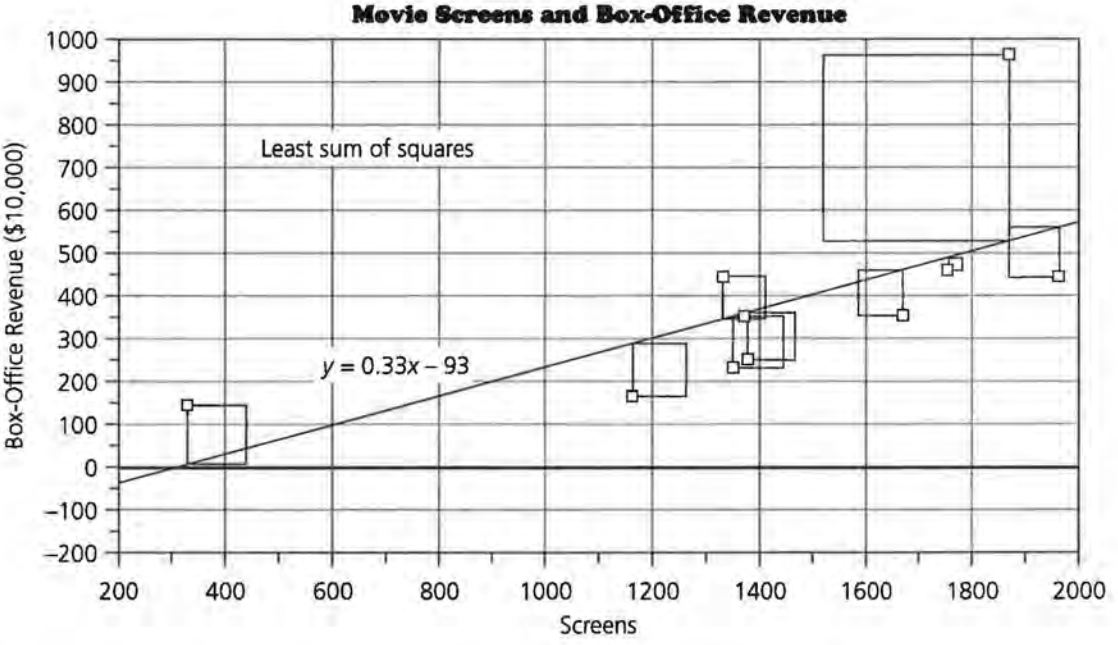
In this lesson, the slope and the intercept were varied simultaneously, and it was determined that the minimum point of the paraboloid occurred at the point with the same slope and intercept found in previous lessons. This assures us that we have found the line which minimizes the sum of the squared residuals. Statisticians refer to this line as the ***least-squares line.*** The least-squares line is the line that minimizes the sum of squared residuals, as in the diagram below. This line can be used as a *best* line to summarize data that appear to be linear.



Practice and Applications

3. Consider the diagrams below. Explain the squares and how they relate to finding a least-squares regression line.





4. Find the least-squares regression line for the BMX dirt-bike data from Lesson 1.

LESSON 7

Quadratic Functions and Their Graphs

How can you find the minimum point of the graph of a quadratic function algebraically?

OBJECTIVES

Find and interpret the x-intercepts of a quadratic equation.

Find a formula to determine the coordinates of the vertex of a parabola.

The quadratic function created by summing the squares of the residuals was used to determine the minimum sum residuals because the graph would always have a minimum point. The goal of previous work was to find the bottom or minimum point of the curve both graphically and numerically.

INVESTIGATE

In this lesson and the next, algebra will be used to find the minimum point. Every straight line can be described by many equivalent equations, and different forms of the equation reveal different characteristics of the line. This lesson investigates the corresponding issues for quadratic functions.

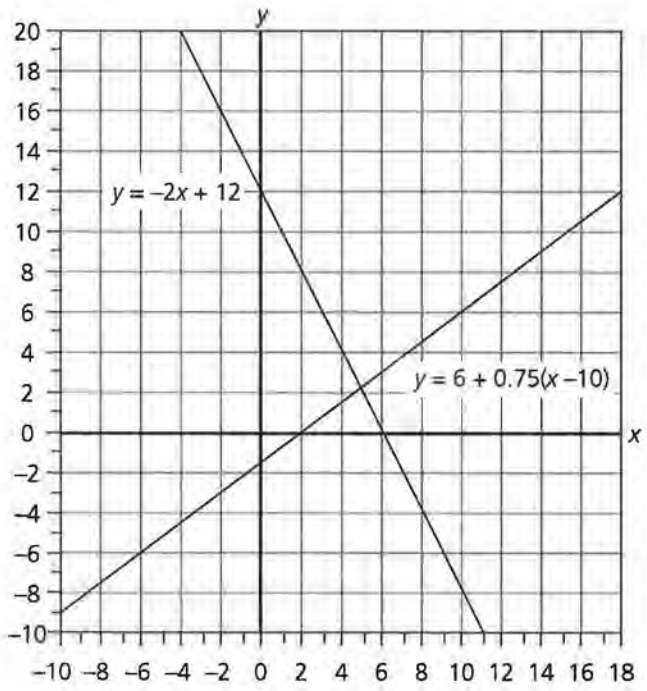
Discussion and Practice

Linear Functions

1. Consider the equation $y = 2x - 8$. What is the significance of the 2 and the –8? How do they relate to the graph?
2. Consider the equation $y = 2(x - 4)$. Graph the line. What is the significance of the point (4, 0)? How does the point relate to the equation and to the graph?

The x-coordinate of the point at which the graph of an equation crosses the x-axis is the x-intercept of the equation. This point is called a *zero* of the equation, since the y-value of the equation is zero at that point.

3. Study the graph and the equation for each line in the plot below.



a. What are the slope and the x-intercept for each equation?

b. Use the graph to explain why the x-intercept is called a *zero*.

Quadratic Functions

The zeros of a quadratic function behave the same as the zeros of a linear function. They make the y-value of the function zero.

4. The zeros of a quadratic can be found in several ways. Consider the equation $y = x^2 - 3x - 10$.

a. Use a spreadsheet or calculator to create a table to help you find the value(s) of x that will make $y = 0$.

b. Graph the equation. Where does the graph cross the horizontal axis?

c. What is the solution to the equation $0 = x^2 - 3x - 10$? How does this solution relate to the graph?

5. Graph the equation $y = (x - 2)(x + 3)$.

a. Describe the graph.

b. What are the zeros of the equation?

6. Graph each equation. Describe the points at which each graph intersects the horizontal axis.

a. $y = x^2 - 4$ **b.** $y = -x^2 + x + 2$

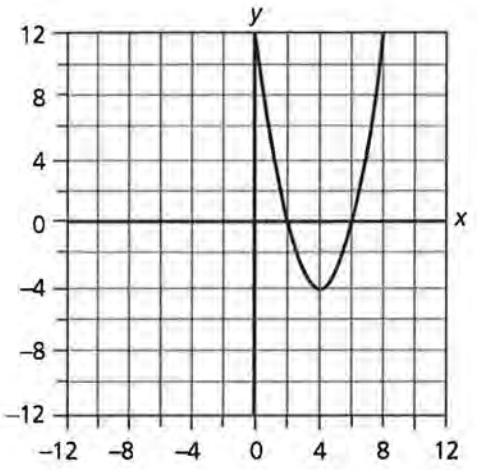
c. $y = x^2 + 2x - 24$ **d.** $y = x^2 - 3x - 5$

e. $y = 2x^2 - x - 2$ **f.** $y = -x^2 - 2$

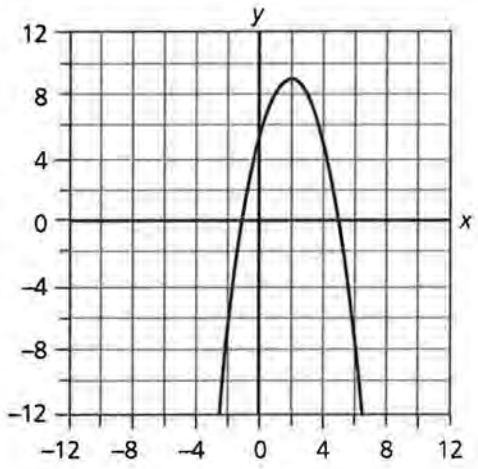
7. The graph of a function may have minimum points, maximum points, both minimum and maximum points, or neither minimum nor maximum points. Estimate what you think might be the x-coordinate of a minimum or maximum point for each graph in Problem 6. Explain why you selected those points and how those points are related to the x-intercepts.

8. Describe each of the following graphs in terms of its x-intercepts, symmetry, and minimum or maximum point.

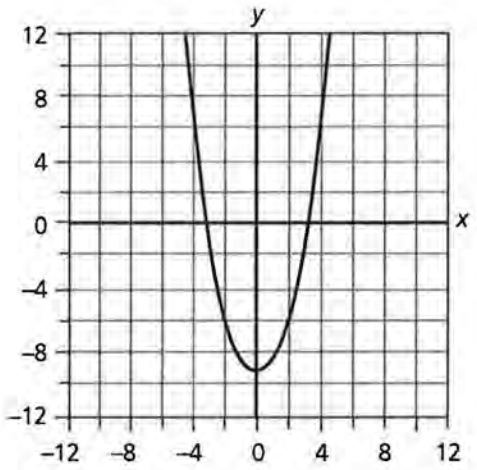
a.



b.



c.



d. How does knowing the x-intercepts, or zeros, of a quadratic equation, along with an understanding of symmetry, help you to find the x-coordinate of a maximum or minimum point?

9. How can you find the minimum point of a quadratic function? Use your technique to find the minimum point for $y = x^2 - 10x + 16$.

10. A parabola is the graph of a quadratic equation of the form $y = ax^2 + bx + c$ where a, b, and c represent constants in the equation and $a \neq 0$. The table below is also on *Activity Sheet 6*.

a. Give the value of a, b, and c; the x-intercepts; and the maximum or minimum point for each equation.

Equation	*a*	*b*	*c*	*x*-Intercepts	Minimum/ Maximum
$y = -2x^2$					
$y = 4x^2$					
$y = x^2 - 10x + 16$					
$y = x^2 - 10x - 11$					
$y = 3x^2 + 13x + 4$					
$y = x^2 - x - 2$					
$y = 8x^2 - 18x + 7$					
$y = x^2 - x + 2$					
$y = 2x^2 - x - 1$					
$y = x^2 - 2x + 1$					
$y = 3x^2 - 2x - 1$					
$y = x^2 - 4$					
$y = 9 - x^2$					

b. Find a pattern or relationship between the x-coordinate of the minimum or maximum point and the coefficients in the equation; that is, tell how the x-coordinate of the minimum or maximum point depends on a, b, and/or c in the equation. Test your conjecture with these two examples: $y = x^2 - 2x - 24$ and $y = 2x^2 - 3x - 5$.

c. What happens in the formula in your conjecture above if $a = 0$?

d. Does the value of c help you find the x-coordinate of the minimum or maximum point?

If the equation is written in the form $y = ax^2 + bx + c$, the x-coordinate of the minimum or maximum point can be found by using the formula $x = \frac{-b}{2a}$. Will that rule always work? Try it with the equations in the table above.

11. A researcher studied the operating cost and speed for commercial jet planes. As a result, the equation $C = 0.2S^2 - 155.9S + 31{,}212$ was produced as a model for C, the cost in dollars per hour of operating an airplane in terms of the airborne speed, S, in miles per hour.

a. Graph the equation. What is the minimum point and what does it mean?

b. Is it realistic to think that the operating costs will be greater for lower speeds?

c. Find the intercepts and determine if they make sense in terms of the situation.

d. Do you think the formula will apply to the Piper Cub, which has an operating cost of \$45 per hour and an airborne speed of 100 miles per hour? Explain why or why not.

12. Consider the equation $y = ax^2 + bx + c$.

a. Find y when $x = \frac{-b}{2a}$.

b. Explain what the value of y represents.

c. What happens when $a = 0$?

Summary

A parabola is any equation of the form $y = ax^2 + bx + c$, $a \neq 0$. The graph of a parabola will be a U-shaped curve. The vertex of the parabola is where the minimum or maximum occurs. If the graph crosses the x-axis, the x-coordinate of the vertex can be found by taking the average of the x-intercepts. A formula can also be used to find the x-coordinate of the vertex. If $a > 0$, the curve opens up, and the minimum point is at $x = \frac{-b}{2a}$.

If $a < 0$, the curve opens down, and the maximum point is at $x = \frac{-b}{2a}$. To find the maximum or minimum y-value, evaluate the expression $y = \frac{-b^2 + 4ac}{4a}$.

LESSON 8

The Least-Squares Line

Can an equation be found that will minimize the sum of the squared residuals?

How do you find the least-squares line for a set of data?

OBJECTIVE

Understand the mathematics behind the least-squares line.

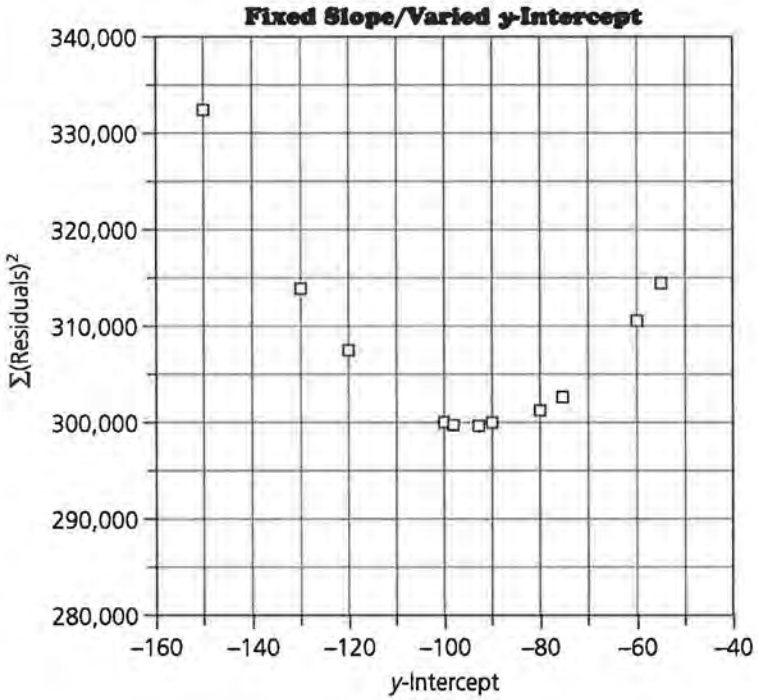
Return to the original problem of trying to find an equation to minimize the sum of squared residuals. Remember that a residual is the difference between the observed y-value and the predicted y-value for a given x-value. The less the sum of squared residuals, the less the root mean squared error will be in the predictions. Finding an equation that will give this minimum requires finding the slope of that equation and a point through which the graph of the equation passes.

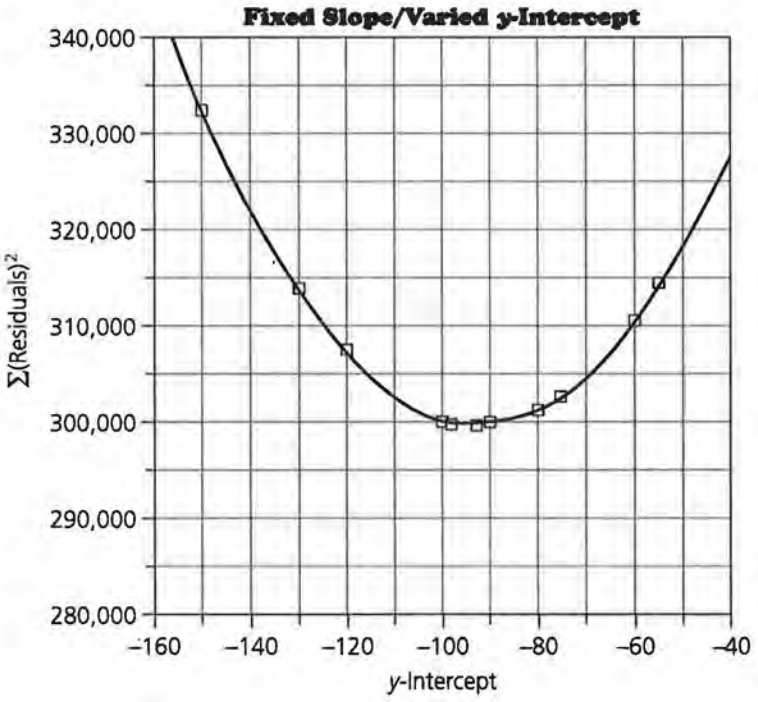
INVESTIGATE

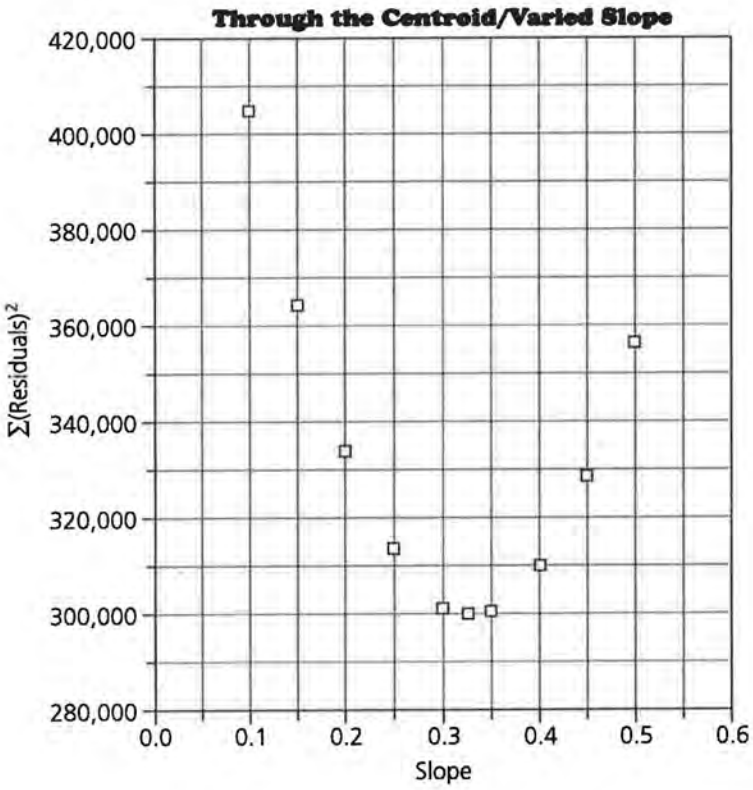
Earlier investigations that explored slopes and intercepts to find the line that gives the minimum sum of squared residuals resulted in a parabolic function. The minimum value of the sum of squared residuals occurs at the vertex of the parabola. In this lesson, you will use the mathematics you just studied to find an equation for the least-squares line for the residuals.

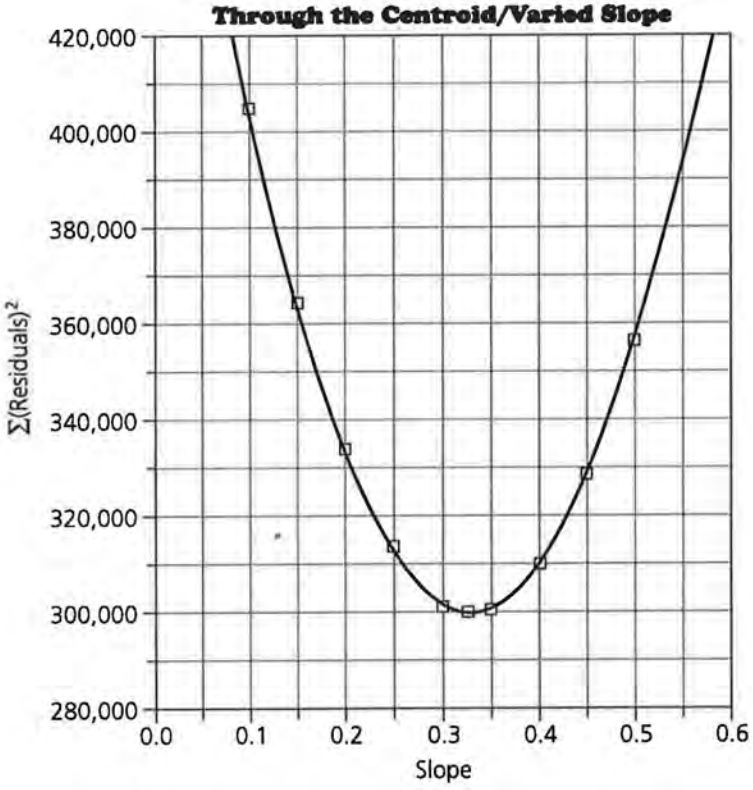
Discussion and Practice

Study the four graphs that follow.






Through the Centroid/Varied Slope
420,000
400,000
380,000
360,000
340,000
320,000
300,000
280,000
Σ(Residuals)²
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Slope


Through the Centroid/Varied Slope
420,000
400,000
380,000
360,000
340,000
320,000
300,000
280,000
Σ(Residuals)²
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Slope

The goal is to find the equation of a line that has the least sum of squared residuals. In order to do this, you must have a point and the slope of that line. The point that gave the minimum residual was the centroid, $(\bar{x}, \bar{y})$. Finding the slope of the equation is a challenging task. Investigating many different values, plotting the graph, and finding the coordinates of the minimum point led to the slope. A better way is to find a formula for the slope using algebra and the characteristics of a quadratic function.

Option I: Generalizing the (Number of Screens, Box-Office Revenue) Data

Remember the steps you used in earlier lessons. Calculate the difference between the actual revenue for the movies that week and the amount of money to be earned predicted by the equation of the line. Square each difference, or residual, and calculate the total sum for all the movies. Considering one movie at a time, you can find a formula. Call the slope s. Use the actual data for each movie to determine an equation with the slope s as a variable.

There were 1878 screens showing *Wayne's World,* and the actual income was 964 ten thousand dollars, or $9,640,000. Using the averages for the number of screens and the income (1418.2, 375.1) as the base point for the line gives the following equation to predict how much a movie should earn:

$y = 375.1 + s(x - 1418.2)$, where s represents slope.

Thus, to find the square of the residual for *Wayne's World,* (1878, 964), you would use the revenue of 964 ten thousand dollars minus the predicted income for the 1,878 screens on which *Wayne's World* was shown or:

$$\begin{aligned}\text{residual squared} &= (\text{observed} - \text{predicted})^2 \\ &= \{y - [375.1 + s(x - 1418.2)]\}^2.\end{aligned}$$

Using the information for *Wayne's World,* substitute 1878 for x and 964 for y, and the expression becomes

$\{964 - [375.1 + s(1878 - 1418.2)]\}^2$.

This expression can be simplified:

$$\begin{aligned}&\{964 - [375.1 + s(1878 - 1418.2)]\}^2 \\ &\quad = \{964 - [375.1 + 459.8s)\}^2 \\ &\quad = (964 - 375.1 - 459.8s)^2 \\ &\quad = (588.9 - 459.8s)^2\end{aligned}$$

Squaring the binomial yields:

$$
\begin{aligned}
&(588.9 - 459.8s)^2 \\
&= (588.9 - 459.8s)(588.9 - 459.8s) \\
&= 588.9(588.9 - 459.8s) - 459.8s(588.9 - 459.8s) \\
&= 588.9 \cdot 588.9 - 588.9 \cdot 459.8s - 459.8s \cdot 588.9 + \\
&\quad 459.8s \cdot 459.8s \\
&= 346{,}803.21 - 541{,}552.44s + 211{,}416.04s^2
\end{aligned}
$$

1. What does s represent?
2. Consider the graph of $y = 346{,}803.21 - 541{,}552.44s + 211{,}416.04s^2$.
   a. How do you know the graph is a parabola?
   b. Find the minimum point and indicate what that point represents.
   c. Find the squared residual when s is 2.
3. To help you investigate the sum of squared residuals, find the corresponding expression for each of the other movies by completing a table like the following, or use *Activity Sheet 7*. Be careful with the quantities and squaring terms.

Movie	Screens	Income ($)	Predicted Income ($) $s(x - 1418.2) + 375.1$	Squared Residual	Quadratic-Error Expression
Wayne's World	1878	964	$s(1878 - 1418.2) + 375.1$	$[964 - (459.8s + 375.1)]^2$	$346{,}803.21 - 541{,}552.44s + 211{,}416.04s^2$
Memoirs of an Invisible Man	1753	460	_____	_____	_____
Stop or My Mom Will Shoot	1963	448	_____	_____	_____
Fried Green Tomatoes	1329	436	_____	_____	_____
Medicine Man	1363	353	_____	_____	_____
The Hand That Rocks the Cradle	1679	352	_____	_____	_____
Final Analysis	1383	230	_____	_____	_____
Beauty and the Beast	1346	212	_____	_____	_____
Mississippi Burning	325	150	_____	_____	_____
The Prince of Tides	1163	146	_____	_____	_____

a. What is the expression for the squared residual for *The Prince of Tides?*

b. Describe the graph of (slope, squared residual) for each movie.

Each calculation gives a formula for squared residuals for an individual movie based on slope. The result was a formula for that movie. To find the sum of the squared residuals for all of the movies, you can add the values of the individual movies.

4. Recall your earlier work combining functions in Lesson 3.

   a. What kind of graph do you have when you add three quadratic expressions?

   b. If you combine the individual formulas for each movie, what kind of function will you have? Describe its graph.

5. Use the expressions in the last column of the table for the movies.

   a. Find the sum of all of the constant terms in that column.

   b. Find the sum of all of the linear terms in that column.

   c. Find the sum of all of the quadratic terms in that column.

6. Use the results of Problem 5 for the following.

   a. Write an equation for the sum of squared differences between the amount of revenue given and the amount predicted for each movie.

   b. Explain what your result represents when $s = 0.2$.

7. When you were estimating lines in the earlier section, you had a slope of 0.45 and found the sum of squared differences to be 327,886. How does this compare to the results you will get using your formula?

8. Graph the equation you found in Problem 6.

   a. A quadratic has the form $y = ax^2 + bx + c$, $a \neq 0$. Find a, b, and c in the equation for Problem 6.

   b. Use the formula from Lesson 5 to find the minimum point. What does this point represent?

9. Compare the graph of the curve generated by the equation with the graph you obtained by selecting different values for s and plotting the resulting sum of squared differences. Describe how you made your comparison.

Summary

You have studied the relationship between the number of screens for a given movie and the amount of money that movie earned in a week. In attempting to find a line that seems to best summarize the relationship, you found the least-squared sum of residuals for the actual money the movie earned and the amount of money predicted by that line. If you assume that the *best* line passes through the average number of movie screens and the average amount of money earned, the equation in terms of the slopes of the line is a quadratic function. This equation has a minimum point for which a slope will give the least sum of squared differences. For this set of data, the point turns out to be (0.33, 299,894). For a slope of 0.33, the least sum of squared differences is 299,894. Thus, for a given prediction, the average root mean squared error would be \$10,000 times the square root of $\frac{299{,}894}{10}$, approximately 173 ten thousand dollars, or \$1,730,000, which represents the average difference between the mean and the amount of money a movie earned.

10. Suppose a movie had been shown on 1700 screens.

- **a.** Use the line you found, the least-squares line, to predict how much income the theater owners would have expected the movie to earn that week.
- **b.** Explain how the root mean squared error affects your prediction.

Extension

11. Find a line that seems to minimize the sum of the absolute residuals.

Option II: The General Formula

Often, a mathematical formula can be found to generalize a situation. A formula may be created in which s, the slope, is the variable. An example follows.

For each individual movie, you calculated the squared residuals, $\{y_i - [375.1 + s(x_i - 1418.2)]\}^2$, between the actual revenue and the revenue predicted by the line. To find the sum of squared residuals, you added the squared residuals for all of the movies.

In symbols, this is what you calculated:

$$\Sigma[y_i - (375.1 + s(x_i - 1418.2))]^2$$
$$= \Sigma[(y_i - 375.1) - s(x_i - 1418.2)]^2$$

Simplify the expression above to

$$= \Sigma[(y_i - 375.1)^2 - 2s(y_i - 375.1)(x_i - 1418.2) + s^2(x_i - 1418.2)^2].$$

Because $\Sigma(a_i + b_i) = \Sigma a_i + \Sigma b_i$, the expression above is

$$= \Sigma(y_i - 375.1)^2 - \Sigma 2s(y_i - 375.1)(x_i - 1418.2) + \Sigma s^2(x_i - 1418.2)^2.$$

The slope s is a common factor and is not the variable for the summation, so $2s$ can be factored out of the second term of the summation expression and s^2 out of the third term. This gives

$$\Sigma(y_i - 375.1)^2 - 2s\Sigma(y_i - 375.1)(x_i - 1418.2) + s^2\Sigma(x_i - 1418.2)^2.$$

Therefore,

$$\Sigma\{y_i - [375.1 + s(x_i - 1418.2)]\}^2 = \Sigma(y_i - 375.1)^2 - 2s\Sigma(y_i - 375.1)(x_i - 1418.2) + s^2\Sigma(x_i - 1418.2)^2.$$

12. Thus, the sum of squared residuals is a quadratic equation in which s, the slope, is the variable.

a. Identify the constant term in this expanded equation.

b. Which term is the linear term in this equation?

The minimum y-value for the quadratic can be found using the formula $y = \frac{-b}{2a}$.

$$y = \frac{-b}{2a} = -\frac{-2\Sigma(y_i - 375.1)(x_i - 1418.2)}{2\Sigma(x_i - 1418.2)^2}$$
$$= \frac{2\Sigma(y_i - 375.1)(x_i - 1418.2)}{2\Sigma(x_i - 1418.2)^2}$$
$$= \frac{\Sigma(y_i - 375.1)(x_i - 1418.2)}{\Sigma(x_i - 1418.2)^2}$$

Thus, this calculation gives the slope s that minimizes this quadratic. This is the slope of the least-squares line. The point (1418.2, 375.1) used in these calculations is the centroid $(\bar{x}, \bar{y})$ of the data.

13. The average number of screens and the average income can be expressed with the general expression using the centroid $(\bar{x}, \bar{y})$.

a. Write the general rule for finding the slope of the line that minimizes the sum of squared residuals between the observed value and the value predicted by the line.

b. Write an equation for the least-squares line.

In your search for a line that minimized the sum of squared residuals, you found the slope numerically by using a spreadsheet or calculator and algebraically using an iterative process. The section above develops an algebraic formula for the slope:

$$s = \frac{\Sigma(y_i - \bar{y})(x_i - \bar{x})}{\Sigma(x_i - \bar{x})^2}$$

Summary

To find the slope of the least-squares regression line, you could find the centroid and repeat either the numerical or the algebraic development. Fortunately, however, graphing calculators and statistical software for computers have the formula developed here programmed into their operating systems. You can access the least-squares line by selecting the **STAT** calculate menu and choosing **LinReg** in the TI calculators and by following the instructions provided when using other types of technology.

Regression is any algorithm used to predict y from a given x. In linear regression, the predicted y-values are a linear function of x. The particular technique investigated in this unit is called ***least-squares linear regression,*** since it gives a linear-regression equation obtained through the least squares approach of minimizing the sum of squared residuals.

Practice and Applications

14. Enter the number of screens and box-office revenue data into your calculator.

a. Use **LinReg** to find the least-squares linear-regression line.

b. What is the slope of the LinReg line and how does it compare to the one you found earlier in this module?

c. Verify that the LinReg line contains the centroid (1418.2, 375.1).

15. Explain how the movie producers might use the least-squares line.

16. Use the BMX dirt-bike data from earlier lessons for each of the following.

    a. Find the slope of the line that minimizes the sum of squared residuals.

    b. Write an equation of the least-squares line.

LESSON 9

Using the Least-Squares Linear-Regression Line

When do you want to fit a line to a set of data?

What advantage is there to having a line to describe the relationship between variables?

INVESTIGATE

In this lesson, you will practice using the least-squares regression line that was developed in earlier lessons.

Discussion and Practice

The data on page 60 are aircraft-operating statistics from the Air Transport Association of America.

OBJECTIVE

Find and interpret the least-squares linear-regression line.

Aircraft-Operating Statistics

Aircraft	Number of Seats	Speed Airborne (mi/hr)	Flight Length (mi)*	Fuel Consumption (gal/hr)	Operating Cost ($/hr)
B747-100	405	519	3149	3529	6132
L-1011-100/200	296	498	1631	2215	3885
DC-10-10	288	484	1410	2174	4236
A300 B4	258	460	1221	1482	3526
A310-300	240	473	1512	1574	3484
B767-300	230	478	1668	1503	3334
B767-200	193	475	1736	1377	2887
B757-200	188	449	984	985	2301
B727-200	148	427	688	1249	2247
MD-80	142	416	667	882	1861
B737-300	131	413	605	732	1826
DC-9-50	122	378	685	848	1830
B727-100	115	422	626	1104	2031
B737-100/200	112	388	440	806	1772
F-100	103	360	384	631	1456
DC-9-30	102	377	421	804	1778
DC-9-10	78	376	394	764	1588

* Distance a plane can travel on a full tank of fuel Source: Air Transport Association of America

1. For a B767-300, the trip from Milwaukee, Wisconsin, to Washington, D.C., takes about 1 hour and 45 minutes.

   a. How much are the operating costs to fly to Washington? What might be included in the cost per hour?

   b. If fuel costs about $2.10 per gallon, how much will the fuel cost for the trip?

   c. Approximately how far is it between Washington, D.C., and Milwaukee, Wisconsin? Explain how you arrived at your answer.

2. Plot the ordered pairs (f, c), representing the cost of operating the plane as a function of fuel used in gallons per hour.
   a. Find an equation of the least-squares linear-regression line and graph it on your plot. What is the slope and what does it tell you about the relationship between fuel and cost?
   b. Find the root mean squared error.
   c. Use your line to predict how much it would cost to operate a plane if the plane used 1000 gallons of fuel per hour. How will the root mean squared error affect your prediction?
   d. Find the number of gallons of fuel that would give a predicted cost of $3000 to operate the plane.
   e. Verify that your least-squares line contains the centroid.
3. Plot the ordered pairs (s, c), representing the cost per hour as a function of the number of seats.
   a. Find the least-squares regression line and graph it on your plot. What is its slope and what does it tell you about the relationship between the number of seats and the operating cost per hour?
   b. How well does the line seem to describe the relationship?
   c. Calculate the cost per hour per seat for each plane. What does this tell you?

Summary

In this lesson, the least-squares regression line was used as a tool to describe the relationship between variables.

Practice and Applications

The following data are from the records of the yearly passing leaders in the National Conference of the National Football League.

Passing Leaders, National Conference of NFL, 1960–1995

Passing Leaders	Attempts	Completions	Yards Earned	Touch-downs	Year
Milt Plum (CL)	250	151	2297	21	1960
Milt Plum (CL)	302	177	2416	18	1961
Bart Starr (GB)	285	178	2438	12	1962
YA Tittle (NYG)	367	221	3145	36	1963
Bart Starr (GB)	272	163	2144	15	1964
Rudy Bukich (CH)	312	176	2641	20	1965
Bart Starr (GB)	251	156	2257	14	1966
Sonny Jurgenson (WA)	508	288	3747	31	1967
Earl Morrall (BA)	317	182	2909	26	1968
Sonny Jurgenson (WA)	442	274	3102	22	1969
John Brodie (SF)	378	223	2941	24	1970
Roger Staubach (DA)	211	126	1882	15	1971
Norm Snead (NYG)	325	196	2307	17	1972
Roger Staubach (DA)	286	179	2428	23	1973
Sonny Jurgenson (WA)	167	107	1185	11	1974
Fran Tarkington (MN)	425	273	2294	25	1975
James Harris (LA)	158	91	1460	8	1976
Roger Staubach (DA)	361	210	2620	18	1977
Roger Staubach (DA)	413	231	3190	25	1978
Roger Staubach (DA)	461	267	3586	27	1979
Ron Jaworski (PH)	451	257	3529	27	1080
Joe Montana (SF)	488	311	3565	19	1981
Joe Thiesmann (WA)	252	161	2033	13	1982
Steve Bartkowski (AT)	423	274	3167	22	1983
Joe Montana (SF)	432	279	3630	28	1984
Joe Montana (SF)	494	303	3653	27	1985
Tommy Kramer (MN)	372	208	3000	24	1986
Joe Montana (SF)	398	266	3054	31	1987
Wade Wilson (MN)	332	204	2746	15	1988
Joe Montana (SF)	386	271	3521	26	1989
Phil Simms (NYG)	311	184	2284	15	1990
Steve Young (SF)	279	180	2517	17	1991
Steve Young (SF)	402	268	3465	25	1992
Steve Young (SF)	462	314	4023	29	1993
Steve Young (SF)	461	324	3969	35	1994
Bret Favre (GB)	570	359	4413	38	1995

Source: *World Almanac and Book of Facts,* 1997

4. Use a spreadsheet or your calculator for the following. Explain how you made your choice in each case.

   a. Which player has the greatest number of yards earned per pass?

   b. Which player was the best in terms of the number of touchdowns he made by passing?

5. Use the data on page 62 to answer the following.

   a. Is it true that the more passes the pass leaders attempt, the more they will complete? What do you think the relationship will be between the number of attempts and the number of completions? Graph the data for (attempts, completions).

   b. Find the least-squares linear-regression line for (attempts, completions) and write a paragraph describing how the data, the equation, and the graph are related. Include in your paragraph a description of the slope and the x- and y-intercepts. Then indicate whether either intercept makes sense in terms of the data.

   c. Investigate the relationship between the number of yards earned and the number of completions. If the relationship appears linear, find the least-squares line and explain how it applies to the data.

6. Find data on the number of passes attempted and the number of completions made by at least 10 college quarterbacks. Compare these results to those given here and use all of your information to answer this question: *Is it true that the more passes attempted the more passes completed?*

7. The table below shows per-capita income (in dollars) in the United States over the period 1971 through 1991. It also shows the suggested retail price for a basic Ford Mustang for those years.

Year	Selected Per-Capita Income ($)	Cost of Ford Mustang ($)
1971	4,302	3,783
1973	5,184	3,723
1975	6,053	4,906
1977	7,269	4,814
1979	9,032	5,339
1981	11,021	7,581
1983	12,216	8,466
1985	14,170	8,441
1987	15,655	9,948
1989	17,705	11,145
1991	19,133	11,873

Source: U.S. Bureau of Census, *Survey of Current Business*, April, 1992

Use these data for the following problems.

a. Plot (year, income) and observe the pattern. Fit a regression line that could be used to summarize the relationship between income and year. Interpret the slope of this line, making sure to use the proper units.

b. Plot Ford Mustang prices over the years and observe the pattern. Fit a regression line to these data. Interpret the slope of this line. How well do you think your line fits the data?

c. Calculate the percent of per-capita income required to purchase a Mustang for each year. Plot (year, percent) and observe the pattern. Fit a regression line to these data and interpret the slope.

d. Considering the three plots and their regression lines, which of the three lines do you think best fits its data? Why?

LESSON 10

Correlation

Is there any general way to measure the *strength* of a linear relationship between two variables?

What is the correlation coefficient?

How well can you predict *y* when you *use* the *x*-variable?

How well can you predict *y* when you *do not use* the *x*-variable?

INVESTIGATE

In the previous lessons, you learned how to find the *best* linear relationship between two variables. Is there any general way to measure the *strength* of the relationship? For example, how closely is the amount of sugar in a cup of cereal related to the number of calories? There exists a numerical value created to do just that, that is, to measure the strength of the linear relationship between two variables. This number is called the ***correlation coefficient.*** This lesson investigates how the correlation coefficient is defined, what it tells you about the relationship between two variables, and what it does not tell you. Consider the data on page 66, taken from *Consumer Reports,* November, 1992, on breakfast cereals. The data are given for a serving size of one cup.

OBJECTIVE

Find and interpret the correlation coefficient.

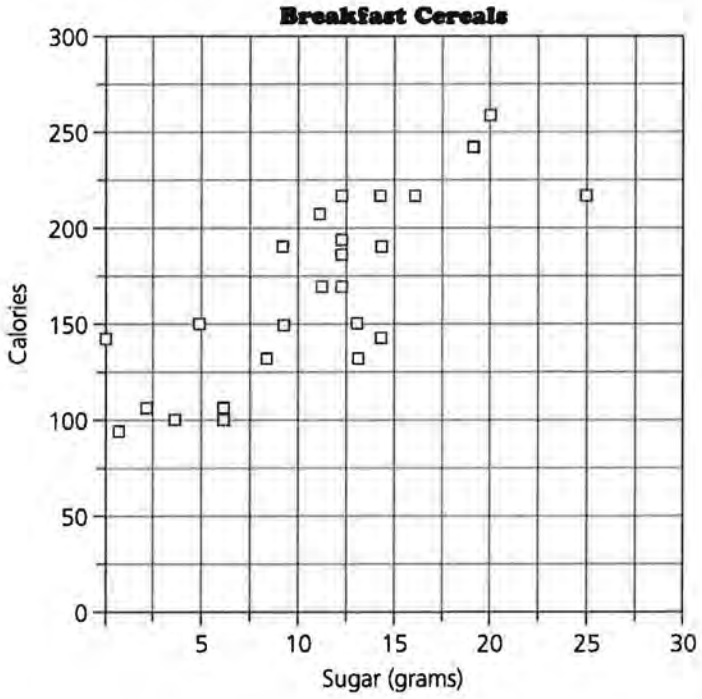
Ready-to-Eat Cereal	Calories	Sodium (mg)	Sugar (g)	Sugar (percent)
Shredded Wheat Spoon Size	140	5	0	0
Common Sense Oat Bran	130	330	8	21
Frosted Mini-Wheats	130	0	8	21
Grape-Nut Flakes	110	160	6	18
Whole Grain Wheat Chex	150	350	5	11
Whole Grain Wheaties	100	200	3	11
Total Raisin Bran	140	190	14	33
Raisin Nut Bran	220	280	16	28
Raisin Squares	180	0	12	21
Oatios with Extra Oat Bran	110	0	2	6
Nutri-Grain Almond Raisin	210	330	11	18
Crispy Wheats 'N' Raisins	130	180	13	35
Life	150	230	9	21
Multi-Grain Cheerios	100	220	6	21
Oatmeal Squares	220	270	12	21
Mueslix Crispy Blend	240	220	19	31
Cheerios	90	230	1	4
Cinnamon Oatmeal Squares	220	250	14	25
Clusters	220	280	25	14
100% Natural Whole Grain with Raisins (Low Fat)	220	30	14	24
Honey Bunches of Oats with Almonds	180	240	9	21
Low-Fat Granola with Raisins	180	90	14	29
Basic 4	170	310	11	12
Just Right with Fruit & Nuts	190	250	12	24
Apple Cinnamon Cheerios	150	240	13	35
Honey Nut Cheerios	150	330	13	35
Oatmeal Raisin Crisp	260	340	20	29
Nut & Honey Crunch	170	300	12	28

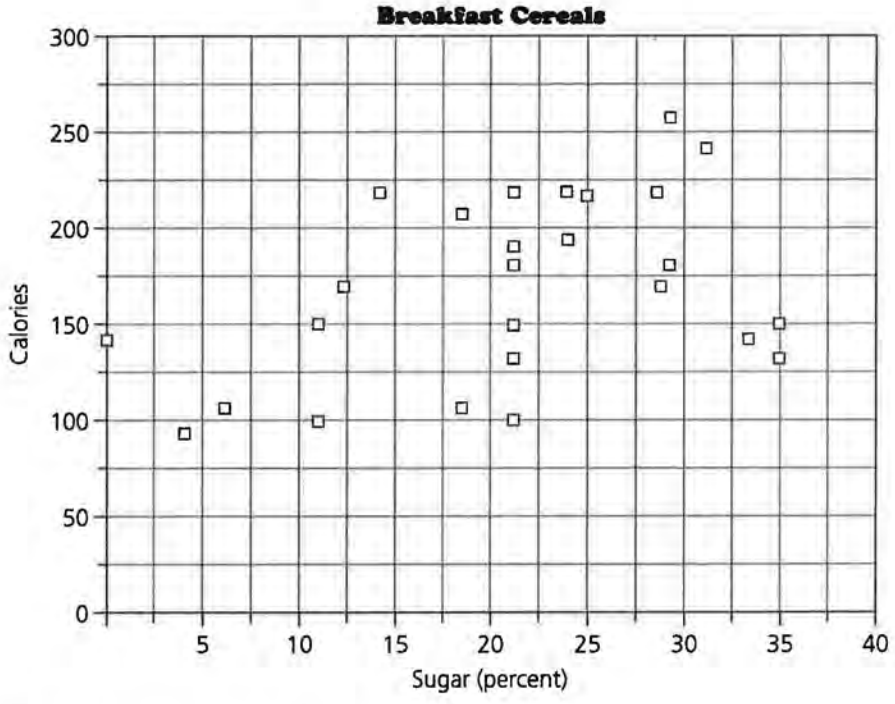
Source: *Consumer Reports,* November, 1992
(Note: The column labeled "Sugar (percent)" is computed based on the weight in grams of one cup of the specific cereal.)

Discussion and Practice

1. Use the data in the table above for the following.
   a. Describe Honey Nut Cheerios in terms of the information in the table.
   b. What is the difference in the number of grams of sugar and the percent of sugar for a cereal?
   c. Estimate the average number of calories in a serving.
   d. Which cereal seems to be the most healthful in terms of the information you have? Explain your choice.

2. The plots below represent different relationships between the amount of sugar and the calories in the cereals.





a. What are the differences in the plots?

b. For which plot does the linear association seem to be stronger? How did you make your decision?

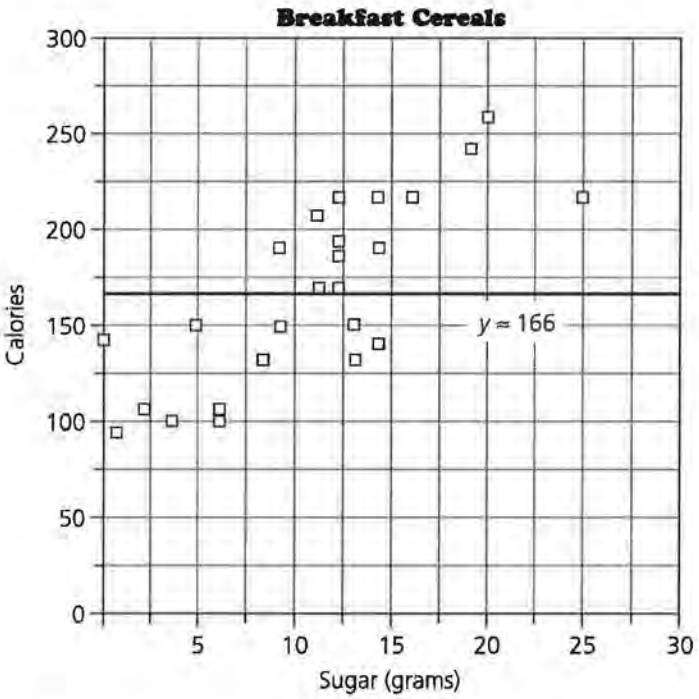
The Correlation Coefficient

A scatter plot gives a good overall impression about the relationship between two variables. But, as the two previous plots show, it is difficult to decide exactly how strong a relationship might be or to precisely compare the strength of association in two different plots. A numerical measure of association would help. The least-squares line summarizes the linear relationship between two variables and can also be used to develop a numerical measure of association. To do so, consider two questions:

1. How well can you predict y when you *use* the x-variable?
2. How well can you predict y when you *do not use* the x-variable? For example, would it be just as accurate to estimate the y-variable based on $\bar{y}$?

Will using x make a difference? If using x does not improve the prediction of y very much, it makes sense to say that x and y are not associated very strongly. If using x greatly improves the prediction of y, then it makes sense to say that x and y are strongly associated. A statistic called the ***correlation coefficient,*** which you may have encountered in different forms in earlier work, is used to turn this idea into a specific numerical measure of association.

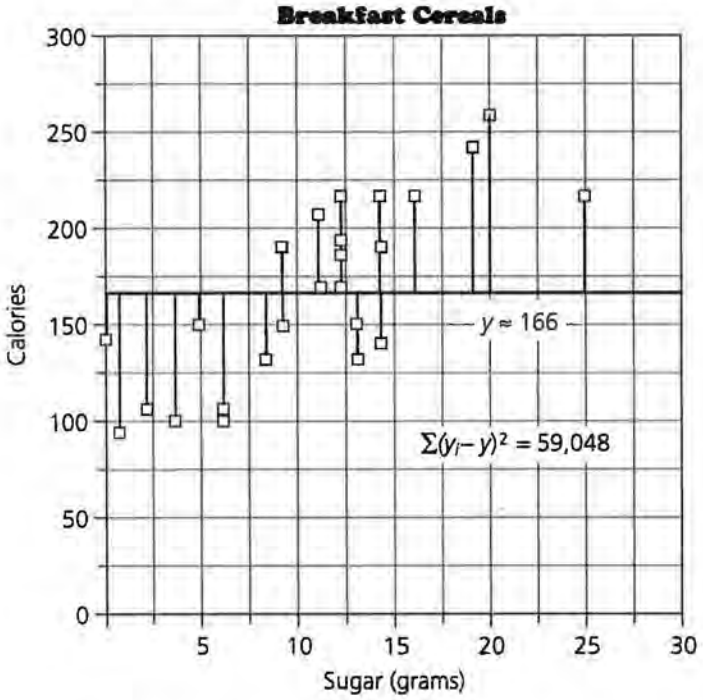
Suppose you want to predict the number of calories in a new breakfast cereal, based on the preceding data. Using nothing other than the number of calories about the cereal—that is, not using the amounts of sodium or sugar—the best prediction is the mean number of calories from the cereals in the data, about 166.429 calories. That is, mean $y = \bar{y} \approx 166$ calories.



How accurate would you expect the prediction to be? What was the actual data point? The observed number of calories for Shredded Wheat Spoon Size is 140 calories, and the residual using the mean to predict would be (140 – 166). Nut & Honey Crunch has 170 calories, so the residual using the mean to predict is (170 – 166). But a measure of total error should involve all the cereals in the data, not just these two. This problem is similar to an earlier problem in this module, when you needed to find an overall measure of error from a line. There you eventually settled on the sum of squared residuals as a reasonable overall measure of error.

It makes sense to use a similar measure of error here. The only difference is that now the prediction for each cereal is mean $y \approx 166$. Thus, for Shredded Wheat Spoon Size, square the residual, giving $(140 - 166)^2$. For Nut & Honey Crunch, the squared residual is $(170 - 166)^2$. The total squared error using the mean calories for each prediction is found by summing these squares over all cereals in the data.

$$\Sigma(\text{observed } y_i - \text{mean } y)^2 = \Sigma(y_i - 166)^2 = 59{,}048$$

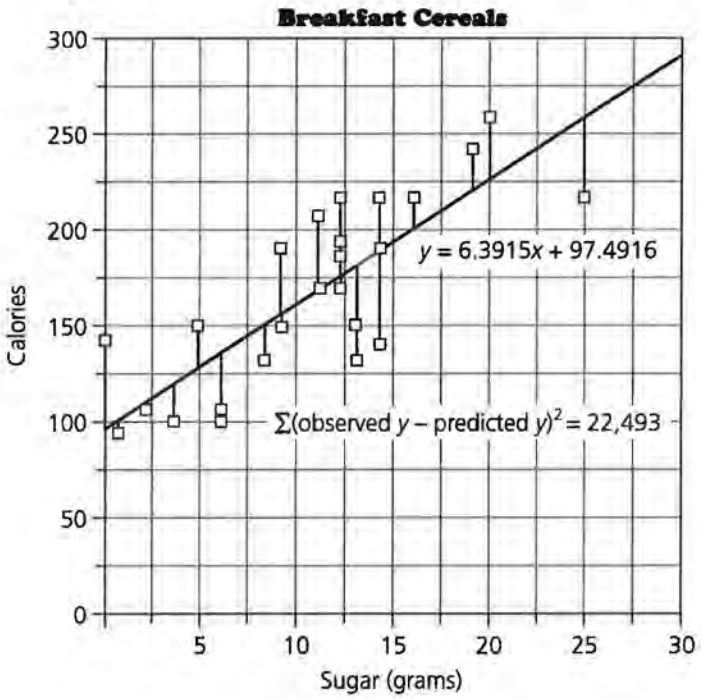


Next, suppose you use grams of sugar (the x-variable) to predict the calories (the y-variable). This is the sort of problem you have investigated throughout this module. You now know that the least-squares line gives the least sum of squared residuals. Furthermore, the total error is the sum of squared residuals from using the least-squares line for prediction, which is

Σ(observed y – predicted y using the least-squares line)^2.

In this example, for grams of sugar (x) and calories (y) the equation of the least-squares line is

$y = 6.3915x + 97.4916$, and
Σ(observed calories – predicted calories using the least-squares line from sugar)^2 = 22,493.



To summarize, a measure of how strongly x is associated with y can be investigated by working with the two measures of total error first derived. To predict y without using x, the square of the error is

(observed y – mean y)^2 = 59,048, or $(y - \bar{y})^2 = 59{,}048$.

To predict y by using x and the least-squares line, the square of the error is

(observed y – predicted y from least-squares line)^2 = 23,493, or $(y - \hat{y})^2 = 23{,}493$.

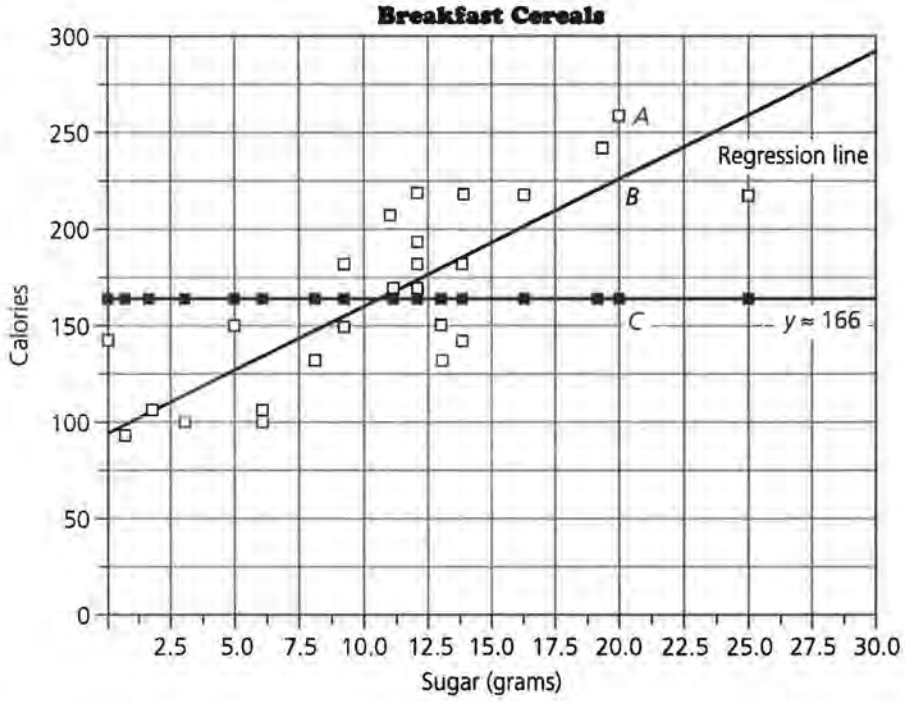
There is one more important fact to note about these two expressions for total error. You know that the least-squares line gives the least sum of squared residuals for any possible line. But prediction using mean $y \approx 166$ is equivalent to using the horizontal line with y-intercept approximately 166 and slope zero, a possible line. Thus, it will always be the case that

Σ(observed y – predicted y using least squares line)^2 $\leq \Sigma$(observed y – mean y)^2, or $\Sigma(y - \hat{y})^2 \leq \Sigma(y - \bar{y})^2$.

That is, (total error in prediction *using* x) $\leq$ (total error in prediction *not using* x).

Finally, these pieces can be put together to give a measure of the strength of association between x and y. The reduction in total error is

(total error in prediction not using x – the total error in prediction using x) or, symbolically, $\Sigma(y - \bar{y})^2 - \Sigma(y - \hat{y})^2$.



Segment AC is the unexplained error strictly using $\bar{y}$ as the predictor of y. Segment BC is the part of the error accounted for, or explained, by using the regression line as the predictor of y and segment AB, the residual, is the error not explained by using the line to predict y. Thus, $AC - AB$ is the error explained by using the line, and $\frac{AC - AB}{AC}$ equals the proportion, or percent, of error explained by using the line to predict y. In other words, the proportion of reduction in total error when using x compared to not using x is

$$\frac{\left(\begin{array}{c}\text{total error in prediction} \\ \text{not using } x\end{array}\right) - \left(\begin{array}{c}\text{total error in prediction} \\ \textit{using } x\end{array}\right)}{\text{total error in prediction not using } x}$$

$$= \frac{\Sigma(\text{observed } y - \text{mean } y)^2 - \Sigma(\text{observed } y - \text{predicted } y \text{ using least-squares line})^2}{\Sigma(\text{observed } y - \text{mean } y)^2}$$

$$= \frac{\Sigma(y - \bar{y})^2 - \Sigma(y - \hat{y})^2}{\Sigma(y - \bar{y})^2} = \frac{59{,}048 - 22{,}493}{59{,}048}$$

$$= 0.619$$

$$= r^2.$$

This value is the coefficient of determination, denoted by r^2. The number r^2 represents the percent of the sum of the squared residuals that can be attributed to a linear relationship.

In the cereal data example, $r^2 \approx 0.62$ for calories and grams of sugar. This means that 62% of the variation in the amount of calories in a cup of cereal can be attributed to a linear relationship between calories and the grams of sugar in a cup of cereal. If you use the grams of sugar, you can predict the calories for the cereal more accurately than if you do not know anything about the sugar. In fact, in the precise mathematical sense described, you can do 62% better.

The square root of r^2, $\pm r$, is called the ***correlation coefficient.*** The correlation coefficient r is a number between 1 and –1 and is a measure of linear association or the way the data points cluster around the least-squares regression line. The square of the correlation coefficient, r^2 expresses the proportion of variability in the y-variable that is explained by a change in the x-variable in the least-squares regression line. In the example above, the correlation coefficient $r = 0.78$. The formulas for finding r and r^2 have been programmed into your calculator, and your calculator can be used to find r quickly and easily for any pair of variables you have entered.

Range of r^2 and r

Notice that if all data points fall on a line, the x-variable predicts the y-variable perfectly, and each residual is zero. Thus,

$$\Sigma(\text{observed } y - \text{predicted } y \text{ using least-squares line})^2 = \Sigma(y - \hat{y})^2 = 0.$$

So in this case,

$$r^2 = \frac{\Sigma(\text{observed} - \text{mean})^2 - \Sigma(\text{observed} - \text{predicted})^2}{\Sigma(\text{observed} - \text{mean})^2}$$

$$= \frac{\Sigma(y - \bar{y})^2 - \Sigma(y - \hat{y})^2}{\Sigma(y - \bar{y})^2}$$

$$= \frac{\Sigma(\text{observed} - \text{mean})^2 - 0}{\Sigma(\text{observed} - \text{mean})^2} = \frac{\Sigma(y - \bar{y})^2 - 0}{\Sigma(y - \bar{y})^2} = 1.$$

And this is the greatest value for r^2, since the numerator is less than or equal to the denominator. So r^2 is always less than or equal to 1.

Now consider the least value that r^2 could have. Refer to the formula for r^2. It shows that

> Σ(observed y – mean y)^2 – Σ(observed y – predicted y using least-squares line)$^2 \geq 0$ or, symbolically, $\Sigma(y - \bar{y})^2 - \Sigma(y - \hat{y})^2 \geq 0$.

The value of r^2 is zero only when the least-squares line for the data is horizontal. This says that the slope of the least-squares line is zero, and the best prediction of y amounts to simply using the mean y regardless of the value of x.

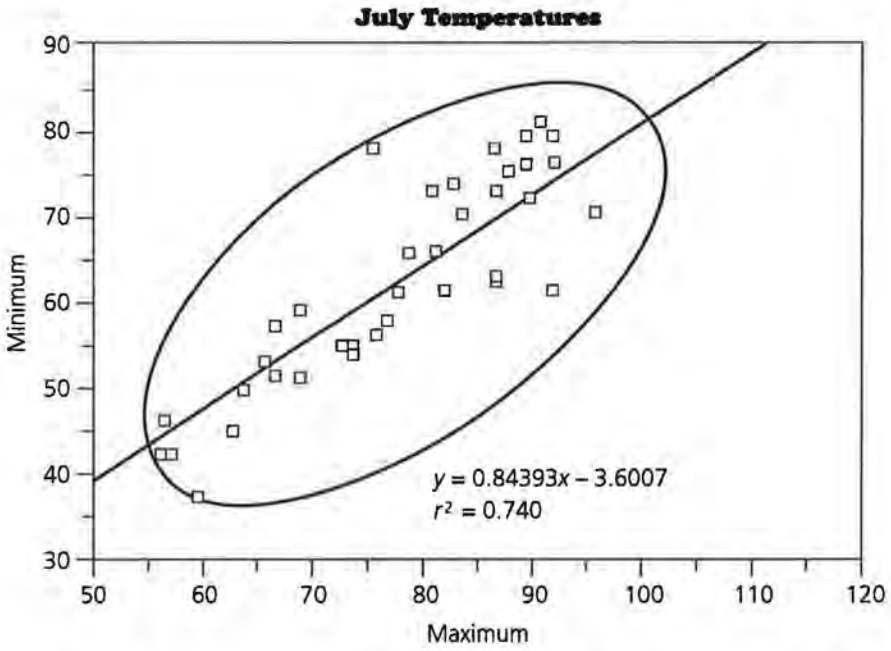
The expression above is the numerator of r^2, and it is positive; the denominator of r^2 is also positive. So, $r^2 \geq 0$, and $0 \leq r^2 \leq 1$.

Summary

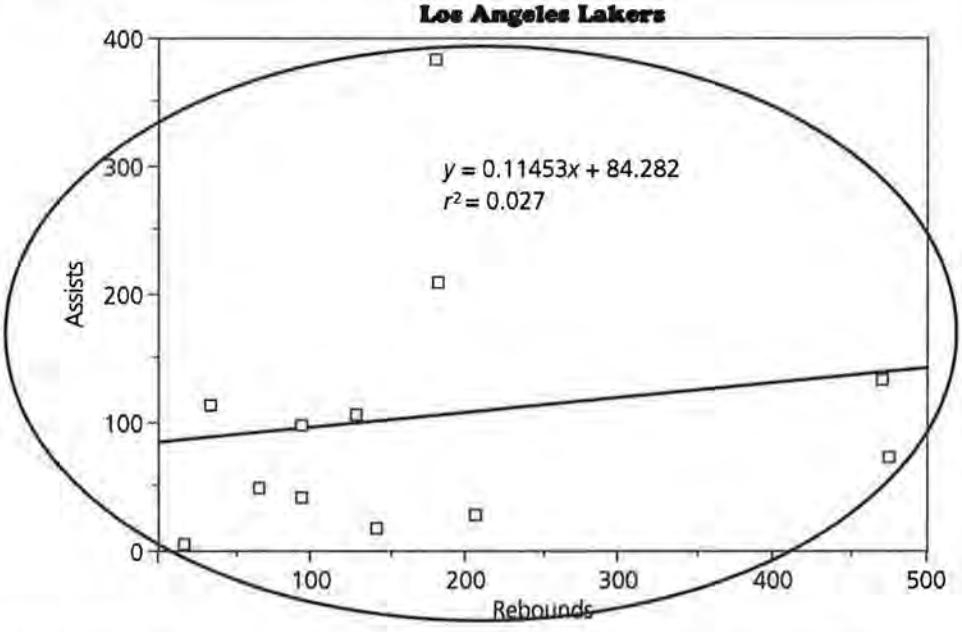
To summarize, if the data points all fall on a line, then $r^2 = 1$. If the slope of the line is positive, $r = 1$; if the slope of the line is negative, $r = -1$. One way to think of r^2 is in terms of a sliding scale from 0 to 1, where 1 is the case in which 100% of the change in y is determined by a change in x and 0 is the case in which 0% of the change in y can be explained by a change in x. Numbers between 1 and 0 indicate some amount of correlation that can be captured using words such as *strong* and *weak;* but determining exactly what amount of correlation is *strong* is quite subjective. For the calories and grams of sugar in the cereal example, an r^2 of 0.61—or correlation, r, of $\pm$ 0.78 (Use the positive value since the data show an increasing relationship.)—would usually be considered strong.

To interpret r^2 and r, it also helps to keep in mind the range of possible values and the types of scatter plots and the extreme situations to which they correspond. The value of r^2 must always satisfy $0 \leq r^2 \leq 1$. It follows that $-1 \leq r \leq 1$, since the square root can be either positive or negative. Determine which sign to use by the dependence shown in the data. The value of r^2 is 1 only when all the data points fall on a line. The value of r^2 is zero only when the least-squares line for the data is horizontal. In a case like this, the slope of the least-squares line is zero, and the best prediction of y amounts to simply using the mean of the y-values regardless of the value of x.

Look at the two plots below to get a better feeling for what r and r^2 indicate about the relationship of the data points to the least-squares regression line. For r close to 1, the points form a narrow elliptical cloud close to the line. For r close to zero, the points form a wider, or "fatter"-appearing, cloud.



In the graph above, note that the points cluster around the line; $r^2 = 0.740$ and $r = 0.86$. There is a high degree of association between the maximum and minimum July temperatures for cities. Knowing the maximum July temperature does help predict the minimum July temperature.



Note that the points do not cluster close to the line; $r^2 = 0.027$ and $r = 0.16$. There is little association between rebounds and assists for the Lakers basketball team. Knowing the number of rebounds for a player does not help to predict the number of assists for that player.

Units of r^2

The units that go along with the r^2 formula involve two different summation expressions. For Σ(observed y – mean y)^2, each difference has a unit of calories, so this entire summation has a unit of calories squared. Similarly, for Σ(observed y – predicted y using least-squares line)^2, each difference has a unit of calories, so the entire summation has a unit of calories squared.

Putting these together and showing the units in the formula for r^2 gives the following expression:

$$\frac{\Sigma(\text{observed} - \text{mean})^2\text{calories}^2 - \Sigma(\text{observed} - \text{predicted})^2\text{calories}^2}{\Sigma(\text{observed} - \text{mean})^2\text{calories}^2}$$

Both numerator and denominator have units calories squared. In the ratio the units reduce, and the ratio does not have any units. Thus, the number r^2 is without a unit.

r^2 Measures the Strength of a *Linear* Relationship

You have learned that the least-squares line is the line that minimizes the sum of squared residuals among all possible lines. The number r^2 is defined for this line. Because the r^2 formula uses the sum of squared residuals from this line, r^2 is a measure of how strongly the data points follow a *linear* relationship.

Consider the relationship between grade-point average and the number of hours students study. Grade-point averages may vary from 0.0 to 4.0 on a four-point scale. Suppose the correlation, r, is around 0.7. This means that the correlation is fairly strong, and the points lie in a cloud close to the least-squares line. Calculating r^2 gives $0.7^2 = 0.49$, so 49% of the variability in grade-point averages can be explained by how much students study and the least-squares line. The remaining 51%, however, is due to other factors, such as difficulty of classes, amount and quality of homework, and differences among individual students.

The relationship between grade-point averages and the number of hours you sleep is a different story. Suppose the correlation is 0.2. There is little association; the points lie in a wide ellipse, and the least-squares line does not tightly summarize the data. Then $r^2 = 0.04$, which indicates that only 4% of the variability in grades would be attributed to the number of hours students sleep. More than 95% of the variation in grade-point averages is due to other factors.

Because r is calculated using the mean y-value and the residuals from the least-squares regression line, the numerical value of r can be misleading. It is important to remember that r measures the strength of a linear relationship. It does not measure the existence of any other pattern in the data, and an r near zero does not mean another pattern might not exist. The value r is very sensitive to outliers because of the way it is calculated. Outliers can make the correlation seem strong when, in fact, little exists. Likewise, outliers can make the correlation seem weak, when, in fact, the relationship is very linear. To be sure you understand what the correlation coefficient indicates about the relationship, look at the plot of the data and check for patterns and outliers. This will help to ensure that the correlation you find makes sense in terms of the data.

3. A positive value of r corresponds to a line with positive slope. A negative value of r corresponds to a line with negative slope.

   a. Sketch a line to illustrate each case.

   b. If r^2 is 0.81, what are possible values for r?

4. Consider the range of values for r and r^2.

   a. Is it possible for r^2 to be greater than 1? Can r ever be greater than 1? Explain.

   b. Can r ever be less than –1? Explain.

   c. If $r^2 = 1$, must all the data points fall on a line? Explain.

5. Suppose $r^2 = 0$.

   a. What can you conclude about the two summations in the numerator of the definition of r^2?

   b. What can you conclude about the least-squares line for these data?

   c. If the slope of the least-squares line is zero, what can you conclude about the value of r^2?

   d. Suppose the (x, y) data points fall on two parallel lines symmetrically distributed. Would you say that these x- and y-variables are strongly associated? What is r^2?

6. Draw an example of a scatter plot for each situation.

   a. The correlation is close to 1.

   b. The correlation is close to zero.

7. Describe the correlation you would expect to get from looking at a plot of each of the following.

   a. The amount of rain and the percent of sun for a set of cities.

   b. The amount of rain and percent of cloudy days for a set of cities.

   c. The amount of rain and the temperature for a set of cities.

8. For each of the following data sets, determine the value of r and discuss how it relates to the data.

   a. {(2, 1), (2.3, 1.5), (2.5, 2), (0.3, 1.4), (2.6, 2.3), (1.8, 1), (1.9, 0.2), (0.7, 0.8), (1, 2.6), (0.2, 2.1)}

   b. {(1, 2), (1.5, 2.3), (2, 2.5), (1.4, 0.3), (2.3, 2.6), (1, 1.8), (0.2, 1.9), (0.8, 0.7), (2.6, 1), (2.1, 0.2), (30, 45)}

   c. {(–2, 4), (–1, 1), (0, 0), (1, 1), (2, 4)}

   d. {(–3, –14), (–1, –11), (1, –2), (3, 4), (7, 16), (10, 32)}

There are several different kinds of correlation and different procedures for finding correlation between variables, depending on the kind of data with which you are working. The correlation coefficient described above is called ***Pearson's r.*** It is widely used and concentrates on the degree of linearity in the relationship.

Summary

It would be useful to have some measure of association that

1. is free of units (such as grams and calories);
2. does not depend on the scale of measurement (such as grams or milligrams); and
3. would always help you judge the strength of the association between two variables.

Unfortunately there is no such measure.

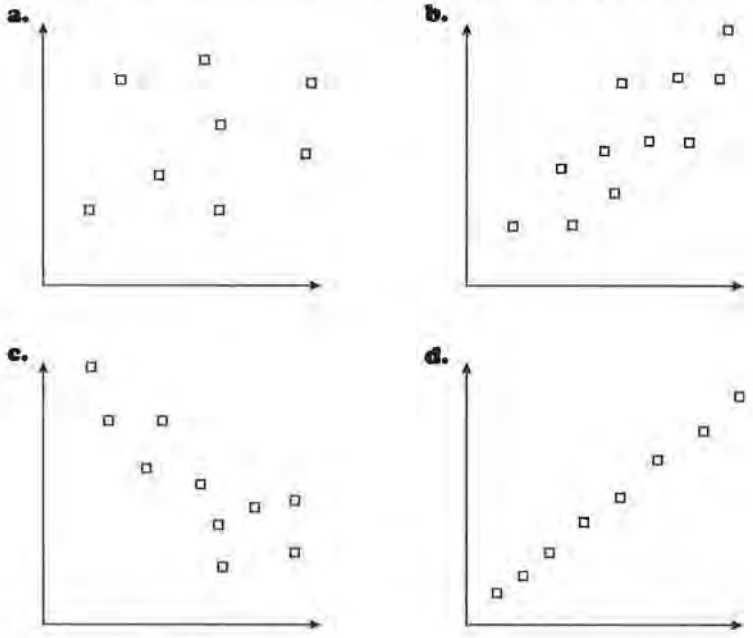
The correlation coefficient is a measure of association that meets the first two of these criteria but not the third. It measures the strength of the ***linear relationship*** between two variables in conjunction with the least-squares line. The linear association between two variables is measured by a number r called the ***correlation coefficient.*** For perfect positive correlation, $r = 1$; and for perfect negative correlation, $r = -1$. A positive correlation indicates that as one variable increases, the other tends to increase also; while a negative correlation indicates that as one increases, the other tends to decrease. If r is close to zero, then it is usually difficult to determine as one variable increases whether the second variable either increases or decreases.

The number r^2 indicates the proportion of variation in y that can be explained by using the least-squares regression line. The closer r^2 is to 1, the more valuable x is for predicting y. There are some important facts to remember about correlation.

- The correlation coefficient measures linear association only, rather than association in general. There may be a clear pattern in a set of data, but if it is not linear, the correlation may be close to zero. An example is the graph of a parabola.
- Correlation is a number without any units attached. Therefore, it does not depend on the units chosen for either variable.
- Many software packages calculate r automatically when they find the coefficients of the regression line. It is important, however, to look at the plot to determine whether the underlying relation is actually linear.

Practice and Applications

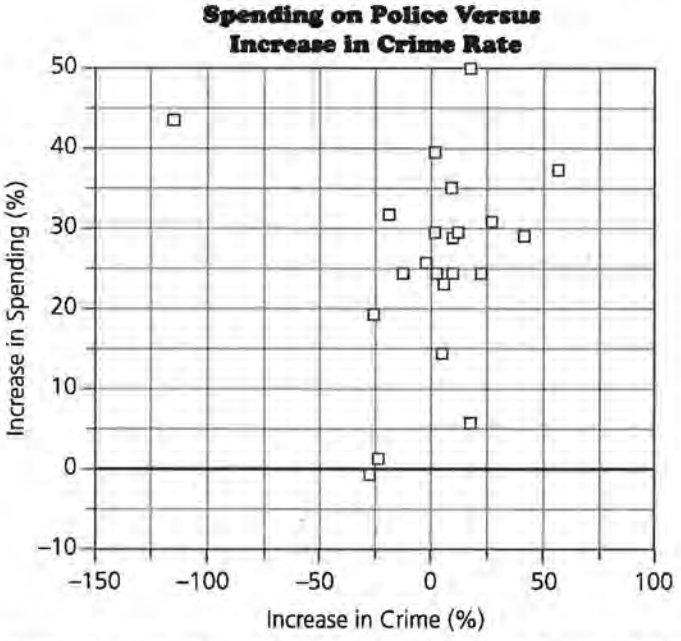
9. Match each correlation r and r^2 with the appropriate graph.



i. $r = -0.8$, $r^2 = 0.64$ **ii.** $r = 0.8$, $r^2 = 0.64$

iii. $r = 0.99$, $r^2 = 0.9801$ **iv.** $r = 0.15$, $r^2 = 0.0225$

10. The following plot shows how, for a set of suburban communities, spending on police increased from 1988 to 1992 and how the crime rate changed during that same time.



Source: Wisconsin Office of Justice Assistance, Wisconsin Taxpayers Alliance

a. Give the coordinates of a point where both the increase in crime and the increase in spending are high.

b. Give the coordinates of a point where the increase in crime is low and the increase in spending is high.

c. r^2 for (% increase in crime, % increase in spending) is 0.003. What does this mean? What is r?

d. If you knew the increase in crime for a given community was 20% between 1988 and 1992, how well do you think you could predict the change in spending based on the plot?

11. Comment on this statement: *If there were no correlation, the best way to predict* y *from an* x *is just to use the mean or average* y *without any regard for the* x *with which it might be associated.*

12. Refer to the plot of the calories and grams of sugar in the beginning of the lesson.

a. Enter the data (grams of sugar, calories) into your calculator. Find the least-squares regression line and the correlation coefficient.

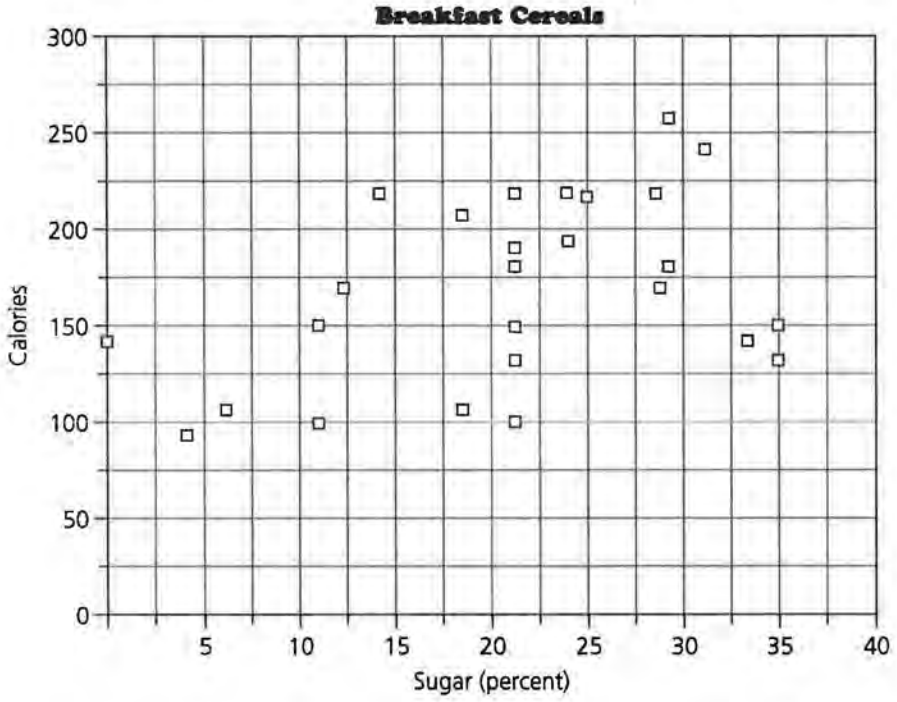
b. How well do you think knowing something about the number of grams of sugar will help you predict the number of calories? What did you consider in arriving at your answer?

13. Reverse your axes for (calories, grams of sugar). Make the plot.

a. Calculate the correlation coefficient. How does it compare to the correlation coefficient for (grams of sugar, calories)?

b. What is the least-squares line? How does it compare to the line for (grams of sugar, calories)? Try to explain any observations.

14. The plot below is of calories and percent sugar from the beginning of this lesson. An equation of the least-squares line for the number of calories, c, as a function of the percent sugar, s, is $c = 128 + 1.80s$. Use the first plot on *Activity Sheet 8* for this problem.



a. What is the slope of the equation? What does it tell you about the calories and the percent of sugar?

b. Suppose you know the equation for the least-squares line, as above. What can you anticipate about the value of r?

c. Sketch the least-squares line on the plot (percent sugar, calories) and the line $y \approx 166$ for the mean number of calories. In general, describe the difference between making predictions if you were to use the least-squares line and if you were to use the line representing the mean number of calories.

d. $r^2 = 0.132$ for the data. What is r and what does this tell you about the association between the percent of sugar and the number of calories in breakfast cereal?

15. Use the cereal data at the beginning of this lesson to plot the data for the number of milligrams of sodium as a function of the percent of sugar (percent sugar, mg sodium). Use the second grid on *Activity Sheet 8*.

a. How strongly do you think the variables are related?

b. Use a calculator to find the correlation coefficient. What is r^2 and what does it tell you?

c. Find the least-squares regression line for the data. What is the slope of the regression line?

d. Comment on this statement: *The slope of the regression line measures the same thing as the correlation coefficient.*

e. What does the correlation coefficient tell you about the relationship between the amount of sugar and the amount of sodium?

Correlation and Cause and Effect

People often confuse correlation with cause and effect. *Just because two variables are correlated does not mean that one causes the other.*

- The two variables could both be a function of some other cause,
- the supposed cause could be the effect, or
- the relationship could be purely coincidental.

Consider the relationship between overall grade-point averages and grades in English. The association may be strong, but English grades alone do not cause high grade-point averages; other courses also contribute. The association between grade-point averages and hours of study is high, and it is reasonable to assume that the time spent studying is a primary cause of grade-point averages. In some sense, however, higher grades could also cause one to study more. The correlation between grade-point averages and SAT scores is strong, but neither variable causes the other. A good SAT score does not cause high grade-point averages.

Sometimes the relationship is purely coincidental. It could be that the correlation between grade-point averages and distance from school was strong. But it seems unlikely that all the good students live the same distance from school. Much more reasonable is the assumption that the connection is coincidental, and that there is no real link between distance and grade point.

16. Suppose you had scatter plots for data dealing with the following situations:

Number of cigarettes smoked and number of deaths due to cancer

Number of basketball fouls committed and number of points scored in field goals

Years of Latin and SAT scores

Weight of a car and the amount of fuel used

Years of schooling and yearly income

Foot length and reading level

a. What do you think each graph might look like?

b. Do you think there is positive correlation, negative correlation, or no correlation for each?

c. For those that have fairly good correlation, does one cause the other? Explain.

17. Think of an example involving two variables with a positive correlation in which one variable does not cause the other.

18. Toy prices from a fall catalog are given in the table below.

Toy	Price ($)
Stacking rings	10.99
Curious George books (3)	8.99
Popcorn popper	14.95
Stuffed animal	18.50
Baby All Gone	19.99
Infant rocking horse	25.78
Barbie doll	21.99
Lego set	42.49

In December, the company offered a reduced price of $3.00 off every item as part of a holiday sale.

a. What does the plot (old, new) for the prices look like?

b. Estimate the correlation between the new and the old prices. Calculate the correlation and compare it to your estimate.

c. Suppose the company slashed every price by 10%. Recalculate the correlation between the new and old prices. Explain any differences.

19. The data below are the number of minutes played, number of rebounds, number of assists, and number of points made by the Chicago Bulls in the 1995–1996 NBA season in which they won their fourth NBA championship.

Chicago Bulls, 1995–1996

Players	Minutes	Rebounds	Assists	Points
Jordan	3090	543	352	2491
Pippen	2825	496	452	1496
Kukoc	2103	323	287	1065
Longley	1641	318	119	564
Kerr	1919	110	192	688
Harper	1886	213	208	594
Rodman	2088	952	160	351
Wennington	1065	174	46	376
Salley*	673	140	54	85
Buechler	740	111	56	278
Simpkins	685	156	38	216
Brown	671	66	73	185

Source: *World Almanac and Book of Facts,* 1997
* Played for more than one team

a. Find the correlation coefficient between the following pairs of variables and discuss what each coefficient tells you.

i. Number of rebounds and number of points

ii. Number of assists and number of rebounds

iii. Minutes played and number of points

iv. Minutes played and number of rebounds

b. For which pair of variables is the correlation the strongest? Do you think there is a cause-and-effect relationship?

c. For which pair of variables is the correlation the weakest?

20. The following quote appeared in a suburban Milwaukee newspaper article entitled "Spending More on Police Doesn't Reduce Crime."

A CNI study of crime statistics and police department budgets over the last four years reveals there really is no correlation between what a community spends on law enforcement and its crime rate.

a. Make a sketch of what you think a plot of the data above would look like.

b. Use the data in the table below about the suburban crime rate and the per-capita spending on police. Plot the data and find the correlation coefficient.

Community	Suburban Crime Rate per 1,000 Residents	Per-Capita Spending on Police ($)
Glendale	74.39	222.25
West Allis	50.43	164.47
Greendale	50.25	123.43
Greenfield	48.68	143.59
Wauwatosa	48.52	150.34
South Milwaukee	43.20	110.64
Brookfield	42.16	131.42
Cudahy	41.47	137.12
St. Francis	41.32	144.84
Shorewood	40.84	156.39
Oak Creek	40.84	160.41
Brown Deer	37.09	150.57
Germantown	31.16	125.86
Menomonee Falls	29.73	159.28
Hales Corners	27.04	155.40
New Berlin	25.45	125.33
Franklin	23.09	94.92
Elm Grove	21.86	191.64
Whitefish Bay	21.28	120.34
Muskego	17.00	105.35
Fox Point	15.07	132.85
Mequon	11.31	136.39

Source: *Hub*, November 4, 1993

c. Does the correlation coefficient support the conclusions in the paragraph?

d. What does r^2 indicate about the relationship between spending and the crime rate?

21. The following article was taken from the *Milwaukee Journal* on Friday, November 20, 1992.

> **Yes, But How Come Those Hee Haw People Smile?**
>
> The mournful lyrics of country music may lament that "I'm So Lonesome I Could Cry." Now a statistical study claims such songs bring out more than tears in beer—they may lead to an increase in suicide. The study, co-authored by Steven Stack of Wayne State University in Detroit and Auburn University sociologist Jim Gundlach, was published in the September issue of *Social Forces,* a journal of sociology. Gundlach said Tuesday that the survey found a correlation between suicides in America and listening to country music, known for its often plaintive sounds and themes of loss and loneliness. Gundlach said the survey was based on the radio market share for country music in the nation's 49 leading music markets and the incidence of suicides in those areas. He said the study, based on 1985 statistics, found that for every 1% increase in country music's share of the market, there was a corresponding increase in the number of suicides.

Comment on the use of the statistics in this article. Describe the kind of data used, what a plot might look like, and what the various terms mean.

LESSON 11

Which Model When?

Is the least-squares regression line always the *best* line to use?

What considerations should be kept in mind when using the least-squares model?

OBJECTIVE

Recognize the need for different linear models and the impact of outliers on the least-squares regression line.

INVESTIGATE

In some earlier work, the median-fit line or another model may have been used for summarizing the relationship between two variables. This lesson centers on using the least-squares regression line to describe that relationship.

Discussion and Practice

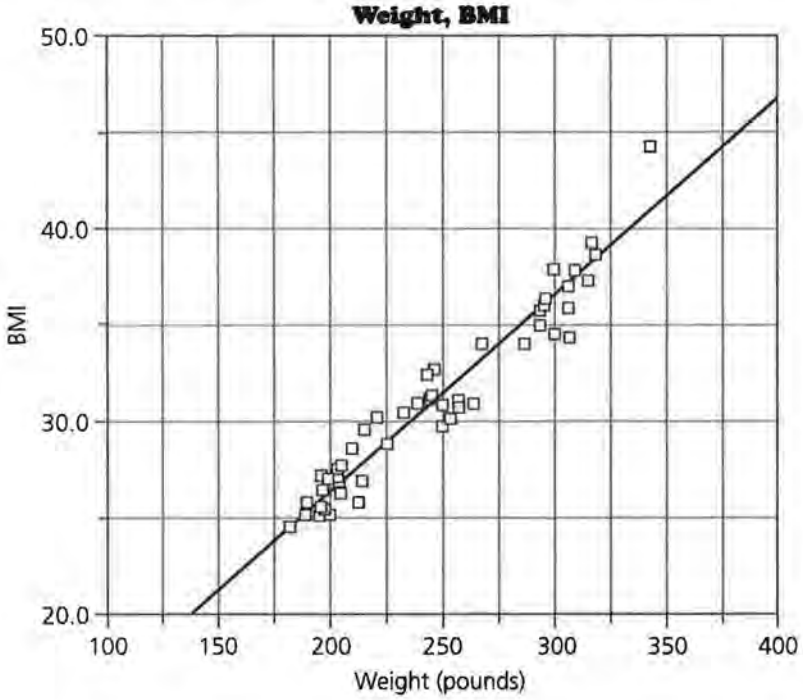
The following table lists team members of the 1997 NFC champion Green Bay Packers. It gives position, height, weight, and body-mass index for each player.

Player	Position	Height (ft-in.)	Weight (lb)	Body-Mass Index
Robert Brooks	WR	6-0	180	24.5
Ryan Longwell	K	6-0	185	25.1
Craig Hentrich	P	6-3	200	25.1
Doug Evans	CB	6-1	190	25.1
Bill Schroeder	WR	6-2	198	25.5
Antonio Freeman	WR	6-1	194	25.6
Steve Bono	QB	6-4	212	25.9
Don Beebe	WR	5-11	185	25.9
Darren Sharper	CB/S	6-2	205	26.4
Eugene Robinson	S	6-0	197	26.8
Roderick Mullen	CB/S	6-1	204	27.0
Doug Pederson	QB	6-3	216	27.1
LeRoy Butler	S	6-0	200	27.2
Tyrone Williams	CB	5-11	195	27.3
Terry Mickens	WR	6-0	201	27.3
Mike Prior	S	6-0	203	27.6

Player	Position	Height (ft-in.)	Weight (lb)	Body-Mass Index
Derrick Mayes	WR	6-0	205	27.9
Chris Darkins	RB	6-0	210	28.5
Brett Favre	QB	6-2	225	28.9
Aaron Hayden	RB	6-0	216	29.4
Jeff Thomason	TE	6-5	250	29.7
Travis Jervey	RB	6-0	222	30.2
Mark Chmura	TE	6-5	253	30.1
Dorsey Levens	RB	6-1	230	30.4
Tyrone Davis	TE	6-4	255	31.1
Lamont Hollinquest	LB	6-3	250	31.3
Seth Joyner	LB	6-2	245	31.5
Brian Williams	LB	6-1	240	31.7
Paul Frase	DE	6-5	267	31.7
Bernardo Harris	LB	6-2	247	31.8
Keith McKenzie	LB/DE	6-3	255	31.9
George Koonce	LB	6-1	243	32.1
William Henderson	FB	6-1	249	32.9
Rob Davis	LS	6-3	271	33.9
Santana Dotson	DT	6-5	285	33.9
John Michels	T	6-7	304	34.3
Jeff Dellenbach	C/G	6-6	300	34.7
Gabe Wilkins	DE	6-5	295	35.1
Marco Rivera	G	6-4	295	36.0
Adam Timmerman	G	6-4	295	36.0
Bob Kuberski	DT	6-4	295	36.0
Reggie White	DE	6-5	304	36.1
Jermaine Smith	DT	6-3	289	36.2
Ross Verba	G/T	6-4	299	36.5
Bruce Wilkerson	T	6-5	310	36.8
Aaron Taylor	G	6-4	305	37.2
Frank Winters	C	6-3	300	37.6
Darius Holland	DT	6-5	320	38.0
Earl Dotson	T	6-4	315	38.4
Joe Andruzzi	G	6-3	313	39.2
Gilbert Brown	DT	6-2	345	44.4

Source: *Milwaukee Journal Sentinel,* November 18,1997

Body-mass index, or *BMI,* is a commonly used guideline to gauge obesity and is based on height and weight. A BMI of 27.3 or above for women and 27.8 or above for men are considered obese. To calculate BMI, you must first convert your weight and height into kilograms and meters. Divide weight in pounds by 2.2 for weight in kilograms. Multiply height in inches by 0.0254 for height in meters. BMI is weight in kilograms divided by (height in meters) squared. A plot of the weight and BMI with the least-squares regression line is shown below.



1. Use the data about the BMI for each of the following problems.

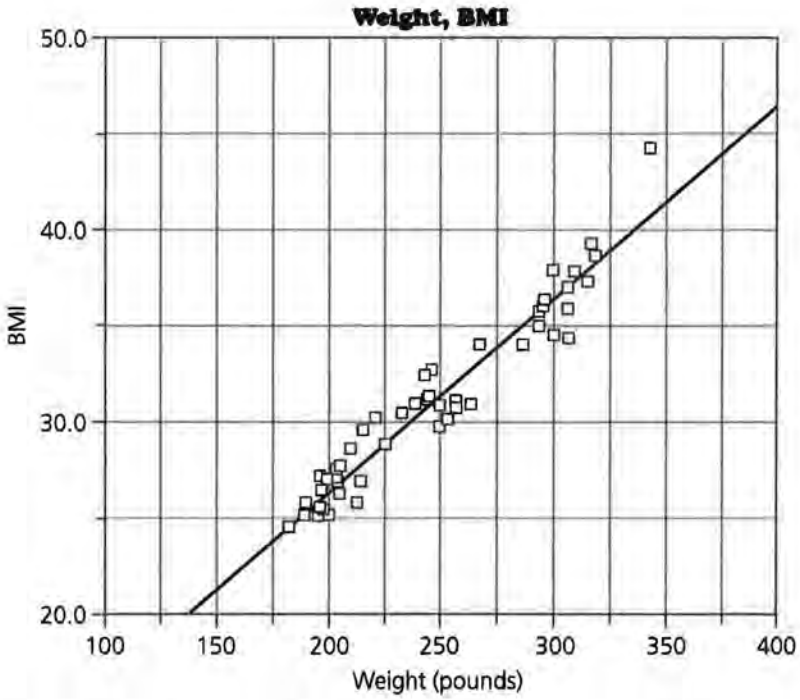
a. One player weighs more than all others and has a high BMI. How does the line seem to reflect the impact of that point?

b. The correlation coefficient, r, is 0.971. What does this tell you about predicting the BMI from the weight?

c. The equation of the line is BMI = 0.09884 × weight + 6.8401. Find the sum of the squared errors and use it to find the root mean squared error.

d. Use your information to predict the BMI for a person that weighs 140 pounds. What effect will the error have on your prediction?

2. A median-fit line has been plotted for the data in the graph below. The equation of the median-fit line is BMI = 0.098 × weight + 6.898.



   a. How does the graph of the median-fit line differ from the graph of the least-squares regression line?

   b. Find a measure of the typical error in prediction using the median-fit line. How do you define *typical* error?

   c. Use the median-fit line to identify the BMI for a player who weighs 140 pounds.

   d. Which line do you think will be a better predictor, the median-fit line or the least-squares line? Justify your answer.

3. Now delete the point that seems to be an outlier and analyze the data again.

   a. Find an equation of the least-squares line without using that point.

   b. What is the new correlation? What does that tell you about predicting the BMI from the weight?

   c. How does the new equation compare to the equations from Problems 1 and 2?

   d. Do you think removing a point before you do your analysis is justified? Explain your answer.

4. The four data sets below were constructed by Frank Anscombe ("Graphs in Statistical Analysis," *The American Statistician,* February, 1973). Complete parts a and b for each set. Then complete part c.

   a. Plot the data. Plot x on the horizontal axis and y on the vertical axis.

   b. Find the least-squares regression line and the correlation coefficient.

   c. What conclusions can you make after you have investigated all four data sets?

Data Set 1		Data Set 2		Data Set 3		Data Set 4	
x	y	x	y	x	y	x	y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	9	12.50
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Source: *Statistics for Business: Data Analysis and Modeling,* Jonathan D. Cryer/ Robert B. Miller

Summary

Finding a good model for a set of data involves much more than *number crunching.* It is important that you use all of the information, including a scatter plot of the data, to find a model that represents the data well. Both the least-squares regression line and the correlation coefficient are sensitive to extreme values. Patterns that are not linear are not captured by the correlation coefficient. Some nonlinear patterns have a high correlation. This makes it even more important for you to study a scatter plot of the data as a first step in your analysis.

Practice and Applications

Many times there are relationships between variables involved in sports situations. The following is from an article in *USA Today.*

Center Admitted to an Elite Group

The National Hockey League's 500-goal club added its 18th member Saturday night when Los Angeles Kings center Jari Kurri scored an empty-net goal in an 8–6 victory against the Boston Bruins. The list includes three "active" players.

Player	Primary or Current Team	Seasons	Games	Goals
Gordie Howe	Detroit	26	1767	801
Wayne Gretzky*	Los Angeles	13	999	749
Marcel Dionee	Los Angeles	18	1348	731
Phil Esposito	Boston	18	1282	717
Bobby Hull	Chicago	16	1063	610
Mike Bossy	N.Y. Islanders	10	752	573
Guy Lafleur	Montreal	17	1126	560
John Bucyk	Boston	23	1540	556
Maurice Richard	Montreal	18	978	544
Mike Gartner	N.Y. Rangers	14	1011	542
Stan Mikita	Chicago	22	1394	541
Frank Mahovlich	Toronto	18	1181	533
Bryan Trottier	N.Y. Islanders	17	1238	520
Gil Perreault	Buffalo	17	1191	512
Michel Goulet*	Chicago	14	976	511
Jean Beliveau	Montreal	20	1125	507
Lanny McDonald	Toronto	16	1111	500
Jari Kurri *	Los Angeles	12	833	500

* Active player Source: *USA Today*

5. Make scatter plots of (seasons, games) and (seasons, goals).

a. If there is a linear relationship in either plot, find a model and use it to predict the number of games or goals for a future hockey player who scores at least 500 goals and plays for 19 seasons.

b. Which model did you use and why? How well do you think your model will predict?

c. Wayne Gretzky was still an active player at the time of this article. Use your model to predict how many games or goals he will have if he plays for 28 seasons. Do you think this is reasonable?

6. The following data came from a poster on the door in a school lunchroom.

"LITE" JUNK FOOD ... How Healthy Is It Really?

Item	Fats	Calories	Calories from Fat	Percent of Calories from Fat
McDonald's				
McLean Deluxe Sandwich	10	320	90	28.13%
Filet O' Fish Sandwich	18	370	162	43.78%
McLean Deluxe and Small Fries	22	540	198	36.67%
Burger King				
Weight Watchers Fettucini Broiled Chicken	11	298	99	33.23%
Fried Chicken Sandwich	40	685	360	52.56%
Wendy's				
Chicken Sandwich	20	450	180	40.00%
Baked Potato with Broccoli & Cheese	24	550	216	39.28%
Taco Salad	23	530	216	39.28%
Kentucky Fried Chicken				
Skinfree Crispy Breast	17	293	153	52.22%
Chicken Sandwich	27	482	243	50.41%
Hardee's				
Oat Bran Muffin	18	440	162	36.82%
Real Lean Deluxe and Small Fries	24	570	116	37.90%

Source: *Health & Healing*

a. Make a scatter plot of (fats, calories). Find both a median-fit line and a least-squares regression line for the data. How do the two lines compare?

b. Find the correlation coefficient. What does this tell you about the relationship between fats and calories in "lite" junk food?

c. Which model seems to summarize the relationship better?

d. Predict how many calories would be in a "lite" food that has 30 grams of fat. How well do you think your line predicts? Why?

7. Use the data in Problem 6 for the following.

a. What do you think the plot of calories and calories from fat should look like? Plot the data and find a line to summarize the relationship.

b. Find the correlation coefficient.

c. If you haven't already done so, look closely at the plot. There is something unusual about the data. Can you find a reasonable explanation for this observation?

d. Based on the conclusions you drew in part c above, how would you adjust your model?

In Problems 8–10, plot each set of data. Decide whether a median-fit line or a least-squares regression line will give a better description of the relationship between the variables. Explain how you made your choice and why you selected the one you did.

8. The amount of waste for glass and plastic in the United States since 1960 and projected until 2010 is given in the table. Plot (glass, plastic).

Year	Glass (million tons)	Plastic (million tons)
1960	49	94
1970	70	124
1980	80	150
1990	96	172
2000	115	200
2010	125	230

Source: *World Almanac and Book of Facts,* 1992

9. The following data are the overall scores and the price of VCRs as rated in *Consumer Reports,* October, 1997. Plot (price, score).

Brand and Model	Price ($)	Overall Score
Sony SLV-775HF	300	82
Panasonic PV-7662	290	75
Samsung VR8807	220	75
Toshiba M-683	250	75
Hitachi VT-FX624A	270	70
Sharp VC-H978U	230	68
RCA VR626FH	290	68
JVC HR-VP644U	260	65
RCA VR631HF	230	65
Mitsubishi HS-U580	380	62
Quasar VHQ760	170	60
GE VG4261	180	50
Philips Magnavox VRX362AT	240	45

10. The following list gives median 1995 prices of homes in metropolitan areas of the United States along with estimated percents of annual growth of these prices. Plot (price, growth).

Metro Area	Median Home Price 1995 ($1000s)	Annual Growth to Year 2000 (percent)
Atlanta	97.7	4.3
Baltimore	111.4	1.3
Boston	178.2	2.3
Chicago	147.4	3.2
Cincinnati	102.6	4.7
Cleveland	103.3	4.4
Dallas	96.3	3.9
Denver	126.2	5.4
Detroit	98.5	5.9
Houston	79.4	3.1
Kansas City	91.1	3.8
Los Angeles	176.9	0.4
Miami	106.5	3.2
Milwaukee	114.3	4.1
Minneapolis	107.3	4.7
New York City	169.6	1.7
Philadelphia	117.0	1.7
Phoenix	96.3	4.5
Pittsburgh	80.6	1.8
San Francisco	255.3	2.8
San Diego	172.0	2.1
Seattle	158.5	3.2
St. Louis	87.4	3.8
Tampa-St. Petersburg	77.8	3.2
Washington, D.C.	155.8	1.1

Source: *Consumer Reports*, May, 1996

Dale Seymour Publications® is a leading publisher of K–12 educational materials in mathematics, thinking skills, science, language arts, and art education.

DALE
SEYMOUR
PUBLICATIONS®

ISBN 1-57232-245-4
21182