Statway™

A statistics pathway for college students

Version 1.0

A resource from
The Charles A. Dana Center at
The University of Texas at Austin

July
2011

The Charles A. Dana Center
The University of Texas at Austin
1616 Guadalupe Street, Suite 3.206
Austin, TX 78701-1222
Fax: 512-232-1855
dana-txshop@utlists.utexas.edu
www.utdanacenter.org

The Carnegie Foundation for the Advancement of Teaching
51 Vista Lane
Stanford, California, 94305

Phone: 650-566-5110
pathways@carnegiefoundation.org
www.carnegiefoundation.org

About the development of this resource

The content for this full version of Statway was developed under a November 30, 2010, agreement by a team of faculty authors and reviewers contracted and managed by the Charles A. Dana Center at the University of Texas at Austin with funding from the Carnegie Foundation for the Advancement of Teaching.

This resource was produced in Microsoft Word 2008 and 2011 for the Mac. The content of these 12 modules was developed and produced (that is, written, reviewed, edited, and laid out) by the Charles A. Dana Center at The University of Texas at Austin and delivered by the Dana Center to the Carnegie Foundation for the Advancement of Teaching on June 30, 2011.

Some issues to be aware of:

- PDF files need to be viewed with Adobe Acrobat for full functionality. If viewed through Preview, which is the default on some computers, the URLs may not be correct.
- The file names indicate the lesson number and whether the document is the instructor or student version or the out-of-class experience.

The Dana Center is engaged in a process of revising and improving these materials to create the Dana Center Statistics Pathway. We welcome feedback from the community as part of our course revision process. If you would like to discuss these materials or learn more about the Dana Center's plans for this course, contact us at **mathways@austin.utexas.edu**.

About the Charles A. Dana Center at The University of Texas at Austin

The Dana Center collaborates with local and national entities to improve education systems so that they foster opportunity for all students, particularly in mathematics and science. We are dedicated to nurturing students' intellectual passions and ensuring that every student leaves school prepared for success in postsecondary education and the contemporary workplace—and for active participation in our modern democracy.

The Center was founded in 1991 in the College of Natural Sciences at The University of Texas at Austin. Our original purpose—which continues in our work today—was to raise student achievement in K–16 mathematics and science, especially for historically underserved populations. We carry out our work by supporting high standards and building system capacity; collaborating with key state and national organizations to address emerging issues; creating and delivering professional supports for educators and education leaders; and writing and publishing education resources, including student supports.

Our staff of more than 80 researchers and education professionals has worked intensively with dozens of school systems in nearly 20 states and with 90 percent of Texas's more than 1,000 school districts. As one of the College's largest research units, the Dana Center works to further the university's mission of achieving excellence in education, research, and public service. We are committed to ensuring that the accident of where a student attends school does not limit the academic opportunities he or she can pursue.

For more information about the Dana Center and our programs and resources, see our homepage at **www.utdanacenter.org.** To access our resources (many of them free) please see our products index at **www.utdanacenter.org/products.** To learn about Dana Center professional development sessions, see our professional development site at **www.utdanacenter.org/pd**.

Acknowledgments

The original version of this work was created by The Charles A. Dana Center at The University of Texas at Austin under sponsorship of the Carnegie Foundation for the Advancement of Teaching.

Carnegie Corporation of New York, The Bill & Melinda Gates Foundation, The William and Flora Hewlett Foundation, Lumina Foundation, and The Kresge Foundation joined in partnership with the Carnegie Foundation for the Advancement of Teaching in this work.

Leadership—Charles A. Dana Center at the University of Texas at Austin

Uri Treisman, director
Susan Hudson Hull, program director of mathematics national initiatives

Leadership—Carnegie Foundation for the Advancement of Teaching

Anthony S. Bryk, president
Bernadine Chuck Fong, senior managing partner
Louis Gomez, senior fellow
Paul LeMahieu, senior fellow
James Stigler, senior fellow
Uri Treisman, senior fellow
Guadalupe Valdés, senior fellow

Statway Project Leads

Kristen Bishop, former team lead for the New Mathways Project, the Charles A. Dana Center at the University of Texas at Austin
Thomas J. Connolly, project lead, Statway, the Charles A. Dana Center at the University of Texas at Austin
Karon Klipple, director of Statway, the Carnegie Foundation for the Advancement of Teaching
Jane Muhich, director of Quantway, the Carnegie Foundation for the Advancement of Teaching

Project Staff—Charles A. Dana Center at the University of Texas at Austin

Richard Blount, advisor
Kathi Cook, project director, online services team
Jenna Cullinane, research associate
Steve Engler, lead editor and production editor
Amy Getz, team lead for the New Mathways Project
Susan Hudson Hull, program director of mathematics national initiatives
Joseph Hunt, graduate research assistant
Rachel Jenkins, consulting editor
Erica Moreno, program coordinator
Carol Robinson, administrative associate
Cathy Seeley, senior fellow
Rachele Seifert, administrative associate
Lilly Soto, senior administrative associate
Phil Swann, senior designer
Laura Torres, graduate research assistant
Thomas Wiegel, freelance formatter and proofreader

Authors Contracted by the Dana Center

Roxy Peck, professor emerita of statistics, California Polytechnic State University, San Luis Obispo, California
Beth Chance, professor of statistics, California Polytechnic State University, San Luis Obispo, California
Robert C. delMas, associate professor of educational psychology, University of Minnesota, Minneapolis, Minnesota
Scott Guth, professor of mathematics, Mt. San Antonio College, Walnut, California
Rebekah Isaak, graduate research student, University of Minnesota, Minneapolis, Minnesota
Leah McGuire, assistant professor, University of Minnesota, Minneapolis, Minnesota
Jiyoon Park, graduate research student, University of Minnesota, Minneapolis, Minnesota
Brian Kotz, associate professor of mathematics, Montgomery College, Germantown, Maryland
Chris Olsen, assistant professor of mathematics and statistics, Grinnell College, Grinnell, Iowa
Mary Parker, professor of mathematics, Austin Community College, Austin, Texas
Michael A. Posner, associate professor of statistics, Villanova University, Villanova, Pennsylvania
Thomas H. Short, professor, John Carroll University, University Heights, Ohio
Penny Smeltzer, teacher of statistics, Westwood High School, Austin, Texas
Myra Snell, professor of mathematics, Los Medanos College, Pittsburg, California
Laura Ziegler, graduate research student, University of Minnesota, Minneapolis, Minnesota

Reviewers Contracted by the Dana Center

Michelle Brock, American River College, Sacramento, California
Thomas J. Connolly, the Charles A. Dana Center at the University of Texas at Austin
Andre Freeman, Capital Community College, Hartford, Connecticut
Karon Klipple, the Carnegie Foundation for the Advancement of Teaching
Roxy Peck, professor emerita of statistics, California Polytechnic State University, San Luis Obispo, California
Jim Smart, Tallahassee Community College, Tallahassee, Florida
Myra Snell, Los Medanos College, Pittsburg, California

Committee for Statistics Learning Outcomes

Rose Asera, formerly of the Carnegie Foundation for the Advancement of Teaching
Kristen Bishop, formerly of the Charles A. Dana Center at the University of Texas at Austin
Richelle (Rikki) Blair, American Mathematical Association of Two-Year Colleges (AMATYC); Lakeland Community College, Ohio
David Bressoud, Mathematical Association of America (MAA); Macalester College, Minnesota
John Climent, American Mathematical Association of Two-Year Colleges (AMATYC); Cecil College, Maryland
Peg Crider, Lone Star College, Tomball, Texas
Jenna Cullinane, the Charles A. Dana Center at the University of Texas at Austin
Robert C. delMas, Consortium for the Advancement of Undergraduate Statistics Education (CAUSE); University of Minnesota, Minneapolis, Minnesota
Bernadine Chuck Fong, the Carnegie Foundation for the Advancement of Teaching
Karen Givvin, the University of California, Los Angeles
Larry Gray, American Mathematical Society (AMS); University of Minnesota
Susan Hudson Hull, the Charles A. Dana Center at the University of Texas at Austin
Rob Kimball, American Mathematical Association of Two-Year Colleges (AMATYC); Wake Technical Community College, North Carolina
Dennis Pearl, Consortium for the Advancement of Undergraduate Statistics Education (CAUSE); The Ohio State University
Roxy Peck, American Statistical Association (ASA); Consortium for the Advancement of Undergraduate Statistics Education (CAUSE); California Polytechnic State University, San Luis Obispo, California

Myra Snell, American Mathematical Association of Two-Year Colleges (AMATYC); Los Medanos College, Pittsburg, California
Jim Stigler, the Carnegie Foundation for the Advancement of Teaching; the University of California, Los Angeles
Daniel Teague, Mathematical Association of America (MAA); North Carolina School of Science and Mathematics, Durham
Uri Treisman, the Carnegie Foundation for the Advancement of Teaching; the Charles A. Dana Center at the University of Texas at Austin

Version 1.0 of Statway was developed in collaboration with faculty from the following colleges, the "Collaboratory," who advised on the development of the course. These Collaboratory colleges are:

California

American River College, Sacramento, California
Foothill College, Los Altos Hills, California
Mt. San Antonio College, Walnut, California
Pierce College, Woodland Hills, California
San Diego City College, San Diego, California

California State University System

CSU Northridge
Sacramento State University
San Jose State University

Connecticut

Capital Community College, Hartford, Connecticut
Gateway Community College, New Haven, Connecticut
Housatonic Community College, Bridgeport, Connecticut
Naugatuck Valley Community College, Waterbury, Connecticut

Florida

Miami Dade College, Miami, Florida
Tallahassee Community College, Tallahassee, Florida
Valencia Community College, Orlando, Florida

Texas

Austin Community College, Austin, Texas
El Paso Community College, El Paso, Texas
Houston Community College, Houston, Texas
Northwest Vista College, San Antonio, Texas
Richland College, Dallas, Texas

Washington

Seattle Central Community College, Seattle, Washington
Tacoma Community College, Tacoma, Washington

Statway, Full Version 1.0, July 2011

Table of Contents

Initiating Lesson 3.1.1: Introduction to Scatterplots and Bivariate Relationships

Estimated number of 50-minute class sessions: 2

Learning Goals

S.2. Distributional Thinking Goal: Students will demonstrate the use of distributional thinking to reason about the data in order to describe and summarize distributions of data, identify trends and patterns, judge the fit of a model to a distribution, and describe similarities and differences in comparing distributions.

The series of tasks in this introductory lesson are designed to motivate an initial and informal understanding of concepts related to interpreting scatterplots.

Specifically, you want students to gain a preliminary and informal *understanding* that

- each point on the scatterplot represents a single observation consisting of measurements on two variables.
- an overall downward trend in the data indicates that small values for *x* tend to correspond to large values for *y*. Larger values for *x* tend to correspond to smaller values for *y*. Also with a similar understanding of overall upward trends.
- the accuracy in a prediction is related to the variability (scatter) in the data. Variability can be explained by increases in *x* or by additional factors having influence on *y*.

Students should begin to learn *how* to

- interpret the meaning of particular points on the scatterplot.
- without the benefit of formal development of the statistical concepts of association, correlation, and regression, recognize directional trends in the distribution of bivariate data and use these trends to make predictions.
- assess the strength of the relationship informally by looking at the degree of scatter.
- develop plausible explanations for the variability seen in the data.

In this series of tasks, you are not formally developing the statistical concepts of correlation or regression. Rather, you are working to build students' ability to see associative trends through the noise of real data and to make decisions in the face of variability. After this lesson, students are introduced more formally to the following concepts:

- form of a relationship as linear or nonlinear,
- correlation coefficient as a measure of the strength of a linear association, and
- least squares regression line as a way of describing central tendency of a bivariate distribution (much like the mean describes the central tendency of a univariate distribution).

Developmental Math Connections

Since this is an introduction to working with bivariate data, the learning goals are focused on interpreting scatterplots and motivating an initial and informal understanding of concepts related to distributional thinking with bivariate data that underlies modeling data with a linear function. In future lessons, these ideas become more explicit.

You might wonder why the authors did not begin with a discussion of the Cartesian plane or an exercise in which students construct a scatterplot, as you might in an elementary algebra course. The rationale is

multifold. First, in Statway, you have a more heterogeneous group of students since students typically found in community college elementary algebra and intermediate algebra are in the same class. Therefore, you need to find ways to level the playing field so that some students are not bored and others lost. Of course, this level playing field means that all students must be able to bring something to the task that puts them into play, so to speak. Second, the authors build some of these skills directly into the lesson. For example, students plot points in Task 2, but the plotting of points is connected to constructing a scatterplot in which no association exists between the variables. Point plotting is in the service of the learning goal of interpreting trends in bivariate data. Third, research suggests that students need to struggle with meaningful tasks before a careful explication of concepts helps them construct deeply held understandings. The lesson is constructed to provide an opportunity for productive struggle, followed by carefully constructed tasks that the authors think are sufficient for all students to begin to make sense of the concepts at hand.

If you anticipate that some students, due to deficits in developmental math skills, may have trouble at specific points in the lesson, observe students work to see if your expectations are founded. Remember, the authors' hypothesis that *productive* struggle is a key ingredient in facilitating conceptual learning. You will need to use your own judgment about when to intercede to work with students individually. Report your observations back to the Lesson Study group. This will be invaluable information for improving the lesson.

Lesson Structure

This lesson has the following components:

- Introduction to the context of the lesson (15–20 minutes)
- Part I: Students work on a rich task, wrap-up, and transition to Part II (20–25 minutes)
- Part II: Scaffolded conceptual tasks (via discussion/group work/lecture) and wrap-up (40–50 minutes)
- Part III: Homework (done outside of class) (5 minutes)

Initiating Lesson 3.1.1: Introduction to Scatterplots and Bivariate Relationships

Introduction to the Context of the Lesson

(**Note:** To orient students to the data set, distribute the following handout and go over the introductory material before giving them a few minutes to choose the most nutritious cereal.)

Student Handout

In the United States, the Food and Drug Administration requires that most prepared food be labeled with nutrition information. Today part of the data set for the lesson was gathered from the list of required nutrition facts that appear on boxes of breakfast cereals.

Below are the nutrition facts for three cereals. There is a lot of information given. An entire math lesson could probably be designed around reading and understanding the information given on cereal labels. For now though, you will look over the labels fairly quickly, like you might do in the grocery store. Use your initial impressions to pick the cereal that you would rate as the most nutritious (the best for your health). You have a few minutes to do this.

Nutrition Facts
Serving Size 1.0 cup (1 NLEA serving)

Amount Per Serving	
Calories 117	Calories from Fat 9
	% Daily Value *
Total Fat 1g	1%
Saturated Fat 0g	1%
Polyunsaturated Fat 0g	
Monounsaturated Fat 1g	
Cholesterol 0mg	0%
Sodium 171mg	7%
Total Carbohydrate 26g	9%
Dietary Fiber 1g	3%
Sugars 14g	
Protein 1g	

Vitamin A 0% • Vitamin C 10%
Calcium 10% • Iron 30%

* Percent Daily Values are based on a 2,000 calorie diet. Your daily value may be higher or lower depending on your calorie needs:

	Calories:	2,000	2,500
Total Fat	Less than	65g	80g
Sat Fat	Less than	20g	25g
Cholesterol	Less than	300mg	300mg
Sodium	Less than	2,400mg	2,400mg
Total Carbohydrate		300g	375g
Dietary Fiber		25g	30g

Calories per gram:
Fat 9 • Carbohydrate 4 • Protein 4

Nutrition Facts
Serving Size 1.0 cup (1 NLEA serving)

Amount Per Serving	
Calories 101	Calories from Fat 1
	% Daily Value *
Total Fat 0g	0%
Saturated Fat 0g	0%
Polyunsaturated Fat 0g	
Monounsaturated Fat 0g	
Cholesterol 0mg	0%
Sodium 202mg	8%
Total Carbohydrate 24g	8%
Dietary Fiber 1g	3%
Sugars 3g	
Protein 2g	

Vitamin A 0% • Vitamin C 10%
Calcium 0% • Iron 54%

* Percent Daily Values are based on a 2,000 calorie diet. Your daily value may be higher or lower depending on your calorie needs:

	Calories:	2,000	2,500
Total Fat	Less than	65g	80g
Sat Fat	Less than	20g	25g
Cholesterol	Less than	300mg	300mg
Sodium	Less than	2,400mg	2,400mg
Total Carbohydrate		300g	375g
Dietary Fiber		25g	30g

Calories per gram:
Fat 9 • Carbohydrate 4 • Protein 4

Nutrition Facts
Serving Size 1.0 cup (1 NLEA serving)

Amount Per Serving	
Calories 111	Calories from Fat 16
	% Daily Value *
Total Fat 2g	3%
Saturated Fat 0g	2%
Polyunsaturated Fat 1g	
Monounsaturated Fat 1g	
Cholesterol 0mg	0%
Sodium 213mg	9%
Total Carbohydrate 22g	7%
Dietary Fiber 4g	14%
Sugars 1g	
Protein 4g	

Vitamin A 0% • Vitamin C 10%
Calcium 12% • Iron 69%

* Percent Daily Values are based on a 2,000 calorie diet. Your daily value may be higher or lower depending on your calorie needs:

	Calories:	2,000	2,500
Total Fat	Less than	65g	80g
Sat Fat	Less than	20g	25g
Cholesterol	Less than	300mg	300mg
Sodium	Less than	2,400mg	2,400mg
Total Carbohydrate		300g	375g
Dietary Fiber		25g	30g

Calories per gram:
Fat 9 • Carbohydrate 4 • Protein 4

(**Note:** These cereals are [from left to right] General Mills Cocoa Puffs, Kellogg's Corn Flakes, and General Mills Cheerios. Do not tell students the name of the cereals until after they have made their choices because the names might divert their attention from using the data to make their decision.)

Initiating Lesson 3.1.1: Introduction to Scatterplots and Bivariate Relationships

Discussion

After giving students a few minutes to make a decision, take a quick tally to see which cereal students chose as most nutritious. Then call on a few students to describe which ingredients they focused on. The goal in this short interaction with the class is to highlight the following three points to get students thinking about how ingredient amounts might be associated with nutritional ratings:

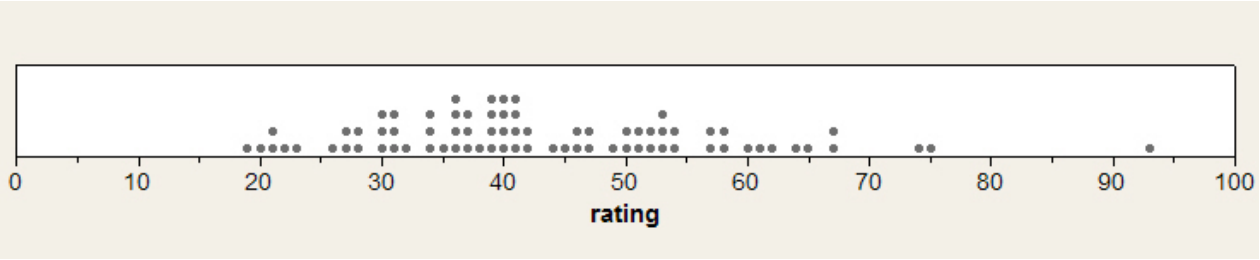
- Different people might devise different ways to rate the nutrition of cereals by focusing on different ingredients.
- You could use one ingredient or more than one ingredient to determine a nutrition rating.
- In a rating system, there is an association between the ingredient and the nutrition rating. For example, large amounts of sugar give lower ratings. Large amounts of fiber might give higher ratings.

Now transition to the *Consumer Reports* ratings by using the following information on the student handout to discuss *Consumer Reports*.

Part I [Student Handout]

Instead of making up your own rating system, you are going to investigate the *Consumer Reports* nutritional ratings for 77 breakfast cereals. *Consumer Reports* is published by a nonprofit organization called the Consumers Union, whose mission it is to work for a fair, just, and safe marketplace for all consumers and to empower consumers to protect themselves. *Consumer Reports* rates products based on its own criteria and testing. It prides itself on producing objective results. *Consumer Reports* maintain its objectivity by not allowing advertising within their publications and not allowing use of their results for commercial gain. (Retrieved from www.consumerreports.org/cro/aboutus/mission/overview/index.htm)

Consumer Reports uses a rating system with a scale of 0 to 100. Here is the distribution of *Consumer Reports* ratings for 77 cereals:

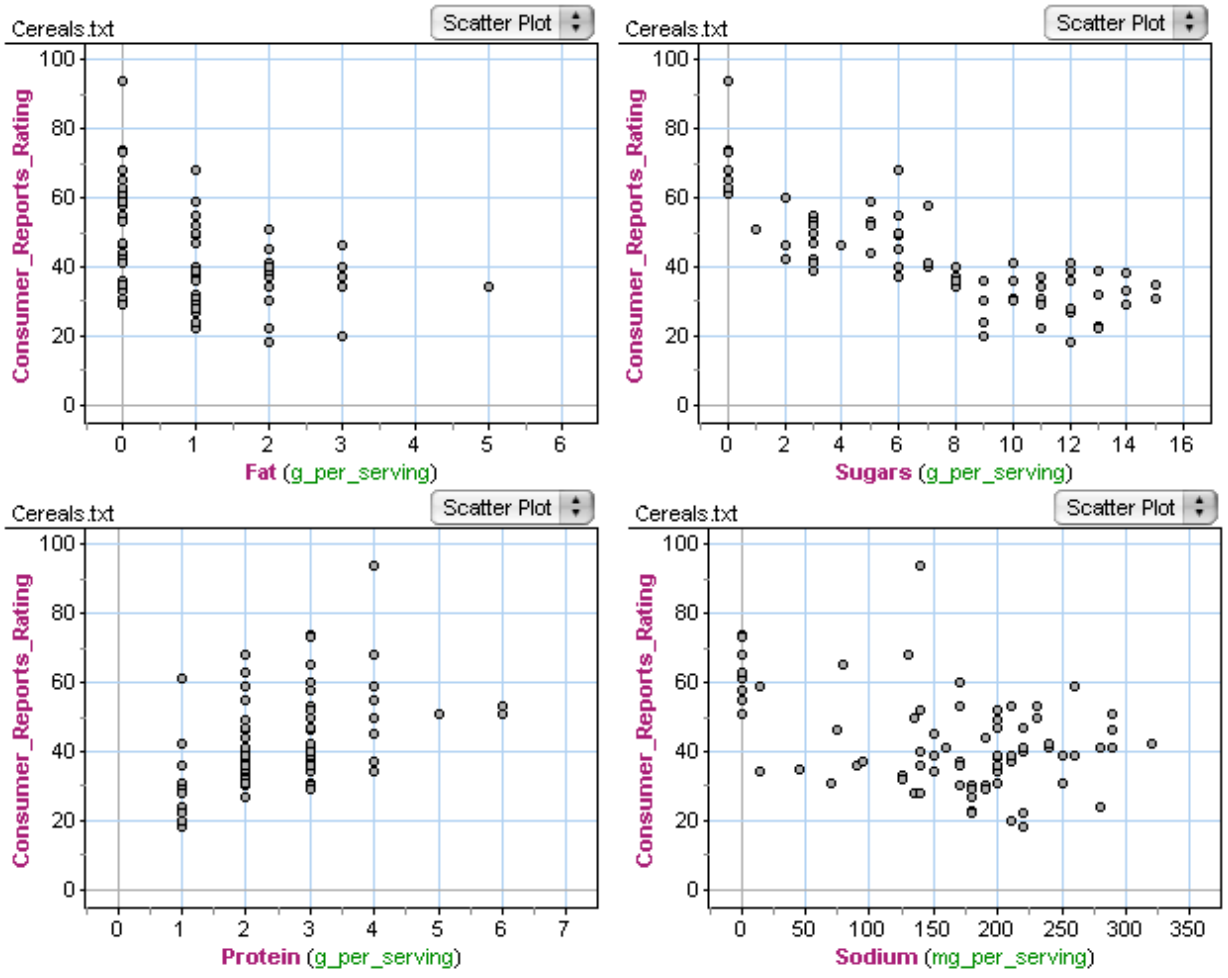


- What does each dot represent in this distribution?
- For this distribution, what seems to be an average rating?
- How would you describe the variability in ratings?
- How would you describe the shape of this distribution? What does the shape suggest about the rating system?

Initiating Lesson 3.1.1: Introduction to Scatterplots and Bivariate Relationships

(**Note:** Students should be able to describe shape, center, and spread of this distribution based on what they have learned in Module 2. Answers can vary. Reasonable responses include the following: The dots in the distribution are cereals; The average rating is about 40 [actual mean is 44.0, median 41.0]; Ratings fall between about 18 and 94 on a scale of 0 to 100, with a reasonable estimate for the standard deviation being 10 rating points [actual standard deviation is 13.8, Q1 34, Q3 52.9].)

What you cannot tell from the dotplot is how the cereal ingredients (such as sugar or fat) are related to the ratings. You need a new type of graph, called a scatterplot, to investigate how two variables relate to each other. The scatterplots below show the amount of an ingredient in a serving of cereal and the *Consumer Reports* rating for 77 breakfast cereals.



The *Consumer Reports* rating formula is not made public. So, you do not know which ingredients are used in its rating formula. In this lesson, you will try to "crack their code" in a sense. Use the data to figure out which ingredients *Consumer Reports* may, or may not, use in their rating formula. The only clues you have are these scatterplots.

Initiating Lesson 3.1.1: Introduction to Scatterplots and Bivariate Relationships

The first step in this investigation is to answer the following two questions AND write down enough of your reasoning that someone can follow your thinking.

Two new cereals are being rated by Consumer Reports*. Cereal A has 10.5 grams of sugar in a serving and Cereal B has 2.5 grams of protein in a serving.*

(1) Based on the data shown, predict the *Consumer Reports* rating for the two cereals.

(2) For which cereal do you think your prediction is probably more accurate (more likely to be closer to the actual Consumer Report rating)? Why?

Note to Instructors

Give students about 3 minutes to work on these questions alone and then in small groups for 5–10 minutes (depending on your sense of whether productive conversations are occurring).

Part I provides students the opportunity to struggle with important ideas (like interpreting scatterplots and seeing patterns that relate to a question at hand). So, at this point in the lesson, you do not need to guide students to discover canonical ideas, such as correlation, or even correct their misconceptions or fix their errors. This is an informal introduction to distributional thinking with bivariate data. While students work, listen to how students are reasoning as they discuss the task. In the wrap-up, you have the opportunity to talk with students in general terms about making predictions and using a visual sense of the variability in the data to determine which ingredient is a more accurate predictor of ratings. In the wrap-up, you can refer to what you observed as students worked, giving praise and noting interesting aspects of their conversations that are relevant to learning goals for the lesson.

The next segment of the lesson, Part II, is designed with more explicit attention to developing the skills and understandings described in the learning outcomes. It is in Part II that you make connections as well correct errors and address misconceptions as appropriate.

Wrap-Up/Direction Instruction About Statistical Concepts

You will need to project the scatterplots for this discussion. To see if students are using the patterns in the data to make predictions, ask them to determine if the following predictions for ratings are reasonable or unreasonable (perhaps ask students to show a thumbs up for *reasonable prediction* and thumbs down for *unreasonable prediction*): Cereal A: 10, 30, 60; Cereal B: 10, 30, 60.

Plot each of these predictions on the scatterplot and highlight how the prediction fits the pattern in the data or deviates from the pattern.

Cereal A: 10 (not reasonable), 30 (reasonable), 60 (not reasonable); Cereal B: 10 (not reasonable), 30 (reasonable), 60 (reasonable).

Discuss the following questions through a brief minilecture or class discussion:

- What is a range of reasonable predictions for ratings of Cereal A? of Cereal B?
- Which ingredient, sugar or protein, is a more accurate predictor of *Consumer Reports* ratings?

Reasonable Answers: For Cereal A, there is a narrower range of reasonable predictions for the *Consumer Reports* rating (approximately 20 to 40), so the amount of sugar is a fairly accurate predictor of the rating. Draw a vertical line segment at $x = 10.5$, $20 < y < 40$ to visually represent this reasonable range of predictions. For Cereal B, there is a wider range of reasonable predictions for the *Consumer Reports* rating (approximately 25 to 70), so the amount of protein is not as accurate a predictor of the rating. Represent this range with a vertical line segment. There are other factors influencing the rating besides the amount of protein. So, it is more likely that your prediction of the rating for Cereal A is closer to the actual rating than your prediction for Cereal B.

Part II: Scaffolded Conceptual Tasks

Task 1: Reading and Interpreting Scatterplots

Introduction *[Student Handout]*

In this task, you are going to take a short detour from your investigation into which ingredients are the best predictors of *Consumer Reports* ratings. Here, you will work on interpreting scatterplots just to make sure everyone is comfortable with reading this new type of graph.

Activities *[Student Handout]*

(3) Captain Crunch has the lowest *Consumer Reports* rating of the 77 cereals in the data set. How much fat is in a serving of Captain Crunch?

(4) In this set of 77 cereals, Product 19 has the most sodium in a serving. What is the rating for Product 19?

(5) All-Bran Extra Fiber is the cereal with the highest rating. How much sugar, fat, and sodium are in a serving of All-Bran Extra Fiber?

Wrap-Up/Direct Instruction About Statistical Concepts

If students have been working in groups, you will have some sense of where they had difficulty. For the subsequent whole-class discussion, address areas of difficulty and answer questions in the context of the wrap-up questions. You will not have time to go over the answers to the scaffolded questions.

To bring closure, ask students to discuss the following questions or deliver a short minilecture that answers these questions:

> When a statistician reads a scatterplot, he or she asks herself two questions: (1) Who or what is described by the data? (i.e., What does a dot represent?) and (2) What measurements were made? (i.e., What are the variables?) Pick a scatterplot, and answer these two questions.

Initiating Lesson 3.1.1: Introduction to Scatterplots and Bivariate Relationships

Conceptual Task 2: Seeing Patterns and Relationships in Scatterplots

Introduction *[Student Handout]*

Now you will continue your detective work with *Consumer Reports* ratings. Try to identify ingredients that are good predictors of ratings and ingredients that are not. More importantly, focus on how patterns in the data are related to identifying ingredients that are good predictors.

(**Note:** If students are working in groups, for most of these activities, the authors suggest that you intervene if students have answers that are wildly off base. The goal is to foster distributional thinking with bivariate data. So, when you intervene, refrain from correcting; instead get students to talk about what they are seeing. Make observations or ask questions to nudge them in the right direction. Remember that in the wrap-up, you will provide direct instruction relative to the learning goals for the lesson, so you do not have to fix everything during group work.)

Activities *[Student Handout]*

(6) There are four cereals that have 3 grams of fat in a serving. Estimate the ratings for these four cereals. What might explain the variability in the ratings?

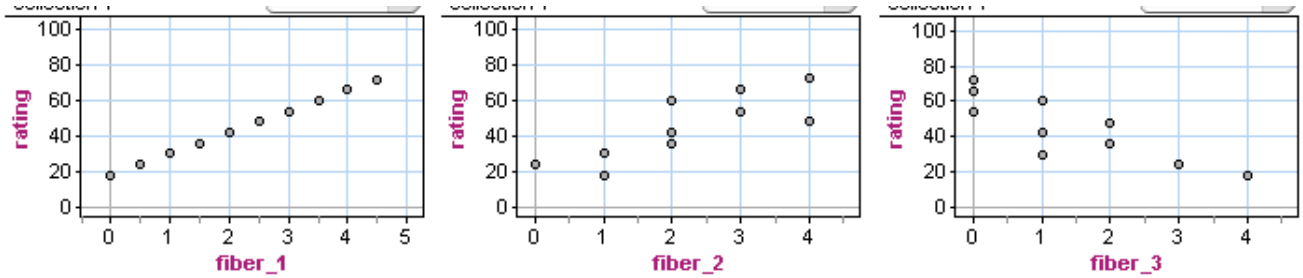
(**Answers:** Estimates might vary: 20, 34, 37, 46. Intervene and correct if students are misreading the scatterplots. Variability is explained by the impact of other ingredients on the ratings.)

(7) Imagine changing the recipe for a cereal that has 0 grams of fat in a serving and a rating of 60. Increase the amount of fat to 3 grams in a serving. Do you think the rating will probably increase or decrease or remain about the same? Or do you think that it is impossible to use the scatterplot to predict the impact of this change on the rating? How does the pattern in the data support your decision?

(**Answers:** The rating will probably decrease. You see a downward trend in the data that suggests that larger amounts of fat in a serving tend to be associated with lower ratings. If students think that it is impossible to predict how the rating will change, ask them to say more about why this is so. The goal is to get students to begin to develop distributional thinking for bivariate data, so nudge them to think about trends and patterns. It is reasonable to be unsure about how the rating will change because of the noise in the data. One pattern they could describe to support the "cannot tell" response is the large amount of variability in ratings for cereals with the same amount of fat.)

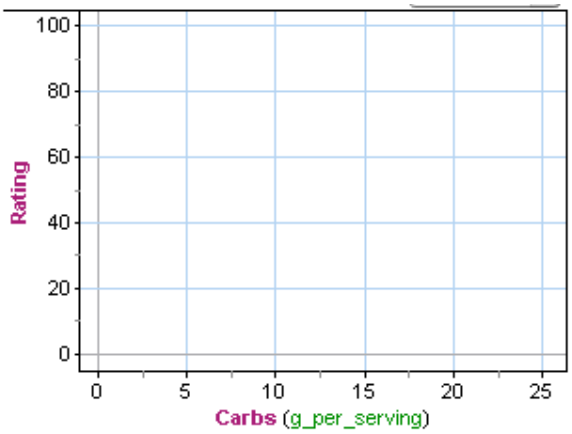
Initiating Lesson 3.1.1: Introduction to Scatterplots and Bivariate Relationships

(8) Think about how the amount of fiber in a cereal might relate to the *Consumer Reports* rating. Here are three scatterplots with make-believe data from 10 made-up cereals. Which scatterplot do you think displays a pattern similar to what you may see in the actual data? Why?



(**Answer:** The middle graph. Intervene, if students pick another graph. Get them to talk about what they are seeing. Do they know that fiber is a good thing? Why would you expect some variability in ratings for cereals with the same amount of fiber?)

(9) Suppose that carbohydrates are not used in the *Consumer Reports* rating formula. Sketch a scatterplot with make-believe data from 10 make-believe cereals to illustrate what the data might look this situation.



(**Answer:** There are many possibilities here. If students do not have 10 dots in their graph, ask them why. Are they thinking that more than one cereal has the same measurements, so that dots are on top of each other? Or is there some confusion about what a dot represents or about the instructions. Also intervene if there is a strong association between the variables shown in their graph. With a strong association you might point out that they can predict the ratings with some confidence using their graph.)

Wrap-Up/Direct Instruction About Statistical Concepts

In this wrap-up, return to the issue of cracking the code on the *Consumer Reports* ratings by determining which ingredients appear to be good predictors of ratings. The real goal is to focus on issues related to building distributional thinking in this new setting of bivariate data. By the end of this wrap-up, the following should be clear to students:

- Upward and downward trends can help them make predictions even with very noisy data.
- The more scatter (variability) in the data, the less accurate their predictions probably are.

- When there is a lot of variability in the ratings for a fixed amount of ingredient, the ingredient is not a good predictor of the rating. Other factors are affecting the rating.

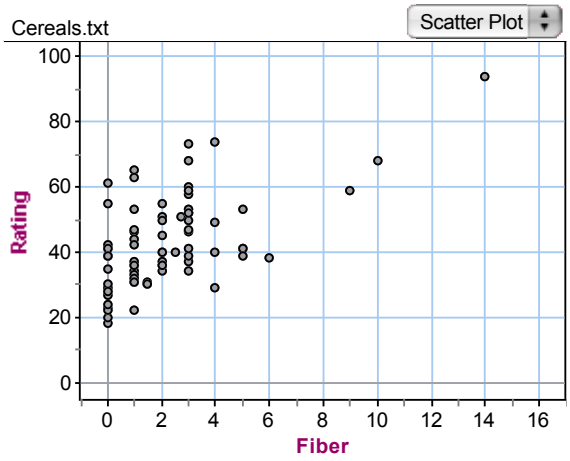
Compare and contrast fat-ratings and sugar-ratings scatterplots to address issues related to the previous activities:

Both scatterplots show a downward trend that indicates that cereals with smaller amounts of fat (or sugar) tend to have higher ratings and cereals with larger amounts of fat (or sugar) tend to have lower ratings. Draw a summarizing line with a negative slope or a highlight a downward diagonal region to highlight the trend. The downward trends indicate that increasing the amount of fat (or sugar) in a serving will tend to decrease the *Consumer Reports* rating.

The downward trends suggest that both fat and sugar are used in the *Consumer Reports* rating, but sugar is a better predictor of the ratings than fat because there is less variability in the ratings for similar cereals. Here you are reasoning in the same way that you did previously. For example, when you look at cereals with 1 gram of fat in a serving, you see a wide range of ratings falling between 20 and 70. You tend to see a similar wide range of ratings when you compare cereals with the same amount of fat. If you compare cereals with the same amount of sugar, you usually see less variability in their ratings. For example, cereals with 6 grams of sugar in a serving have ratings that differ, but the range of variability is less (about 35 to 65). The variability that we have been discussing is due to the impact of other ingredients on the rating.

Discussion of the Fiber-Ratings Relationship

Poll student responses to fiber-ratings problem (Question 8). Look at actual fiber data and compare and contrast it with the hypothetical data pattern that the majority chose.



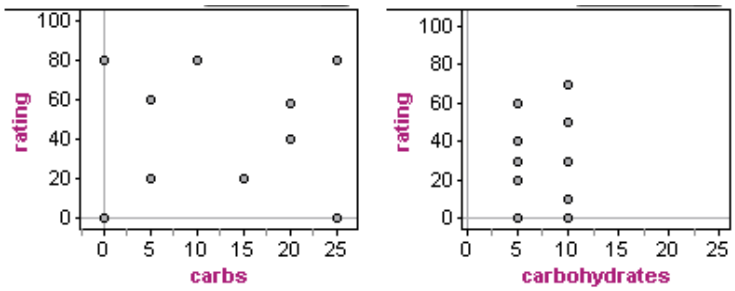
In what ways does the scatterplot of the real data look like what they expected? You might have expected the upward trend in the data since smaller amounts of fiber would probably result in lower ratings and larger amounts of fiber in higher ratings. Highlight this upward trend with a line or a diagonal region.

In what ways is this graph surprising?
Again, you see a surprising amount of variability in the ratings for cereals with the same amount of fiber. This suggests that fiber is used in the rating formula but does not have as strong an impact as sugar. You might also be surprised that there are only a few cereals with more than 6 grams of fiber in a serving. These cereals have pretty high ratings.

Initiating Lesson 3.1.1: Introduction to Scatterplots and Bivariate Relationships

Discussion of Graphs Where You See No Relationship Between Carbohydrates and Ratings

Here are two graphs that illustrate no relationship between carbohydrates and ratings. In the graph on the left, there is no discernable pattern that can be used to make predictions. In the graph on the right, there is a vertical stripe pattern, but knowing the amount of carbohydrate does not help predict the rating.



Initiating Lesson 3.1.1: Introduction to Scatterplots and Bivariate Relationships

Part III: Homework [Student Handout]

(10) Summarize what you feel you learned today.

(11) The average *Consumer Reports* rating for these 77 cereals is 44. What is the largest amount of sugar per serving in a cereal that has above average ratings?

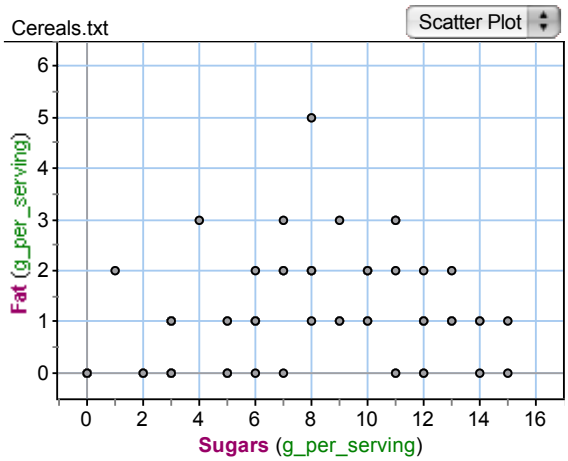
(**Answer:** 7 grams per serving.)

(12) Which is a better predictor of the Consumer Report ratings: sugar or sodium? Explain how the scatterplots support your answer.

(**Answer:** Sugar is a better predictor of ratings. There is a clear downward trend that makes it easier to predict the ratings based on a given amount of sugar. The sodium-ratings scatterplot has a lot of variability in ratings for cereals with similar amounts of sodium.)

(13) A friend says that she only pays attention to sugar amounts, even though she is also concerned by fat. Her reasoning is that low levels of sugar signal that the food also has low amounts of fat. Similarly, high levels of sugar signal that the food also has high amounts of fat. Does this appear to be true for breakfast cereals? Explain how the scatterplot supports your answer.

(**Answer:** This is not true for the breakfast cereals in this data set. The pattern described is an upward trend, but this graph does not have an upward trend. In this graph the cereals with highest amounts of sugar [14–16 grams in a serving], have low amounts of fat, 0–1 grams in a serving.)



Initiating Lesson 3.1.1: Introduction to Scatterplots and Bivariate Relationships

In the United States, the Food and Drug Administration requires that most prepared food be labeled with nutrition information. Today part of the data set for the lesson was gathered from the list of required nutrition facts that appear on boxes of breakfast cereals.

Below are the nutrition facts for three cereals. There is a lot of information given. An entire math lesson could probably be designed around reading and understanding the information given on cereal labels. For now though, you will look over the labels fairly quickly, like you might do in the grocery store. Use your initial impressions to pick the cereal that you would rate as the most nutritious (the best for your health). You have a few minutes to do this.

Nutrition Facts
Serving Size 1.0 cup (1 NLEA serving)

Amount Per Serving	
Calories 117	Calories from Fat 9
	% Daily Value *
Total Fat 1g	1%
Saturated Fat 0g	1%
Polyunsaturated Fat 0g	
Monounsaturated Fat 1g	
Cholesterol 0mg	0%
Sodium 171mg	7%
Total Carbohydrate 26g	9%
Dietary Fiber 1g	3%
Sugars 14g	
Protein 1g	

Vitamin A 0% • Vitamin C 10%
Calcium 10% • Iron 30%

* Percent Daily Values are based on a 2,000 calorie diet. Your daily value may be higher or lower depending on your calorie needs:

	Calories:	2,000	2,500
Total Fat	Less than	65g	80g
Sat Fat	Less than	20g	25g
Cholesterol	Less than	300mg	300mg
Sodium	Less than	2,400mg	2,400mg
Total Carbohydrate		300g	375g
Dietary Fiber		25g	30g

Calories per gram:
Fat 9 • Carbohydrate 4 • Protein 4

Nutrition Facts
Serving Size 1.0 cup (1 NLEA serving)

Amount Per Serving	
Calories 101	Calories from Fat 1
	% Daily Value *
Total Fat 0g	0%
Saturated Fat 0g	0%
Polyunsaturated Fat 0g	
Monounsaturated Fat 0g	
Cholesterol 0mg	0%
Sodium 202mg	8%
Total Carbohydrate 24g	8%
Dietary Fiber 1g	3%
Sugars 3g	
Protein 2g	

Vitamin A 0% • Vitamin C 10%
Calcium 0% • Iron 54%

* Percent Daily Values are based on a 2,000 calorie diet. Your daily value may be higher or lower depending on your calorie needs:

	Calories:	2,000	2,500
Total Fat	Less than	65g	80g
Sat Fat	Less than	20g	25g
Cholesterol	Less than	300mg	300mg
Sodium	Less than	2,400mg	2,400mg
Total Carbohydrate		300g	375g
Dietary Fiber		25g	30g

Calories per gram:
Fat 9 • Carbohydrate 4 • Protein 4

Nutrition Facts
Serving Size 1.0 cup (1 NLEA serving)

Amount Per Serving	
Calories 111	Calories from Fat 16
	% Daily Value *
Total Fat 2g	3%
Saturated Fat 0g	2%
Polyunsaturated Fat 1g	
Monounsaturated Fat 1g	
Cholesterol 0mg	0%
Sodium 213mg	9%
Total Carbohydrate 22g	7%
Dietary Fiber 4g	14%
Sugars 1g	
Protein 4g	

Vitamin A 0% • Vitamin C 10%
Calcium 12% • Iron 69%

* Percent Daily Values are based on a 2,000 calorie diet. Your daily value may be higher or lower depending on your calorie needs:

	Calories:	2,000	2,500
Total Fat	Less than	65g	80g
Sat Fat	Less than	20g	25g
Cholesterol	Less than	300mg	300mg
Sodium	Less than	2,400mg	2,400mg
Total Carbohydrate		300g	375g
Dietary Fiber		25g	30g

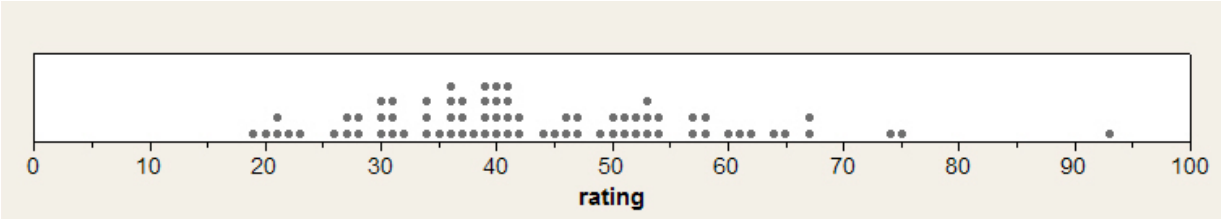
Calories per gram:
Fat 9 • Carbohydrate 4 • Protein 4

Part I

Instead of making up your own rating system, you are going to investigate the *Consumer Reports* nutritional ratings for 77 breakfast cereals. *Consumer Reports* is published by a nonprofit organization called the Consumers Union, whose mission it is to work for a fair, just, and safe marketplace for all consumers and to empower consumers to protect themselves. *Consumer Reports* rates products based on its own criteria and testing. It prides itself on producing objective results. *Consumer Reports* maintain its objectivity by not allowing advertising within their publications and not allowing use of their results for commercial gain. (Retrieved from www.consumerreports.org/cro/aboutus/mission/overview/index.htm)

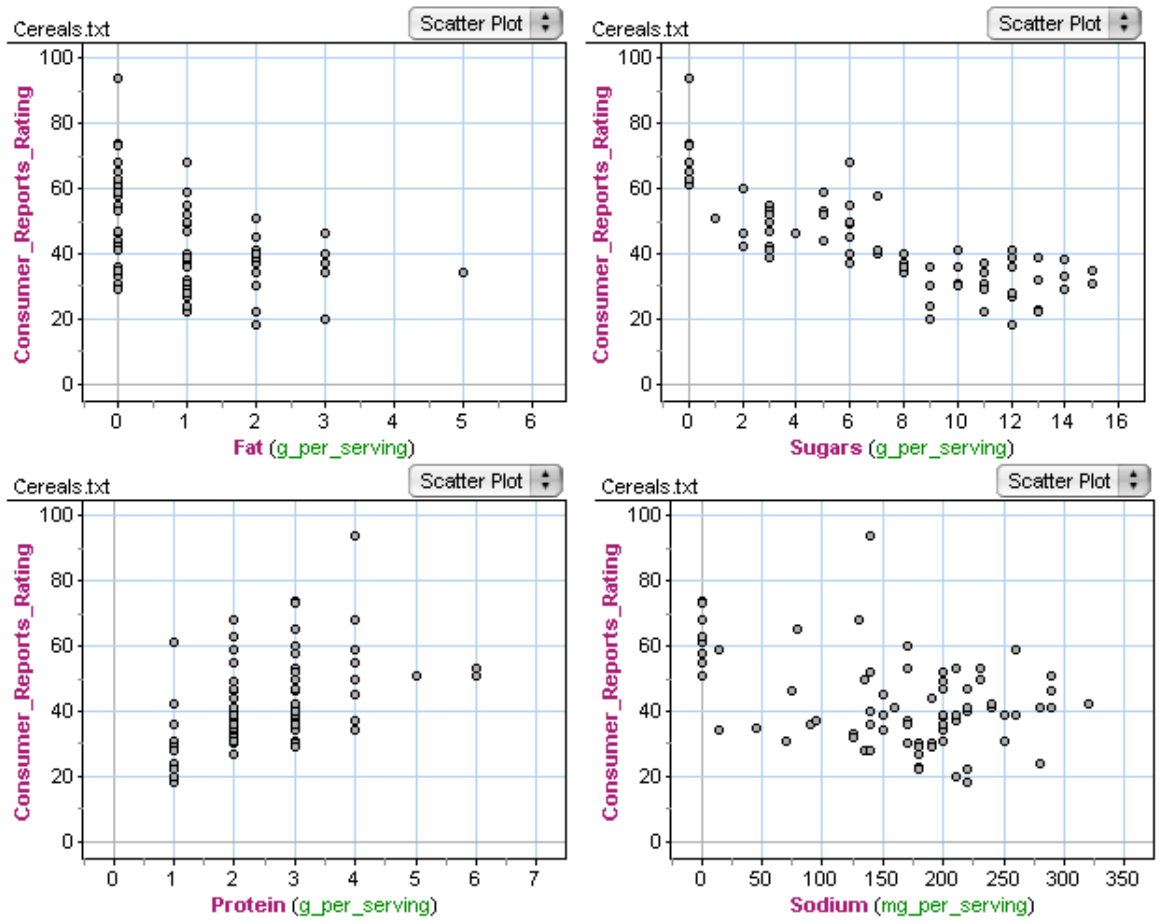
Consumer Reports uses a rating system with a scale of 0 to 100. Here is the distribution of *Consumer Reports* ratings for 77 cereals:

Initiating Lesson 3.1.1: Introduction to Scatterplots and Bivariate Relationships



- What does each dot represent in this distribution?
- For this distribution, what seems to be an average rating?
- How would you describe the variability in ratings?
- How would you describe the shape of this distribution? What does the shape suggest about the rating system?

What you cannot tell from the dotplot is how the cereal ingredients (such as sugar or fat) are related to the ratings. You need a new type of graph, called a scatterplot, to investigate how two variables relate to each other. The scatterplots below show the amount of an ingredient in a serving of cereal and the *Consumer Reports* rating for 77 breakfast cereals.



Initiating Lesson 3.1.1: Introduction to Scatterplots and Bivariate Relationships

The *Consumer Reports* rating formula is not made public. So, you do not know which ingredients are used in its rating formula. In this lesson, you will try to "crack their code" in a sense. Use the data to figure out which ingredients *Consumer Reports* may, or may not, use in their rating formula. The only clues you have are these scatterplots.

The first step in this investigation is to answer the following two questions AND write down enough of your reasoning that someone can follow your thinking.

Two new cereals are being rated by Consumer Reports*. Cereal A has 10.5 grams of sugar in a serving and Cereal B has 2.5 grams of protein in a serving.*

(1) Based on the data shown, predict the *Consumer Reports* rating for the two cereals.

(2) For which cereal do you think your prediction is probably more accurate (more likely to be closer to the actual Consumer Report rating)? Why?

Initiating Lesson 3.1.1: Introduction to Scatterplots and Bivariate Relationships

Part II: Scaffolded Conceptual Tasks

Task 1: Reading and Interpreting Scatterplots

In this task, you are going to take a short detour from your investigation into which ingredients are the best predictors of *Consumer Reports* ratings. Here, you will work on interpreting scatterplots just to make sure everyone is comfortable with reading this new type of graph.

(3) Captain Crunch has the lowest *Consumer Reports* rating of the 77 cereals in the data set. How much fat is in a serving of Captain Crunch?

(4) In this set of 77 cereals, Product 19 has the most sodium in a serving. What is the rating for Product 19?

(5) All-Bran Extra Fiber is the cereal with the highest rating. How much sugar, fat, and sodium are in a serving of All-Bran Extra Fiber?

Initiating Lesson 3.1.1: Introduction to Scatterplots and Bivariate Relationships

Conceptual Task 2: Seeing Patterns and Relationships in Scatterplots

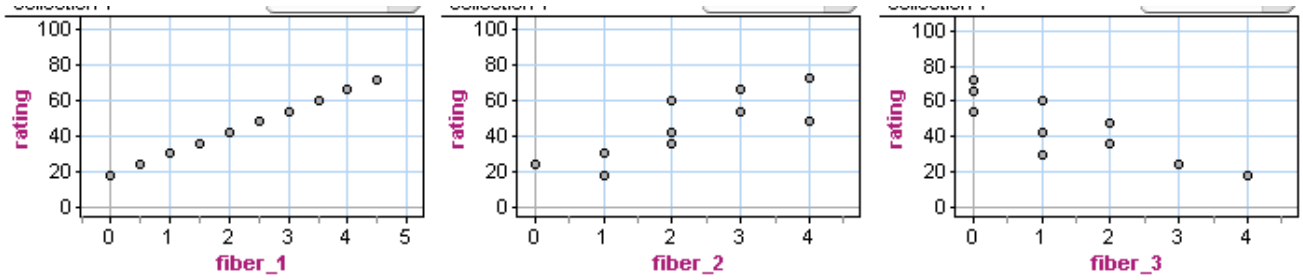
Now you will continue your detective work with *Consumer Reports* ratings. Try to identify ingredients that are good predictors of ratings and ingredients that are not. More importantly, focus on how patterns in the data are related to identifying ingredients that are good predictors.

(6) There are four cereals that have 3 grams of fat in a serving. Estimate the ratings for these four cereals. What might explain the variability in the ratings?

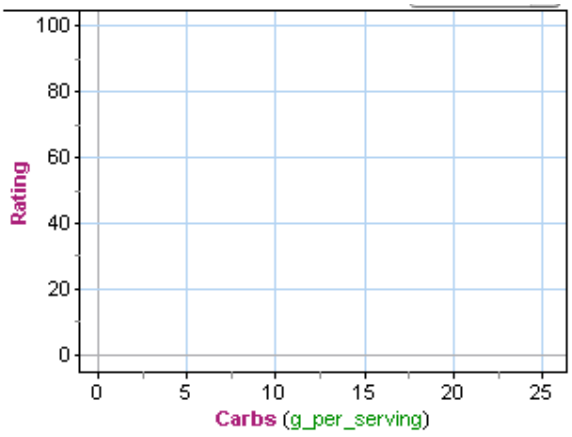
(7) Imagine changing the recipe for a cereal that has 0 grams of fat in a serving and a rating of 60. Increase the amount of fat to 3 grams in a serving. Do you think the rating will probably increase or decrease or remain about the same? Or do you think that it is impossible to use the scatterplot to predict the impact of this change on the rating? How does the pattern in the data support your decision?

Initiating Lesson 3.1.1: Introduction to Scatterplots and Bivariate Relationships

(8) Think about how the amount of fiber in a cereal might relate to the *Consumer Reports* rating. Here are three scatterplots with make-believe data from 10 made-up cereals. Which scatterplot do you think displays a pattern similar to what you may see in the actual data? Why?



(9) Suppose that carbohydrates are not used in the *Consumer Reports* rating formula. Sketch a scatterplot with make-believe data from 10 make-believe cereals to illustrate what the data might look this situation.



Initiating Lesson 3.1.1: Introduction to Scatterplots and Bivariate Relationships

Part III: Homework

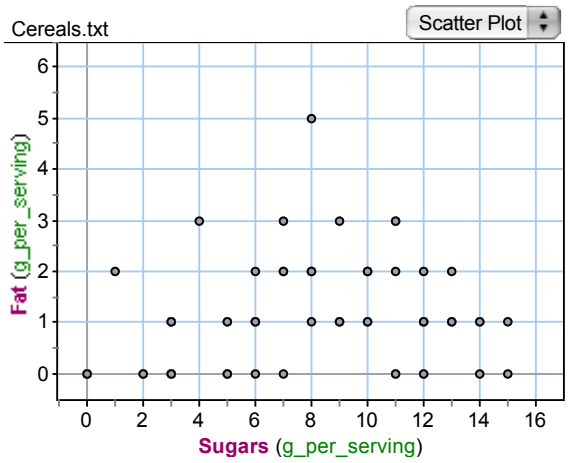
(10) Summarize what you feel you learned today.

(11) The average *Consumer Reports* rating for these 77 cereals is 44. What is the largest amount of sugar per serving in a cereal that has above average ratings?

(12) Which is a better predictor of the *Consumer Reports* ratings: sugar or sodium? Explain how the scatterplots support your answer.

Initiating Lesson 3.1.1: Introduction to Scatterplots and Bivariate Relationships

(13) A friend says that she only pays attention to sugar amounts, even though she is also concerned by fat. Her reasoning is that low levels of sugar signal that the food also has low amounts of fat. Similarly, high levels of sugar signal that the food also has high amounts of fat. Does this appear to be true for breakfast cereals? Explain how the scatterplot supports your answer.



Supporting Lesson 3.1.2: Developing an Intuitive Sense of Form, Direction, and Strength of the Relationship Between Two Measurements

Estimated number of 50-minute class sessions: 1

Learning Goals

Students will understand that

- each point on the scatterplot represents a single observation consisting of measurements on two variables.
- the overall pattern in a scatterplot can be described in terms of the direction, form, and strength of the relationship between the two measurements.
- the linear relationship between two measurements is positive if smaller values of *x* tend to correspond to smaller values of *y* and larger values of *x* tend to correspond to larger values of *y*.
- the linear relationship between two measurements is negative if smaller values of *x* tend to correspond to larger values of *y* and larger values of *x* tend to correspond to smaller values of *y*.

Students will be able to distinguish between

- linear and nonlinear relationships,
- strong and weak relationships, and
- positive and negative linear relationships.

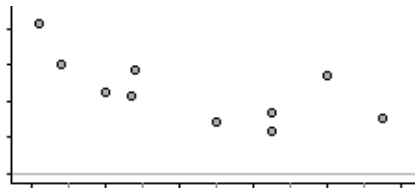
Introduction [Student Handout]

In this lesson, you will compare and contrast a variety of scatterplots with the goal of thinking about how to describe relationships you see in the data. At the end of the lesson, you will discuss ways that statisticians describe these relationships.

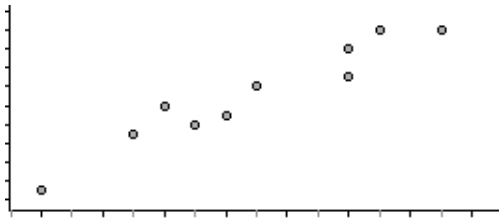
Tasks [Student Handout, 15 minutes]

(1) Match each set of measurements to a scatterplot, and briefly explain your reasoning.

Scatterplot 1



Scatterplot 2



(a) *x* = city miles per gallons and *y* = highway miles per gallon for 10 cars

(b) *x* = sodium (milligrams/serving) and *y* = *Consumer Reports* quality rating for 10 salted peanut butters

(**Answer:** Scatterplot 1: b, Scatterplot 2: a)

(2) For each scatterplot in Question 1, describe what a dot represents.

(**Answer:** Scatterplot 1: Each dot is a peanut butter. Scatterplot 2: Each dot is a car.)

Supporting Lesson 3.1.2: Developing an Intuitive Sense of Form, Direction, and Strength of the Relationship Between Two Measurements

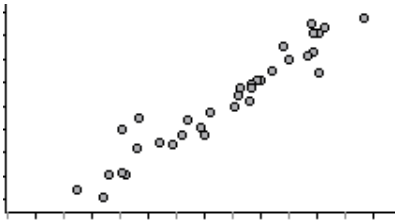
(3) These scatterplots show body measurements for 34 adults who are physically active. Some measurements are a *girth*, which is a measure of length around a body part. Match each description to a scatterplot. Briefly explain your reasoning.

(a) x = forearm girth (centimeters), y = bicep girth (cm)

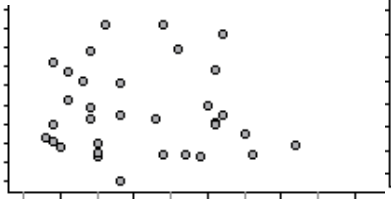
(b) x = calf girth (cm), y = bicep girth (cm)

(c) x = age (years), y = bicep girth (cm)

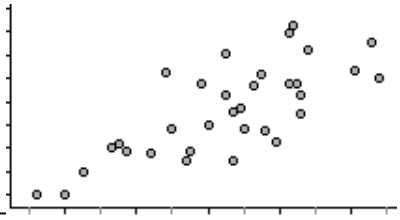
Scatterplot 1



Scatterplot 2



Scatterplot 3



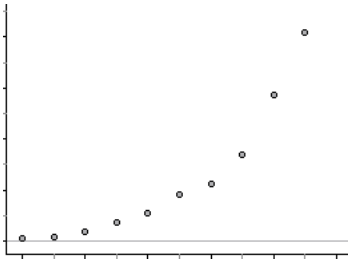
(**Answer:** Scatterplot 1: a, Scatterplot 2: c, Scatterplot 3: b)

(4) What does a dot represent in each scatterplot in Question 3?

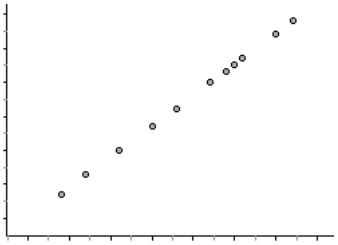
(**Answer:** In all of the scatterplots a dot represents an adult who is physically active.)

(5) Match each set of measurements to a scatterplot. Briefly explain your reasoning.

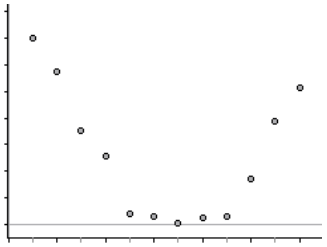
Scatterplot 1



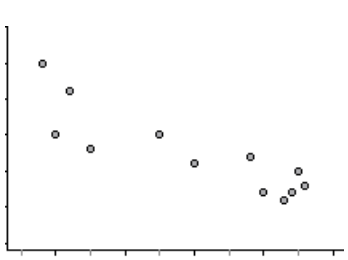
Scatterplot 2



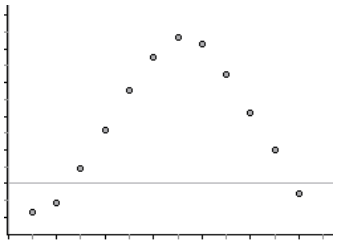
Scatterplot 3



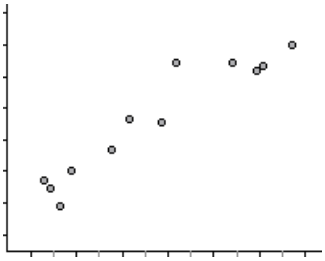
Scatterplot 4



Scatterplot 5



Scatterplot 6



(a) x = month number (January = 1) and y = rainfall (inches) in Napa, California. Napa has several months of drought each summer.

(b) x = month number (January = 1) and y = average temperature in Boston, Massachusetts. Boston has cold winters and hot summers.

(c) x = year (from 1970) in five-year increments and y = Medicare expenditures ($). The yearly increase in Medicare costs has been getting bigger over time. Costs are predicted for 2015.

(d) x = average temperature (°C) and y = average temperature (°F) each month in San Francisco, California.

(e) x = chest girth (cm) and y = shoulder girth (cm) for a sample of men.

(f) x = engine displacement (in liters) and y = city miles per gallon for a sample of cars. Engine displacement is roughly a measurement of the size of the engine. Large engines use more gas.

(**Answer:** Scatterplot 1: c, Scatterplot 2: d, Scatterplot 3: a, Scatterplot 4: f, Scatterplot 5: b, Scatterplot 6: e)

(6) What does a dot represent in each scatterplot in Question 5?

(**Answer:** Scatterplot 1: Each dot is a year. Scatterplot 2: Each dot is a month. Scatterplot 3: Each dot is a month. Scatterplot 4: Each dot is a car. Scatterplot 5: Each dot is a month. Scatterplot 6: Each dot is a man.)

Wrap-Up Questions/Direct Instruction About Statistical Concepts [20 minutes]

Highlight the following concepts through discussion or lecture.

Linear vs. Nonlinear Form

Some relationships between the measurements in these scatterplots can be summarized well by a line. Other relationships clearly have a nonlinear form. In Question 5, identify the scatterplots for which a line is a good summary of the relationship in the data. Sketch a line on each scatterplot to illustrate the linear form of the data. In Question 5, sketch curves to summarize the relationships that are nonlinear. Tell students that in Module 12, they will develop mathematical models or formulas with curves that are similar to those you drew. For now, you just want to be able to determine whether a line is a good summary of the relationship in the data.

Strong vs. Weak Relationships

Some relationships between measurements are strong and others are weak. If the relationship is strong, the line or the curve does a good job summarizing the relationship between the measurements. This means there is less scatter in the data about the line or curve. In Question 3, sketch a line to summarize the relationship in each scatterplot. If students are concerned that everyone has a different line, reassure them that in future lessons you will discuss how statisticians determine the best summary line. For now, encourage them to just sketch a line that cuts through the cloud of data in a way that illustrates the linear form of the relationship.

Supporting Lesson 3.1.2: Developing an Intuitive Sense of Form, Direction, and Strength of the Relationship Between Two Measurements

In Question 3, which is the weakest and which is the strongest linear relationship? In Question 5, which scatterplot shows the strongest linear relationship? In Question 5, do the scatterplots with nonlinear relationships have a strong or weak association between the variables?

Positive vs. Negative Linear Relationships

You will spend some time discussing linear relationships in future lessons in this module. For now, you want to focus in general on the fact that a linear relationship has direction. When a linear relationship is positive, smaller values of x tend to correspond to smaller values of *y* and larger values of *x* tend to correspond to larger values of *y*. In Question 1, which scatterplot shows a positive association?

When a linear relationship is negative, smaller values of *x* tend to correspond to larger values of *y* and larger values of *x* tend to correspond to smaller values of *y*. In Question 5, which scatterplot shows a negative association?

(**Note:** If you have time after the wrap-up, get students started on the second homework problem. It is open-ended and requires students to think more precisely about a variable as a measurement, so they will benefit from discussing this problem with their peers.)

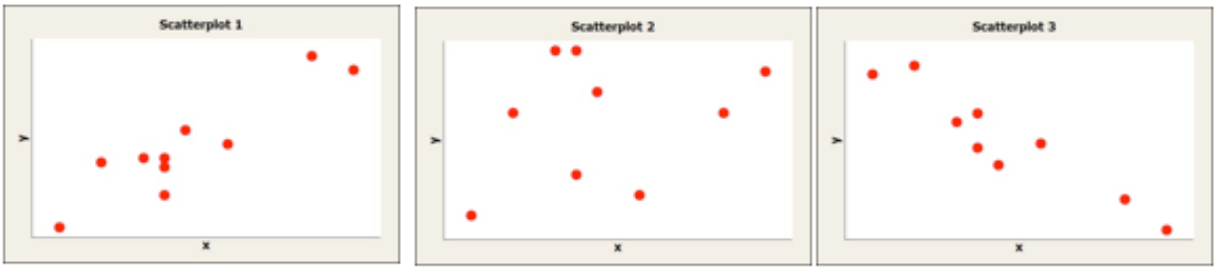
Homework [Student Handout]

(7) Match each set of measurements to a scatterplot. Then describe what a dot represents in each graph.

(a) *x* = average outdoor temperature and *y* = heating costs for a residence for 10 winter days

(b) *x* = height (inches) and *y* = shoe size for 10 adults

(c) *x* = height (inches) and *y* = IQ for 10 teenagers



(**Answer:** a: Scatterplot 3, b: Scatterplot 1, c: Scatterplot 2)

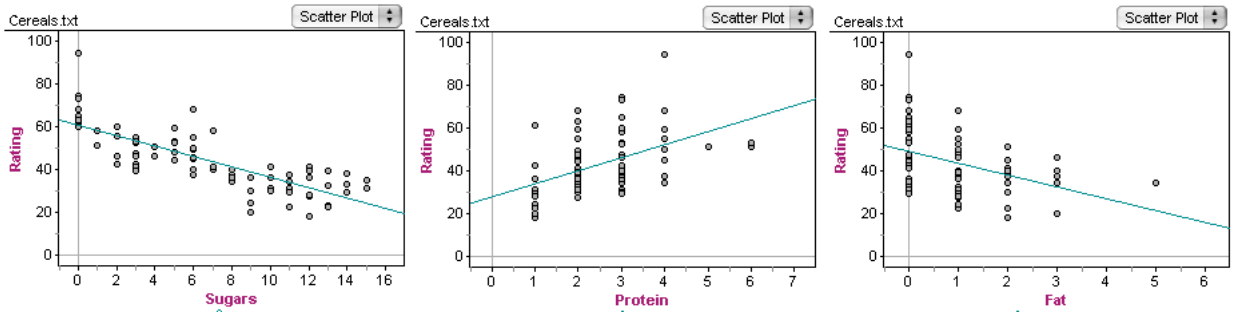
(8) Lines have been added to some of the scatterplots used in the Lesson 3.1.1 to summarize the relationship between the ingredient and the *Consumer Reports* rating for breakfast cereals. You will learn more about summary lines in future lessons.

(a) Which ingredients (sugar, protein, and/or fat) are negatively associated with ratings? (**Answer:** sugar, fat)

Supporting Lesson 3.1.2: Developing an Intuitive Sense of Form, Direction, and Strength of the Relationship Between Two Measurements

(b) Which is more strongly associated with ratings: sugar or fat? (**Answer:** sugar)

(c) How is the idea of strength related to whether an ingredient is a good predictor of ratings? (**Answer:** The stronger the relationship, the less scatter there is about the line [or curve]. You get better predictions when there is a stronger relationship.)



(9) Suppose you gathered the following information from students at a local high school:

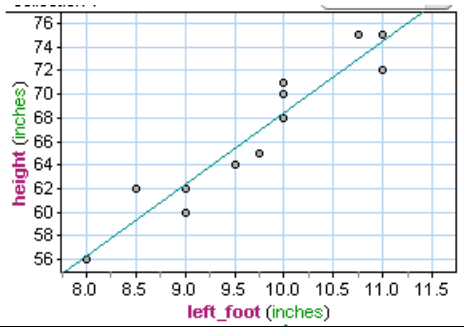
- GPA (grade point average),
- average weekly hours spent working at a job,
- average weekly hours spent doing homework,
- average hours of sleep a night,
- hourly wage,
- height,
- weight,
- length of the left foot,
- age of the oldest child in the student's immediate family,
- number of children in the student's immediate family,
- gender,
- race, and
- age.

(a) From the list, choose two variables that you think will show a positive linear association, two variables you think will show a negative linear association, and two variables you think will not show an association in a scatterplot. You may reuse variables. Briefly explain your reasoning.

(b) Sketch a scatterplot with 12 students to illustrate each of the three relationships. You will have three scatterplots. If there is an association, sketch a line to highlight the association. For each scatterplot, label the axes of each graph with the name of the variable. Scale the graph with realistic numbers for the variable.

(**Answers will vary.** The authors suggest that that you provide one possible solution to part of this problem and discuss its merit with students. There are many issues that will arise as students attempt this problem that have not been directly discussed until this point. For example, some variables are categorical. Scatterplots look at the relationship between quantitative variables. This is

an important point that can be underscored in the class discussion. Another issue that may arise is which variable to put on the *x*-axis. Students will not know to distinguish between explanatory and response variables, so they may generate graphs in which the explanatory variable is inappropriately put on the *y*-axis. At this point in their learning, this is acceptable. Later, you will discuss this point explicitly. Right now, you want students to focus on reading scatterplots that they and their peers create and thinking about whether the measurements and patterns make sense. After the class discussion of the solution, let students discuss their solutions with each other, so that their discussions are informed by the example you provided. If you want to grade the problem, consider letting students revise their initial solutions based on the class and group discussions.)

Sample Solution	What makes this a good answer?
I think there will be a positive linear association between length of the left foot and height. Students with shorter feet tend to be shorter in height and students with longer feet tend to be taller. This might happen because boys and girls are mixed together in the data, and boys tend to be bigger. Also freshmen and seniors could be mixed together, and older students may be bigger.	The answer tells us the two variables and the type of association expected. The explanation says what positive association means for these two variables. An extra special aspect of this answer is the inclusion of the gender and age issues to help the reader understand why the association might be positive for high school students. This is above and beyond what you might expect in an answer, but it shows good thinking about who is described by the data.
	The data has an upward trend, and you see a line that highlights the positive linear association. The measurements are reasonable for the variables. Both axes are labeled with the name of the variable, and units are given. Scales are clearly marked and consistent across the axis. (Students will have hand-drawn sketches, which is fine.)

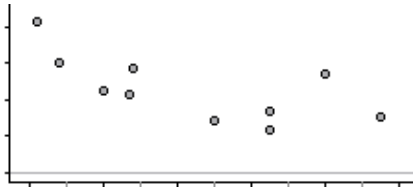
Supporting Lesson 3.1.2: Developing an Intuitive Sense of Form, Direction, and Strength of the Relationship Between Two Measurements

In this lesson, you will compare and contrast a variety of scatterplots with the goal of thinking about how to describe relationships you see in the data. At the end of the lesson, you will discuss ways that statisticians describe these relationships.

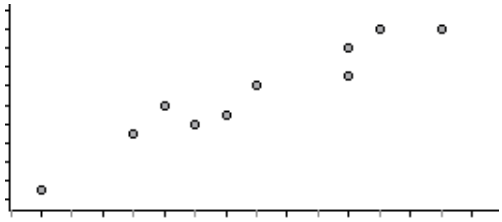
Tasks

(1) Match each set of measurements to a scatterplot, and briefly explain your reasoning.

Scatterplot 1



Scatterplot 2



(a) x = city miles per gallons and y = highway miles per gallon for 10 cars

(b) x = sodium (milligrams/serving) and y = *Consumer Reports* quality rating for 10 salted peanut butters

(2) For each scatterplot in Question 1, describe what a dot represents.

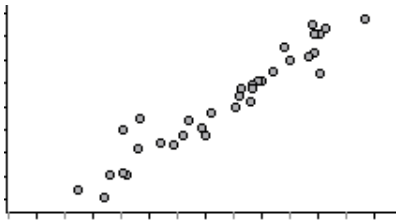
(3) These scatterplots show body measurements for 34 adults who are physically active. Some measurements are a *girth*, which is a measure of length around a body part. Match each description to a scatterplot. Briefly explain your reasoning.

(a) x = forearm girth (centimeters), y = bicep girth (cm)

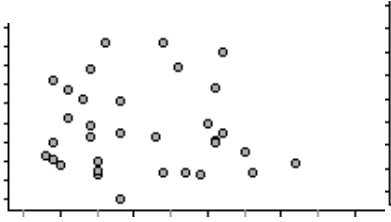
(b) x = calf girth (cm), y = bicep girth (cm)

(c) x = age (years), y = bicep girth (cm)

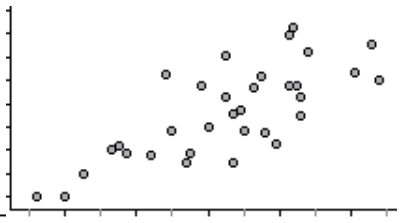
Scatterplot 1



Scatterplot 2



Scatterplot 3

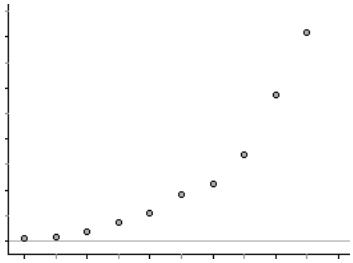


Supporting Lesson 3.1.2: Developing an Intuitive Sense of Form, Direction, and Strength of the Relationship Between Two Measurements

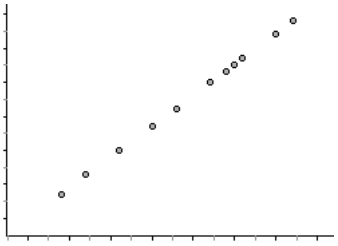
(4) What does a dot represent in each scatterplot in Question 3?

(5) Match each set of measurements to a scatterplot. Briefly explain your reasoning.

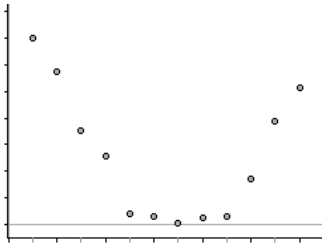
Scatterplot 1



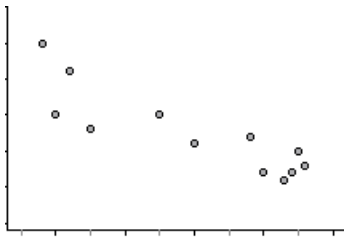
Scatterplot 2



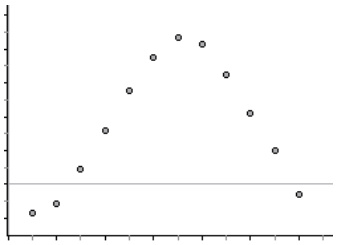
Scatterplot 3



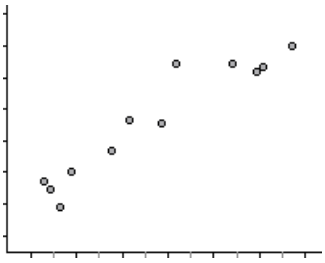
Scatterplot 4



Scatterplot 5



Scatterplot 6



(a) x = month number (January = 1) and y = rainfall (inches) in Napa, California. Napa has several months of drought each summer.

(b) x = month number (January = 1) and y = average temperature in Boston, Massachusetts. Boston has cold winters and hot summers.

(c) x = year (from 1970) in five-year increments and y = Medicare expenditures ($). The yearly increase in Medicare costs has been getting bigger over time. Costs are predicted for 2015.

(d) x = average temperature (°C) and y = average temperature (°F) each month in San Francisco, California.

(e) x = chest girth (cm) and y = shoulder girth (cm) for a sample of men.

(f) x = engine displacement (in liters) and y = city miles per gallon for a sample of cars. Engine displacement is roughly a measurement of the size of the engine. Large engines use more gas.

(6) What does a dot represent in each scatterplot in Question 5?

Supporting Lesson 3.1.2: Developing an Intuitive Sense of Form, Direction, and Strength of the Relationship Between Two Measurements

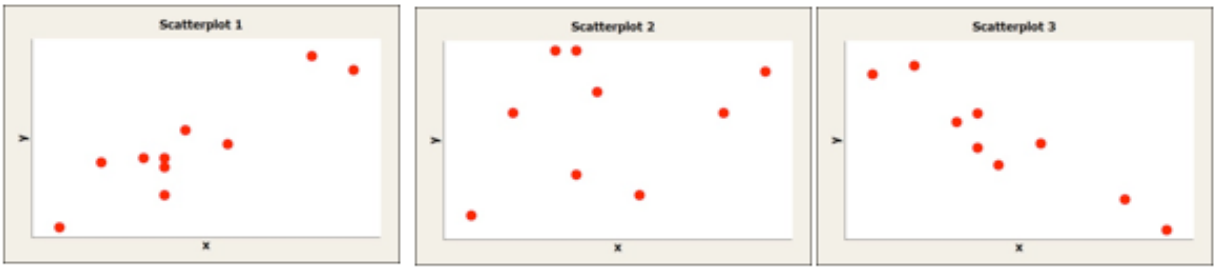
Homework

(7) Match each set of measurements to a scatterplot. Then describe what a dot represents in each graph.

(a) x = average outdoor temperature and y = heating costs for a residence for 10 winter days

(b) x = height (inches) and y = shoe size for 10 adults

(c) x = height (inches) and y = IQ for 10 teenagers

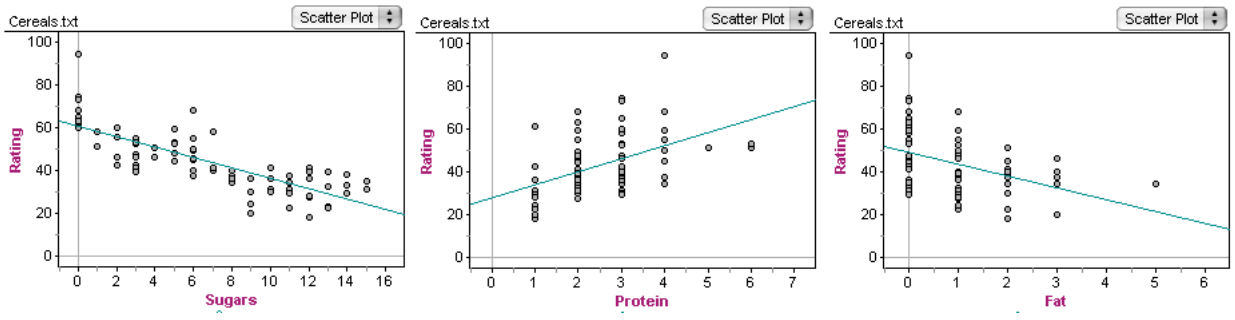


(8) Lines have been added to some of the scatterplots used in the Lesson 3.1.1 to summarize the relationship between the ingredient and the *Consumer Reports* rating for breakfast cereals. You will learn more about summary lines in future lessons.

(a) Which ingredients (sugar, protein, and/or fat) are negatively associated with ratings?

(b) Which is more strongly associated with ratings: sugar or fat?

(c) How is the idea of strength related to whether an ingredient is a good predictor of ratings?



Supporting Lesson 3.1.2: Developing an Intuitive Sense of Form, Direction, and Strength of the Relationship Between Two Measurements

(9) Suppose you gathered the following information from students at a local high school:

- GPA (grade point average),
- average weekly hours spent working at a job,
- average weekly hours spent doing homework,
- average hours of sleep a night,
- hourly wage,
- height,
- weight,
- length of the left foot,
- age of the oldest child in the student's immediate family,
- number of children in the student's immediate family,
- gender,
- race, and
- age.

(a) From the list, choose two variables that you think will show a positive linear association, two variables you think will show a negative linear association, and two variables you think will not show an association in a scatterplot. You may reuse variables. Briefly explain your reasoning.

(b) Sketch a scatterplot with 12 students to illustrate each of the three relationships. You will have three scatterplots. If there is an association, sketch a line to highlight the association. For each scatterplot, label the axes of each graph with the name of the variable. Scale the graph with realistic numbers for the variable.

Supporting Lesson 3.1.3: Introduction to the Correlation Coefficient and Its Properties

Estimated number of 50-minute class sessions: 1–1.5 (depending on technology use)

Learning Goals

Students will understand that

- the correlation coefficient is a numerical measure of the strength of a *linear* relationship with a value *r* where $-1 \le r \le 1$. The sign of *r* gives information about the direction of the linear relationship.
- if a relationship is nonlinear, the value of *r* does not give information about form or strength.

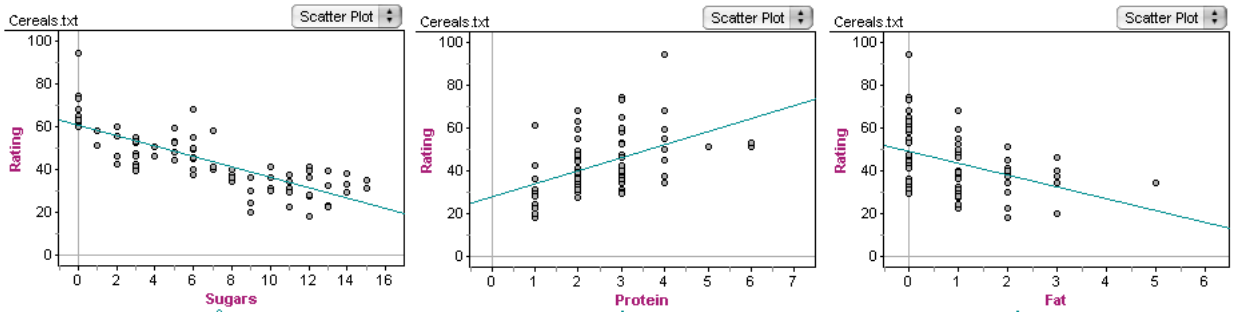
Students will be able to

- assess the strength and direction of a linear relationship using the correlation coefficient.

Introduction [Student Handout, about 5 minutes]

(**Note**: Students applied the concepts of form, direction, and strength to the cereal data in a homework problem from Lesson 3.1.2.)

When you were trying to determine which cereal ingredients influenced the *Consumer Reports* ratings, you saw a lot of variability in the data. This variability made it difficult to predict ratings using some of the ingredients. Using a line to summarize the relationship between the ingredient and the rating makes it easier to see positive and negative association in the data. Sugar and fat are negatively associated with ratings. Protein is positively associated with ratings.



A line can also help you evaluate the strength of the relationship between the ingredient and the rating. Some ingredients are more strongly related to the rating, like sugar. You see this in a linear pattern with not much scatter around a line. A small amount of scatter around the line means that the ingredient is a good predictor of the rating. For other ingredients, like fat, the relationship with rating was not as strong. You see this as more scatter about the line. This means that fat is not as good of a predictor of ratings.

In this lesson, you will investigate a measurement of variability in scatterplots called the *correlation coefficient*, which is denoted with letter *r*. In the next lesson, you will work with the formula that is used to compute the correlation coefficient, but in this lesson you will use technology to compute the value of *r.* The goal in this lesson is to determine the properties of the statistic *r.*

Supporting Lesson 3.1.3: Introduction to the Correlation Coefficient and Its Properties

Task 1: Investigating the Properties of *r* [Student Handout, about 15 minutes, longer if you are teaching students to use technology]

(**Note:** If students have access to technology in class, demonstrate how to find the correlation coefficient for a simple data set before students do this activity. If students do not have access to technology, provide the *r*-values for each scatterplot.)

Activities

(1) Which of the nine graphs on the following page show a positive association between *x* and *y*? A negative association? No association?

(**Note:** At this point, students should be able to do this easily. If this is not the case, review the concept of positive and negative association with individuals as necessary and correct mistakes.)

(2) Use technology to find the correlation coefficient (*r*) for each graph. The data are given in the table following the scatterplots. (If you do not have access to technology in class, your instructor will provide the *r*-value for each scatterplot.)

(3) Now look for patterns by comparing scatterplots and *r*-values. How does the value of *r* seem to be related to the patterns you see in the scatterplots?

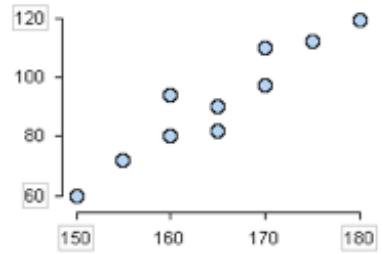
(**Note:** Let students grapple here. Most will notice that scatterplots with positive association have positive *r*-values and scatterplots with negative association have negative *r*-values. If they are not making this connection, nudge them with questions like, "Some of the *r*-values are positive, and some are negative. What do you think that the sign of *r* tells you?" Many will also notice that the closer the data are to 1 or –1, the closer the data are to a perfect linear relationship. If students are not making this connection, suggest that they put the *r*-values in order from smallest to largest, and then look for more patterns. Intervene if the entire group is making mistakes with ordering decimal numbers and not self-correcting.)

(4) Prepare for the class discussion by discussing the following questions in your group:

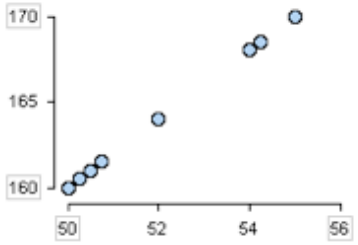
- What do you think the correlation coefficient (*r*) measures?
- Is there a largest possible value for *r,* or can it have larger and larger values without limit? What makes you think so?
- Is there a smallest possible value for *r,* or can it have smaller and smaller values without limit? What makes you think so?

(**Note:** Let students grapple here and do not intervene. Wrap-up is focused on these questions.)

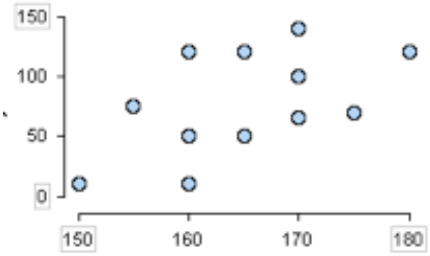
Supporting Lesson 3.1.3: Introduction to the Correlation Coefficient and Its Properties



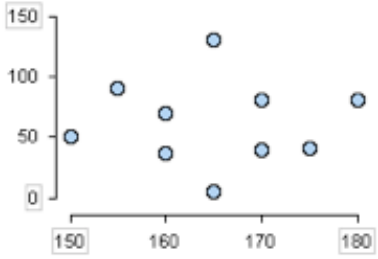
Scatterplot 1



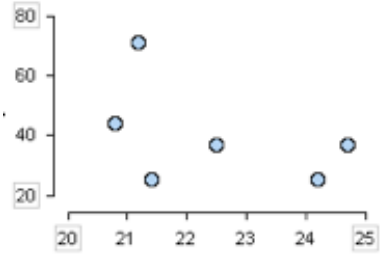
Scatterplot 2



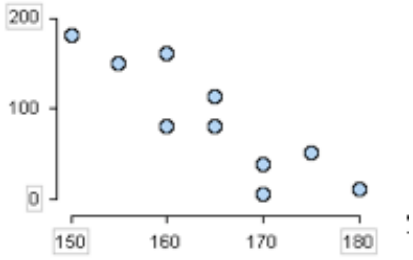
Scatterplot 3



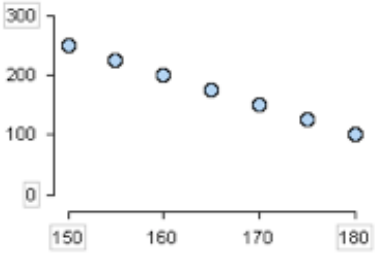
Scatterplot 4



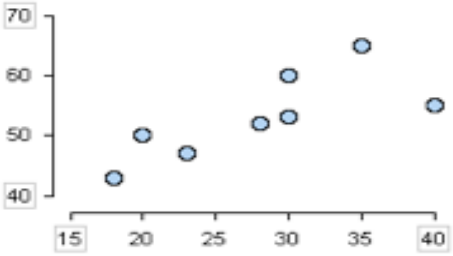
Scatterplot 5



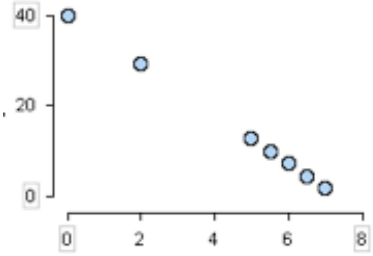
Scatterplot 6



Scatterplot 7



Scatterplot 8



Scatterplot 9

Supporting Lesson 3.1.3: Introduction to the Correlation Coefficient and Its Properties

Data for Calculating the Correlation Coefficient[1]

*x*1	*y*1	*x*2	*y*2	*x*3	*y*3	*x*4	*y*4	*x*5	*y*5
150	60	50.00	160.0	150	10	150	50	20.8	44
155	72	50.25	160.5	155	75	155	90	21.4	25
160	94	50.50	161.0	160	10	160	36	21.2	71
160	80	50.75	161.5	165	50	160	70	24.2	25
165	82	52.00	164.0	170	140	165	5	24.7	37
165	90	54.00	168.0	175	70	165	130	22.5	37
170	97	54.25	168.5	180	120	170	39		
170	110	55.00	170.0	165	120	170	80		
175	112			170	65	175	40		
180	119			160	120	180	80		
				170	100				
				160	50				

*x*6	*y*6	*x*7	*y*7	*x*8	*y*8	*x*9	*y*9
150	180	150	250	35	65	0.0	40.00
155	150	155	225	20	50	2.0	29.00
160	160	160	200	30	60	5.0	12.50
160	80	165	175	40	55	6.0	7.00
165	80	170	150	23	47	5.5	9.75
165	112	175	125	30	53	6.5	4.25
170	38	180	100	18	43	7.0	1.50
170	5			28	52		
175	50						
180	10						

[1]Scatterplot 1 is *x*1 versus *y*1, Scatterplot 2 is *x*2 versus *y*2, and so on.

Wrap-Up Questions/Direction Instruction About Statistical Concepts [about 15 minutes]

Facilitate a discussion of the questions in Question 4 that students discussed in their groups or do a minilecture on the properties of *r.*

The following are properties of the correlation coefficient *r* :

- *r* measures the direction and strength of a linear association.
- *r* is positive if there is a positive linear association and negative if there is a negative linear association.
- $-1 \le r \le 1$.
- The association is perfectly linear when $r = 1$ or $r = -1$. The closer *r* is to 1 or –1, the closer the data are to having a perfect linear association (assuming, of course, that you see linear association in the scatterplot).

Supporting Lesson 3.1.3: Introduction to the Correlation Coefficient and Its Properties

Go back to the cereal scatterplots shown at the beginning of the lesson. Tell students that the *r*-values for these three scatterplots are –0.76, –0.40, and 0.48. Ask them to match the *r*-values to the scatterplots.

You can also use the movie at www.causeweb.org/repository/statjava/CorrMovieApplet.html to illustrate how correlation relates to direction and strength of the linear relationship. As the movie plays, call out the correlation coefficient values.

Task 2: Linear Correlation with Nonlinear Scatterplots [Student Handout, about 5 minutes]

(5) Based on your current experience with *r*, answer the following two questions:

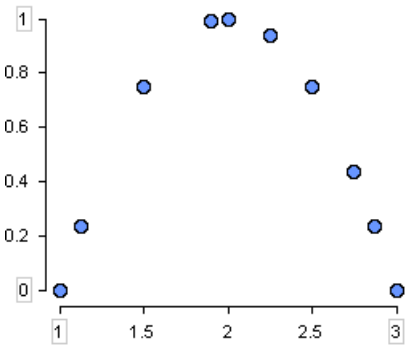
(a) If *r* is close to zero, can there be a strong relationship between the variables?

(b) If *r* is close to one, can the relationship between the variables be nonlinear?

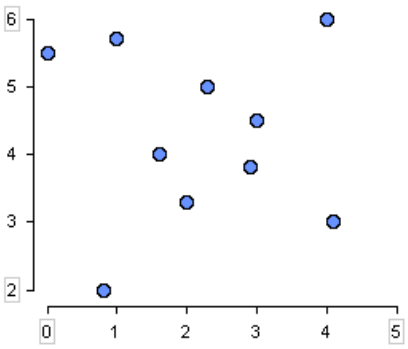
(6) Examine the following scatterplots.

(a) Two of these scatterplots have an *r*-value close to 0. Determine which two without calculating the *r*-value.

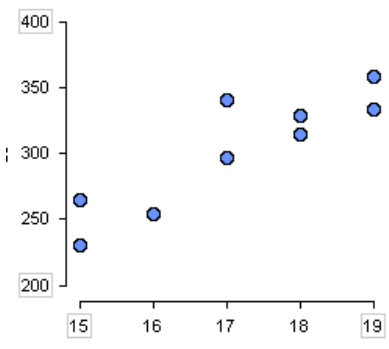
(b) Does this activity make you rethink your answer to one of the two questions from Question 5? Explain your reasoning.



Scatterplot 1



Scatterplot 2



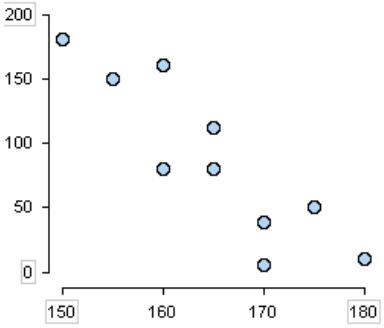
Scatterplot 3

(7) Examine the following scatterplots.

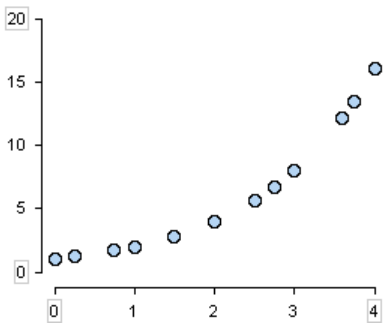
(a) Two of these scatterplots have an *r*-value close to 0.94. Determine which two without calculating the *r*-value.

(b) Does this activity make you rethink your answer to one of the two questions from Question 5? Explain your reasoning.

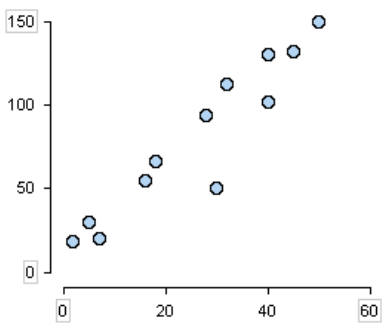
Supporting Lesson 3.1.3: Introduction to the Correlation Coefficient and Its Properties



Scatterplot 1



Scatterplot 2



Scatterplot 3

Wrap-Up Questions/Direction Instruction About Statistical Concepts [about 5 minutes]

Ask students to identify the two scatterplots that have a nonlinear form. It may be helpful to have students trace the curves. In both graphs there is a strong relationship between *x* and *y*; you could accurately predict *y* based on *x*. However, the relationship is summarized best by a curve instead of a line. In one case the value of *r* is close to zero in the other the value of *r* is close to one. The correlation coefficient *r* is a measure of the strength and direction of a *linear* relationship, but the value of *r* alone cannot tell you if a relationship is linear and a value of *r* close to zero does not mean that there is no relationship. The important lesson here is that before interpreting a correlation coefficient, it is important to look at the scatterplot.

Homework [Student Handout]

(8) Go to http://istics.net/stat/Correlations. Match the values of the correlation coefficient with the corresponding scatterplot using what you know about strength and direction of linear relationships. Click Answers to check your work. Click New Plots for a new set of scatterplots. Just below the plots, the applet keeps a running count of how many correct matches you have made. Continue matching scatterplot and correlation coefficients until you have accumulated at least 50 correct matches. (If you know what you are doing, this takes about 10 minutes.)

(9) Go to http://www.rossmanchance.com/applets/guesscorrelation/GuessCorrelation.html. Click on the New Sample button, which generates a scatterplot. Enter your guess for the correlation in the box called Correlation Guess and hit Enter. The applet then reveals the actual value of the correlation coefficient.

It is not easy to guess the value of the correlation coefficient exactly, so if a guess is within 0.1 of the actual value, it is a pretty good guess. (For example, if you guess 0.7 and the actual value is anything between 0.6 and 0.8, you have a pretty good guess.)

Click New Sample and estimate the correlation as many times as it takes for you to be comfortable with your ability to estimate the value of the correlation coefficient within 0.1.

Supporting Lesson 3.1.3: Introduction to the Correlation Coefficient and Its Properties

(10) Go to http://www.causeweb.org/repository/statjava/CorrCreateApplet.html. In this applet, you use the mouse to add points to a scatterplot by clicking on the scatterplot wherever you want to add a point. Try to create scatterplots that have a correlation coefficient that is close (within 0.1) to each of the following *r*-values:

$$r = +0.8,\ r = -0.9,\ r = 0.4,\ r = 0.7,\ r = -0.2$$

Tip: If you have difficulty comparing decimal numbers, envision the same number of decimal places in the numbers you are comparing.

Need help deciding if your correlation coefficient is accurate enough? Study the following examples.

- ***Example 1:*** What values are within 0.1 of your first target $r = +0.8$?

  Answer: Any correlation value between 0.7 and 0.9 is within 0.1 of your first target $r = +0.8$.

  Which of the following r-values are within 0.1 of $r = +0.8$?

  $r = 0.74$, $r = 0.91$, $r = 0.86$

  Answer: Since the *r*-values have two decimal places, compare them to 0.70 and 0.90. Only $r = 0.74$ and $r = 0.86$ are between 0.70 and 0.90 (i.e., within 0.1 of the target $r = +0.8$).

- ***Example 2:*** What values are within 0.1 of your second target $r = -0.9$?

  Answer: Any correlation value between –1.0 and –0.8 is within 0.1 of our second target $r = -0.9$.

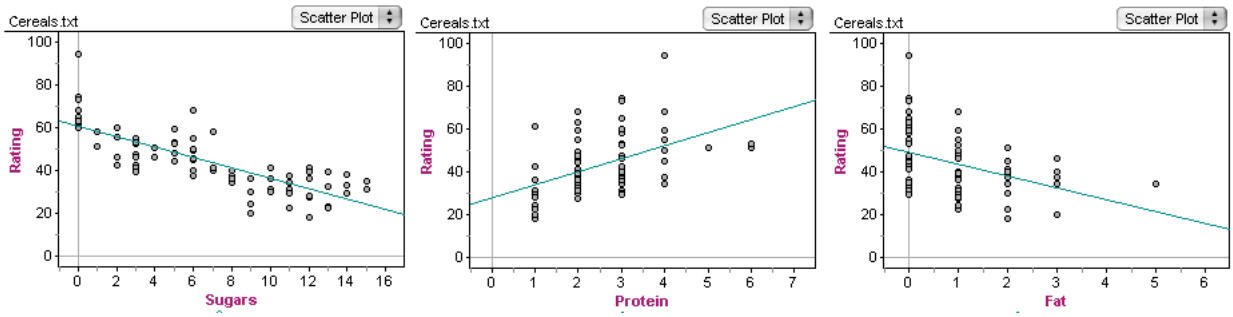
  Which of the following *r*-values are within 0.1 of $r = -0.9$?

  $r = -0.740$, $r = -0.905$, $r = -0.856$

  Answer: Since the *r*-values have three decimal places, compare them to –1.000 and –0.800. Only $r = -0.905$ and $r = -0.856$ are between –1.000 and –0.800 (i.e., within 0.1 of the target $r = -0.9$).

Supporting Lesson 3.1.3: Introduction to the Correlation Coefficient and Its Properties

When you were trying to determine which cereal ingredients influenced the *Consumer Reports* ratings, you saw a lot of variability in the data. This variability made it difficult to predict ratings using some of the ingredients. Using a line to summarize the relationship between the ingredient and the rating makes it easier to see positive and negative association in the data. Sugar and fat are negatively associated with ratings. Protein is positively associated with ratings.



A line can also help you evaluate the strength of the relationship between the ingredient and the rating. Some ingredients are more strongly related to the rating, like sugar. You see this in a linear pattern with not much scatter around a line. A small amount of scatter around the line means that the ingredient is a good predictor of the rating. For other ingredients, like fat, the relationship with rating was not as strong. You see this as more scatter about the line. This means that fat is not as good of a predictor of ratings.

In this lesson, you will investigate a measurement of variability in scatterplots called the *correlation coefficient*, which is denoted with letter *r*. In the next lesson, you will work with the formula that is used to compute the correlation coefficient, but in this lesson you will use technology to compute the value of *r.* The goal in this lesson is to determine the properties of the statistic *r.*

Supporting Lesson 3.1.3: Introduction to the Correlation Coefficient and Its Properties

Task 1: Investigating the Properties of *r*

(1) Which of the nine graphs on the following page show a positive association between *x* and *y*? A negative association? No association?

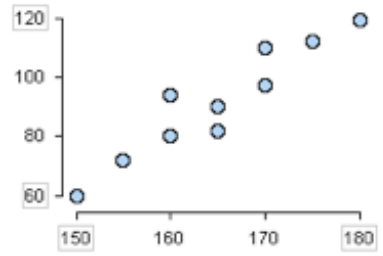
(2) Use technology to find the correlation coefficient (*r*) for each graph. The data are given in the table following the scatterplots. (If you do not have access to technology in class, your instructor will provide the *r*-value for each scatterplot.)

(3) Now look for patterns by comparing scatterplots and *r*-values. How does the value of *r* seem to be related to the patterns you see in the scatterplots?

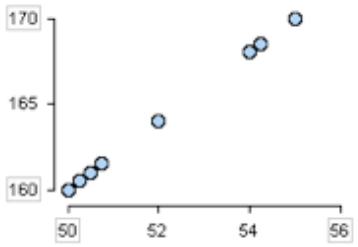
(4) Prepare for the class discussion by discussing the following questions in your group:

- What do you think the correlation coefficient (*r*) measures?
- Is there a largest possible value for *r,* or can it have larger and larger values without limit? What makes you think so?
- Is there a smallest possible value for *r,* or can it have smaller and smaller values without limit? What makes you think so?

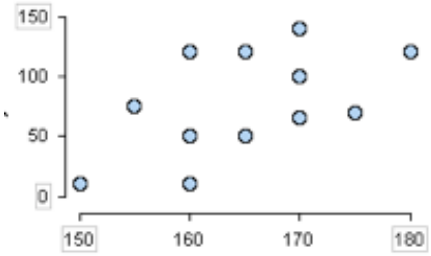
Supporting Lesson 3.1.3: Introduction to the Correlation Coefficient and Its Properties



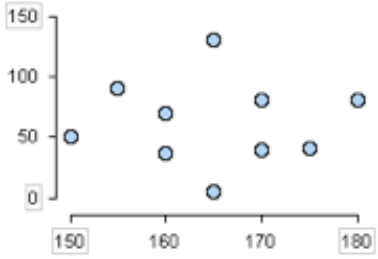
Scatterplot 1



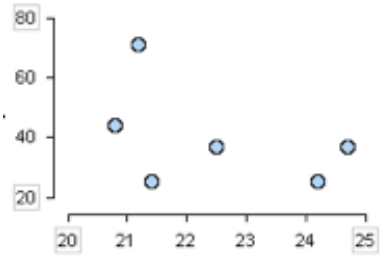
Scatterplot 2



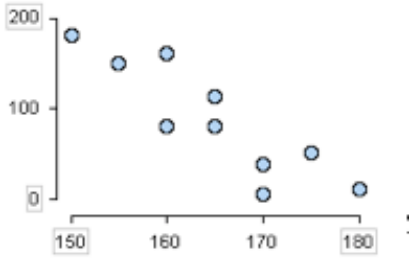
Scatterplot 3



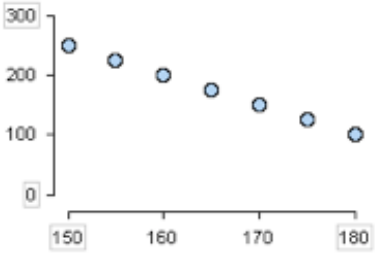
Scatterplot 4



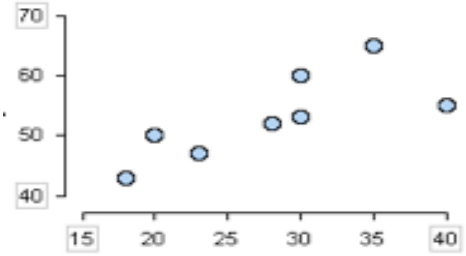
Scatterplot 5



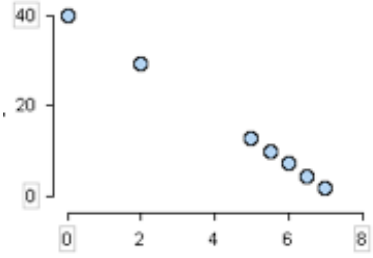
Scatterplot 6



Scatterplot 7



Scatterplot 8



Scatterplot 9

Supporting Lesson 3.1.3: Introduction to the Correlation Coefficient and Its Properties

Data for Calculating the Correlation Coefficient[1]

*x*1	*y*1	*x*2	*y*2	*x*3	*y*3	*x*4	*y*4	*x*5	*y*5
150	60	50.00	160.0	150	10	150	50	20.8	44
155	72	50.25	160.5	155	75	155	90	21.4	25
160	94	50.50	161.0	160	10	160	36	21.2	71
160	80	50.75	161.5	165	50	160	70	24.2	25
165	82	52.00	164.0	170	140	165	5	24.7	37
165	90	54.00	168.0	175	70	165	130	22.5	37
170	97	54.25	168.5	180	120	170	39		
170	110	55.00	170.0	165	120	170	80		
175	112			170	65	175	40		
180	119			160	120	180	80		
				170	100				
				160	50				

*x*6	*y*6	*x*7	*y*7	*x*8	*y*8	*x*9	*y*9
150	180	150	250	35	65	0.0	40.00
155	150	155	225	20	50	2.0	29.00
160	160	160	200	30	60	5.0	12.50
160	80	165	175	40	55	6.0	7.00
165	80	170	150	23	47	5.5	9.75
165	112	175	125	30	53	6.5	4.25
170	38	180	100	18	43	7.0	1.50
170	5			28	52		
175	50						
180	10						

[1]Scatterplot 1 is *x*1 versus *y*1, Scatterplot 2 is *x*2 versus *y*2, and so on.

Supporting Lesson 3.1.3: Introduction to the Correlation Coefficient and Its Properties

Task 2: Linear Correlation with Nonlinear Scatterplots

(5) Based on your current experience with *r*, answer the following two questions:

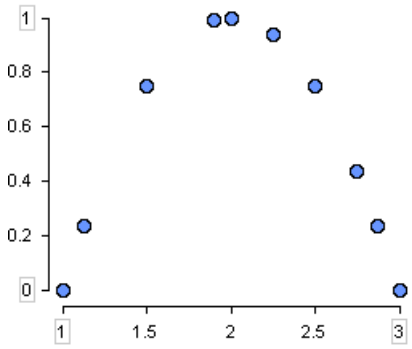
(a) If *r* is close to zero, can there be a strong relationship between the variables?

(b) If *r* is close to one, can the relationship between the variables be nonlinear?

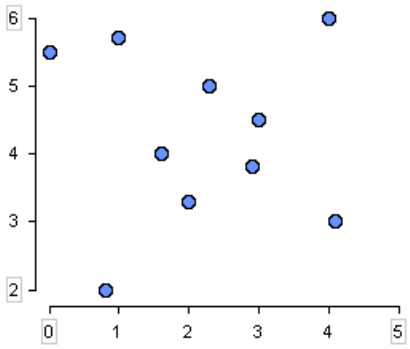
(6) Examine the following scatterplots.

(a) Two of these scatterplots have an *r*-value close to 0. Determine which two without calculating the *r*-value.

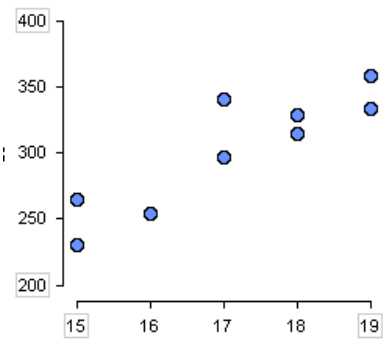
(b) Does this activity make you rethink your answer to one of the two questions from Question 5? Explain your reasoning.



Scatterplot 1



Scatterplot 2

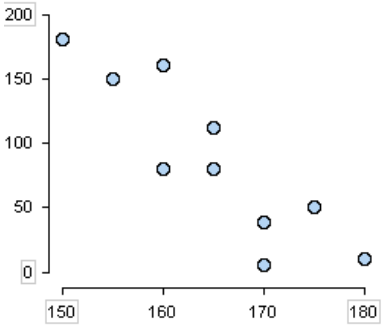


Scatterplot 3

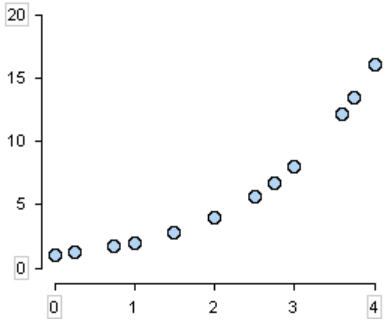
(7) Examine the following scatterplots.

(a) Two of these scatterplots have an *r*-value close to 0.94. Determine which two without calculating the *r*-value.

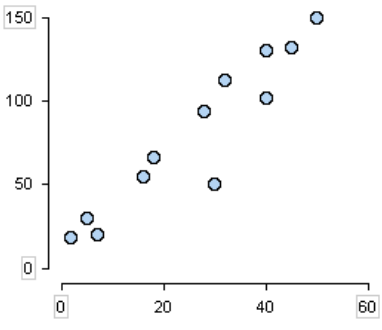
(b) Does this activity make you rethink your answer to one of the two questions from Question 5? Explain your reasoning.



Scatterplot 1



Scatterplot 2



Scatterplot 3

Supporting Lesson 3.1.3: Introduction to the Correlation Coefficient and Its Properties

Homework

(8) Go to http://istics.net/stat/Correlations. Match the values of the correlation coefficient with the corresponding scatterplot using what you know about strength and direction of linear relationships. Click Answers to check your work. Click New Plots for a new set of scatterplots. Just below the plots, the applet keeps a running count of how many correct matches you have made. Continue matching scatterplot and correlation coefficients until you have accumulated at least 50 correct matches. (If you know what you are doing, this takes about 10 minutes.)

(9) Go to http://www.rossmanchance.com/applets/guesscorrelation/GuessCorrelation.html. Click on the New Sample button, which generates a scatterplot. Enter your guess for the correlation in the box called Correlation Guess and hit Enter. The applet then reveals the actual value of the correlation coefficient.

It is not easy to guess the value of the correlation coefficient exactly, so if a guess is within 0.1 of the actual value, it is a pretty good guess. (For example, if you guess 0.7 and the actual value is anything between 0.6 and 0.8, you have a pretty good guess.)

Click New Sample and estimate the correlation as many times as it takes for you to be comfortable with your ability to estimate the value of the correlation coefficient within 0.1.

(10) Go to http://www.causeweb.org/repository/statjava/CorrCreateApplet.html. In this applet, you use the mouse to add points to a scatterplot by clicking on the scatterplot wherever you want to add a point. Try to create scatterplots that have a correlation coefficient that is close (within 0.1) to each of the following *r*-values:

$$r = +0.8,\ r = -0.9,\ r = 0.4,\ r = 0.7,\ r = -0.2$$

Tip: If you have difficulty comparing decimal numbers, envision the same number of decimal places in the numbers you are comparing.

Need help deciding if your correlation coefficient is accurate enough? Study the following examples.

- ***Example 1:*** What values are within 0.1 of your first target $r = +0.8$?

  Answer: Any correlation value between 0.7 and 0.9 is within 0.1 of your first target $r = +0.8$.

  Which of the following r-values are within 0.1 of $r = +0.8$?

  $r = 0.74$, $r = 0.91$, $r = 0.86$

  Answer: Since the *r*-values have two decimal places, compare them to 0.70 and 0.90. Only $r = 0.74$ and $r = 0.86$ are between 0.70 and 0.90 (i.e., within 0.1 of the target $r = +0.8$).

- ***Example 2:*** What values are within 0.1 of your second target $r = -0.9$?

  Answer: Any correlation value between –1.0 and –0.8 is within 0.1 of our second target $r = -0.9$.

Supporting Lesson 3.1.3: Introduction to the Correlation Coefficient and Its Properties

Which of the following *r*-values are within 0.1 of $r = -0.9$?

$r = -0.740$, $r = -0.905$, $r = -0.856$

Answer: Since the *r*-values have three decimal places, compare them to –1.000 and –0.800. Only $r = -0.905$ and $r = -0.856$ are between –1.000 and –0.800 (i.e., within 0.1 of the target $r = -0.9$).

Supporting Lesson 3.1.4: Correlation Formula

Estimated number of 50-minute class sessions: 2

Learning Goals

Students will understand that

- the correlation coefficient is a numerical measure defined by $\frac{\sum\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)}{n-1}$. It is roughly an average of the product of *Z*-scores.
- if a linear association is positive, *x*-values below $\bar{x}$ (negative *Z*-scores relative to the mean of *x*) tend to be associated with *y*-values below $\bar{y}$ (negative *Z*-scores relative to the mean of *y*) and *x*-values above $\bar{x}$ (positive *Z*-scores relative to the mean of *x*) tend to be associated with *y*-values above $\bar{y}$ (positive *Z*-scores relative to the mean of *y*). Similarly, if a linear association is negative, *x*-values below $\bar{x}$ (negative *Z*-scores relative to the mean of *x*) tend to be associated with *y*-values above $\bar{y}$ (positive *Z*-scores relative to the mean of *y*) and *x*-values above $\bar{x}$ (positive *Z*-scores relative to the mean of *x*) tend to be associated with *y*-values below $\bar{y}$ (negative *Z*-scores relative to the mean of *y*).

Students will be able to

- identify data values with positive and negative *Z*-scores.
- locate points relative to the mean lines in a scatterplot.
- analyze the formula for the correlation coefficient by relating it to the regions created by mean lines and to standardized scores.

Introduction to the Lesson [about 10 minutes]

In this lesson, you will continue the discussion of correlation. In the previous lesson, you investigated the properties of *r*. Use the following applet to review these properties:

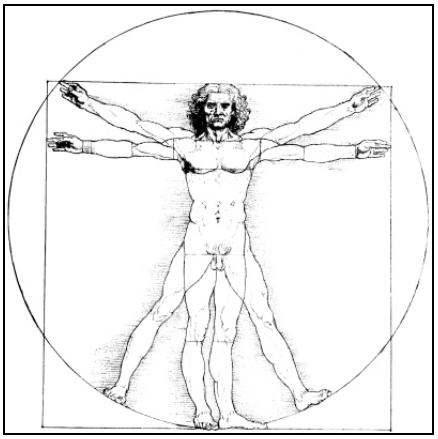
- Go to http://istics.net/stat/Correlations. In this applet, students match the values of the correlation coefficient with the corresponding scatterplot using what they now know about strength and direction of linear relationships. You can use this applet with clickers, poll the class, or pair students off to decide on their answers before taking a class poll. Conduct several rounds, and ask a few students to articulate how they are determining their answers.

After this review, introduce Task 1 by reading the introduction paragraph and give students a few minutes to think about Question 1. Then show them how the location of the man in the square in the picture tells them arm span equals height for the Vitruvian Man. After that give students 3–5 minutes to work together on finishing the task. The goal here is for students to get a chance to check their ability to estimate correlation and also to acquaint them with the context used in the investigation of the correlation formula that follows.

Supporting Lesson 3.1.4: Correlation Formula

Task 1: Check Your Understanding of Correlation [about 5 minutes]

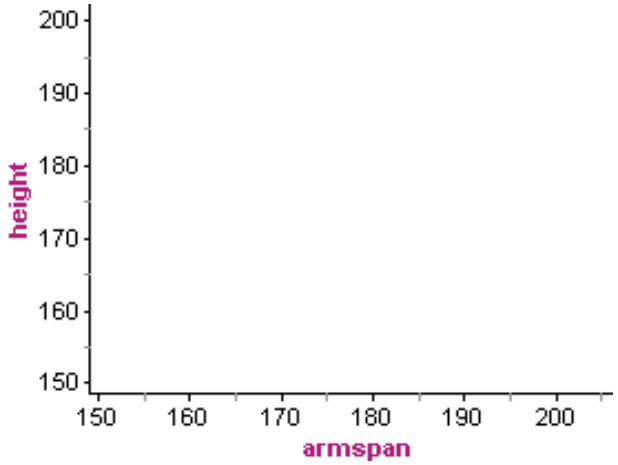
This is a version of a famous drawing of the Vitruvian Man by Leonardo da Vinci in 1487. In this drawing, Leonardo is representing the work of an ancient Roman architect named Vitruvius, who connected the proportions of the male figure to the proportions used in classical architecture.



(1) Vitruvius wrote that a man's arm span is equal to his height. How does Leonardo's drawing communicate this relationship between arm span and height?

(2) Graph the arm span and height measurements of five hypothetical men who fit the proportions of the Vitruvian Man.

(**Answer:** Five dots should form a line at $y = x$.)



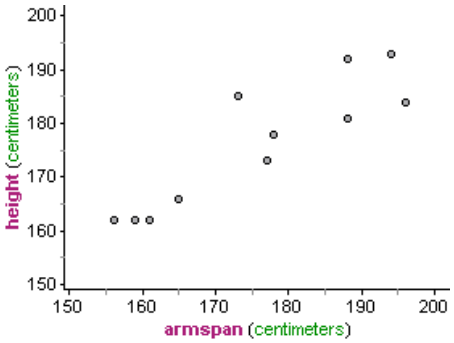
(3) Without doing any calculations, what is the correlation for your set of five men?

(**Answer:** Correlation is 1.)

(4) Here is a scatterplot of real arm span and height measurements for a sample of 11 men. The relationship between arm span and height is roughly linear, though these men are not perfectly proportioned according to Vitruvius. Which of the following values for the correlation coefficient describes the relationship in the scatterplot?

0.90 0.23 0.02 –0.45 –0.78

(**Answer:** 0.90)



Supporting Lesson 3.1.4: Correlation Formula

Wrap-Up [0 minutes]

While students are working, do a quick assessment of whether everyone is able to answer Questions 1–4. Help groups as necessary to correct answers. Since this is a review, there is no need for a class discussion unless your observations suggest otherwise.

Task 2: Introduction to the Correlation Formula [Student Handout, about 20 minutes]

(**Note:** Do this part of the lesson as a class discussion/lecture. The goal is to engage students in analyzing the symbolic form of the correlation formula. Students will work on calculating correlation by hand and/or with technology later in the lesson.

Here is the formula for *r*: $$r = \frac{\sum\left(\frac{x - \bar{x}}{s_x}\right)\left(\frac{y - \bar{y}}{s_y}\right)}{n - 1}$$

Discuss the following questions as a class. Do not forget to build in think time! You might use "think, pair, share" or a similar strategy to give students time to ponder.)

(5) How is the formula for *r* related to things you have seen before?

(**Answers may vary:** Students might see descriptive statistics, like means and standard deviations. They might also recognize standardized scores. Some might see that the overall formula looks like an average of something.)

(6) In the formula for *r*, you see the following terms: x, $\bar{x}$, s_x, y, $\bar{y}$, s_y, $n - 1$.

(a) If you are calculating the correlation for the relationship between *x* = arm span and *y* = height for a sample of 11 men, describe what each of these terms represents.

(b) For your sample of 11 men, which of these terms is fixed in the formula (calculated once and then used as a constant) and which have different values for the different men?

(7) Statisticians may describe the correlation *r* as "an average of the products of the standardized scores for n observations."

(a) Where in the formula do you see a standardized score for *x*? A standardized score for *y*?

(b) If a man has a standardized arm span of 1.2, what do you know about his arm span measurement?

(**Answer:** His arm span is 1.2 standard deviations above the mean arm span for the 11 men.)

(c) If a man has a standardized height of –0.5, what do you know about his height?

(**Answer:** His arm span is half a standard deviation below the mean arm span for the 11 men.)

(d) How can you tell in the formula that *r* is an average of something?

(**Answer:** You see a summation and a division by almost *n*.)

Supporting Lesson 3.1.4: Correlation Formula

(8) In Module 2, you saw that data on a single variable are summarized numerically using measures of center and measures of variability. When you summarized data with a mean, you used variance and standard deviation to measure variability about the mean. In the previous lesson, you saw that when a relationship between two variables is summarized by a line, the correlation is a measure of scatter or variability about the line. In this way, correlation is similar to variance and standard deviation because it is a measure of scatter or variability.

How is the formula for correlation similar to the formula for variance? How is it different?

$$\text{correlation} = \frac{\sum\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)}{n-1} \qquad \text{variance} = \frac{\sum(x-\bar{x})^2}{n-1} = \frac{\sum(x-\bar{x})(x-\bar{x})}{n-1}$$

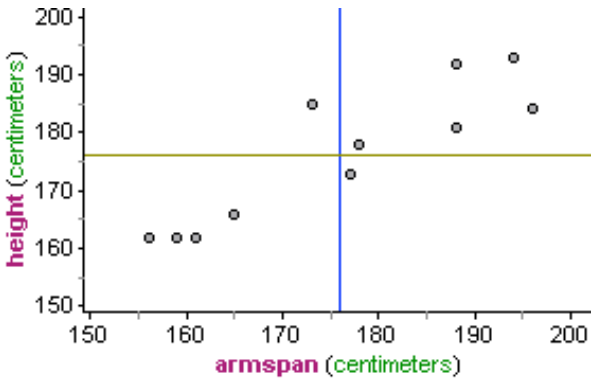
(**Answers will vary.** Do not hesitate to point out similarities in the structure of the formulas. Both statistics are an averaging of something related to distances from the mean. More specifically, both involve an average of the product of distances from the mean. In the case of correlation, you are looking at the product of standardized distances from the mean for both *x* and *y*. In the case of variance, you are looking at the product of the distance of just *x* from the mean. This makes sense because correlation is a measure of variability for bivariate data. Variance is a measure of variability for univariate data.)

(**Note:** Now move into Task 3. Do this task in groups.)

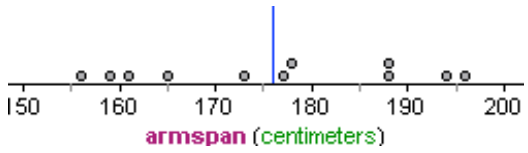
Supporting Lesson 3.1.4: Correlation Formula

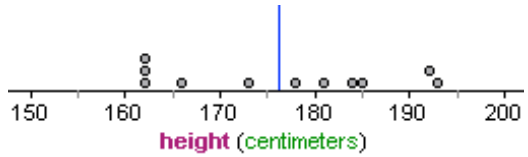
Task 3: Digging into the Correlation Formula [Student Handout, about 25 minutes]

Below, arm span and height measurements for a sample of 11 men are represented in a table, in dotplots, and in a scatterplot. The summary statistics for both measurements are given. The means are marked in each graph.



	armspan	height
1	161 cm	162 cm
2	196 cm	184 cm
3	177 cm	173 cm
4	188 cm	181 cm
5	159 cm	162 cm
6	178 cm	178 cm
7	194 cm	193 cm
8	188 cm	192 cm
9	173 cm	185 cm
10	165 cm	166 cm
11	156 cm	162 cm





	Mean	Standard Deviation
Arm span	175.91	14.36
Height	176.18	11.90

To get familiar with the formula for correlation, look at just a few men.

(9) Locate the male with an arm span of 188 centimeters and a height of 192 centimeters in the table and in each graph.

(a) This man has (*circle one*: above-average, below-average, average) arm span and (*circle one*: above-average, below-average, average) height.

(**Answer:** above average in both measurements)

(b) This man has a (*circle one*: positive, negative, zero) standardized arm span measurement.

(**Answer:** positive)

(c) This man has a (*circle one*: positive, negative, zero) standardized height measurement.

(**Answer:** positive)

Supporting Lesson 3.1.4: Correlation Formula

(d) Calculate each of the following for this man:

$x - \bar{x} =$ _____ $\quad \dfrac{x - \bar{x}}{s_x} =$ _____

$y - \bar{y} =$ _____ $\quad \dfrac{y - \bar{y}}{s_y} =$ _____

$\left\{ \left(\dfrac{x - \bar{x}}{s_x} \right)\left(\dfrac{y - \bar{y}}{s_y} \right) = \text{_____} \right.$

(**Answer:**

$x - \bar{x} = 12.09 \quad \dfrac{x - \bar{x}}{s_x} = 0.84$

$y - \bar{y} = 15.82 \quad \dfrac{y - \bar{y}}{s_y} = 1.33$

$\left\{ \left(\dfrac{x - \bar{x}}{s_x} \right)\left(\dfrac{y - \bar{y}}{s_y} \right) = 1.12 \right.$)

(10) Find a man who has *below-average* arm span and *below-average* height. Locate this man in all three graphs.

(a) This man has a (*circle one:* positive, negative, zero) standardized score for his arm span measurement. (Determine this without making any calculations if you can.)

(**Answer:** negative)

(b) This man has a (*circle one:* positive, negative, zero) standardized score for his height measurement. (Determine this without making any calculations if you can.)

(**Answer:** negative)

(c) Calculate each of the following for this man:

$x - \bar{x} =$ _____ $\quad \dfrac{x - \bar{x}}{s_x} =$ _____

$y - \bar{y} =$ _____ $\quad \dfrac{y - \bar{y}}{s_y} =$ _____

$\left\{ \left(\dfrac{x - \bar{x}}{s_x} \right)\left(\dfrac{y - \bar{y}}{s_y} \right) = \text{_____} \right.$

(**Answers can vary** based on choice of man.)

(d) Locate all of the men in the scatterplot who have a *below-average* arm span and *below-average* height.

(**Answer:** The four men in the lower left mean-quadrant.)

(11) Locate the men in the scatterplot who have both a *below-average* arm span and an *above-average* height. How many men in the sample fit this description? (**Answer:** 1)

For each of these men is the product of his *Z*-scores positive, negative, or zero? (Determine this without making any calculations if you can.) (**Answer:** negative)

Supporting Lesson 3.1.4: Correlation Formula

(12) In the scatterplot, circle all of the men for whom the product of *Z*-scores is positive.

(**Answer:** Circle all men in the lower left and upper right mean-quadrants.)

(13) Complete the table.

	armspan	zscore_armspan	height	zscore_height	product_zscores
1	161 cm	-1.04	162 cm	-1.19	1.24
2	196 cm		184 cm	0.657	
3	177 cm	0.076	173 cm	-0.267	-0.0203
4	188 cm	0.842	181 cm	0.405	0.341
5	159 cm	-1.18	162 cm	-1.19	1.4
6	178 cm	0.146	178 cm	0.153	0.0223
7	194 cm	1.26	193 cm	1.41	1.78
8	188 cm	0.842	192 cm	1.33	1.12
9	173 cm	-0.203	185 cm	0.741	-0.15
10	165 cm	-0.76	166 cm	-0.856	0.65
11	156 cm	-1.39	162 cm		

(**Answers:** No. 2: *Z*-score arm span = 1.40, product *Z*-scores = 0.92; No. 11: *Z*-score height = –1.19, product *Z*-scores = 1 .65)

(14) Write the appropriate expression above the appropriate column in the table.

$$x,\ y,\ \frac{x-\bar{x}}{s_x},\ \frac{y-\bar{y}}{s_y},\ \left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)$$

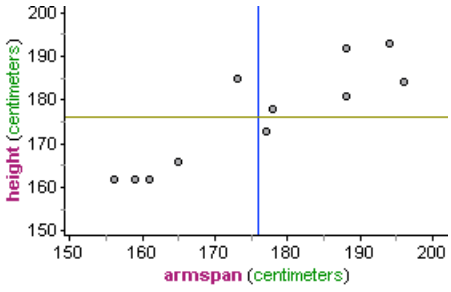
(15) Calculate the correlation by taking the sum of product of the *Z*-scores and dividing by $n - 1$. Double-check that your answer matches your estimate in Task 1.

(**Answer:** 8.953/10 = 0.895)

Wrap-Up [about 25 minutes]

Discuss the following questions using the scatterplot from Task 3:

- What is the sign of the *Z*-scores for data points in each mean-quadrant?
- In which regions is most of the data located when *r* is positive?
- Why does this make sense given the formula for *r*?



Supporting Lesson 3.1.4: Correlation Formula

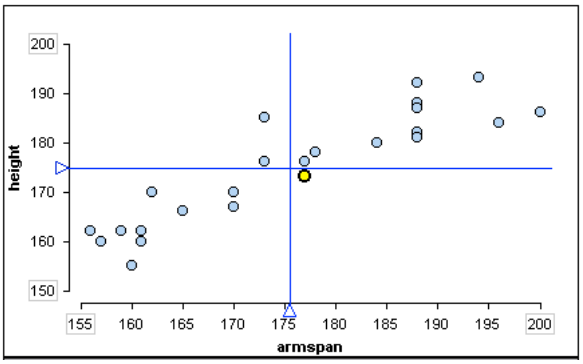
Use the next sequence of questions to check for understanding. You can use some of these questions with clickers or have students write answers anonymously on scrap paper, collect, redistribute, and then poll by reporting what is on the paper they received. Explain items that a significant portion of the class misses. Alternatively, just go over the items one at a time, giving students a minute to think about each item before you discuss it.

- Is the correlation coefficient in this scatterplot positive, negative, or close to zero? (**Answer:** positive)
- In the scatterplot circle the points for which the expression $\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)$ is negative.

  (**Answer:** circle points in the upper left and lower right mean-quadrants)
- If you calculated *r* by hand for this set of data, how many times would you calculate a value for the expression $\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)$? Would most of the values be positive, negative or zero? (**Answers:** 9, negative)

Homework [Student Handout]

(16) Here is a scatterplot of arm span and height measurements (in centimeters) for a sample of men. Of the following statements, what is true about the data point highlighted in the lower right mean-quadrant?



(a) Compared to other men in this sample, this man has an arm span that is above average, but he is shorter than average.

(b) Compared to other men in this sample, this man is unusual because he is tall but has short arms.

(c) Compared to other men in this sample, this man is above average in both arm span and height.

(d) Compared to other men in this sample, this man is small, with a shorter arm span than average and shorter than average height.

(**Answer:** a)

Supporting Lesson 3.1.4: Correlation Formula

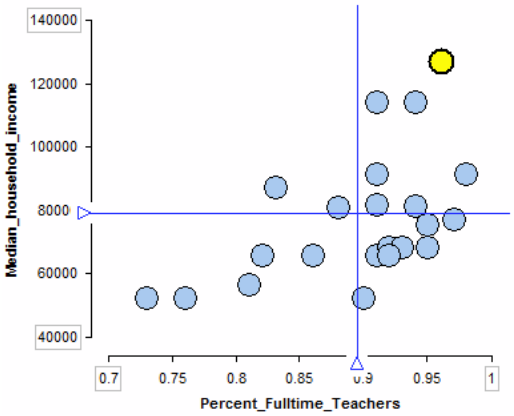
(17) Here you have data from cities in Contra Costa County in California. Each data point shows the percent of full-time teachers at a high school and the median household income for that city. For the data point highlighted in the upper right, three of the four statements are false. Which statement is true?

(a) The school has a below-average percent of full-time teachers but is in a city with an above-average median household income.

(b) The expression $\frac{x - \bar{x}}{s_x}$ is positive.

(c) The expression $\frac{y - \bar{y}}{s_y}$ is negative.

(d) The expression $\left(\frac{x - \bar{x}}{s_x}\right)\left(\frac{y - \bar{y}}{s_y}\right)$ is negative.



(**Answer:** b)

(18) For data with a linear form that have a very strong negative association, you would NOT expect to see many (*x*, *y*) points with the following:

(a) *x*-values below the mean of *x* when *y*-values are above the mean of *y.*

(b) *x*-values below the mean of *x* when *y*-values are below the mean of *y*.

(c) *x*-values above the mean of *x* when *y*-values are below the mean of *y*.

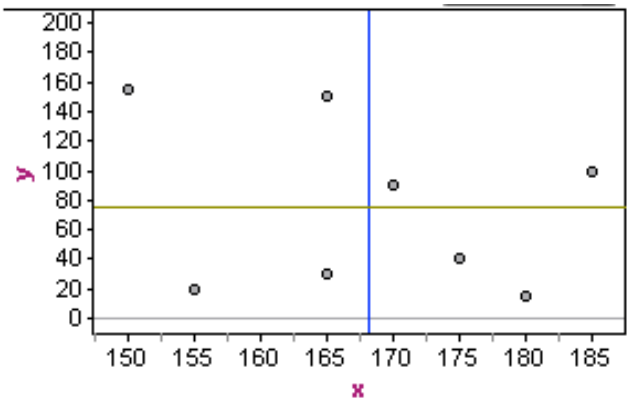
(d) It is impossible to predict this without seeing the scatterplot.

(**Answer:** b)

Supporting Lesson 3.1.4: Correlation Formula

(19) Here is the data set for the scatterplot shown.

x	150	155	165	165	170	175	180	185
y	155	20	150	30	90	40	15	100



(a) Which value is the most reasonable estimate for the correlation coefficient?

–0.28 –0.64 0.73

(**Answer:** –0.28)

(b) Complete the missing parts of the table and then calculate the correlation coefficient using the values in the table. Show or describe what calculation you performed to find *r* (after you complete the table).

	x	y	z_score_for_x	z_score_for_y	product_of_z_scores
1	150	155	-1.519	1.403	-2.131
2	155	20		-0.965	
3	165	150	-0.262		
4	165	30	-0.262	-0.789	0.207
5	170	90	0.157	0.263	0.041
6	175	40	0.576	-0.614	-0.354
7	180	15	0.995	-1.052	-1.047
8	185	100	1.414	0.439	0.620

(**Answers:** No. 2: *Z*-score for $x = -1.1$, product of *Z*-scores = 1.061; No. 3: *Z*-score for $y = 1.316$, product of *Z*-scores = –0.345; to calculate *r* add the values in the last column and divide by 7 to get –1.948/7= –0.28)

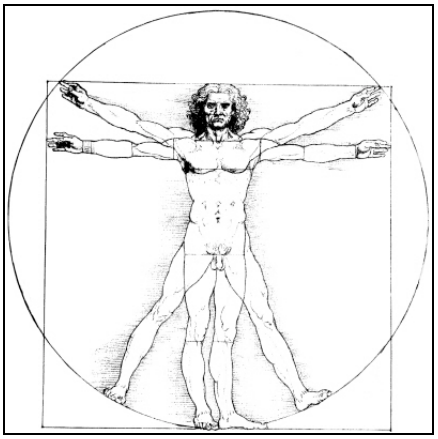
(c) Use technology to calculate the correlation coefficient.

(**Answer:** –0.28)

Supporting Lesson 3.1.4: Correlation Formula

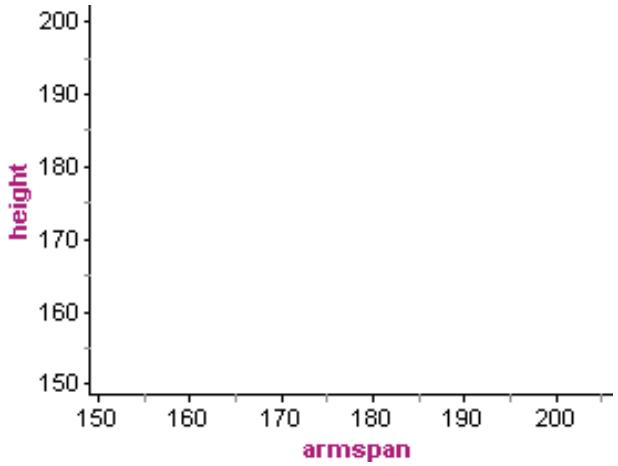
Task 1: Check Your Understanding of Correlation

This is a version of a famous drawing of the Vitruvian Man by Leonardo da Vinci in 1487. In this drawing, Leonardo is representing the work of an ancient Roman architect named Vitruvius, who connected the proportions of the male figure to the proportions used in classical architecture.



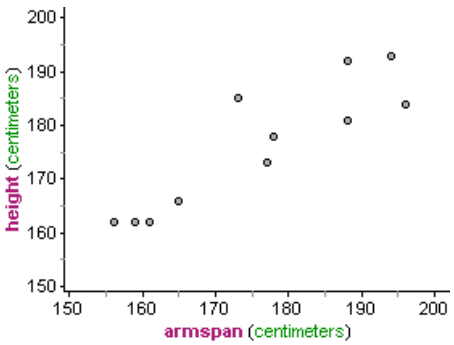
(1) Vitruvius wrote that a man's arm span is equal to his height. How does Leonardo's drawing communicate this relationship between arm span and height?

(2) Graph the arm span and height measurements of five hypothetical men who fit the proportions of the Vitruvian Man.



(3) Without doing any calculations, what is the correlation for your set of five men?

(4) Here is a scatterplot of real arm span and height measurements for a sample of 11 men. The relationship between arm span and height is roughly linear, though these men are not perfectly proportioned according to Vitruvius. Which of the following values for the correlation coefficient describes the relationship in the scatterplot?



0.90 0.23 0.02 –0.45 –0.78

Supporting Lesson 3.1.4: Correlation Formula

Task 2: Introduction to the Correlation Formula

(5) How is the formula for *r* related to things you have seen before?

(6) In the formula for *r*, you see the following terms: x, $\bar{x}$, s_x, y, $\bar{y}$, s_y, $n-1$.

(a) If you are calculating the correlation for the relationship between *x* = arm span and *y* = height for a sample of 11 men, describe what each of these terms represents.

(b) For your sample of 11 men, which of these terms is fixed in the formula (calculated once and then used as a constant) and which have different values for the different men?

(7) Statisticians may describe the correlation *r* as "an average of the products of the standardized scores for *n* observations."

(a) Where in the formula do you see a standardized score for *x*? A standardized score for *y*?

(b) If a man has a standardized arm span of 1.2, what do you know about his arm span measurement?

(c) If a man has a standardized height of –0.5, what do you know about his height?

(d) How can you tell in the formula that *r* is an average of something?

(8) In Module 2, you saw that data on a single variable are summarized numerically using measures of center and measures of variability. When you summarized data with a mean, you used variance and standard deviation to measure variability about the mean. In the previous lesson, you saw that when a relationship between two variables is summarized by a line, the correlation is a measure of scatter or variability about the line. In this way, correlation is similar to variance and standard deviation because it is a measure of scatter or variability.

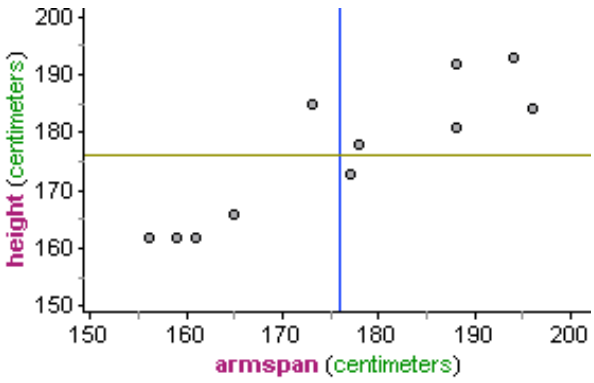
How is the formula for correlation similar to the formula for variance? How is it different?

$$\text{correlation} = \frac{\sum\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)}{n-1} \qquad \text{variance} = \frac{\sum(x-\bar{x})^2}{n-1} = \frac{\sum(x-\bar{x})(x-\bar{x})}{n-1}$$

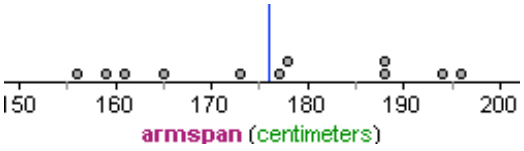
Supporting Lesson 3.1.4: Correlation Formula

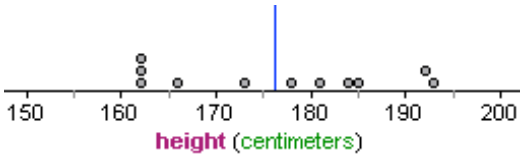
Task 3: Digging into the Correlation Formula

Below, arm span and height measurements for a sample of 11 men are represented in a table, in dotplots, and in a scatterplot. The summary statistics for both measurements are given. The means are marked in each graph.



	armspan	height
1	161 cm	162 cm
2	196 cm	184 cm
3	177 cm	173 cm
4	188 cm	181 cm
5	159 cm	162 cm
6	178 cm	178 cm
7	194 cm	193 cm
8	188 cm	192 cm
9	173 cm	185 cm
10	165 cm	166 cm
11	156 cm	162 cm





	Mean	Standard Deviation
Arm span	175.91	14.36
Height	176.18	11.90

To get familiar with the formula for correlation, look at just a few men.

(9) Locate the male with an arm span of 188 centimeters and a height of 192 centimeters in the table and in each graph.

(a) This man has (*circle one*: above-average, below-average, average) arm span and (*circle one*: above-average, below-average, average) height.

(b) This man has a (*circle one*: positive, negative, zero) standardized arm span measurement.

(c) This man has a (*circle one*: positive, negative, zero) standardized height measurement.

Supporting Lesson 3.1.4: Correlation Formula

(d) Calculate each of the following for this man:

$x - \bar{x} =$ _____ $\quad \dfrac{x - \bar{x}}{s_x} =$ _____

$y - \bar{y} =$ _____ $\quad \dfrac{y - \bar{y}}{s_y} =$ _____

$\left.\begin{array}{l}\\ \\ \\ \\ \end{array}\right\} \left(\dfrac{x - \bar{x}}{s_x}\right)\left(\dfrac{y - \bar{y}}{s_y}\right) =$ _____

(10) Find a man who has *below-average* arm span and *below-average* height. Locate this man in all three graphs.

(a) This man has a (*circle one:* positive, negative, zero) standardized score for his arm span measurement. (Determine this without making any calculations if you can.)

(b) This man has a (*circle one:* positive, negative, zero) standardized score for his height measurement. (Determine this without making any calculations if you can.)

(c) Calculate each of the following for this man:

$x - \bar{x} =$ _____ $\quad \dfrac{x - \bar{x}}{s_x} =$ _____

$y - \bar{y} =$ _____ $\quad \dfrac{y - \bar{y}}{s_y} =$ _____

$\left.\begin{array}{l}\\ \\ \\ \\ \end{array}\right\} \left(\dfrac{x - \bar{x}}{s_x}\right)\left(\dfrac{y - \bar{y}}{s_y}\right) =$ _____

(d) Locate all of the men in the scatterplot who have a *below-average* arm span and *below-average* height.

(11) Locate the men in the scatterplot who have both a *below-average* arm span and an *above-average* height. How many men in the sample fit this description?

For each of these men is the product of his *Z*-scores positive, negative, or zero? (Determine this without making any calculations if you can.)

(12) In the scatterplot, circle all of the men for whom the product of *Z*-scores is positive.

Supporting Lesson 3.1.4: Correlation Formula

(13) Complete the table.

	armspan	zscore_armspan	height	zscore_height	product_zscores
1	161 cm	-1.04	162 cm	-1.19	1.24
2	196 cm		184 cm	0.657	
3	177 cm	0.076	173 cm	-0.267	-0.0203
4	188 cm	0.842	181 cm	0.405	0.341
5	159 cm	-1.18	162 cm	-1.19	1.4
6	178 cm	0.146	178 cm	0.153	0.0223
7	194 cm	1.26	193 cm	1.41	1.78
8	188 cm	0.842	192 cm	1.33	1.12
9	173 cm	-0.203	185 cm	0.741	-0.15
10	165 cm	-0.76	166 cm	-0.856	0.65
11	156 cm	-1.39	162 cm		

(14) Write the appropriate expression above the appropriate column in the table.

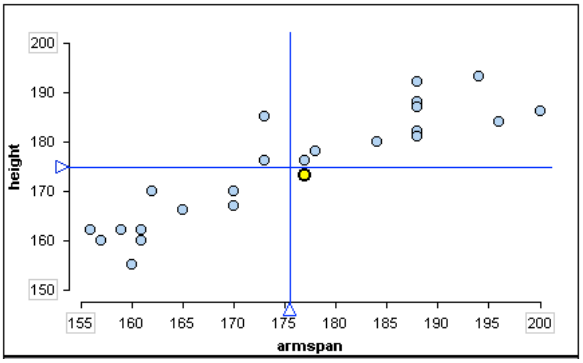
x, y, $\frac{x-\bar{x}}{s_x}$, $\frac{y-\bar{y}}{s_y}$, $\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)$

(15) Calculate the correlation by taking the sum of product of the *Z*-scores and dividing by $n - 1$. Double-check that your answer matches your estimate in Task 1.

Supporting Lesson 3.1.4: Correlation Formula

Homework

(16) Here is a scatterplot of arm span and height measurements (in centimeters) for a sample of men. Of the following statements, what is true about the data point highlighted in the lower right mean-quadrant?



(a) Compared to other men in this sample, this man has an arm span that is above average, but he is shorter than average.

(b) Compared to other men in this sample, this man is unusual because he is tall but has short arms.

(c) Compared to other men in this sample, this man is above average in both arm span and height.

(d) Compared to other men in this sample, this man is small, with a shorter arm span than average and shorter than average height.

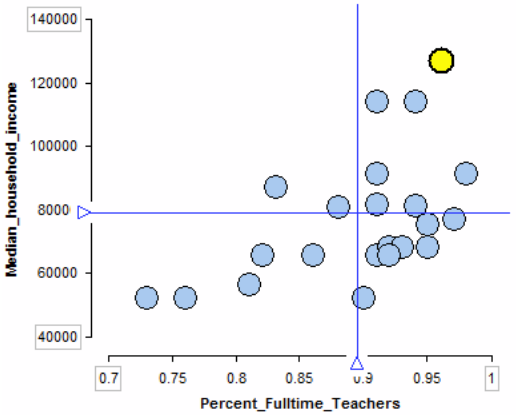
(17) Here you have data from cities in Contra Costa County in California. Each data point shows the percent of full-time teachers at a high school and the median household income for that city. For the data point highlighted in the upper right, three of the four statements are false. Which statement is true?

(a) The school has a below-average percent of full-time teachers but is in a city with an above-average median household income.

(b) The expression $\frac{x - \bar{x}}{s_x}$ is positive.

(c) The expression $\frac{y - \bar{y}}{s_y}$ is negative.

(d) The expression $\left(\frac{x - \bar{x}}{s_x}\right)\left(\frac{y - \bar{y}}{s_y}\right)$ is negative.



Supporting Lesson 3.1.4: Correlation Formula

(18) For data with a linear form that have a very strong negative association, you would NOT expect to see many (*x*, *y*) points with the following:

(a) *x*-values below the mean of *x* when *y*-values are above the mean of *y*.

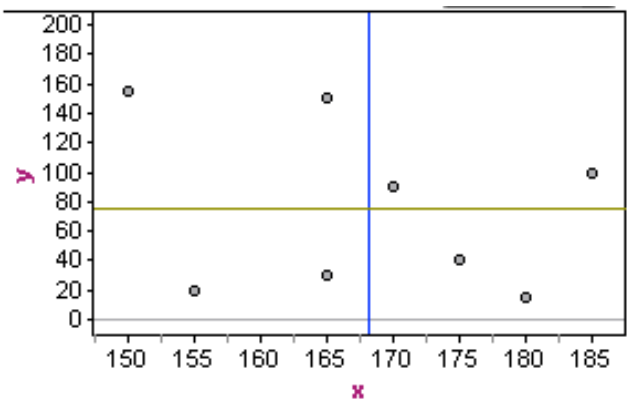
(b) *x*-values below the mean of *x* when *y*-values are below the mean of *y*.

(c) *x*-values above the mean of *x* when *y*-values are below the mean of *y*.

(d) It is impossible to predict this without seeing the scatterplot.

(19) Here is the data set for the scatterplot shown.

x	150	155	165	165	170	175	180	185
y	155	20	150	30	90	40	15	100



(a) Which value is the most reasonable estimate for the correlation coefficient?

–0.28 –0.64 0.73

(b) Complete the missing parts of the table and then calculate the correlation coefficient using the values in the table. Show or describe what calculation you performed to find *r* (after you complete the table).

	x	y	z_score_for_x	z_score_for_y	product_of_z_scores
1	150	155	-1.519	1.403	-2.131
2	155	20		-0.965	
3	165	150	-0.262		
4	165	30	-0.262	-0.789	0.207
5	170	90	0.157	0.263	0.041
6	175	40	0.576	-0.614	-0.354
7	180	15	0.995	-1.052	-1.047
8	185	100	1.414	0.439	0.620

(c) Use technology to calculate the correlation coefficient.

Supporting Lesson 3.1.5: Correlation Is Not Causation

Estimated number of 50-minute class sessions: 1

Learning Goals

Students will understand that

- correlation does not imply causation. Two variables can be strongly associated (as measured by the correlation coefficient) but have no cause-and-effect relationship. Often a confounding variable affecting both measurements may better explain the relationship implied by a strong association (correlation near 1 or –1).

Students will be able to

- explain why association does not imply causation.
- identify explanatory and response variables and plausible confounding variables.

Introduction [about 10 minutes]

The first part of this activity was inspired by a paper by Allan Rossman in the *Journal of Statistics Education* (1994). (See www.amstat.org/publications/jse/v2n2/datasets.rossman.html.)

Introduce this activity by telling students that they will do some estimating as a way to get them thinking about the variables for today's data set.)

- What do you think the life expectancy is for the following countries: Bangladesh, Japan, Mexico, Pakistan, and the United States. (Just give a rough estimate, we will look at some data later.) (**Note:** You may have to explain the term *life expectancy*.)
- Now think about the number of internet users per 1,000 people in a country. Give a rough estimate for each of these five countries.
- Do you think the correlation between life expectancy and the number of internet users per 1,000 people will be positive or negative?
- Do you think the correlation will indicate a strong or weak association between life expectancy and the number of internet users per 1,000 people?

After students have had a few minutes to make estimates and think about association between these two variables, call on a few students to give their estimates and conjectures about the relationship between the number of internet users per 1,000 people and life expectancy. The next task resolves the issue.

Have students work on Questions 1–4 (alone, in pairs, or in groups). Alternatively, use these questions for a whole-class discussion, but give students think time for each question before calling on them to participate in the discussion.

Supporting Lesson 3.1.5: Correlation Is Not Causation

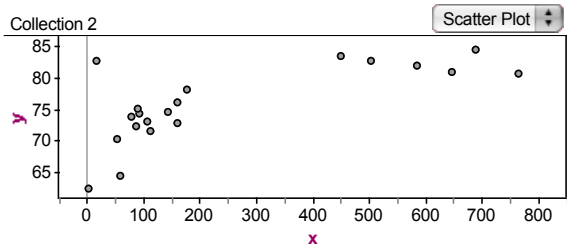
Task 1 (about 15 minutes, including wrap-up)

Activities [Student Handout]

(1) The scatterplot shows

x = internet users per 1,000 people
y = life expectancy (years)

for the 20 countries with the largest population for 2009. (World Almanac Book of Facts, 2009)



Describe what the data point (2, 62.5) tells about Bangladesh.

(**Answer:** In Bangladesh, there are two internet users for every 1,000 people. The life expectancy is 62.5 years.)

country	x	y
Banglade...	2	62.5
Brazil	160	76.1
China	92	74.5
Egypt	79	73.9
France	449	83.5
Germany	583	82.0
India	53	70.4
Indonesia	87	72.4
Iran	113	71.7
Italy	503	82.9
Japan	688	84.7
Mexico	177	78.3
Pakistan	58	64.4
Philippines	105	73.2
Russia	160	72.9
Thailand	143	74.7
Turkey	88	75.2
United Ki...	646	81.1
United St...	765	80.8
Nigeria	17	82.9

(2) In this group of 20 countries does an increase in the density of internet users (i.e., the number of internet users per 1,000 people) tend to be associated with an increase or a decrease in life expectancy?

(**Answer:** increase)

(3) The correlation coefficient is 0.62. How does the value of the correlation coefficient relate to your answer to Question 2?

(**Answer:** A positive correlation suggests that an increase in *x* tends to correspond to an increase in *y*.)

(4) The value of the correlation coefficient indicates a fairly strong positive linear relationship. Based on this observation, someone might suggest that an easy way to increase a country's life expectancy would be to get more people online. Do you think this is a reasonable conclusion? Why or why not?

(**Answer:** Ridiculous! Obviously, getting more people on line will not increase how long people live on average in the country.)

Wrap-Up/Direct Instruction

Ask students to share their conclusions to Question 4. Most likely students will realize the silliness of the suggestion that increasing internet use will increase life expectancy. If they do, name what they have discovered. If today's lesson had a title it might be: **Correlation measures association. But association is not the same as causation.** A strong correlation between two variables is evidence that there is a statistical relationship between the variables. In other words, the variables vary together in a predictable way.

In the unlikely situation that students do not detect that the conclusion in Question 4 is illogical, reassure them that they are not alone. Correlation is often confused with causation in the media.

Ask students to conjecture what might explain the strong correlation between internet use and life expectancy. Use this discussion as an opportunity to review the concept of a confounding variable from Module 1. At this time, also review the terms *explanatory variable* and *response variable*.

Task 2 [about 25 minutes, including wrap-up]

Introduction [Student Handout]

It is easy and fun to construct silly examples of correlations that do not result from causal connections. Here are some examples from John Allen Paulos, a mathematics professor at Temple University who is well known for his popular books on mathematical literacy.

(**Note:** Have students work on Questions 5–7 below (alone, in pairs, or in groups). Alternatively, use these questions for a whole-class discussion, but give students think time for each question before calling on them to participate in the discussion.)

Activities [Student Handout]

(5) Read this excerpt from *A Mathematician Reads the Newspaper*[1] by Paulos. Identify the explanatory, response, and confounding variables in Paulos' examples.

> A more elementary widespread confusion is that between correlation and causation. Studies have shown repeatedly, for example, that children with longer arms reason better than those with shorter arms, but there is no causal connection here. Children with longer arms reason better because they're older! Consider a headline that invites us to infer a causal connection: BOTTLED WATER LINKED TO HEALTHIER BABIES. Without further evidence, this invitation should be refused, since affluent parents are more likely both to drink bottled water and to have healthy children; they have the stability and wherewithal to offer good food, clothing, shelter, and amenities. Families that own cappuccino makers are more likely to have healthy babies for the same reason. Making a practice of questioning correlations when reading about "links" between this practice and that condition is good statistical hygiene.

(6) Paulos also writes a column for ABCNews.com called *Who's Counting?* In his February 1, 2001, column, Paulos discusses the idea that correlation does not imply causation. He points out that the consumption of hot chocolate is negatively correlated with crime rate. Obviously, drinking more hot chocolate does not lower the crime rate.

For this situation assume that the data describe large cities in the United States. What is the explanatory variable? What is the response variable? Identify a plausible confounding variable in this scenario.

[1] Paulos, J.A. (1995). *A mathematician reads the newspaper* (p. 137). New York: Basic Books.

(7) Describe a scenario with two quantitative variables that are probably highly correlated due to a third confounding variable.

Wrap-Up/Direct Instruction

Ask students to share their responses to Questions 5–7. In the discussion emphasize that the confounding variable affects both explanatory and response variables. The idea is that the confounding variable results in arm length and ability to reason changing together in a predictable way, and thus creates a statistical relationship between the two.

Remind students that in Module 1 they learned that only the results of a randomized comparative experiment can establish a causal connection statistically. In practice, it is often difficult, if not impossible, to conduct a randomized comparative experiment. In these situations nonstatistical guidelines can provide a way to determine whether a causal link is reasonable. However, association alone is not enough to establish a causal link.

Homework

(8) Summarize the main point of today's lesson and give an example to illustrate the main point.

(**Answer:** The main point is that association does not imply as causation. Examples will vary.)

Supporting Lesson 3.1.5: Correlation Is Not Causation

Task 1

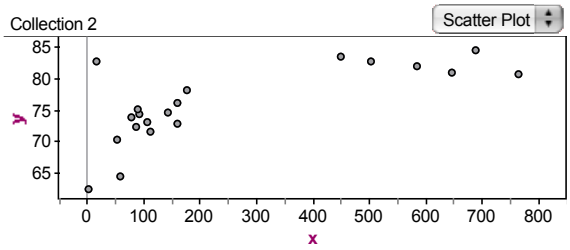
(1) The scatterplot shows

x = internet users per 1,000 people

y = life expectancy (years)

for the 20 countries with the largest population for 2009. (World Almanac Book of Facts, 2009)

Describe what the data point (2, 62.5) tells about Bangladesh.



country	x	y
Banglade...	2	62.5
Brazil	160	76.1
China	92	74.5
Egypt	79	73.9
France	449	83.5
Germany	583	82.0
India	53	70.4
Indonesia	87	72.4
Iran	113	71.7
Italy	503	82.9
Japan	688	84.7
Mexico	177	78.3
Pakistan	58	64.4
Philippines	105	73.2
Russia	160	72.9
Thailand	143	74.7
Turkey	88	75.2
United Ki...	646	81.1
United St...	765	80.8
Nigeria	17	82.9

(2) In this group of 20 countries does an increase in the density of internet users (i.e., the number of internet users per 1,000 people) tend to be associated with an increase or a decrease in life expectancy?

(3) The correlation coefficient is 0.62. How does the value of the correlation coefficient relate to your answer to Question 2?

(4) The value of the correlation coefficient indicates a fairly strong positive linear relationship. Based on this observation, someone might suggest that an easy way to increase a country's life expectancy would be to get more people online. Do you think this is a reasonable conclusion? Why or why not?

Supporting Lesson 3.1.5: Correlation Is Not Causation

Task 2

It is easy and fun to construct silly examples of correlations that do not result from causal connections. Here are some examples from John Allen Paulos, a mathematics professor at Temple University who is well known for his popular books on mathematical literacy.

(5) Read this excerpt from *A Mathematician Reads the Newspaper*[1] by Paulos. Identify the explanatory, response, and confounding variables in Paulos' examples.

> A more elementary widespread confusion is that between correlation and causation. Studies have shown repeatedly, for example, that children with longer arms reason better than those with shorter arms, but there is no causal connection here. Children with longer arms reason better because they're older! Consider a headline that invites us to infer a causal connection: BOTTLED WATER LINKED TO HEALTHIER BABIES. Without further evidence, this invitation should be refused, since affluent parents are more likely both to drink bottled water and to have healthy children; they have the stability and wherewithal to offer good food, clothing, shelter, and amenities. Families that own cappuccino makers are more likely to have healthy babies for the same reason. Making a practice of questioning correlations when reading about "links" between this practice and that condition is good statistical hygiene.

(6) Paulos also writes a column for ABCNews.com called *Who's Counting?* In his February 1, 2001, column, Paulos discusses the idea that correlation does not imply causation. He points out that the consumption of hot chocolate is negatively correlated with crime rate. Obviously, drinking more hot chocolate does not lower the crime rate.

For this situation assume that the data describe large cities in the United States. What is the explanatory variable? What is the response variable? Identify a plausible confounding variable in this scenario.

(7) Describe a scenario with two quantitative variables that are probably highly correlated due to a third confounding variable.

Homework

(8) Summarize the main point of today's lesson and give an example to illustrate the main point.

[1] Paulos, J.A. (1995). *A mathematician reads the newspaper* (p. 137). New York: Basic Books.

Initiating Lesson 3.2.1: Using Lines to Make Predictions

Estimated number of 50-minute class sessions: 1

Learning Goals and Concept Flow

S.2. Distributional Thinking Goal: Students will demonstrate the use of distributional thinking to reason about the data in order to describe and summarize distributions of data, identify trends and patterns, judge the fit of a model to a distribution, and describe similarities and differences in comparing distributions.

Students will understand that

- in a statistical relationship two variables tend to vary together in a predictable way. When a line is a good summary of a statistical relationship, the line can be used to predict the value of a response variable given a value of the explanatory variable (the predictor), but only part of the variation in *y* is explained by changes in *x*.
- predictions are more accurate when the relationship is strong.
- making predictions based on extrapolating outside the range of the data can be risky and should be done with caution.

Students will be able to

- given a scenario involving a bivariate numerical data set, identify the response variable (dependent variable) and the explanatory variable (predictor variable).
- given a regression line and a value for the predictor variable, predict the value of the response variable using both the graph of the line and its equation.
- explain the danger of extrapolation in a regression setting.

Developmental Math Connections

The focus of this lesson is on whether a line seems to be a reasonable summary of the bivariate relationship and, if it is, using a line to predict *y* based on *x*. Students will probably begin to sketch lines onto the scatterplots to make their predictions before this idea is formally introduced in the lesson. In this lesson, students use both the graph and the equation of the least squares regression line to make predictions. In subsequent lessons in this topic, you will review the concepts of slope and *y*-intercept explicitly. You will also delve into the idea of "best fit" based on the sum of the squares of residuals, so it is not necessary to teach these concepts here. However, take opportunities if they arise from students' comments to connect to students' knowledge of linear functions.

Part I [Student Handout, 15–25 minutes total]

Introduction [5 minutes]

Statistical methods are used in forensics to identify human remains based on the measurements of bones. In the 1950s, Dr. Mildred Trotter and Dr. Goldine Gleser measured skeletons of people who had died in the early 1900s. From these measurements they developed statistical methods for predicting a person's height based on the lengths of various bones. These formulas were first used to identify the remains of U.S. soldiers who died in World War II and were buried in unmarked graves in the Pacific zone. Modern forensic scientists have made adjustments to the formulas developed by Trotter and Gleser to account the differences in bone length and body

Initiating Lesson 3.2.1: Using Lines to Make Predictions

proportions of people living now. You will not use Trotter and Gleser's formulas in this problem, but you will use a similar process.

(**Note:** For information on the Terry skeleton collection, see http://anthropology.si.edu/cm/terry.htm. For a more recent example of how forensic scientists are still building on the work of Trotter and Gleser, see the following:

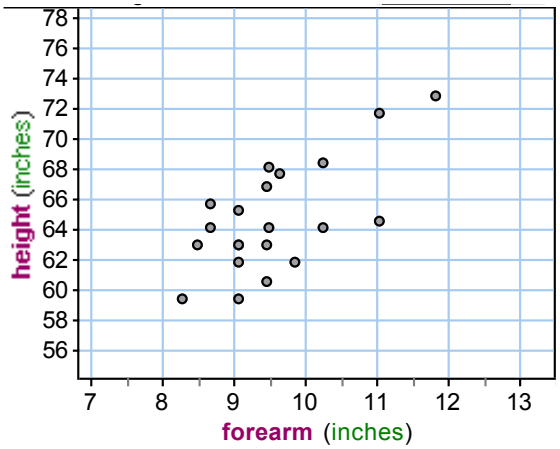
- Jantz, R. L. (1993). Modification of the Trotter and Gleser female stature estimation formulae. *Journal of Forensic Science*. *38*(4), 758–63.)

To illustrate the type of data analysis done in forensics, let's see if you can identify a female student based on the length of her forearm. The mystery student has a forearm measurement of 10 inches. (She is alive and healthy!)

Height and weight measurements for three female college students are given in the table.

	Jane Doe 1	**Jane Doe 2**	**Jane Doe 3**
Age	18	23	33
Gender	Female	Female	Female
Height	5 feet, 5 inches	5 feet, 2 inches	6 feet
Weight	128 pounds	120 pounds	155 pounds

Your task is to determine if the mystery student could be one of these three students.



First, you need data that relates forearm length to either height or weight for females. The scatterplot is a graph of height versus forearm length for 21 female college students taking Introductory Statistics at Los Medanos College in Pittsburg, California, in 2009.

(**Note:** Let students work for a few minutes alone, and then compare responses with a neighbor or group. Students may feel uncomfortable making a prediction since there is clearly no "right" answer. This is acceptable. Remind them that in statistics they are constantly making decisions in the face of variability in the data. Perhaps also give an obviously unreasonable prediction, such as a height of 56 or 76, ask students if the prediction seems reasonable, and then encourage them to give a better prediction.)

Initiating Lesson 3.2.1: Using Lines to Make Predictions

Task [5–10 minutes]

(1) Based on the scatterplot, what is a reasonable prediction for the height of the mystery student? Briefly explain or show how you made your prediction.

(2) The variability in the data makes it difficult to determine if one of these students is the mystery student. Could any of the three students be eliminated as a possibility of being the mystery student? Explain your reasoning.

Wrap-Up [5–10 minutes]

Based on the data in the scatterplot, ask students to determine if the following predictions for the height of the mystery student are reasonable or unreasonable (perhaps ask students to show a thumbs up for *reasonable prediction* and thumbs down for *unreasonable prediction*): 62 inches (reasonable), 66 inches (reasonable), 74 inches (unreasonable).

Plot each of these predictions on the scatterplot and highlight how the prediction fits the pattern in the data or deviates from the pattern. Call on a few students to give their predictions and plot these as well.

Indicate with a vertical line segment a reasonable range of predictions for $x = 10$ (approximately 62–68 inches). Based on this range of reasonable predictions, it looks like the mystery student is probably not Jane Doe 3, who is 72 inches tall.

Most likely students have already started eyeballing lines to fit the data, so you may want to elicit this idea before moving into Part II. You could do this by asking a few students to explain how they determined their prediction or just point out that the association is positive and somewhat linear, and then ask if anyone used a line to help summarize the data to make a prediction. If so, congratulate them for thinking like a statistician!

Transition to Part II

In Part II, students are introduced to making predictions using both the equation and graph of a line. The concepts in this lesson will probably not be difficult for most students, so the wrap-up for this segment of the lesson is an assessment item.

Decide whether you will use the problems in Part II as group work, to frame a whole-class discussion, or as a basis for a lecture. Orient students accordingly. Since the wrap-up does not include a summary, if you are conducting the lesson using group work, address individual difficulties as students are working.

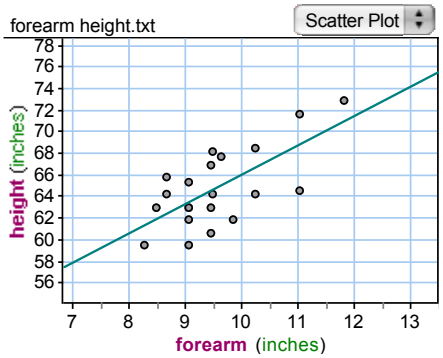
Initiating Lesson 3.2.1: Using Lines to Make Predictions

Part II: Using a Line to Make Predictions [Student Handout, 20–30 minutes total, including about 6 minutes for assessment item at the end of the lesson]

(3) The scatterplot has a positive linear association. The correlation is 0.68, which is pretty strong. So, it makes sense to use a linear model to summarize the relationship between the forearm and height measurements. There is one line that is considered the best description of how height and forearm length are related. You will learn more about how to find this line in future lessons. For now, you will use technology to find the equation of this line.

(a) Use the graph of the best-fit line to predict the height of the mystery student.

(**Answer:** 66 inches)



(b) The equation of this line is approximately

predicted height = 2.7(forearm length) + 39.

$$\hat{y} = 2.7x + 39$$

(Notice that when you use letters to represent variables in the prediction line, you put a "hat" on the y and write $\hat{y}$ instead of y. The hat is a signal that the variable is *predicted* values, not actual data values.)

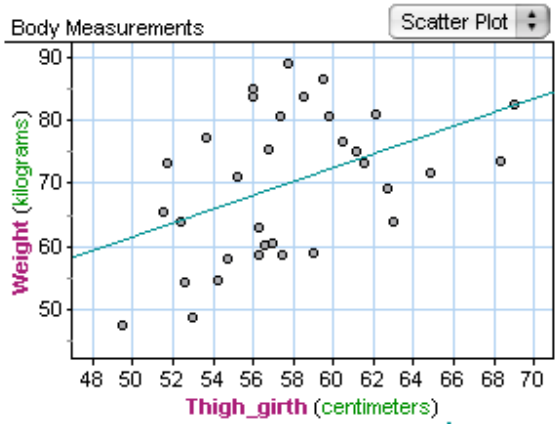
Use the equation to predict the height of the mystery person.

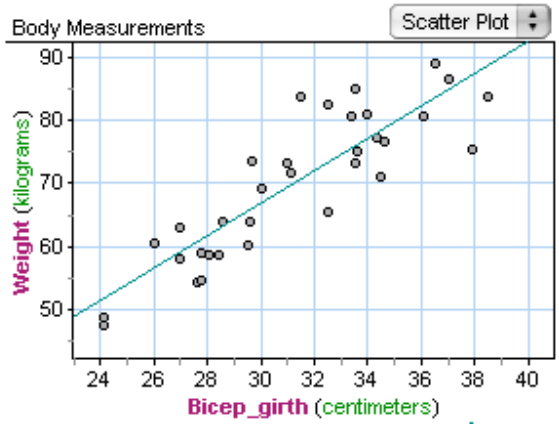
(**Answer:** 66 inches)

(c) Is the height of Jane Doe 1, 2, or 3 closest to the predicted height of the mystery student given by the line? (Of course, this does not guarantee that you have correctly identified the mystery student, but it suggests that one student's height, together with the 10-inch forearm measurement, fits the linear pattern in the data better than the other students.)

(**Answer:** Jane Doe 1)

(4) The scatterplots below are graphs of body measurements in centimeters for 34 adults who are physically active. These data are a random sample taken from a larger nonrandom data set gathered by researchers investigating the relationship of various body measurements and weight. Girth is the measurement around a body part. (Retrieved from www.amstat.org/publications/jse/v11n2/datasets.heinz.html)





Initiating Lesson 3.2.1: Using Lines to Make Predictions

(a) Based on these data, which do you think is a better predictor of an adult's weight: thigh girth or bicep girth? Why?

(**Answer:** bicep girth because there is less scatter about the line)

(b) Adriana has a thigh girth of 57 centimeters and a bicep girth of 25 centimeters. Predict her weight using the measurement that you think will give the most accurate prediction, and then plot Adriana on the scatterplot that you used to make her weight prediction.

(**Answer:** Use the bicep girth. 54 kilograms is a reasonable prediction from line.)

(c) The equations of the two lines shown are

weight = 6.3 + 1.1(thigh girth) weight = –10.5 + 2.6(bicep girth)

Predict Adriana's weight using the equation that you think best predicts weight.

(**Answer:** 54.5 kilograms)

(d) Of course, you do not really know Adriana's weight. How accurate do you think the line's prediction of Adriana's weight is? Choose the option that is the most reasonable and explain your thinking.

- very accurate (within a range of plus or minus 1 kilogram)
- somewhat accurate (within a range of plus or minus 5 kilograms)
- not very accurate (within a range of plus or minus 10 kilograms)

(**Answer:** Somewhat accurate. One way to see this is to shade a region parallel to the line with width ±5 kilograms to show that most of the data falls within this range of the predicted values.)

(5) In previous lessons, you studied the concept of correlation to describe the strength and direction of the linear association between two quantitative variables. Now you are working on predicting the value of one variable based on the other. Are these two ideas related? Explain your reasoning.

(**Answer:** Yes. If the relationship looks linear, a strong association, indicated by r close to 1 or –1, indicates that there is not a lot of scatter about the line. So predictions will probably be more accurate.)

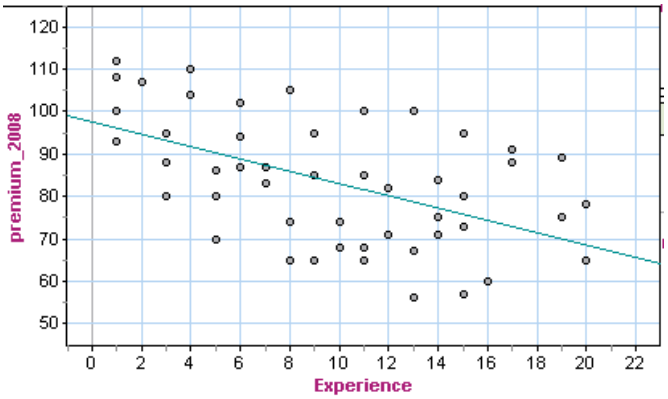
Initiating Lesson 3.2.1: Using Lines to Make Predictions

Wrap-Up [Student Handout, 6 minutes]

(**Note:** Give students about three minutes to do the following exercise. Assess class performance in aggregate protecting anonymity by using clickers or have students write their answer anonymously on a piece of paper, collect, redistribute, and tally responses with students reporting the answer on the paper they receive. Class results will guide you in determining whether more explanation is necessary.)

In 2008, a statistics student gathered data on monthly car insurance premiums paid by students and faculty at Los Medanos College. Relating monthly car insurance premiums to years of driving experience, she found a linear relationship and used statistical methods to get the following equation:

predicted monthly car insurance premium = 97 – 1.45(years of driving experience)



(8) Predict the monthly car insurance premium paid by someone who has been driving 12 years.

(9) Which of the following methods can be used to make the prediction?

(a) Find 12 on the horizontal axis, trace up to the line, and read off the corresponding value on the *y*-axis.

(b) Substitute 12 in the equation, and calculate the predicted premium.

(c) Look at the data and find a person who has been driving 12 years. Report the premium paid by this person.

(d) Both a and b.

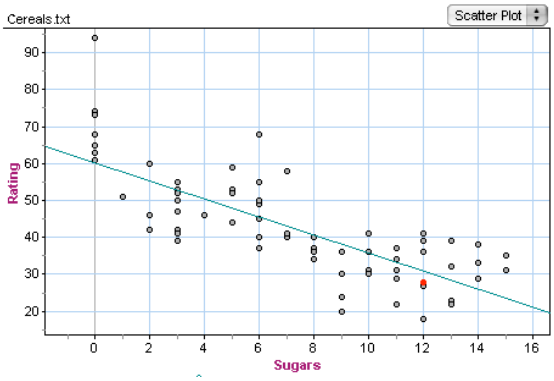
(e) Both b and c.

(**Answer:** $80; d)

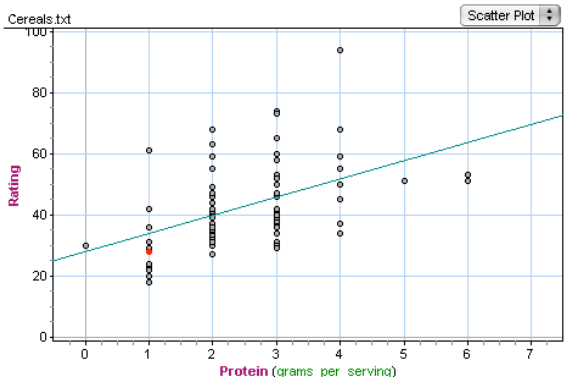
Initiating Lesson 3.2.1: Using Lines to Make Predictions

Homework

(10) Here you return to the data set for the 77 breakfast cereals you investigated at the beginning of Module 3.



ratings = 60 – 2.43(sugars)



ratings = 28 + 5.96(protein)

Two new cereals are being rated by *Consumer Reports*. Cereal A has 10.5 grams of sugar in a serving and Cereal B has 2.5 grams of protein in a serving.

(a) Predict the *Consumer Reports* rating for the two cereals using the best-fit lines.

(b) For which cereal do you think your prediction is probably more accurate (more likely to be closer to the actual *Consumer Reports* rating)? Why?

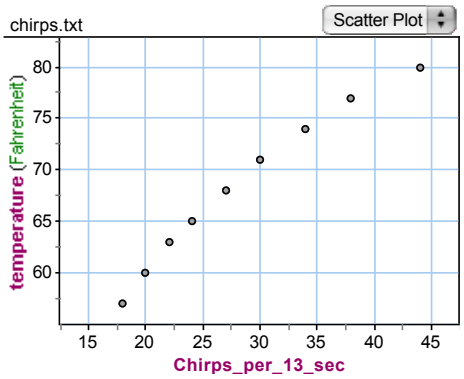
(**Answers:** Cereal A: 34.485, Cereal B: 22.9. Note that the line does not give whole number ratings. Students may round to indicate a realistic rating value. The prediction for Cereal A is probably more accurate because there is less scatter about the line.)

(11) Can the rate that crickets chirp be used to predict the temperature?

According to Tom Walker, an entomologist with the University of Florida, all crickets are pretty good thermometers because they chirp at a rate that is related to the temperature. The chirping noise results when the cricket rubs its wings together. A cricket studied by Walker, the snowy tree cricket (*Oecanthus* fultoni), chirps at a rate that is slow enough to count. These crickets also synchronize their wing rubbing so determining the chirp rate easier. The snowy tree cricket is found throughout the United States. To hear the snowy tree cricket go to http://entnemdept.ufl.edu/walker/buzz/585a.htm.

(a) The scatterplot is a graph of data from the June 1995 issue of *Outside* magazine. Use the scatterplot to predict the temperature when the snowy tree crickets are chirping at a rate of 40 chirps every 13 seconds.

(**Answer:** 77°F ±1°F is a reasonable estimate.)



(b) How accurate do you think your prediction is? Choose the option that is most reasonable and briefly explain your thinking.

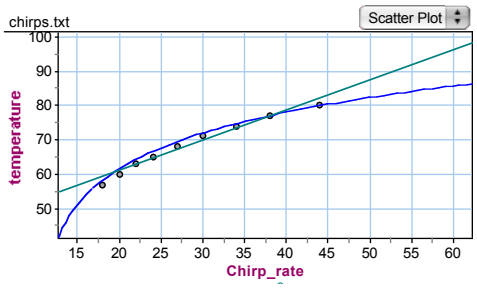
- very accurate (within a range of plus or minus 1 degree)
- somewhat accurate (within a range of plus or minus 5 degrees)
- not very accurate (within a range of plus or minus 10 degrees)

(**Answer:** Very accurate. The association is so strong you can sketch a curve that has essentially no scatter about it.)

(c) This is the same data graphed in two different windows. The data has been zoomed out by expanding both axes. The line pictured is the best-fit line:

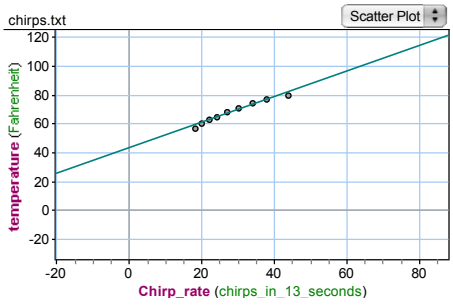
temperature = 0.88(chirp rate) + 43

For some chirp rates, this line gives very accurate predictions of the temperature. However, the data are actually slightly curved, so that for chirp rates above 50 a nonlinear model might give more accurate predictions. One possible nonlinear model is also shown.



The line also has limitations in that some chirp rates are meaningless and should not be used to make predictions.

In statistics, *extrapolation* is the process of using a statistical model (like a line) to make predictions outside the range of the available data. To use a statistical model to make a prediction for an explanatory variable value that is outside the range of values in the data set requires that we make the assumption that the pattern observed in the data continues outside this range. If this is not the case, predictions are unreliable and may be very far off from the actual response variable values. You should be very cautious in doing this.



Illustrate the concept of extrapolation by identifying a point on the line that gives either meaningless results or unreliable results. Explain how this point illustrates the concept of extrapolation.

(**Answers can vary:** Answers need to use chirp rates outside the range of the data to illustrate extrapolation. For example, using the linear model, a chirp rate of 60 gives a prediction of 95.8°F. However, the curve in the data suggests that a prediction of 90 might be more accurate. Another example is using a chirp rate of –10, which is a nonsensical value, and predicting a temperature of 30°F, a temperature at which crickets are probably dead.)

Initiating Lesson 3.2.1: Using Lines to Make Predictions

(12) **A Note About Statistical Vocabulary:** A variable that is used to predict the value of another variable is called the *predictor variable*, also known as the *independent variable* or *explanatory variable*. The other variable, whose values you are predicting, is called the *response variable*, also known as the *dependent variable*.

(a) The introductory problem in this lesson has forearm lengths and heights for 21 female college students. In this situation, which variable is the predictor?

(**Answer:** forearm lengths)

(b) The cereal data has the amount of sugar in a serving and the *Consumer Reports* rating. In this situation, which variable is the predictor?

(**Answer:** sugar)

(c) When graphing bivariate data, you put the predictor variable on the (*choose one:* horizontal axis, vertical axis).

(**Answer:** horizontal axis)

(d) Using measurements of temperature (°F) and the chirp rate of the snowy tree cricket (measured in number of chirps in 13 seconds), students use technology to find a best-fit line. However, some students use temperature as the predictor variable, and others use chirp rate as the predictor variable. For which of the two lines below is temperature treated as the predictor variable?

$$\text{temperature} = 0.88(\text{chirp rate}) + 43$$

$$\text{chirp rate} = 1.1(\text{temperature}) - 47$$

[**Answer:** chirp rate = 1.1(temperature) – 47]

Initiating Lesson 3.2.1: Using Lines to Make Predictions

Part I

Statistical methods are used in forensics to identify human remains based on the measurements of bones. In the 1950s, Dr. Mildred Trotter and Dr. Goldine Gleser measured skeletons of people who had died in the early 1900s. From these measurements they developed statistical methods for predicting a person's height based on the lengths of various bones. These formulas were first used to identify the remains of U.S. soldiers who died in World War II and were buried in unmarked graves in the Pacific zone. Modern forensic scientists have made adjustments to the formulas developed by Trotter and Gleser to account the differences in bone length and body proportions of people living now. You will not use Trotter and Gleser's formulas in this problem, but you will use a similar process.

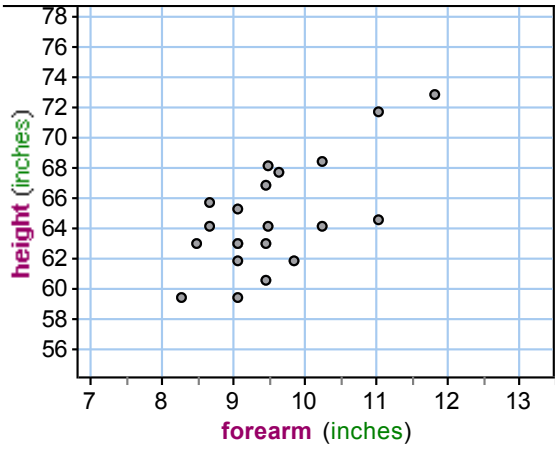
To illustrate the type of data analysis done in forensics, let's see if you can identify a female student based on the length of her forearm. The mystery student has a forearm measurement of 10 inches. (She is alive and healthy!)

Height and weight measurements for three female college students are given in the table.

	Jane Doe 1	**Jane Doe 2**	**Jane Doe 3**
Age	18	23	33
Gender	Female	Female	Female
Height	5 feet, 5 inches	5 feet, 2 inches	6 feet
Weight	128 pounds	120 pounds	155 pounds

Your task is to determine if the mystery student could be one of these three students.

First, you need data that relates forearm length to either height or weight for females. The scatterplot is a graph of height versus forearm length for 21 female college students taking Introductory Statistics at Los Medanos College in Pittsburg, California, in 2009.



(1) Based on the scatterplot, what is a reasonable prediction for the height of the mystery student? Briefly explain or show how you made your prediction.

(2) The variability in the data makes it difficult to determine if one of these students is the mystery student. Could any of the three students be eliminated as a possibility of being the mystery student? Explain your reasoning.

Initiating Lesson 3.2.1: Using Lines to Make Predictions

Part II: Using a Line to Make Predictions

(3) The scatterplot has a positive linear association. The correlation is 0.68, which is pretty strong. So, it makes sense to use a linear model to summarize the relationship between the forearm and height measurements. There is one line that is considered the best description of how height and forearm length are related. You will learn more about how to find this line in future lessons. For now, you will use technology to find the equation of this line.

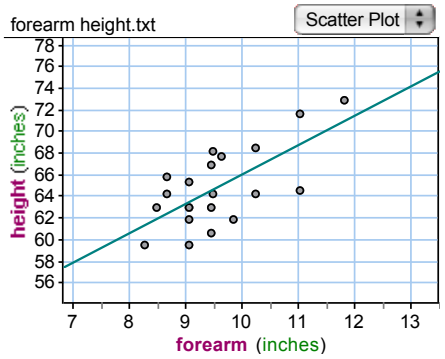
(a) Use the graph of the best-fit line to predict the height of the mystery student.

(b) The equation of this line is approximately

predicted height = 2.7(forearm length) + 39.

$\hat{y} = 2.7x + 39$

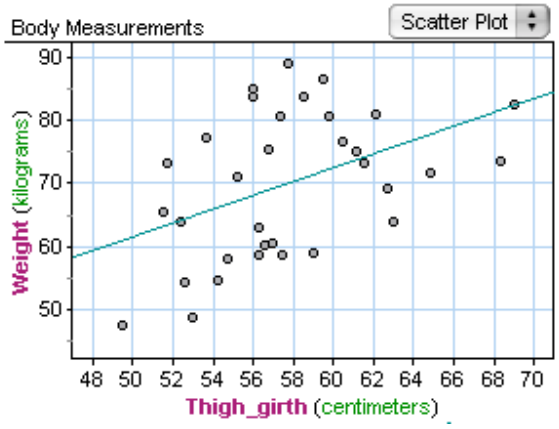
(Notice that when you use letters to represent variables in the prediction line, you put a "hat" on the y and write $\hat{y}$ instead of y. The hat is a signal that the variable is *predicted* values, not actual data values.)

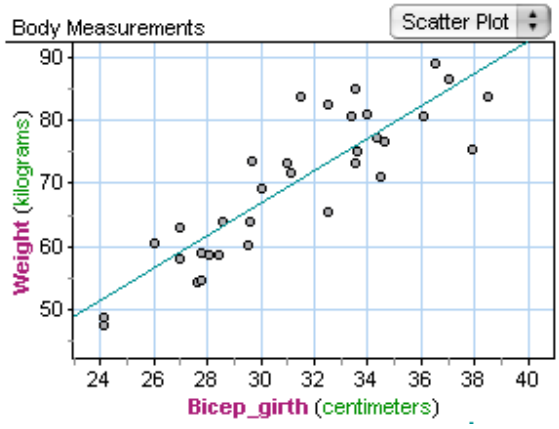


Use the equation to predict the height of the mystery person.

(c) Is the height of Jane Doe 1, 2, or 3 closest to the predicted height of the mystery student given by the line? (Of course, this does not guarantee that you have correctly identified the mystery student, but it suggests that one student's height, together with the 10-inch forearm measurement, fits the linear pattern in the data better than the other students.)

(4) The scatterplots below are graphs of body measurements in centimeters for 34 adults who are physically active. These data are a random sample taken from a larger nonrandom data set gathered by researchers investigating the relationship of various body measurements and weight. Girth is the measurement around a body part. (Retrieved from www.amstat.org/publications/jse/v11n2/datasets.heinz.html)





Initiating Lesson 3.2.1: Using Lines to Make Predictions

(a) Based on these data, which do you think is a better predictor of an adult's weight: thigh girth or bicep girth? Why?

(b) Adriana has a thigh girth of 57 centimeters and a bicep girth of 25 centimeters. Predict her weight using the measurement that you think will give the most accurate prediction, and then plot Adriana on the scatterplot that you used to make her weight prediction.

(c) The equations of the two lines shown are

weight = 6.3 + 1.1(thigh girth) weight = –10.5 + 2.6(bicep girth)

Predict Adriana's weight using the equation that you think best predicts weight.

(d) Of course, you do not really know Adriana's weight. How accurate do you think the line's prediction of Adriana's weight is? Choose the option that is the most reasonable and explain your thinking.

- very accurate (within a range of plus or minus 1 kilogram)
- somewhat accurate (within a range of plus or minus 5 kilograms)
- not very accurate (within a range of plus or minus 10 kilograms)

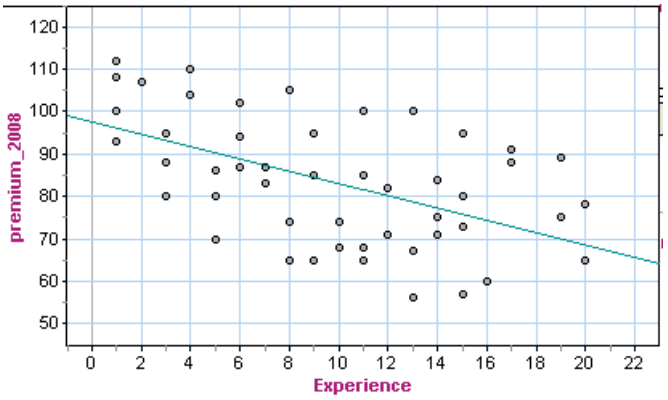
(5) In previous lessons, you studied the concept of correlation to describe the strength and direction of the linear association between two quantitative variables. Now you are working on predicting the value of one variable based on the other. Are these two ideas related? Explain your reasoning.

Initiating Lesson 3.2.1: Using Lines to Make Predictions

Class Discussion

In 2008, a statistics student gathered data on monthly car insurance premiums paid by students and faculty at Los Medanos College. Relating monthly car insurance premiums to years of driving experience, she found a linear relationship and used statistical methods to get the following equation:

predicted monthly car insurance premium = 97 – 1.45(years of driving experience)



(8) Predict the monthly car insurance premium paid by someone who has been driving 12 years.

(9) Which of the following methods can be used to make the prediction?

(a) Find 12 on the horizontal axis, trace up to the line, and read off the corresponding value on the *y*-axis.

(b) Substitute 12 in the equation, and calculate the predicted premium.

(c) Look at the data and find a person who has been driving 12 years. Report the premium paid by this person.

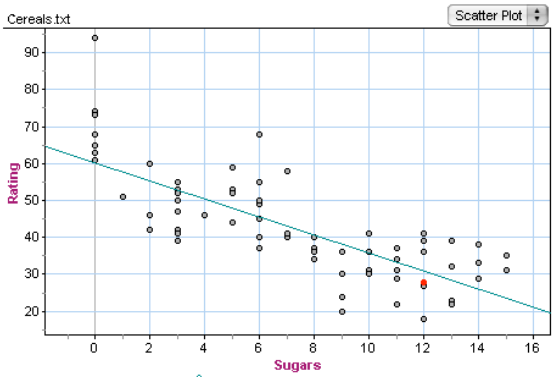
(d) Both a and b.

(e) Both b and c.

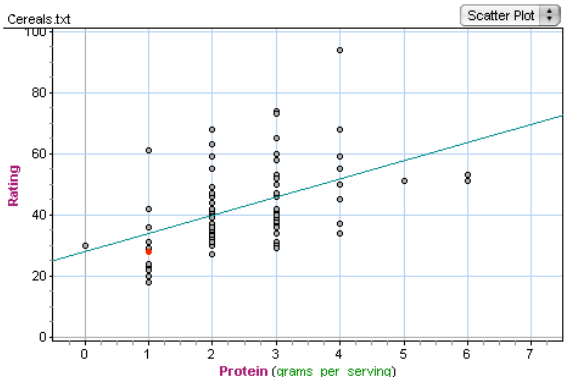
Initiating Lesson 3.2.1: Using Lines to Make Predictions

Homework

(10) Here you return to the data set for the 77 breakfast cereals you investigated at the beginning of Module 3.



ratings = 60 – 2.43(sugars)



ratings = 28 + 5.96(protein)

Two new cereals are being rated by *Consumer Reports*. Cereal A has 10.5 grams of sugar in a serving and Cereal B has 2.5 grams of protein in a serving.

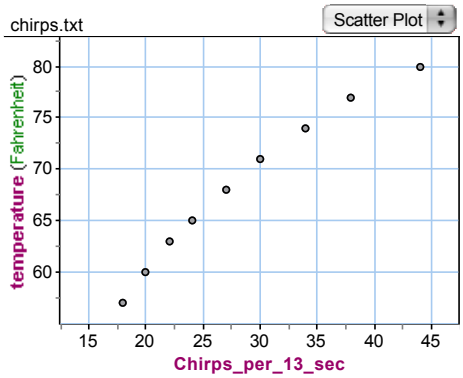
(a) Predict the *Consumer Reports* rating for the two cereals using the best-fit lines.

(b) For which cereal do you think your prediction is probably more accurate (more likely to be closer to the actual *Consumer Reports* rating)? Why?

(11) Can the rate that crickets chirp be used to predict the temperature?

According to Tom Walker, an entomologist with the University of Florida, all crickets are pretty good thermometers because they chirp at a rate that is related to the temperature. The chirping noise results when the cricket rubs its wings together. A cricket studied by Walker, the snowy tree cricket (*Oecanthus* fultoni), chirps at a rate that is slow enough to count. These crickets also synchronize their wing rubbing so determining the chirp rate easier. The snowy tree cricket is found throughout the United States. To hear the snowy tree cricket go to http://entnemdept.ufl.edu/walker/buzz/585a.htm.

(a) The scatterplot is a graph of data from the June 1995 issue of *Outside* magazine. Use the scatterplot to predict the temperature when the snowy tree crickets are chirping at a rate of 40 chirps every 13 seconds.



Initiating Lesson 3.2.1: Using Lines to Make Predictions

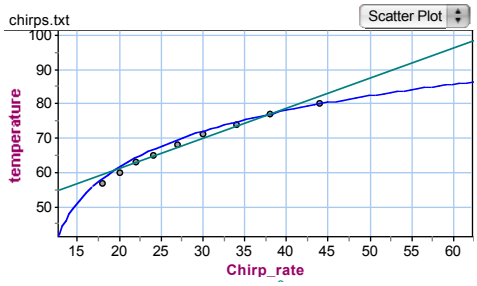
(b) How accurate do you think your prediction is? Choose the option that is most reasonable and briefly explain your thinking.

- very accurate (within a range of plus or minus 1 degree)
- somewhat accurate (within a range of plus or minus 5 degrees)
- not very accurate (within a range of plus or minus 10 degrees)

(c) This is the same data graphed in two different windows. The data has been zoomed out by expanding both axes. The line pictured is the best-fit line:

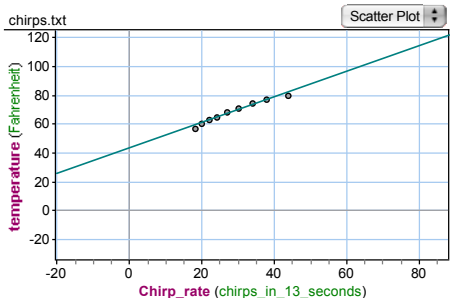
temperature = 0.88(chirp rate) + 43

For some chirp rates, this line gives very accurate predictions of the temperature. However, the data are actually slightly curved, so that for chirp rates above 50 a nonlinear model might give more accurate predictions. One possible nonlinear model is also shown.



The line also has limitations in that some chirp rates are meaningless and should not be used to make predictions.

In statistics, *extrapolation* is the process of using a statistical model (like a line) to make predictions outside the range of the available data. To use a statistical model to make a prediction for an explanatory variable value that is outside the range of values in the data set requires that we make the assumption that the pattern observed in the data continues outside this range. If this is not the case, predictions are unreliable and may be very far off from the actual response variable values. You should be very cautious in doing this.



Illustrate the concept of extrapolation by identifying a point on the line that gives either meaningless results or unreliable results. Explain how this point illustrates the concept of extrapolation.

Initiating Lesson 3.2.1: Using Lines to Make Predictions

(12) **A Note About Statistical Vocabulary:** A variable that is used to predict the value of another variable is called the *predictor variable*, also known as the *independent variable* or *explanatory variable*. The other variable, whose values you are predicting, is called the *response variable*, also known as the *dependent variable*.

(a) The introductory problem in this lesson has forearm lengths and heights for 21 female college students. In this situation, which variable is the predictor?

(b) The cereal data has the amount of sugar in a serving and the *Consumer Reports* rating. In this situation, which variable is the predictor?

(c) When graphing bivariate data, you put the predictor variable on the (*choose one:* horizontal axis, vertical axis).

(d) Using measurements of temperature (°F) and the chirp rate of the snowy tree cricket (measured in number of chirps in 13 seconds), students use technology to find a best-fit line. However, some students use temperature as the predictor variable, and others use chirp rate as the predictor variable. For which of the two lines below is temperature treated as the predictor variable?

$$\text{temperature} = 0.88(\text{chirp rate}) + 43$$

$$\text{chirp rate} = 1.1(\text{temperature}) - 47$$

Supporting Lesson 3.2.2: Least Squares Regression Line as Line of Best Fit

Estimated number of 50-minute class sessions: 2

Goals

Students will understand that

- the least squares regression line is the line that minimizes the sum of the squared vertical deviations from the line.

Students will be able to

- given a bivariate numerical data set, find the equation of the least squares regression line using technology.
- explain the meaning of *least squares* in a regression setting.

Task 1: Comparing Lines for Predicting Textbook Costs [30–45 minutes total]

Introduction (Student Handout, 3 minutes)

In the previous lesson, you predicted the value of the response variable knowing the value of the *explanatory variable* (also known as *predictor variable*) using a best-fit line. In the homework you also investigated the concept of extrapolation, which is the idea that, even with the best line, the predictions based on this line may be unreliable if the value of the explanatory variable is outside the range of the data.

So, how do you identify the line that is the best fit? You will use technology to find the equation of the line, but what does it mean to say that a particular line is the best fit? In this lesson, you will investigate this question with the goal of developing a method for determining which line is the best-fit line.

Activities [15–25 minutes]

(1) Here are the publishers' suggested list prices in 2010 for 12 popular introductory statistics textbooks. The table below gives the descriptive statistics for the price data.

	Min	Q1	Median	Q3	Max	Mean	Standard Deviation
List price	170.95	122.00	150.67	162.55	190.95	147.61	25.72

Stats textbooks

	price
1	150.67
2	122.00
3	149.10
4	166.15
5	107.95
6	181.95
7	158.95
8	151.95
9	122.00
10	150.67
11	190.95
12	118.95

(a) If someone asks you how much an introductory statistics textbook costs, what prediction would you give? Explain your reasoning.

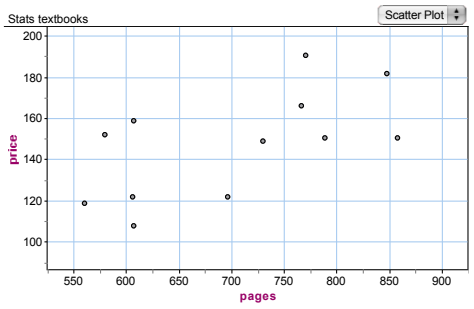
(**Answers will vary:** Reasonable answers are the mean or median, or perhaps a range mean ± stddev or Q1 or Q3.)

(b) What variables might be useful for predicting the cost of an introductory statistics textbook?

(**Answers will vary:** Possibilities include both categorical and quantitative variables, such as type of course, hard cover vs. soft cover vs. e-books, number of pages.)

Supporting Lesson 3.2.2: Least Squares Regression Line as Line of Best Fit

(c) The number of pages in the textbook is one variable you could use to predict price. The scatterplot shows the relationship between pages and price for these 12 textbooks. The data have a somewhat linear form and the correlation coefficient is 0.79, so it makes sense to use a line to summarize the relationship between pages and price. Draw a line that you think is a good summary of the relationship between these two variables. Use the graph of your line to predict the price of a 650-page textbook. Then compare your prediction with a classmate.



(2) Since there are infinitely many lines that you could draw, you need a way to determine which line is the best summary of the relationship between two quantitative variables.

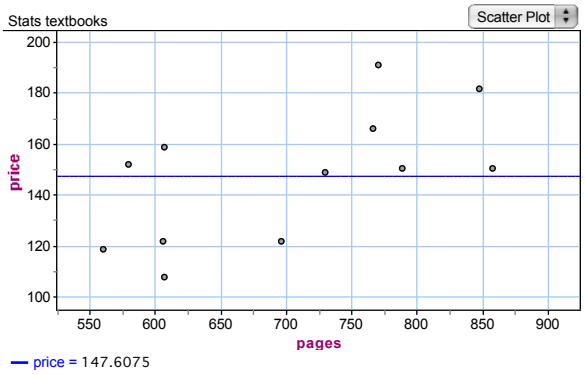
You will begin your investigation of how to define a best-fit line by comparing how well four lines predict the list price of the textbooks based on the number of pages.

Stats textbooks

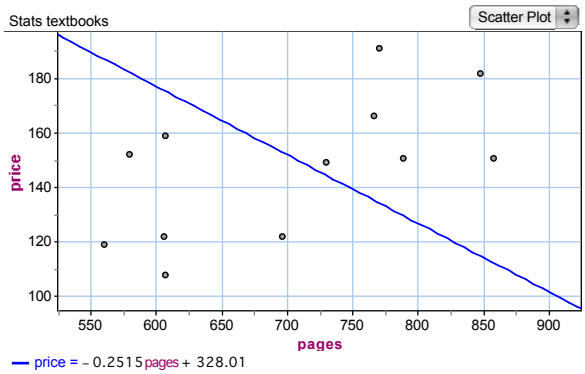
	pages	price	Line_A_predictions	Line_B_predictions	Line_C_predictions	Line_D_predictions
1	560	118.95	147.6075	187.1700	126.3852	109.8900
2	579	151.95	147.6075	182.3915	129.2428	114.6020
3	606	122.00	147.6075	175.6010	133.3036	121.2980
4	607	107.95	147.6075	175.3495	133.4540	121.5460
5	607	158.95	147.6075	175.3495	133.4540	121.5460
6	696	122.00				
7	730	149.10	147.6075	144.4150	151.9532	152.0500
8	766	166.15	147.6075	135.3610	157.3676	160.9780
9	770	190.95	147.6075	134.3550	157.9692	161.9700
10	788	150.67	147.6075	129.8280	160.6764	166.4340
11	847	181.95				
12	857	150.67	147.6075	112.4745	171.0540	183.5460

Supporting Lesson 3.2.2: Least Squares Regression Line as Line of Best Fit

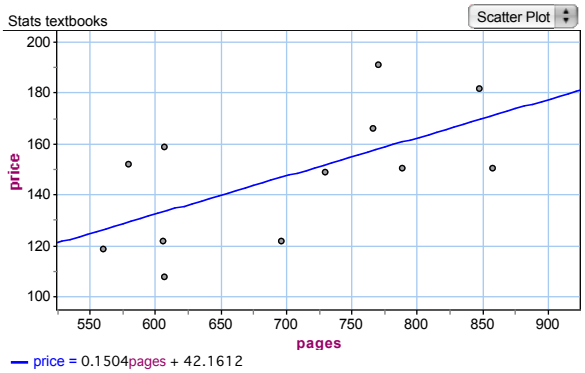
Line A (Mean Price)



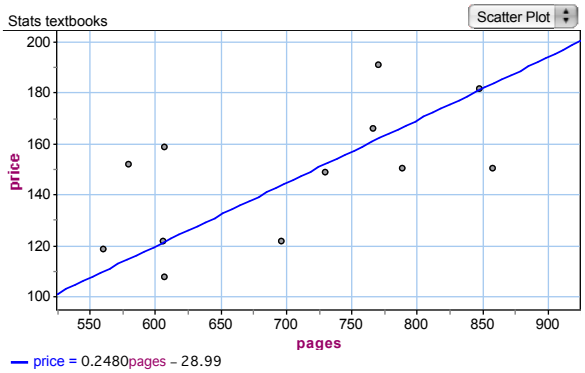
Line B



Line C



Line D



(a) Begin by using the equation for each line to complete the two incomplete rows in the table of predicted values. (You are predicting prices. It makes sense to write prices with two decimal places, such as $147.61 instead of $147.6075 like you see in the table. You might be wondering why you are recording answers to four decimal places. This is because you will need this level of accuracy to develop some ideas later. So, record your answers to four decimal places for these activities.)

(b) Which of the four lines do you think results in the best overall predictions of price? Why? How are you selecting the best line?

Wrap-Up Questions/Direct Instruction About Statistical Concepts [10–15 minutes]

Poll the class to see which lines they choose as the best summary. For each line that receives votes as the best summary, call on a few students to explain why they (or their group) chose that particular line. Students may have difficulty articulating their observations as criteria of why one line seems a better summary, so translate their explanations into criteria. Double-check that you have captured the idea they are trying to articulate. Explain how each criterion is visualized in the scatterplot and how it also relates to numerical information in the table. Do not worry if some of their criteria do not characterize regression lines. You can revisit these criteria at the end of the lesson after students have learned about the least squares criterion.

Supporting Lesson 3.2.2: Least Squares Regression Line as Line of Best Fit

Examples of Criterion from Visual and Numerical Perspectives

	Graph	Table
Possible Criterion	The line should have the same direction as association between the variables. (Slope should be the same sign as the correlation coefficient.)	As the number of pages increases, so does the predicted price.
	The line should come as close as possible to as many points as possible.	The predicted prices are as close as possible to the actual prices. (The differences between predicted price and actual price are as small as possible.)
	The line should go through as many points as possible.	For as many textbooks as possible, the predicted price should equal the actual price.
	The line should have the same number of points above it as below it.	The number of predictions that are greater than the list price equals the number of predictions that are less than the list price.

This table is an example of how to discuss possible criteria. Students might generate other ideas.

Ideas to Discuss as a Transition to Task 2 [5 minutes]

You want a line that minimizes the errors in predictions for individuals in the sample. The reasoning is that if the line is good at predicting the response for the textbooks in the sample, when the response is already known, then it will work well for predicting the response in the future when only the explanatory variable is known.

Now let's get more precise and transform some of the criteria into a numerical measurement that can be used to identify which line gives the best fit.

The idea that you want to develop here is that the line that is the best summary of the relationship between the number of pages and the price of the textbook will give the best predictions for price, but most predictions will have some error. Where possible, tie the criteria generated by the class to the idea of prediction error (e.g., the line gives accurate predictions for points close to it, which means that for a given number of pages, the predicted price is close to the list price).

You can think of the price of the textbook as the price predicted by the linear model plus some error, Price = Prediction + Error. (Some statisticians refer to this idea more generally as Data = Model + Error.) To determine the prediction error, you calculate the difference between price and the predicted price, Price – Prediction = Error.

For the following activities in Task 2, intervene if students need help. Provide guidance as necessary to help students correctly answer the questions about predicted errors.

Supporting Lesson 3.2.2: Least Squares Regression Line as Line of Best Fit

Task 2: Thinking About Prediction Error

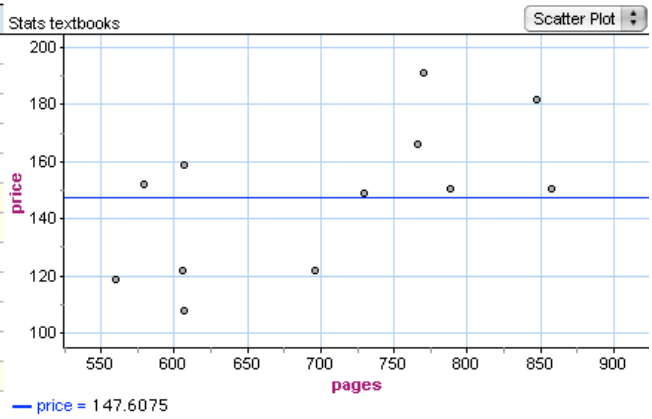
Activities [Student Handout, 15–25 minutes]

(3) For each linear model, complete the missing parts of the table and answer the questions.

(a) Line A (Mean Price)

Stats textbooks_2

	pages	price	Line_A_predictions	Line_A_error
1	560	118.95	147.6075	-28.6575
2	579	151.95	147.6075	4.3425
3	606	122.00	147.6075	-25.6075
4	607	107.95	147.6075	-39.6575
5	607	158.95	147.6075	11.3425
6	696	122.00	147.6075	
7	730	149.10	147.6075	1.4925
8	766	166.15	147.6075	18.5425
9	770	190.95	147.6075	43.3425
10	788	150.67	147.6075	3.0625
11	847	181.95	147.6075	
12	857	150.67	147.6075	3.0625

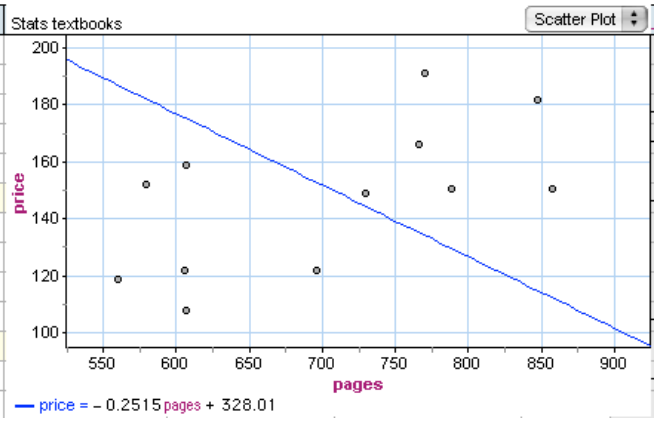


- Identify the following textbooks in the scatterplot and the table:
  - the textbook for which the line comes closest to predicting the list price
  - the textbook for which the prediction is furthest from the list price
- In the scatterplot, circle the textbooks that have a negative prediction error. What does a negative error tell you?

(b) Line B

Stats textbooks

	pages	price	Line_B_predictions	Line_B_error
1	560	118.95	187.1700	-68.2200
2	579	151.95	182.3915	-30.4415
3	606	122.00	175.6010	-53.6010
4	607	107.95	175.3495	-67.3995
5	607	158.95	175.3495	-16.3995
6	696	122.00	152.9660	
7	730	149.10	144.4150	4.6850
8	766	166.15	135.3610	30.7890
9	770	190.95	134.3550	56.5950
10	788	150.67	129.8280	20.8420
11	847	181.95	114.9895	
12	857	150.67	112.4745	38.1955



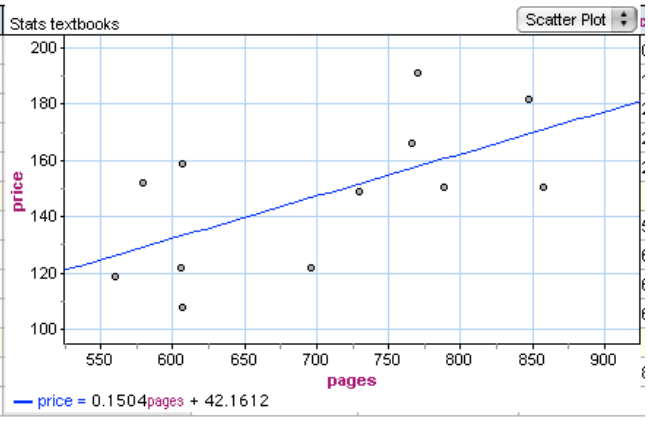
- Identify the following textbooks in the scatterplot and the table:
  - the textbook for which the line comes closest to predicting the list price
  - the textbook for which the prediction is furthest from the list price
- How can you tell by looking at the scatterplot if the prediction error for a textbook is positive or negative?
- Identify a textbook for which Line A predicts too low a price but Line B predicts too high a price.

Supporting Lesson 3.2.2: Least Squares Regression Line as Line of Best Fit

(c) Line C

Stats textbooks

	pages	price	Line_C_predictions	Line_C_error
1	560	118.95	126.3852	-7.4352
2	579	151.95	129.2428	22.7072
3	606	122.00	133.3036	-11.3036
4	607	107.95	133.4540	-25.5040
5	607	158.95	133.4540	25.4960
6	696	122.00	146.8396	
7	730	149.10	151.9532	-2.8532
8	766	166.15	157.3676	8.7824
9	770	190.95	157.9692	32.9808
10	788	150.67	160.6764	-10.0064
11	847	181.95	169.5500	
12	857	150.67	171.0540	-20.3840

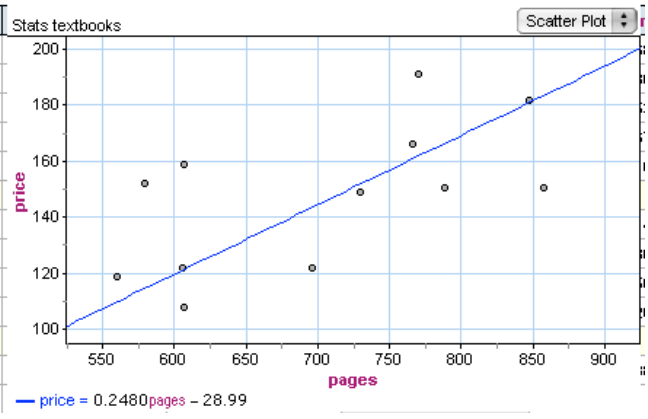


- Identify the following textbooks in the scatterplot and the table:
  - the textbook for which the line comes closest to predicting the list price
  - the textbook for which the prediction is furthest from the list price
  - all the textbooks for which the predicted list price is within $15 of the actual list price
- How can you tell by looking at the scatterplot that the prediction error is positive?

(d) Line D

Stats textbooks

	pages	price	Line_D_predictions	Line_D_error
1	560	118.95	109.8900	9.0600
2	579	151.95	114.6020	37.3480
3	606	122.00	121.2980	0.7020
4	607	107.95	121.5460	-13.5960
5	607	158.95	121.5460	37.4040
6	696	122.00	143.6180	
7	730	149.10	152.0500	-2.9500
8	766	166.15	160.9780	5.1720
9	770	190.95	161.9700	28.9800
10	788	150.67	166.4340	-15.7640
11	847	181.95	181.0660	
12	857	150.67	183.5460	-32.8760



- Identify the following textbooks in the scatterplot and the table:
  - the textbook for which the line comes closest to predicting the list price
  - the textbook for which the prediction is furthest from the list price
  - all the textbooks for which the predicted list price is exceeds the actual list price by $20 or more

(e) The goal is to identify a line that is the best summary of the relationship between pages and price. The best-fit line gives the best predictions of list price, which means that overall it has the least amount of error in the predictions. Rank the four lines from best to worst with the best being the line that gives the best overall predictions of list price. Briefly explain the reasoning behind your rankings.

Wrap-Up Questions/Direct Instruction About Statistical Concepts [20–35 minutes]

Frame this part of the lesson as a discussion of how to use the prediction error for each textbook to define a measure of overall prediction error for the line. You will discuss different possible ways to measure overall prediction error, but ultimately you will only use one measure.

Of course, you want to be able to rank the lines from best to worst using the measurement of overall error. Begin by getting a sense of the class's ranking of the lines. If students worked in groups on the previous tasks, you might take a quick tally on the board and then quickly develop a ranking that reflects (as best as you can) the groups' rankings.

Group	1 = best	2	3	4 = worst
1	D	C	A	B
2	C	D	A	B
etc.				

Begin with a simple way to determine the overall error: just sum the errors. Then discuss whether the sum of the errors helps identify the line that best summarizes the data.

When you add up the errors, you get the following sums:

Line	Sum of Errors
A	0
B	–48.8405
C	0.0404
D	32.7460

Does the sum of the errors help identify the line that makes the best predictions? Why or why not? (**Note:** Even if you are conducting this portion of the lesson as a lecture, let students think about this for a minute before you proceed with the following answer to the question.)

If you use the sum of the errors to identify the best line, you choose Line A as the best because the cumulative error is zero. Line A uses the mean price as the prediction for price of every textbook. However, this line does not appear to be the best line to summarize the data because the flat mean line does not capture the positive association between pages and price. So, the sum of the errors is a poor way to measure overall error.

Why is the sum of the prediction errors from the mean line zero? You know from your work in Module 2 that the sum of the deviations from the mean is always zero (because positive and negative deviations combine to give a sum of zero). This is why you developed a more sophisticated way to measure spread relative to the mean using variance and standard deviation.

You need to do the same type of thinking here. You need to make the errors all positive to get a sense of the total error. There are two obvious ways to do this:

Supporting Lesson 3.2.2: Least Squares Regression Line as Line of Best Fit

- use the absolute value of the errors or
- square each error and then sum.

Use both strategies to calculate the overall error for the four lines and see which (if either) of the methods helps identify the line that best fits the data.

Have students take notes using the following three tables. Keep students focused on the purpose of this investigation: to develop a measure of overall error that can be used to identify the line that best fits the data. Give students a few minutes to complete the tables.

	pages	price	Line_A_predictions	Line_A_error	Absolute_value_of_Line_A_error	Line_A_error_squared
1	560	118.95	147.6075	-28.6575	28.6575	
2	579	151.95	147.6075	4.3425	4.3425	
3	606	122.00	147.6075	-25.6075	25.6075	655.7441
4	607	107.95	147.6075	-39.6575		1572.7173
5	607	158.95	147.6075	11.3425		128.6523
6	696	122.00	147.6075	-25.6075	25.6075	655.7441
7	730	149.10	147.6075	1.4925	1.4925	2.2276
8	766	166.15	147.6075	18.5425	18.5425	343.8243
9	770	190.95	147.6075	43.3425	43.3425	1878.5723
10	788	150.67	147.6075	3.0625	3.0625	9.3789
11	847	181.95	147.6075	34.3425	34.3425	1179.4073
12	857	150.67	147.6075			

	pages	price	Line_C_predictions	Line_C_error	Absolute_Values_of_Line_C_error	Line_C_error_squared
1	560	118.95	126.3852	-7.4352	7.4352	55.2822
2	579	151.95	129.2428	22.7072	22.7072	
3	606	122.00	133.3036	-11.3036	11.3036	
4	607	107.95	133.4540	-25.5040	25.5040	650.4540
5	607	158.95	133.4540	25.4960	25.4960	650.0460
6	696	122.00	146.8396	-24.8396		617.0057
7	730	149.10	151.9532	-2.8532	2.8532	8.1408
8	766	166.15	157.3676	8.7824		77.1305
9	770	190.95	157.9692	32.9808	32.9808	1087.7332
10	788	150.67	160.6764	-10.0064	10.0064	100.1280
11	847	181.95	169.5500	12.4000	12.4000	153.7600
12	857	150.67	171.0540			

Which measures of the total error help you determine how well a line fits the data?			
Line	*Sum of Errors*	*Sum of Absolute Value of Errors (SAE)*	*Sum of Squares of Errors (SSE)*
A	0.0000	239.0600	7,275.7566
B	–48.9605	485.0950	24,774.1494
C	0.0404	204.6924	4,458.5762
D	32.7460	206.3540	5,734.1069

Supporting Lesson 3.2.2: Least Squares Regression Line as Line of Best Fit

(**Note:** Quickly provide answers to the missing parts of the tables with brief explanations as you go.)

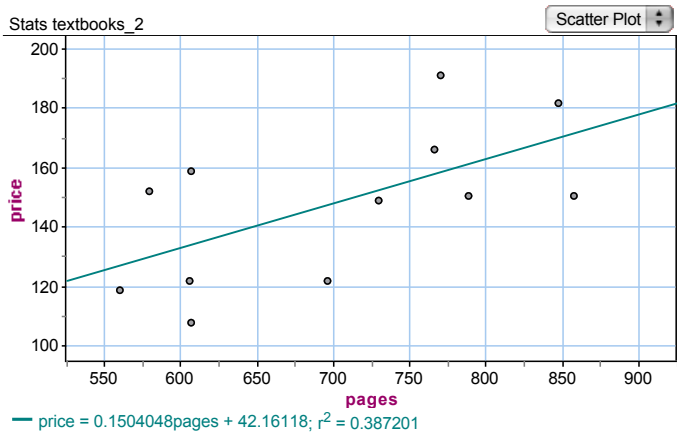
Use these follow-up questions:

- How does the first row of numbers in the last table relate to the previous table for Line A? (**Answer:** These are the sums of the numbers in the previous columns.)
- Describe how you calculate the measures of total error given in the last table for Line D. (**Answer:** Essentially describe how the values in the columns in the table for Line C are calculated.)
- Of the four lines you analyzed, which line is the best summary of the relationship between pages and price if you use the sum of the absolute value of the errors? Which is the second best summary line using this criterion? Third? Fourth? (**Answer:** C, D, A, B)
- Of the four lines you analyzed, which line is the best summary of the relationship between pages and price if you use the sum of the squares of the errors? Rank the lines from best to worst using this criterion. (**Answer:** C, D, A, B)

Statisticians square the errors and then find the line that minimizes the sum of the squared errors. The line that has the smallest sum of the squared errors is called the *least squares regression line*. This line minimizes the sum of the squares of the errors, when compared to **all** other possible lines. You will use technology to find the equation of the least squares regression line. In the next lesson, you will learn more about the distinguishing features of this line.

For these data, Line C is very close to the least squares regression line. Graphically they are indistinguishable. Here is the least squares line shown in the scatterplot.

Now add the least squares line to your table of measurements for total error.



Notice the following when you compare the least squares line to the other lines:

- The sum of the errors is zero for the least squares regression line.
- The sum of the squares of the errors is the smallest for this line (hence the name *least squares*). This is true if you compare the least squares regression line to any other line you created.
- The sum of the absolute value of the errors is not the smallest for the least squares regression line. Line C is a better fit if you use this criterion.

Supporting Lesson 3.2.2: Least Squares Regression Line as Line of Best Fit

Which measures of the total error help determine how well a line fits the data?			
Line	*Sum of Errors*	*SAE*	*SSE*
A	0.0000	239.0600	7,275.7566
B	–48.9605	485.0950	24,774.1494
C	0.0404	204.6924	4,458.5762
D	32.7460	206.3540	5,734.1069
Least squares regression line	0.0000	204.6985	4,458.5761

Now let's return to the criteria you developed to identify best-fit lines at the beginning of the lesson. Given the discussion today, which of the criteria you developed are not true for the least squares line? Which seem to be valid criteria given your work today?

After the discussion/lecture, demonstrate how to find least squares regression lines using technology and/or distribute instructions.

Homework

(4) Here you have data collected from students at Los Medanos College in 2009. The variable *units* gives the number of college course units the student reported he or she was taking that semester. The variable *textbooks* gives the amount that the student reported spending on textbooks or other resources required for their courses that semester.

	units	textbooks
1	3	120.25
2	4	65.95
3	9	465.00
4	12	430.00
5	14	396.50
6	16	475.00
7	8	208.00
8	1	5.00
9	6	49.10
10	15	685.00
11	9	220.00
12	4	172.00
13	12	302.00
14	12	460.12
15	12	530.00

(a) Use technology to find the least squares regression line. (Think carefully about which variable is the explanatory variable.)

(**Answers may vary** slightly based on rounding: textbooks = –42.91 + 38.16 units)

(b) Use the least squares regression line to predict the amount spent on textbooks for a student taking 12 units.

(**Answers may vary** based on rounding in the regression equation: $415.01)

(c) Explain why the least squares regression line is considered the line of best fit.

(**Answer:** The least squares regression line minimizes the sum of the squares of the errors.)

Supporting Lesson 3.2.2: Least Squares Regression Line as Line of Best Fit

(5) With the following applet, you can draw a line that you think fits the data well and compare your line to the least squares regression line.

www.rossmanchance.com/applets/Reg/index.html

Note: In the applet, errors are called *residuals*. This term comes from thinking about a data point as composed of two parts: the part explained by the regression line (the prediction) and the part that is leftover (called the *residual* or *error*).

(a) Instructions

1. Check *Your line* and click *Move line*. Follow directions to move the line so that it fits the data well.
2. Check *Show residuals* and record the SAE for your line in the table below.
3. Check *Show squared residuals* and record the SSE for your line in the table.
4. Check *Regression line*.
5. Check *Show residuals* and record the SAE for the regression line in the table.
6. Check *Show squared residuals* and record the SSE for the regression line in the table.

Line Predicting Height Based on Foot Length	Equation of Line	SAE	SSE
Your line			
Regression line			

(**Answers will vary** for both rows because the applet randomly generates data sets.)

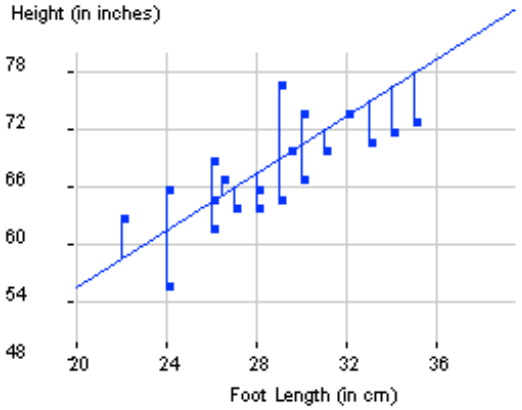
(b) Compare the values of the SAE and SSE for your line with the regression line. What do you notice? Why does this make sense?

(**Answer:** Students should notice that the SSE, and probably the SAE, is less for the regression line. This makes sense because the regression line is designed to minimize the squared errors. It is the best fit when we use the SSE as a criterion of fit.)

(c) When you click *Show residuals*, you see vertical line segments drawn from each data point to the regression line.

Some line segments are long and others are short. Why is this?

(**Answer:** Some vertical line segments are long because the error is large. The line does not predict a value close to the data value. Similarly, some vertical line segments are short because the error is small.)

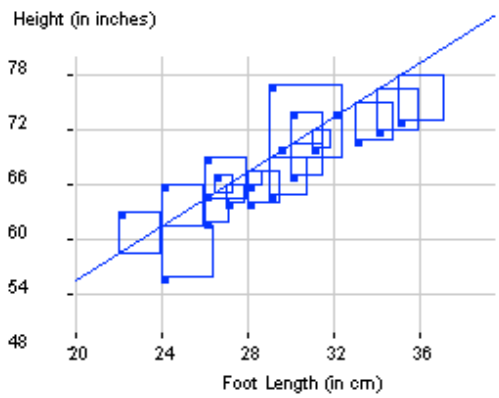


Supporting Lesson 3.2.2: Least Squares Regression Line as Line of Best Fit

Why do you think these vertical line segments are shown?

(**Answer:** These line segments show the size of the error. It is helpful to have a visual way to represent error.)

(d) When you click *Show squared residuals*, you see squares appear.



Some squares are small and others are large. Why is this?

(**Answer:** Some squares are large because the error is large. The squares are drawn based on the length of the vertical line segment representing the error. Similarly, some squares are small because the error is small.)

Why do you think these squares are shown?

(**Answer:** These squares show the size of square of the error. It is helpful to have a visual way to represent squared errors.)

Supporting Lesson 3.2.2: Least Squares Regression Line as Line of Best Fit

Task 1: Comparing Lines for Predicting Textbook Costs

In the previous lesson, you predicted the value of the response variable knowing the value of the *explanatory variable* (also known as *predictor variable*) using a best-fit line. In the homework you also investigated the concept of extrapolation, which is the idea that, even with the best line, the predictions based on this line may be unreliable if the value of the explanatory variable is outside the range of the data.

So, how do you identify the line that is the best fit? You will use technology to find the equation of the line, but what does it mean to say that a particular line is the best fit? In this lesson, you will investigate this question with the goal of developing a method for determining which line is the best-fit line.

(1) Here are the publishers' suggested list prices in 2010 for 12 popular introductory statistics textbooks. The table below gives the descriptive statistics for the price data.

Stats textbooks

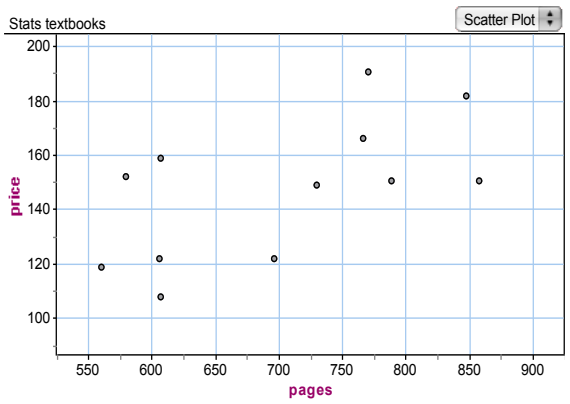
	price
1	150.67
2	122.00
3	149.10
4	166.15
5	107.95
6	181.95
7	158.95
8	151.95
9	122.00
10	150.67
11	190.95
12	118.95

	Min	Q1	Median	Q3	Max	Mean	Standard Deviation
List price	170.95	122.00	150.67	162.55	190.95	147.61	25.72

(a) If someone asks you how much an introductory statistics textbook costs, what prediction would you give? Explain your reasoning.

(b) What variables might be useful for predicting the cost of an introductory statistics textbook?

(c) The number of pages in the textbook is one variable you could use to predict price. The scatterplot shows the relationship between pages and price for these 12 textbooks. The data have a somewhat linear form and the correlation coefficient is 0.79, so it makes sense to use a line to summarize the relationship between pages and price. Draw a line that you think is a good summary of the relationship between these two variables. Use the graph of your line to predict the price of a 650-page textbook. Then compare your prediction with a classmate.



Supporting Lesson 3.2.2: Least Squares Regression Line as Line of Best Fit

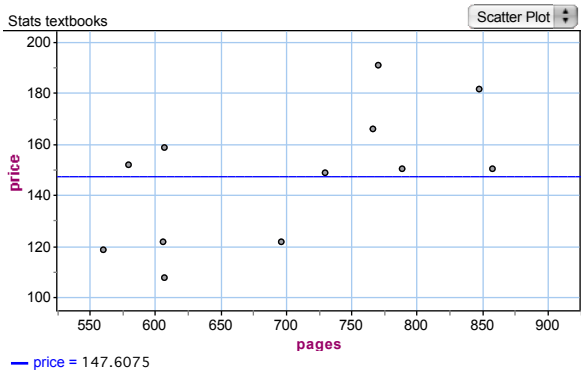
(2) Since there are infinitely many lines that you could draw, you need a way to determine which line is the best summary of the relationship between two quantitative variables.

You will begin your investigation of how to define a best-fit line by comparing how well four lines predict the list price of the textbooks based on the number of pages.

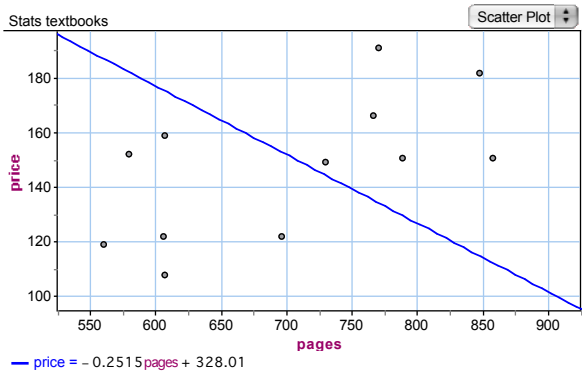
Stats textbooks

	pages	price	Line_A_predictions	Line_B_predictions	Line_C_predictions	Line_D_predictions
1	560	118.95	147.6075	187.1700	126.3852	109.8900
2	579	151.95	147.6075	182.3915	129.2428	114.6020
3	606	122.00	147.6075	175.6010	133.3036	121.2980
4	607	107.95	147.6075	175.3495	133.4540	121.5460
5	607	158.95	147.6075	175.3495	133.4540	121.5460
6	696	122.00				
7	730	149.10	147.6075	144.4150	151.9532	152.0500
8	766	166.15	147.6075	135.3610	157.3676	160.9780
9	770	190.95	147.6075	134.3550	157.9692	161.9700
10	788	150.67	147.6075	129.8280	160.6764	166.4340
11	847	181.95				
12	857	150.67	147.6075	112.4745	171.0540	183.5460

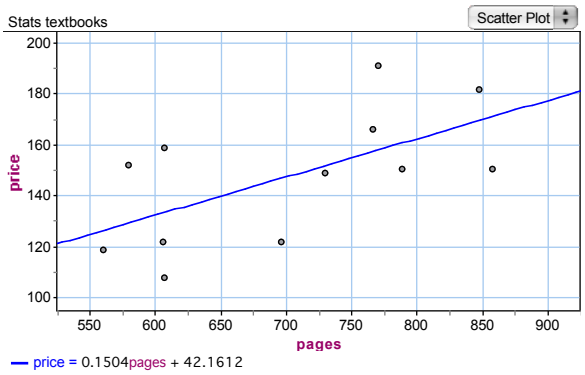
Line A (Mean Price)



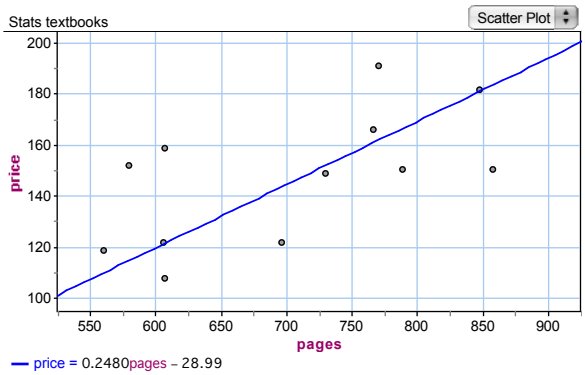
Line B



Line C



Line D



Supporting Lesson 3.2.2: Least Squares Regression Line as Line of Best Fit

(a) Begin by using the equation for each line to complete the two incomplete rows in the table of predicted values. (You are predicting prices. It makes sense to write prices with two decimal places, such as $147.61 instead of $147.6075 like you see in the table. You might be wondering why you are recording answers to four decimal places. This is because you will need this level of accuracy to develop some ideas later. So, record your answers to four decimal places for these activities.)

(b) Which of the four lines do you think results in the best overall predictions of price? Why? How are you selecting the best line?

Supporting Lesson 3.2.2: Least Squares Regression Line as Line of Best Fit

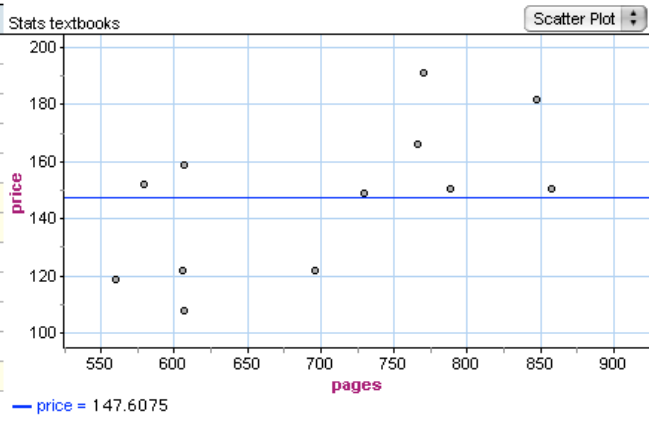
Task 2: Thinking About Prediction Error

(3) For each linear model, complete the missing parts of the table and answer the questions.

(a) Line A (Mean Price)

Stats textbooks_2

	pages	price	Line_A_predictions	Line_A_error
1	560	118.95	147.6075	-28.6575
2	579	151.95	147.6075	4.3425
3	606	122.00	147.6075	-25.6075
4	607	107.95	147.6075	-39.6575
5	607	158.95	147.6075	11.3425
6	696	122.00	147.6075	
7	730	149.10	147.6075	1.4925
8	766	166.15	147.6075	18.5425
9	770	190.95	147.6075	43.3425
10	788	150.67	147.6075	3.0625
11	847	181.95	147.6075	
12	857	150.67	147.6075	3.0625

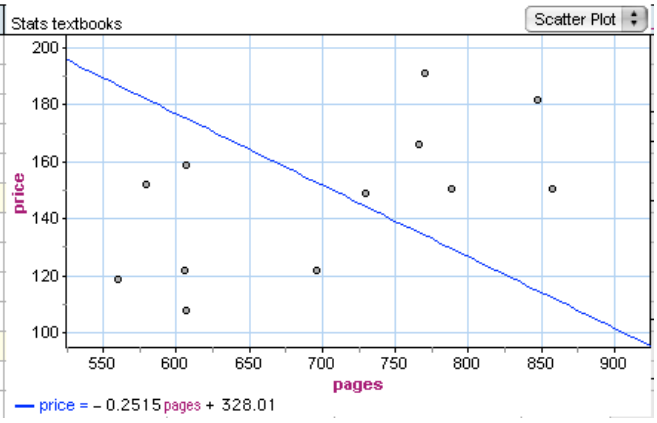


- Identify the following textbooks in the scatterplot and the table:
  - the textbook for which the line comes closest to predicting the list price
  - the textbook for which the prediction is furthest from the list price
- In the scatterplot, circle the textbooks that have a negative prediction error. What does a negative error tell you?

(b) Line B

Stats textbooks

	pages	price	Line_B_predictions	Line_B_error
1	560	118.95	187.1700	-68.2200
2	579	151.95	182.3915	-30.4415
3	606	122.00	175.6010	-53.6010
4	607	107.95	175.3495	-67.3995
5	607	158.95	175.3495	-16.3995
6	696	122.00	152.9660	
7	730	149.10	144.4150	4.6850
8	766	166.15	135.3610	30.7890
9	770	190.95	134.3550	56.5950
10	788	150.67	129.8280	20.8420
11	847	181.95	114.9895	
12	857	150.67	112.4745	38.1955



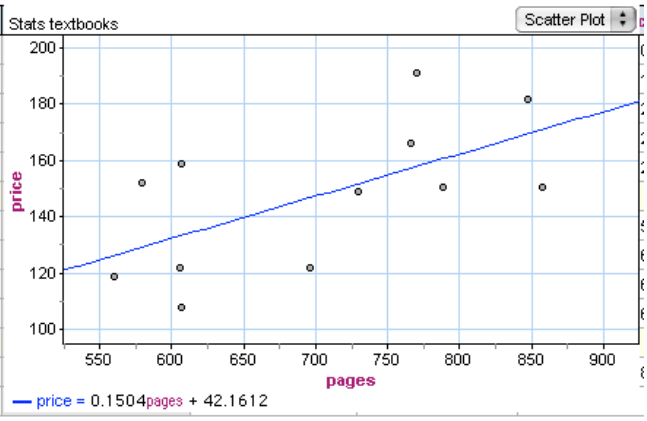
- Identify the following textbooks in the scatterplot and the table:
  - the textbook for which the line comes closest to predicting the list price
  - the textbook for which the prediction is furthest from the list price
- How can you tell by looking at the scatterplot if the prediction error for a textbook is positive or negative?
- Identify a textbook for which Line A predicts too low a price but Line B predicts too high a price.

Supporting Lesson 3.2.2: Least Squares Regression Line as Line of Best Fit

(c) Line C

Stats textbooks

	pages	price	Line_C_predictions	Line_C_error
1	560	118.95	126.3852	-7.4352
2	579	151.95	129.2428	22.7072
3	606	122.00	133.3036	-11.3036
4	607	107.95	133.4540	-25.5040
5	607	158.95	133.4540	25.4960
6	696	122.00	146.8396	
7	730	149.10	151.9532	-2.8532
8	766	166.15	157.3676	8.7824
9	770	190.95	157.9692	32.9808
10	788	150.67	160.6764	-10.0064
11	847	181.95	169.5500	
12	857	150.67	171.0540	-20.3840

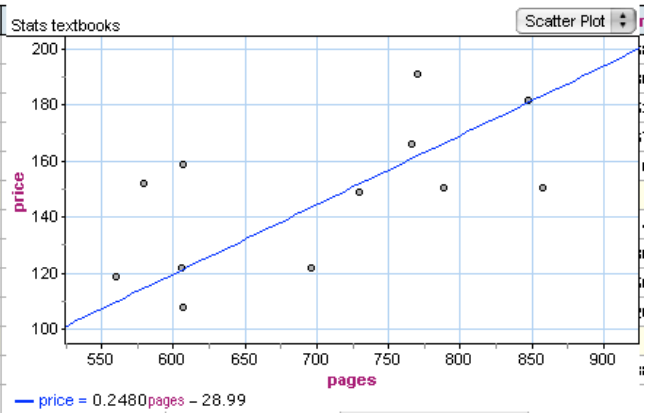


- Identify the following textbooks in the scatterplot and the table:
  - the textbook for which the line comes closest to predicting the list price
  - the textbook for which the prediction is furthest from the list price
  - all the textbooks for which the predicted list price is within $15 of the actual list price
- How can you tell by looking at the scatterplot that the prediction error is positive?

(d) Line D

Stats textbooks

	pages	price	Line_D_predictions	Line_D_error
1	560	118.95	109.8900	9.0600
2	579	151.95	114.6020	37.3480
3	606	122.00	121.2980	0.7020
4	607	107.95	121.5460	-13.5960
5	607	158.95	121.5460	37.4040
6	696	122.00	143.6180	
7	730	149.10	152.0500	-2.9500
8	766	166.15	160.9780	5.1720
9	770	190.95	161.9700	28.9800
10	788	150.67	166.4340	-15.7640
11	847	181.95	181.0660	
12	857	150.67	183.5460	-32.8760



- Identify the following textbooks in the scatterplot and the table:
  - the textbook for which the line comes closest to predicting the list price
  - the textbook for which the prediction is furthest from the list price
  - all the textbooks for which the predicted list price is exceeds the actual list price by $20 or more

(e) The goal is to identify a line that is the best summary of the relationship between pages and price. The best-fit line gives the best predictions of list price, which means that overall it has the least amount of error in the predictions. Rank the four lines from best to worst with the best being the line that gives the best overall predictions of list price. Briefly explain the reasoning behind your rankings.

Supporting Lesson 3.2.2: Least Squares Regression Line as Line of Best Fit

Class Discussion Tables

	pages	price	Line_A_predictions	Line_A_error	Absolute_value_of_Line_A_error	Line_A_error_squared
1	560	118.95	147.6075	-28.6575	28.6575	
2	579	151.95	147.6075	4.3425	4.3425	
3	606	122.00	147.6075	-25.6075	25.6075	655.7441
4	607	107.95	147.6075	-39.6575		1572.7173
5	607	158.95	147.6075	11.3425		128.6523
6	696	122.00	147.6075	-25.6075	25.6075	655.7441
7	730	149.10	147.6075	1.4925	1.4925	2.2276
8	766	166.15	147.6075	18.5425	18.5425	343.8243
9	770	190.95	147.6075	43.3425	43.3425	1878.5723
10	788	150.67	147.6075	3.0625	3.0625	9.3789
11	847	181.95	147.6075	34.3425	34.3425	1179.4073
12	857	150.67	147.6075			

	pages	price	Line_C_predictions	Line_C_error	Absolute_Values_of_Line_C_error	Line_C_error_squared
1	560	118.95	126.3852	-7.4352	7.4352	55.2822
2	579	151.95	129.2428	22.7072	22.7072	
3	606	122.00	133.3036	-11.3036	11.3036	
4	607	107.95	133.4540	-25.5040	25.5040	650.4540
5	607	158.95	133.4540	25.4960	25.4960	650.0460
6	696	122.00	146.8396	-24.8396		617.0057
7	730	149.10	151.9532	-2.8532	2.8532	8.1408
8	766	166.15	157.3676	8.7824		77.1305
9	770	190.95	157.9692	32.9808	32.9808	1087.7332
10	788	150.67	160.6764	-10.0064	10.0064	100.1280
11	847	181.95	169.5500	12.4000	12.4000	153.7600
12	857	150.67	171.0540			

Which measures of the total error help you determine how well a line fits the data?			
Line	*Sum of Errors*	*Sum of Absolute Value of Errors (SAE)*	*Sum of Squares of Errors (SSE)*
A	0.0000	239.0600	7,275.7566
B	–48.9605	485.0950	24,774.1494
C	0.0404	204.6924	4,458.5762
D	32.7460	206.3540	5,734.1069

Supporting Lesson 3.2.2: Least Squares Regression Line as Line of Best Fit

Homework

(4) Here you have data collected from students at Los Medanos College in 2009. The variable *units* gives the number of college course units the student reported he or she was taking that semester. The variable *textbooks* gives the amount that the student reported spending on textbooks or other resources required for their courses that semester.

	units	textbooks
1	3	120.25
2	4	65.95
3	9	465.00
4	12	430.00
5	14	396.50
6	16	475.00
7	8	208.00
8	1	5.00
9	6	49.10
10	15	685.00
11	9	220.00
12	4	172.00
13	12	302.00
14	12	460.12
15	12	530.00

(a) Use technology to find the least squares regression line. (Think carefully about which variable is the explanatory variable.)

(b) Use the least squares regression line to predict the amount spent on textbooks for a student taking 12 units.

(c) Explain why the least squares regression line is considered the line of best fit.

(5) With the following applet, you can draw a line that you think fits the data well and compare your line to the least squares regression line.

www.rossmanchance.com/applets/Reg/index.html

Note: In the applet, errors are called *residuals*. This term comes from thinking about a data point as composed of two parts: the part explained by the regression line (the prediction) and the part that is leftover (called the *residual* or *error*).

(a) Instructions

1. Check *Your line* and click *Move line*. Follow directions to move the line so that it fits the data well.
2. Check *Show residuals* and record the SAE for your line in the table below.
3. Check *Show squared residuals* and record the SSE for your line in the table.
4. Check *Regression line*.
5. Check *Show residuals* and record the SAE for the regression line in the table.
6. Check *Show squared residuals* and record the SSE for the regression line in the table.

Line Predicting Height Based on Foot Length	Equation of Line	SAE	SSE
Your line			
Regression line			

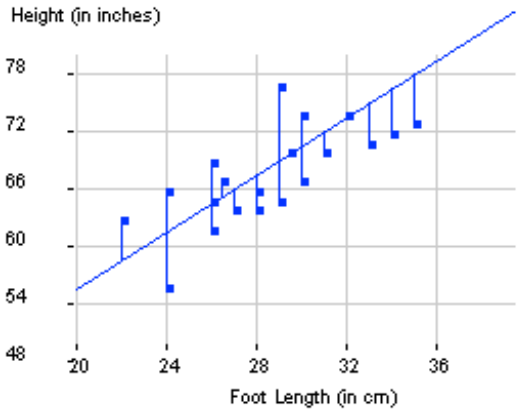
Supporting Lesson 3.2.2: Least Squares Regression Line as Line of Best Fit

(b) Compare the values of the SAE and SSE for your line with the regression line. What do you notice? Why does this make sense?

(c) When you click *Show residuals*, you see vertical line segments drawn from each data point to the regression line.

Some line segments are long and others are short. Why is this?

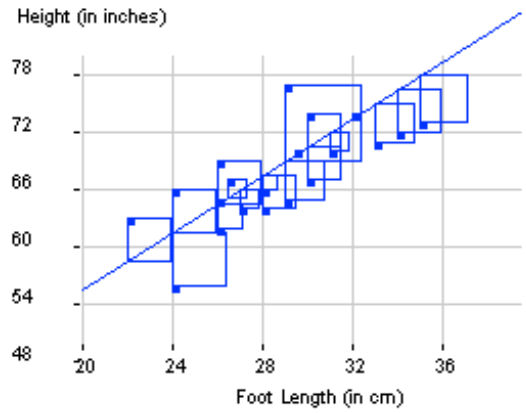
Why do you think these vertical line segments are shown?



(d) When you click *Show squared residuals*, you see squares appear.

Some squares are small and others are large. Why is this?

Why do you think these squares are shown?



Supporting Lesson 3.2.3: Investigating the Meaning of Numbers in the Equation of a Line

Estimated number of 50-minute class sessions: 1

Learning Goals

Students will understand that

- a line is determined by two parameters: its slope and *y*-intercept.
- the slope is described by a ratio of the change in *y* relative to the change in *x* and can be interpreted as the predicted change in *y* associated with a one-unit change in *x*.
- the *y*-intercept is the *y*-value when $x = 0$. Frequently, this point does not have meaning in the context of the data because $x = 0$ lies outside the range of the data. Nevertheless, the initial value is a part of the algebraic structure of the equation of the line.
- the slope of the least squares regression line may be affected by outliers, particularly outliers with extreme values for the predictor variable.

Students will be able to

- interpret (in context) the slope of the least squares line.
- if appropriate, interpret (in context) the y-intercept of the least squares line.
- given a scatterplot, identify observations that are outliers and observations that are potentially influential.

Developmental Math Connections

Instead of the usual approach in algebra of presenting the slope formula and then equations of lines in the form $y = mx + b$, you will discuss the equation of the line from the perspective of predicting *y*. The concepts of slope and *y*-intercept are developed as a way to predict *y* for a given *x*. In Module 12, you will revisit linear functions from the more traditional algebraic perspective.

Introduction

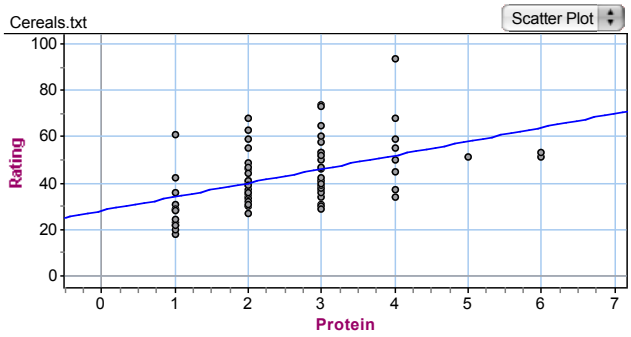
No extensive introduction is required for this lesson. Tell students that the lesson's general goal is to figure out what the numbers in the equation of a line tell them. The first task in the lesson reviews concepts related to the least squares line. Later, tasks focus on understanding and interpreting the numbers in the equation of the line.

Supporting Lesson 3.2.3: Investigating the Meaning of Numbers in the Equation of a Line

Tasks [Estimated time: 15 minutes]

(1) Here you return to the cereal data. Begin with a few questions to review what you know about the least squares line. Here is the least squares line that predicts *Consumer Reports* ratings based on the amount of protein (in grams) in a serving:

- predicted rating = 28 + 6(protein)
- The correlation coefficient is 0.48.



(a) Use the least squares regression line to predict the rating for a cereal containing 2 grams of protein in a serving.

(b) There are two cereals with 6 grams of protein in a serving. Is the predicted rating from the least squares regression line too high or too low for these cereals?

(c) What does the phrase *least squares* tell you about this line?

(**Answers will vary.** This is the line that minimizes the squared errors. By this criterion, it is the line that best fits the data. It is the best summary of the relationship between protein and ratings.)

(d) Based on these data, would you say that protein is

- a very accurate predictor of *Consumer Reports* ratings (errors within a few rating points would be typical)?
- not a very accurate predictor of ratings (errors as large as 10 rating points would not be surprising)?

(**Answer:** The second option is more reasonable due to the large variability in ratings for cereals with the same amounts of protein. Correlation is not strong, 0.48.)

(2) Now you will focus on understanding what information you get from the numbers in the equation of the line. Here we have the equation of the least squares line and a table of protein amounts and predicted ratings from the least squares line.

predicted rating = 28 + 6(protein)

$\hat{y} = 28 + 6x$

x = protein (g/serving)	$\hat{y}$ = predicted rating
0	28
1	34
2	40
3	46
4	52
5	58

How are the 28 and 6 related to the table of values? Be as specific as you can.

(**Note:** Some students may recall the idea of slope and *y*-intercept, but many will not. If students are using these terms and concepts, reinforce the connections that they are making as you work with them. However, many students will not give precise answers to this question. If students are giving cursory answers, such as "28 is in the table," prompt them to be more precise with questions such as the following:

Supporting Lesson 3.2.3: Investigating the Meaning of Numbers in the Equation of a Line

- Is the 28 a protein amount or a predicted rating?
- How much protein is in the cereal if the predicted value is 28?
- When the predicted rating increases by 6, how much more protein is in the cereal?
- If you remade the table so that the protein amount increased by 2s, would you still see the +6 pattern? What pattern would you see in the predicted ratings? By 3s?)

(3) Here you have the least squares equation for predicting *Consumer Reports* ratings based on the amount of sugar (in grams) in a serving. The equation was used to generate the table of predicted values for some sugar amounts.

predicted rating = 60 – 2.4(sugars)

$\hat{y} = 60 - 2.4x$

How are the 60 and –2.4 related to the table of values? Be as specific as you can.

(**Note:** See previous note on intervening as students work.)

***x* = sugar (g/serving)**	**$\hat{y}$ = predicted rating**
0	60
1	57.6
2	55.2
3	52.8
4	50.4
5	48

Wrap-Up/Direct Instruction (Estimated time: 15 minutes)

Discussion of Question 2

***x* = protein (g/serving)**	**$\hat{y}$ = predicted rating**
0	28
1	34
2	40
3	46
4	52
5	58

***x* = protein (g/serving)**	**$\hat{y}$ = predicted rating**
0	28
2	40
4	52
6	64
8	76
5	58

***x* = protein (g/serving)**	**$\hat{y}$ = predicted rating**
0	28
3	46
6	64
9	82
12	100
5	58

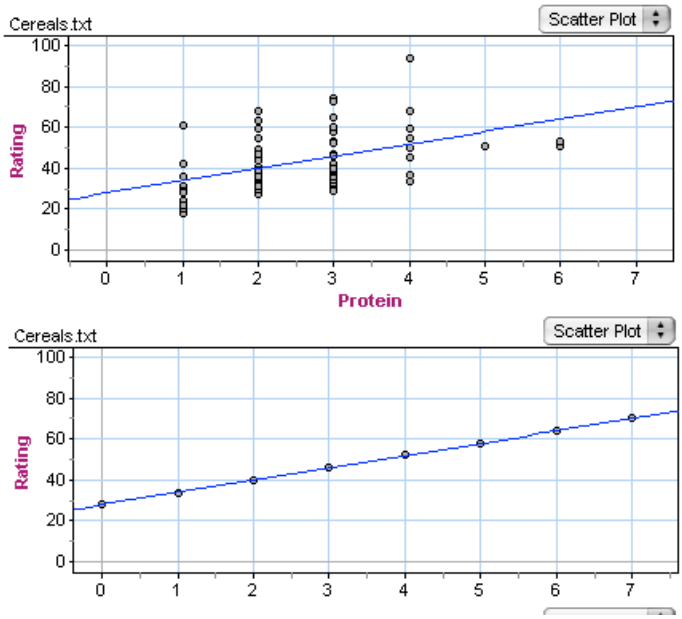
What does the 28 tell you? What does the 6 tell you?

(**Discuss the Values in the Table:** If you think of the line as predicting ratings, you start with an initial rating for a cereal with no protein, which is 28. Then you predict an increase in the rating of 6 for every extra gram of protein put into a serving. So the 6 is the increase in rating *per* gram of protein.

If you use the equation to create a table where you have changed the scale on *x*, the predicted increase in ratings per gram is still the same even though the pattern looks different: increase the rating by 12 for every 2 extra grams of protein, but this is equivalent to saying an increase of 6 in ratings for every gram of protein added to a serving, etc.)

Supporting Lesson 3.2.3: Investigating the Meaning of Numbers in the Equation of a Line

Let's think about what the role that 28 and 6 play when we look at the graph of the line. (**Note:** First make sure that students understand that the two lines shown are the same line. Describe what dots represent in the two graphs. Dots in the first graph represent data from the cereals. Dots in the second graph are values from the table. These dots are not cereals; they do not represent data. However, they show the pattern predicted by the least squares line.



Next show that 28 is the *y*-intercept, the *y*-value when $x = 0$. Note that these graphs are drawn with $x = 0$ slightly to the right of the vertical boundary of the graphing window. This is intentional to get students to focus on the fact that $x = 0$ and not the value that happens to mark the left boundary of the graphing window.

In the second graph illustrate the slope triangle for two different pairs of points in the table [with different Δx] to reiterate the point just made in the discussion of the table about a constant predicted increase of 6 in rating score for every additional gram of protein. Then show how any predicted *y*-value can be obtained by starting at the *y*-intercept of 28 and increasing the rating by 6 for each additional gram of protein added to a serving. For example, if $x = 3$ the slope triangle connecting [0, 28] with [3, 46] shows a change in *y* of 18. The predicted *y*-value is 28 plus the change in *y* of 18.)

Discussion of Question 3

The least squares line is $\hat{y} = 60 - 2.4x$. What do the numbers 60 and –2.4 tell you?

x = Sugar (g/serving)	$\hat{y}$ = predicted rating
0	60
1	57.6
2	55.2
3	52.8
4	50.4

x = Sugar (g/serving)	$\hat{y}$ = predicted rating
0	60
2	55.2
4	50.4
6	45.6
8	40.8

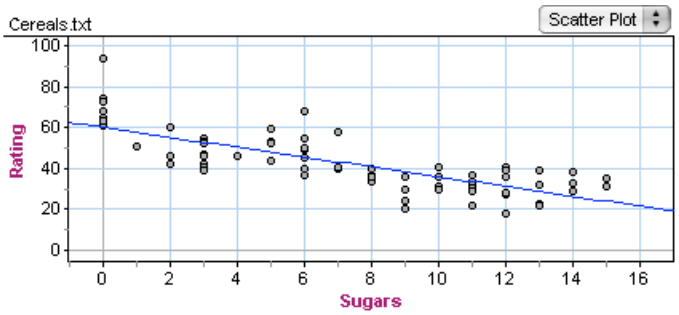
x = Sugar (g/serving)	$\hat{y}$ = predicted rating
0	60
5	48
10	36
15	24
20	12

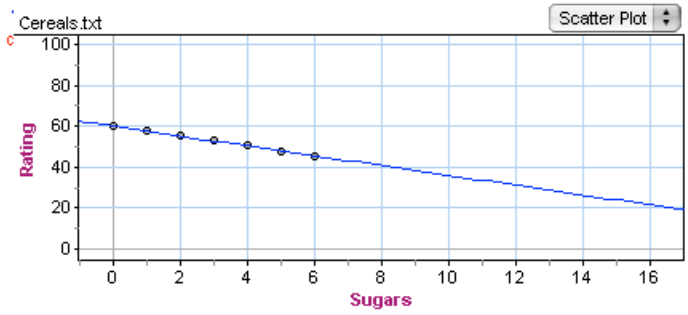
Supporting Lesson 3.2.3: Investigating the Meaning of Numbers in the Equation of a Line

Repeat this same discussion using the sugar-ratings tables and graphs.

If you think of the line as predicting ratings, start with an initial rating for a cereal with no sugar (60), and then decrease the predicted rating by 2.4 for every extra gram of sugar put into a serving. So, the –2.4 is the decrease in rating per gram of sugar.

If you use the equation to create a table where you have changed the scale on *x*, the decrease in ratings per gram is still the same: decrease the predicted rating by 4.8 for every 2 extra grams of sugar, which is equivalent to an increase of 2.4 in predicted ratings for every gram of sugar added to a serving.

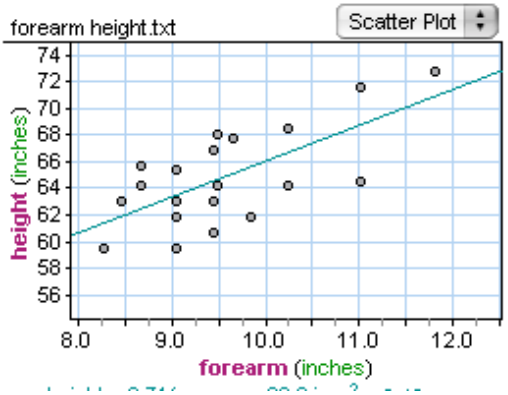


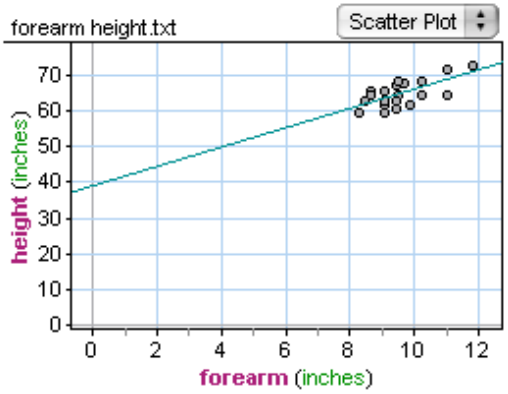


It is also worth noting that interpreting the slope in a graph is easier if you think carefully about how to draw the slope triangle. Choose points on the line that fall on or near the intersection of grid lines to make reading the changes in *x* and *y* easier.)

Discussion of a New Least Squares Line for a Different Scenario

Let's return to another data set you used recently, the forearm and height measurements for 21 female college students taking Introductory Statistics at Los Medanos College in Pittsburg, California, in 2009. The equation of the least squares line is $\hat{y} = 2.7x + 39$, where x = forearm length (inches) and $\hat{y}$ = predicted height (inches). What do the numbers 2.7 and 39 tell you?





(**Note:** Here the line is written with the x-term first, which may cause some initial confusion. This is intentional to get students to realize the terms in the equation can be commuted. Typically, statisticians write lines in the form $y = a + bx$.)

Supporting Lesson 3.2.3: Investigating the Meaning of Numbers in the Equation of a Line

(**Discuss What You Can Tell from the Symbolic Form of the Equation:** In this equation, you are predicting height based on forearm length. The equation predicts a height of 39 inches for a forearm length of 0. Obviously, it does not make sense that you can predict a person's height will be 39 inches when that person has a forearm length of 0. You can see from the graph when $x = 0$ you have extrapolated beyond the range of the data. However, the initial value is an important part of the least squares equation for predicting height, even when the initial value is nonsensical in the context of the data. So, what does the least squares equation tell us? To predict a person's height, start with a height of 39 inches and add 2.7 inches in height for each inch of forearm length.)

In the next set of tasks, you will continue to focus on understanding the numbers in the equation of a line.

(**Note:** The next tasks require that students identify equations by looking at graphs. Some students may work from an understanding that slope is a measure of steepness. This may lead to some confusion since different scales are used on the axes. So, lines that have a smaller slope relative to other lines may visually appear steeper due to the scaling. Encourage all students to apply the idea of a slope triangle. If students are struggling with determining the slope, help them make smart decisions about how to construct slope triangles by looking for places where the lines cross grid marks. This makes determining the slope easier. Treat these tasks as guided practice and do not hesitate to guide. If necessary, call their attention to the scales on the axes.)

Tasks

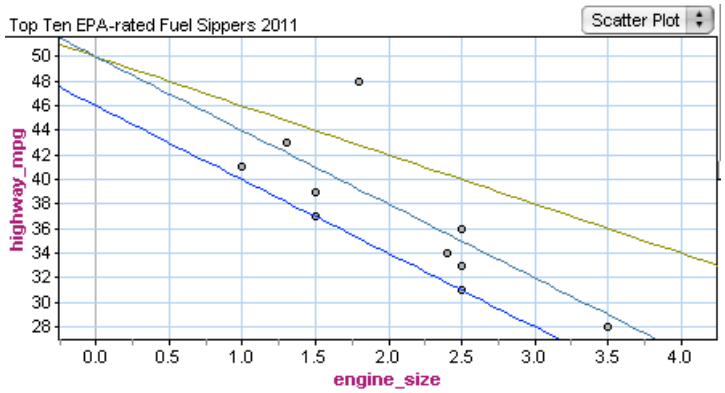
(4) The Environmental Protection Agency picks the 10 most fuel-efficient cars each year. Below is a scatterplot of the highway miles per gallon and the engine size (measured in liters) for the EPA's top 10 for 2011. (Retrieved from www.fueleconomy.gov)

Following are the equations of the three lines shown:

$\hat{y} = 46 - 6x$

$\hat{y} = 50 - 6x$

$\hat{y} = 50 - 4x$



(a) Identify the *equation* of the line that best fits the data. Briefly explain how you made your decision.

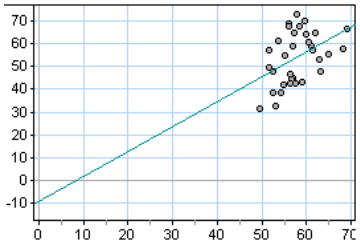
(**Answer:** The middle line does the best job of approximating the data. This line has an initial value of 50 when $x = 0$. A slope triangle through gridlines connects [0, 50] and [1, 44]; when x increases by 1, y decreases by 6. So, the equation is $\hat{y} = 50 - 6x$.)

Supporting Lesson 3.2.3: Investigating the Meaning of Numbers in the Equation of a Line

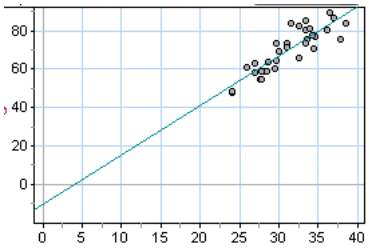
(b) For the line you chose, describe what the numbers in the equation tells you about engine size, highway miles per gallon, and the relationship between the two for these 10 cars.

(**Answer:** The predicted highway mileage decreases by 6 for an increase of 1 liter in engine size. The 50 is the *y*-intercept; it is the highway mileage given by the line for an engine of size 0 liters. This is an example of an initial value that is nonsensical in the context of the data.)

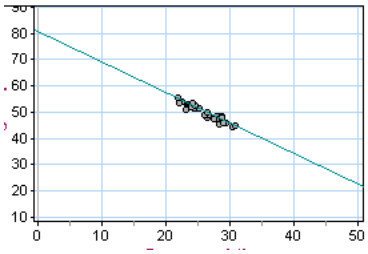
(5) Match the graphs to the least squares equations and *r*-values.



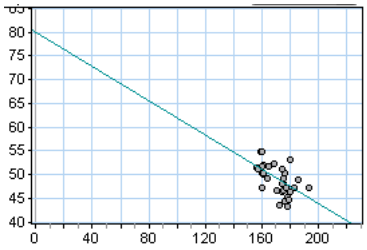
Graph A



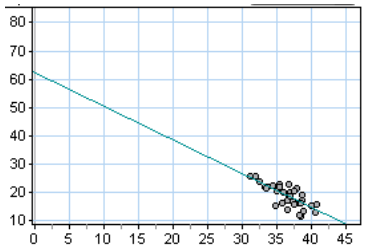
Graph B



Graph C



Graph D



Graph E

Here are *r*-values to choose from:

–0.54 –0.73 0.45 –0.95 0.88

Here are equations to choose from:

$\hat{y} = -10.5 + 2x$ $\hat{y} = 62 + (-1.2)x$ $\hat{y} = -10.5 + 1.1x$ $\hat{y} = 80 + (-1.2)x$ $\hat{y} = 80 + (-0.2)x$

[**Answers:** Graph A ($r = 0.45$, $\hat{y} = 1.1x - 10.5$); Graph B ($r = 0.88$, $\hat{y} = 2.6x - 10.5$); Graph C ($r = -0.95$, $\hat{y} = -1.2x + 80$); Graph D ($r = -0.54$, $\hat{y} = -0.2x + 80$); Graph E ($r = -0.73$, $\hat{y} = -1.2x + 62$)]

Wrap-Up

Since these two tasks require students to apply concepts previously discussed, just go over the answers and field questions that arise. Return as necessary to the idea of *y*-intercept (when $x = 0$) and slope as the amount *y*-hat changes for each one-unit change in *x*.

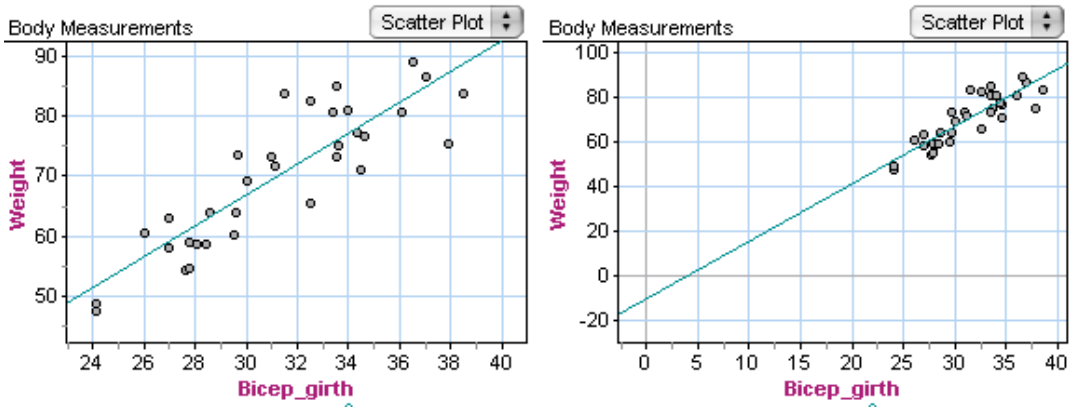
Supporting Lesson 3.2.3: Investigating the Meaning of Numbers in the Equation of a Line

Homework

(6) Based on data from 34 adults who exercise regularly, the least squares line for the relationship between bicep girth and weight is

predicted weight = 2.6(bicep girth) – 10.5

where predicted weight is measured in kilograms and bicep girth is measured in centimeters.



(a) Construct a table or use one of the graphs to explain the meaning of 2.6 in this situation.

(b) The –10.5 is the initial value when the $x = 0$. Does this number have meaning in this scenario? Why or why not?

(c) If you use bicep girth to predict weight, how accurate do you think the predictions will be?

- very accurate (typical prediction error will be within a kilogram)
- somewhat accurate (typical prediction error will be within 10 kilograms)
- not very accurate (prediction errors larger than plus or minus 20 kilograms would not be surprising)

(**Answers:** [6a] The least squares line predicts an increase of 2.6 kilograms of body weight for each additional centimeter in bicep girth. Even if the wording students use is awkward, their answers should demonstrate, in the context of the variables given, an understanding that the slope is a change in predicted *y* relative to a change in *x*. [6b] The initial value here is nonsensical since it does not make sense that the distance around someone's bicep is 0 centimeters. Obviously, you cannot have a weight of –10.5 kilograms. [6c] somewhat accurate)

Supporting Lesson 3.2.3: Investigating the Meaning of Numbers in the Equation of a Line

(7) With the following applet, you will investigate how outliers impact the regression line.

http://www.stat.sc.edu/~west/javahtml/Regression.html

(a) Add points to the scatterplot that are close data points shown. Describe what happens to the regression line.

(**Answer:** The regression line does not change much.)

(b) A data point is *influential* if removing it (or in this case adding it) substantially changes the regression line. Add points to the scatterplot that are outliers relative to the other data points. In other words, add points that are far away from the other data points. Describe what happens to the regression line.

(**Answer:** The regression line changes quite a bit.)

(**Note:** If you have time and think it is important, the National Council of Teachers of Mathematics website has a nice lesson on investigating the impact of influential points on the slope using an applet: http://illuminations.nctm.org/LessonDetail.aspx?ID=L455.)

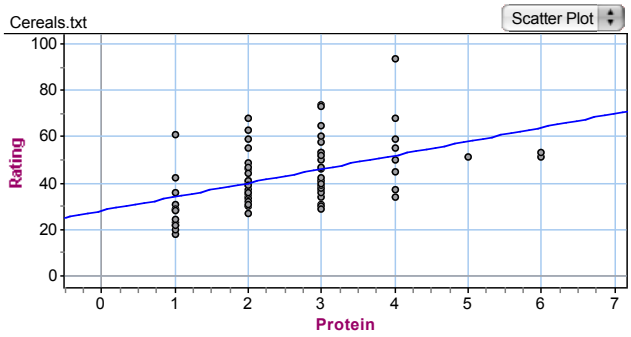
Supporting Lesson 3.2.3: Investigating the Meaning of Numbers in the Equation of a Line

Tasks

(1) Here you return to the cereal data. Begin with a few questions to review what you know about the least squares line. Here is the least squares line that predicts *Consumer Reports* ratings based on the amount of protein (in grams) in a serving:

- predicted rating = 28 + 6(protein)
- The correlation coefficient is 0.48.

(a) Use the least squares regression line to predict the rating for a cereal containing 2 grams of protein in a serving.



(b) There are two cereals with 6 grams of protein in a serving. Is the predicted rating from the least squares regression line too high or too low for these cereals?

(c) What does the phrase *least squares* tell you about this line?

(d) Based on these data, would you say that protein is

- a very accurate predictor of *Consumer Reports* ratings (errors within a few rating points would be typical)?
- not a very accurate predictor of ratings (errors as large as 10 rating points would not be surprising)?

(2) Now you will focus on understanding what information you get from the numbers in the equation of the line. Here we have the equation of the least squares line and a table of protein amounts and predicted ratings from the least squares line.

predicted rating = 28 + 6(protein)

$\hat{y} = 28 + 6x$

How are the 28 and 6 related to the table of values? Be as specific as you can.

x = protein (g/serving)	$\hat{y}$ = predicted rating
0	28
1	34
2	40
3	46
4	52
5	58

Supporting Lesson 3.2.3: Investigating the Meaning of Numbers in the Equation of a Line

(3) Here you have the least squares equation for predicting *Consumer Reports* ratings based on the amount of sugar (in grams) in a serving. The equation was used to generate the table of predicted values for some sugar amounts.

predicted rating = 60 – 2.4(sugars)

$\hat{y} = 60 - 2.4x$

How are the 60 and –2.4 related to the table of values? Be as specific as you can.

***x* = sugar (g/serving)**	**$\hat{y}$ = predicted rating**
0	60
1	57.6
2	55.2
3	52.8
4	50.4
5	48

Supporting Lesson 3.2.3: Investigating the Meaning of Numbers in the Equation of a Line

Class Discussion Outline

Discussion of Question 2

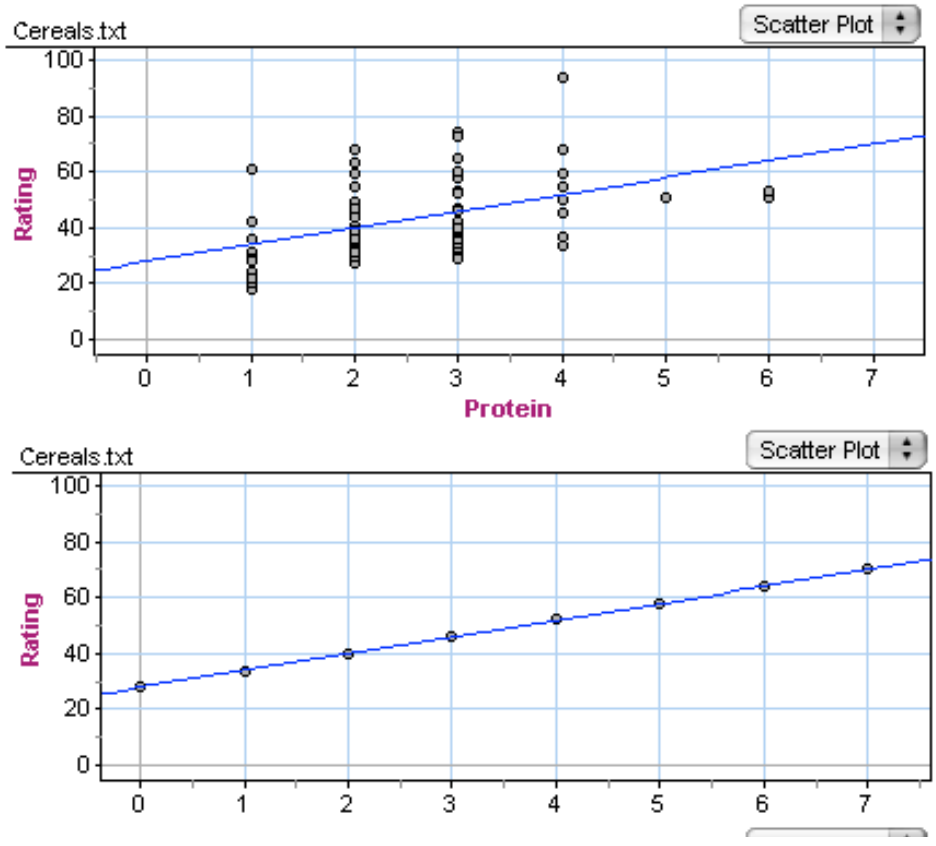
x = protein (g/serving)	$\hat{y}$ = predicted rating
0	28
1	34
2	40
3	46
4	52
5	58

x = protein (g/serving)	$\hat{y}$ = predicted rating
0	28
2	40
4	52
6	64
8	76
5	58

x = protein (g/serving)	$\hat{y}$ = predicted rating
0	28
3	46
6	64
9	82
12	100
5	58

What does the 28 tell you? What does the 6 tell you?

Let's think about what the role that 28 and 6 play when we look at the graph of the line.



Supporting Lesson 3.2.3: Investigating the Meaning of Numbers in the Equation of a Line

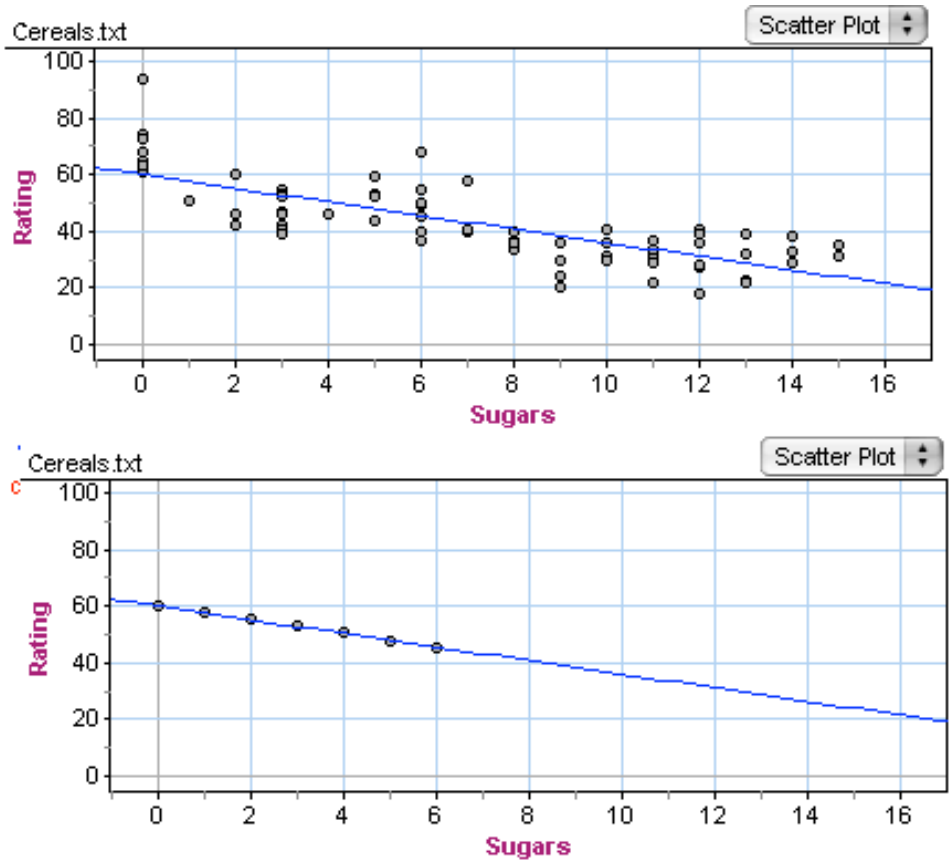
Discussion of Question 3

The least squares line is $\hat{y} = 60 - 2.4x$. What do the numbers 60 and –2.4 tell you?

x = Sugar (g/serving)	$\hat{y}$ = predicted rating
0	60
1	57.6
2	55.2
3	52.8
4	50.4

x = Sugar (g/serving)	$\hat{y}$ = predicted rating
0	60
2	55.2
4	50.4
6	45.6
8	40.8

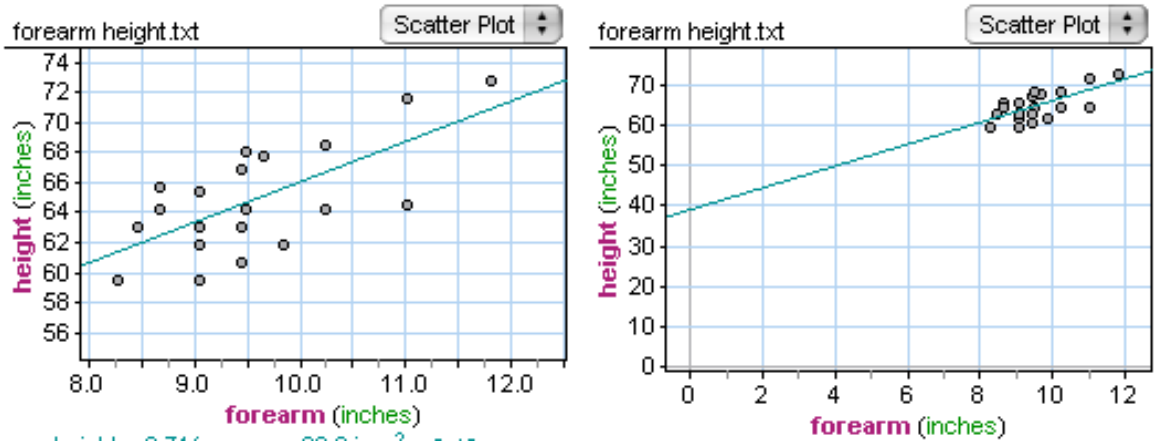
x = Sugar (g/serving)	$\hat{y}$ = predicted rating
0	60
5	48
10	36
15	24
20	12



Supporting Lesson 3.2.3: Investigating the Meaning of Numbers in the Equation of a Line

Discussion of a New Least Squares Line for a Different Scenario

Let's return to another data set you used recently, the forearm and height measurements for 21 female college students taking Introductory Statistics at Los Medanos College in Pittsburg, California, in 2009. The equation of the least squares line is $\hat{y} = 2.7x + 39$, where x = forearm length (inches) and $\hat{y}$ = predicted height (inches). What do the numbers 2.7 and 39 tell you?

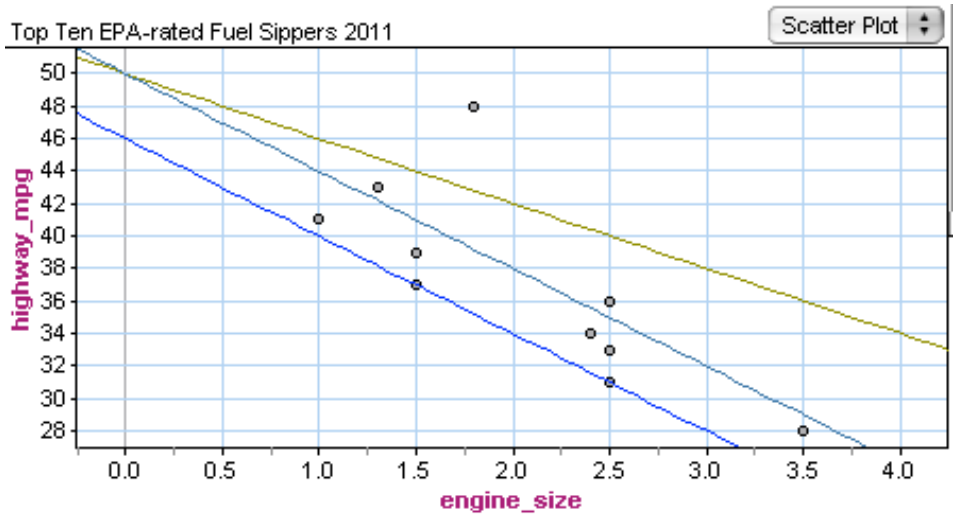


In the next set of tasks, you will continue to focus on understanding the numbers in the equation of a line.

Supporting Lesson 3.2.3: Investigating the Meaning of Numbers in the Equation of a Line

Tasks

(4) The Environmental Protection Agency picks the 10 most fuel-efficient cars each year. Below is a scatterplot of the highway miles per gallon and the engine size (measured in liters) for the EPA's top 10 for 2011. (Retrieved from www.fueleconomy.gov)



Following are the equations of the three lines shown:

$\hat{y} = 46 - 6x$

$\hat{y} = 50 - 6x$

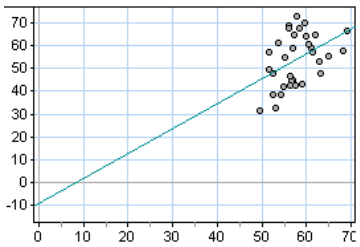
$\hat{y} = 50 - 4x$

(a) Identify the *equation* of the line that best fits the data. Briefly explain how you made your decision.

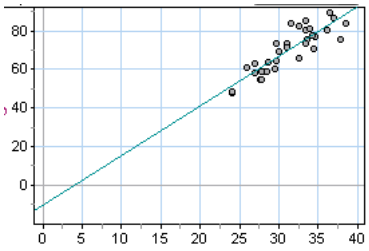
(b) For the line you chose, describe what the numbers in the equation tells you about engine size, highway miles per gallon, and the relationship between the two for these 10 cars.

Supporting Lesson 3.2.3: Investigating the Meaning of Numbers in the Equation of a Line

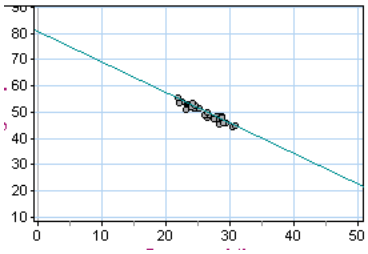
(5) Match the graphs to the least squares equations and *r*-values.



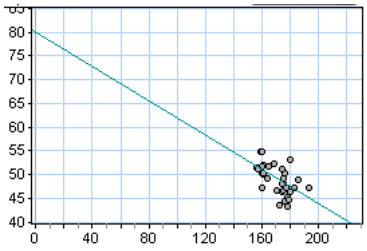
Graph A



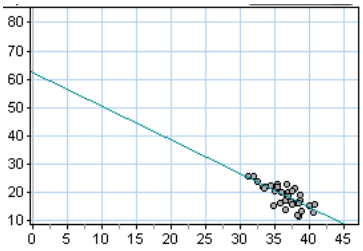
Graph B



Graph C



Graph D



Graph E

Here are *r*-values to choose from:

–0.54 –0.73 0.45 –0.95 0.88

Here are equations to choose from:

$\hat{y} = -10.5 + 2x$ $\hat{y} = 62 + (-1.2)x$ $\hat{y} = -10.5 + 1.1x$ $\hat{y} = 80 + (-1.2)x$ $\hat{y} = 80 + (-0.2)x$

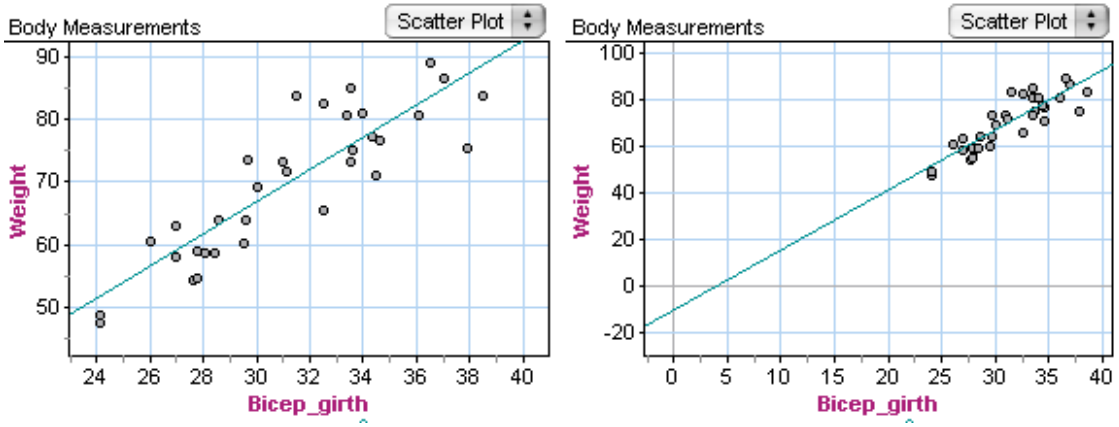
Supporting Lesson 3.2.3: Investigating the Meaning of Numbers in the Equation of a Line

Homework

(6) Based on data from 34 adults who exercise regularly, the least squares line for the relationship between bicep girth and weight is

predicted weight = 2.6(bicep girth) – 10.5

where predicted weight is measured in kilograms and bicep girth is measured in centimeters.



(a) Construct a table or use one of the graphs to explain the meaning of 2.6 in this situation.

(b) The –10.5 is the initial value when the $x = 0$. Does this number have meaning in this scenario? Why or why not?

(c) If you use bicep girth to predict weight, how accurate do you think the predictions will be?

- very accurate (typical prediction error will be within a kilogram)
- somewhat accurate (typical prediction error will be within 10 kilograms)
- not very accurate (prediction errors larger than plus or minus 20 kilograms would not be surprising)

(7) With the following applet, you will investigate how outliers impact the regression line.

http://www.stat.sc.edu/~west/javahtml/Regression.html

(a) Add points to the scatterplot that are close data points shown. Describe what happens to the regression line.

(b) A data point is *influential* if removing it (or in this case adding it) substantially changes the regression line. Add points to the scatterplot that are outliers relative to the other data points. In other words, add points that are far away from the other data points. Describe what happens to the regression line.

Supporting Lesson 3.2.4: Special Properties of the Least Squares Regression Line

Estimated number of 50-minute class sessions: 1

Learning Goals

Students will understand that

- the least squares regression lines contains the point $(\bar{x}, \bar{y})$.
- The slope of the least squares regression line is related to the correlation in that a change of one standard deviation in *x* produces a fractional change of *r* standard deviations in *y*.
- the equation of the least squares line can be generated from summary statistics for *x* and *y*. The slope of the least squares line is $\frac{r \bullet s_y}{s_x}$ and its *y*-intercept is $\bar{y} - slope \bullet \bar{x}$.

Students will be able to

- use the fact that the slope of the least squares line is $\frac{r \bullet s_y}{s_x}$ and the *y*-intercept is $\bar{y} - slope \bullet \bar{x}$ to generate the equation of the least squares line from summary statistics.

Developmental Math Connections

This lesson requires students to understand that if a point is on a line, its *x*- and *y*-coordinates satisfy the relationship defined by the line. This idea is used implicitly in the investigations laid out in this lesson. If students are having trouble following the investigation, this may be at the heart of their difficulty. Therefore, be ready to discuss this issue if you diagnose that students are not understanding this.

Introduction

(**Note:** This lesson is designed as an interactive lecture. Your goal is to develop the following special properties of the least squares line:

- the least squares line contains the point $(\bar{x}, \bar{y})$. This means that when $x = \bar{x}$, the least squares line predicts $\hat{y} = \bar{y}$.
- When *x* increases from the mean of *x* by one standard deviation in *x*, the least squares line predicts that *y* increases by *r* standard deviations in *y*. This means that the least squares line has a slope of $\frac{r \bullet s_y}{s_x}$.
- The initial value of the least squares line is $\bar{y} - \frac{rs_y}{s_x}\bar{x}$

This interactive lecture investigates three questions. The investigation of each question is broken down into smaller questions that you can pose to students. Remember to build in think time. Students will be filling in a discussion outline as you lecture, so you also need to build in time for students to take notes. Consider allowing time periodically for students to compare notes with a classmate, summarize, formulate questions, and so on.)

Supporting Lesson 3.2.4: Special Properties of the Least Squares Regression Line

In this lesson, you are interested in understanding special properties of the least squares line. You will investigate and answer the following three questions about the least squares line in this lesson:

(1) What does the least squares line predict for *y* when *x* is equal to the mean of the *x* values in the sample? Here is the same question written with symbols: When $x = \bar{x}$, $\hat{y} = ?$

(2) What does the least squares line predict for *y* when *x* is one standard deviation above the mean. Here is the same question written with symbols: When $x = \bar{x} + s_x$, $\hat{y} = ?$

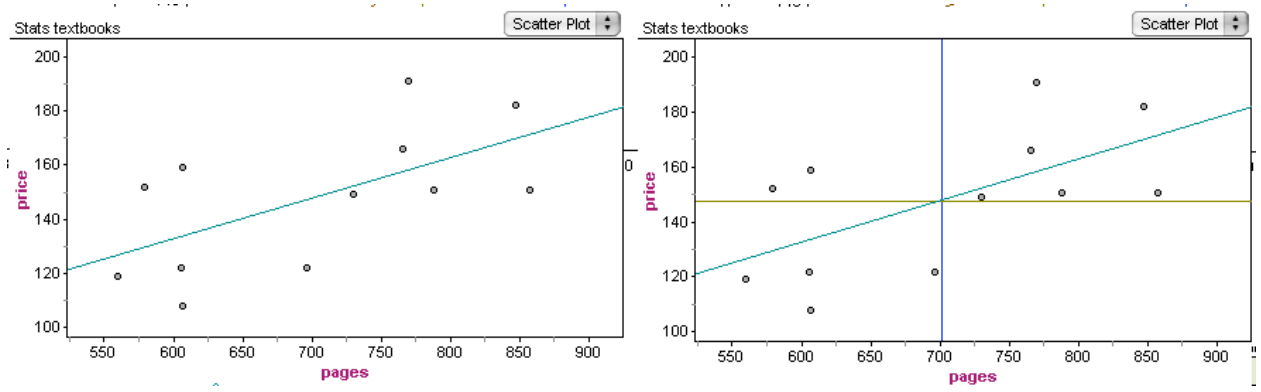
(3) What is the connection between the equation of the least squares line and the summary statistics for *x* and *y*, such as means, standard deviations, and the correlation coefficient?

You will investigate these three questions using data on 12 statistics textbooks.

Here you have the summary statistics describing the number of pages and the publisher's list price for 12 popular statistics textbooks.

	Mean	Standard Deviation
Pages	$\bar{x} = 701.083$	$s_x = 106.402$
Price	$\bar{y} = 147.608$	$s_y = 25.718$

Below on the left, there is a scatterplot of the data and the least squares line $\hat{y} = 42.16 + 0.1504(\text{pages})$. The correlation coefficient is 0.62. On the right, there is the same graph with the addition of the two mean lines.



Investigation and Discussion of Question 1

What does the least squares line predict for y when x is the mean? Here is the same question written with symbols: When $x = \bar{x}$, $\hat{y} = ?$

Can you figure this out from the given information before doing any calculations?

(**Note:** Give students a minute to think about this before calling a student to answer the question or answering the question yourself. **Answer:** The least squares line goes through $(\bar{x}, \bar{y})$. This means that when $x = \bar{x}$, $\hat{y} = \bar{y}$. Tell students that this is always true for least squares lines.)

If you substitute $x = 701.083$ into the given equation of the least squares line, what is the predicted value for the price? Can you figure this before you do any calculations?

[**Answer:** This is really a check for understanding. You should get the mean of *y*, 147.608. You get $\hat{y} = 42.16 + 0.1504(701.083) = 147.603$ when you do the calculation; there is a bit of error due to rounding.]

Investigation and Discussion of Question 2

What does the least squares line predict for y when x is one standard deviation above the mean? Here is the same question written with symbols: When $x = \bar{x} + s_x$, $\hat{y} = ?$

(**Note:** This is a sophisticated question, though students might volunteer the conjecture that $\hat{y} = \bar{y} + s_y$. If they do not, suggest this as a reasonable conjecture to investigate. Then pose the question, "How could you test your conjecture?" Give students a few minutes to think about this. Then proceed with the guided investigation. Alternatively, you could call on students for ideas and build the investigation from student responses.)

To answer this question, let's start by determining whether *y* is one standard deviation above the mean for *y* when *x* is one standard deviation above the mean for *x*. Here is this idea written in symbols: When $x = \bar{x} + s_x$, does $\hat{y} = \bar{y} + s_y$?

The expression $x = \bar{x} + s_x$ tells you that you are imagining looking at a textbook where the number of pages is one standard deviation above the mean number of pages for the books in the sample. How many pages is this?

(**Answer:** $\bar{x} + s_x = 701.083 + 106.402 = 807.485$, approximately 807)

What is the predicted price of a textbook with this many pages? (**Answer:** $\hat{y} = 42.16 + 0.1504(807) \approx 163.5$)

Let's see if this prediction is one standard deviation in price above the mean price for the books in the sample. (**Answer:** $\bar{y} + s_y = 147.608 + 25.718 = 173.326$, approximately 173)

So is your conjecture right? (**Answer:** No, the predicted price is about $10 cheaper than one standard deviation above the mean price.)

Let's summarize what is known so far. When the number of pages is one standard deviation above the mean number of pages, the predicted price increases (*choose one:* less than, more than) one standard deviation above the mean price.

Let's continue to investigate Question 2 and see if you can be more specific. Remember that you are trying to determine the predicted value when $x = \bar{x} + s_x$. You will try to answer the following questions to be more specific:

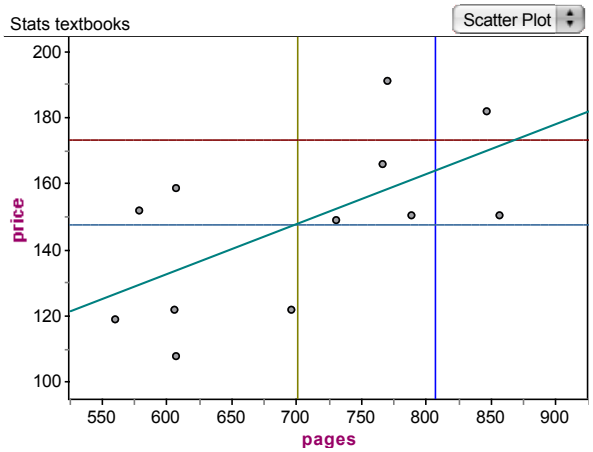
- What fraction of a standard deviation in price does the price increase? Half of a standard deviation? A third of a standard deviation? Some other fractional part of a standard deviation?
- Can you give an answer that is true for all least squares lines?

Supporting Lesson 3.2.4: Special Properties of the Least Squares Regression Line

Let's think about these questions using a graph before you do any more calculations. In the graph shown here, you have added lines marking one standard deviation above each mean. Let's think graphically about what you have already established in your investigation.

Label the following on the graph:

- the least squares line,
- the point $(\bar{x}, \bar{y})$,
- the point $(\bar{x} + s_x, \bar{y} + s_y)$, and
- the point $(\bar{x} + s_x, \hat{y})$.



Hint: The calculations you did previously can help you label these points.

(**Note:** Emphasize how the graph shows that $\hat{y}$ for $\bar{x} + s_x$ is less than $\bar{y} + s_y$. Show this on the slope triangle connecting $(\bar{x}, \bar{y})$ and $(\bar{x} + s_x, \hat{y})$. Show how the change in predicted price, represented by a vertical distance on the slope triangle, is less than a standard deviation in price, represented by a vertical distance longer than the side of the slope triangle. Refer to the contextualized meaning of these symbols. For example, you know that when pages increase by one standard deviation from the mean number of pages, the predicted price increases less than a standard deviation from the mean price.)

Is the increase in predicted price (*choose one:* more than or less than) *half* a standard deviation in price, $0.5\, s_y$? How can you tell?

[**Answer:** You can see that the predicted price is a little more than half a price standard deviation. Show students that the vertical distance of the slope triangle, is a little more than half the vertical distance between $\bar{y}$ and $\bar{y} + s_y$. You can also calculate the increase in predicted price, $\hat{y} - \bar{y} = 163.5 - 147.6 = 15.9$, and compare it to $0.5\, s_y$, $0.5(25.7) = 12.85$.]

Which of the expressions below best describes the increase in the predicted price? How can you tell?

- about $0.3\, s_y$
- about $0.6\, s_y$
- about $0.9\, s_y$

(**Answer:** About $0.6\, s_y$. Explain this using the same geometric estimate as outlined just above. You can also calculate the options and compare to 15.9.)

Now let's actually calculate what fraction of a standard deviation the price increases. The standard deviation in price is _____ and the increase in the predicted price from the mean line is $\hat{y} - \bar{y}$ =_____. So what fraction of a price standard deviation did $\hat{y}$ increase from $\bar{y}$?

(**Answers:** The standard deviation in price is 25.7; increase from mean line is 15.9; 15.9/25.7 = 0.62.)

Does this number have significance when you look back at the descriptive statistics for the data?

(**Answer:** This is the value of the correlation coefficient, *r. Wow.* This will always be true for the least squares line.)

[**Note:** Now go back and highlight the general measurements for the slope triangle connecting $(\bar{x}, \bar{y})$ and $(\bar{x} + s_x, \hat{y})$: horizontal measurement is s_x, the vertical measurement is $r s_y$.]

So now you have an answer to the second question: *What does the least squares line predict for y when x is one standard deviation above the mean? When* $x = \bar{x} + s_x$, $\hat{y} =$ _____.

(**Answer:** $\hat{y} = \bar{y} + rs_y$)

Investigation and Discussion of Question 3

What is the connection between the equation of the least squares line and the summary statistics for x and y, such as means, standard deviations, and the correlation coefficient?

You know from your previous work that the equation of a line has the form *y = initial value + slope • x*. From the slope triangle you have drawn, you can derive an important characteristic about the slope of the least squares line. The slope is the change in *y* divided by the change in *x*.

How can you determine the slope of the least squares line using summary statistics?

[**Answer:** Use the idea that the slope is the ratio of the change in *y* to the corresponding change in *x*, so the slope of the least squares line is $\frac{rs_y}{s_x}$. Alternatively, if your students are familiar with the formula for slope, you could find the slope using the points that they now know are on the line $(\bar{x}, \bar{y})$ and $(\bar{x} + s_x, \bar{y} + rs_y)$. However, this may easier to communicate using a slope triangle than manipulating the algebraic expressions.]

You also know that the least squares line contains the point $(\bar{x}, \bar{y})$. So, you have enough information to write an expression for the y-intercept of the least squares line in terms of summary statistics.

(**Note:** Demonstrate how the initial value can be written in terms of summary statistics. This line of reasoning involves a bit of algebra, so some students may have difficulty following the demonstration. Ultimately, you just want students to understand that the equation of the least squares line can be derived from summary statistics even if they cannot reproduce the development of the formula for initial value.)

Demonstration of the Development of the Formula for Initial Value

For any line *y* = *y*-intercept + slope • *x*

The least squares line contains $(\bar{x}, \bar{y})$, so for the least squares line you can write $\bar{y}$ = *y*-intercept + slope • $\bar{x}$.

You can rewrite this equation as a formula for initial value:

$$y\text{-intercept} = \bar{y}\text{-slope} \bullet \bar{x}$$

Rewriting the slope formula for the least squares line, you have the initial value expressed in terms of summary statistics:

$$y\text{-intercept} = \bar{y} - \left(\frac{r \bullet s_y}{s_x} \right) \bullet \bar{x})$$

Homework

(4) At Los Medanos College, a statistics instructor posted the following information on her office door at the end of the semester:

Statistics FA 2010	Mean	Standard Deviation	Correlation
Prefinal exam average	75	8	0.7
Final exam score	78	12	

(a) Final course grades have not been posted. So Karen wants to predict her final exam score based on this information. She has an 82 prefinal exam average. What does the least squares line predict for Karen's final exam score?

(**Answer:** slope of regression line is $\frac{rs_y}{s_x} = 1.05$. y-intercept of the regression line is

$y\text{-intercept} = \bar{y} - \left(\frac{r \bullet s_y}{s_x} \right) \bullet \bar{x} = -0.75$. Karen's predicted final exam score is 85.35.)

(b) What statement is the most accurate advice you could give Karen?

- Before using the least squares line to predict your final exam score, you need to know that the relationship between prefinal exam average and final exam scores is linear. You are not given enough information to determine if the relationship is linear. Therefore, proceed with caution in using a line to make predictions in this situation.
- The correlation coefficient is 0.7. This tells you that the relationship between prefinal exam average and final exam score is fairly linear. However, there will be some error in the prediction, so do not be surprised if your final exam score differs from your prediction.

(**Answer:** The first statement is the best advice. Always look at the data to make sure a linear model is appropriate. Lesson 3.1.3 demonstrates that r-values alone do not guarantee that the data are linear.)

Supporting Lesson 3.2.4: Special Properties of the Least Squares Regression Line

(5) In this lesson, you learned that for any least squares regression line if $x = \bar{x}$, then $\hat{y} = \bar{y}$. Explain in words what this means.

(**Answer:** Start with the mean of the data's *x*-values. Plug this into the regression equation. The result is the mean of the data's *y*-values. … Of course, students may have a less polished way of communicating this idea ☺.)

(6) You conjectured that for the least squares line if $x = \bar{x} + s_x$, then $\hat{y} = \bar{y} + s_y$.

(a) Explain in words what this means.

(**Answer:** This conjecture says that if *x* increases one standard deviation above its mean, then the least squares regression line predicts that the same is true for *y*. In other words, *y* is predicted to increase by one standard deviation from its mean.)

(b) This conjecture was not true for the statistics textbook data. Would this conjecture ever be true? If so, describe the relationship you would see in the data. If not, explain why this will never be true for any least squares line.

(**Answer:** The conjecture is only true when $r = 1$. The line has to predict an increase of one standard deviation in *y* for one standard deviation in *x*. Students saw in class that when *x* increases one standard deviation, *y* increases a fraction of a standard deviation. The change in *y* is rs_y. So *r* has to be 1 if *y* increases a full standard deviation.)

Supporting Lesson 3.2.4: Special Properties of the Least Squares Regression Line

Class Discussion

In this lesson, you are interested in understanding special properties of the least squares line. You will investigate and answer the following three questions about the least squares line in this lesson:

(1) What does the least squares line predict for *y* when *x* is equal to the mean of the *x* values in the sample? Here is the same question written with symbols: When $x = \bar{x}$, $\hat{y} = ?$

(2) What does the least squares line predict for *y* when *x* is one standard deviation above the mean. Here is the same question written with symbols: When $x = \bar{x} + s_x$, $\hat{y} = ?$

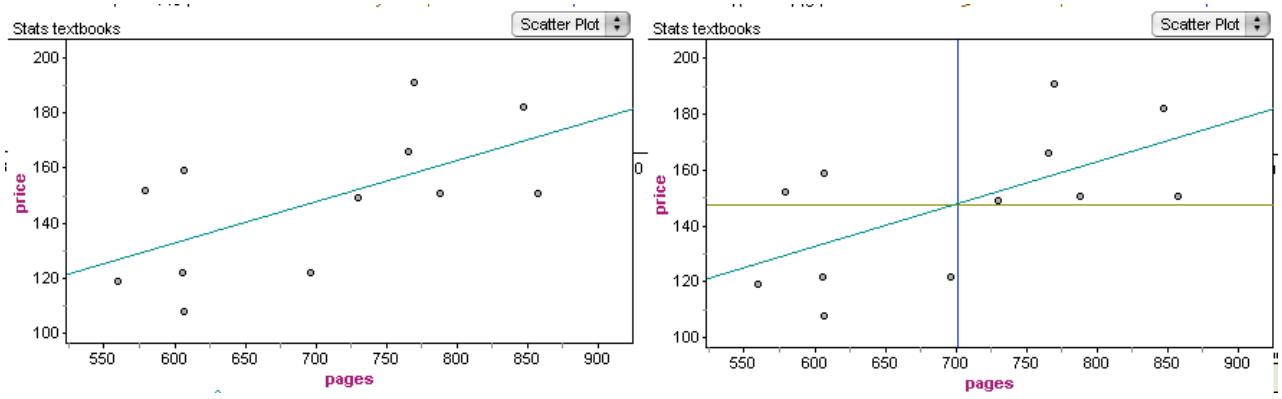
(3) What is the connection between the equation of the least squares line and the summary statistics for *x* and *y*, such as means, standard deviations, and the correlation coefficient?

You will investigate these three questions using data on 12 statistics textbooks.

Here you have the summary statistics describing the number of pages and the publisher's list price for 12 popular statistics textbooks.

	Mean	Standard Deviation
Pages	$\bar{x} = 701.083$	$s_x = 106.402$
Price	$\bar{y} = 147.608$	$s_y = 25.718$

Below on the left, there is a scatterplot of the data and the least squares line $\hat{y} = 42.16 + 0.1504(\text{pages})$. The correlation coefficient is 0.62. On the right, there is the same graph with the addition of the two mean lines.



Investigation and Discussion of Question 1

What does the least squares line predict for y when x is the mean? Here is the same question written with symbols: When $x = \bar{x}$, $\hat{y} = ?$

Can you figure this out from the given information before doing any calculations?

If you substitute $x = 701.083$ into the given equation of the least squares line, what is the predicted value for the price? Can you figure this before you do any calculations?

Supporting Lesson 3.2.4: Special Properties of the Least Squares Regression Line

Investigation and Discussion of Question 2

What does the least squares line predict for y when x is one standard deviation above the mean? Here is the same question written with symbols: When $x = \bar{x} + s_x$, $\hat{y} = ?$

To answer this question, let's start by determining whether *y* is one standard deviation above the mean for *y* when *x* is one standard deviation above the mean for *x*. Here is this idea written in symbols: When $x = \bar{x} + s_x$, does $\hat{y} = \bar{y} + s_y$?

The expression $x = \bar{x} + s_x$ tells you that you are imagining looking at a textbook where the number of pages is one standard deviation above the mean number of pages for the books in the sample. How many pages is this?

What is the predicted price of a textbook with this many pages? Let's see if this prediction is one standard deviation in price above the mean price for the books in the sample. So is your conjecture right?

Let's summarize what is known so far. When the number of pages is one standard deviation above the mean number of pages, the predicted price increases (*choose one:* less than, more than) one standard deviation above the mean price.

Let's continue to investigate Question 2 and see if you can be more specific. Remember that you are trying to determine the predicted value when $x = \bar{x} + s_x$. You will try to answer the following questions to be more specific:

- What fraction of a standard deviation in price does the price increase? Half of a standard deviation? A third of a standard deviation? Some other fractional part of a standard deviation?
- Can you give an answer that is true for all least squares lines?

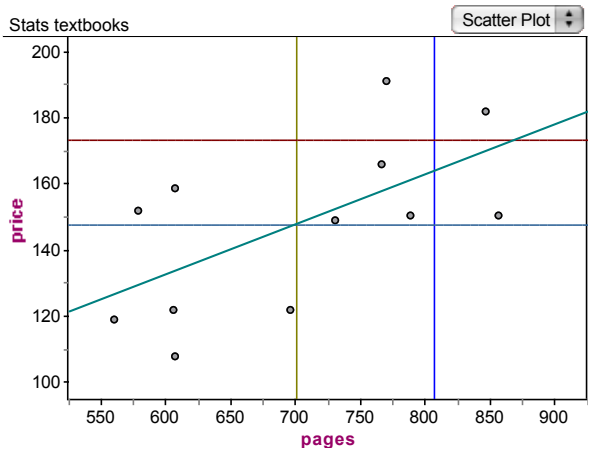
Let's think about these questions using a graph before you do any more calculations. In the graph shown here, you have added lines marking one standard deviation above each mean. Let's think graphically about what you have already established in your investigation.

Label the following on the graph:

- the least squares line,
- the point $(\bar{x}, \bar{y})$,
- the point $(\bar{x} + s_x, \bar{y} + s_y)$, and
- the point $(\bar{x} + s_x, \hat{y})$.

Hint: The calculations you did previously can help you label these points.

Is the increase in predicted price (*choose one:* more than or less than) *half* a standard deviation in price, $0.5\,s_y$? How can you tell?



Supporting Lesson 3.2.4: Special Properties of the Least Squares Regression Line

Which of the expressions below best describes the increase in the predicted price? How can you tell?

- about 0.3 s_y
- about 0.6 s_y
- about 0.9 s_y

Now let's actually calculate what fraction of a standard deviation the price increases. The standard deviation in price is _____ and the increase in the predicted price from the mean line is $\hat{y}$ – $\bar{y}$ =_____. So what fraction of a price standard deviation did $\hat{y}$ increase from $\bar{y}$?

Does this number have significance when you look back at the descriptive statistics for the data?

So now you have an answer to the second question: What does the least squares line predict for *y* when *x* is one standard deviation above the mean? When $x = \bar{x} + s_x$, $\hat{y}$ = _____.

Investigation and Discussion of Question 3

What is the connection between the equation of the least squares line and the summary statistics for x and y, such as means, standard deviations, and the correlation coefficient?

You know from your previous work that the equation of a line has the form *y = initial value + slope • x*. From the slope triangle you have drawn, you can derive an important characteristic about the slope of the least squares line. The slope is the change in *y* divided by the change in *x*.

How can you determine the slope of the least squares line using summary statistics?

You also know that the least squares line contains the point $(\bar{x}, \bar{y})$. So, you have enough information to write an expression for the *y*-intercept of the least squares line in terms of summary statistics.

Demonstration of the Development of the Formula for Initial Value

For any line *y* = *y*-intercept + slope • *x*

The least squares line contains $(\bar{x}, \bar{y})$, so for the least squares line you can write $\bar{y}$ = *y*-intercept + slope • $\bar{x}$.

You can rewrite this equation as a formula for initial value:

$$y\text{-intercept} = \bar{y}\text{-slope} \bullet \bar{x}$$

Rewriting the slope formula for the least squares line, you have the initial value expressed in terms of summary statistics:

$$y\text{-intercept} = \bar{y} - \left(\frac{r \bullet s_y}{s_x} \right) \bullet \bar{x})$$

Supporting Lesson 3.2.4: Special Properties of the Least Squares Regression Line

Homework

(4) At Los Medanos College, a statistics instructor posted the following information on her office door at the end of the semester:

Statistics FA 2010	Mean	Standard Deviation	Correlation
Prefinal exam average	75	8	0.7
Final exam score	78	12	

(a) Final course grades have not been posted. So Karen wants to predict her final exam score based on this information. She has an 82 prefinal exam average. What does the least squares line predict for Karen's final exam score?

(b) What statement is the most accurate advice you could give Karen?

- Before using the least squares line to predict your final exam score, you need to know that the relationship between prefinal exam average and final exam scores is linear. You are not given enough information to determine if the relationship is linear. Therefore, proceed with caution in using a line to make predictions in this situation.
- The correlation coefficient is 0.7. This tells you that the relationship between prefinal exam average and final exam score is fairly linear. However, there will be some error in the prediction, so do not be surprised if your final exam score differs from your prediction.

(5) In this lesson, you learned that for any least squares regression line if $x = \bar{x}$, then $\hat{y} = \bar{y}$. Explain in words what this means.

(6) You conjectured that for the least squares line if $x = \bar{x} + s_x$, then $\hat{y} = \bar{y} + s_y$.

(a) Explain in words what this means.

(b) This conjecture was not true for the statistics textbook data. Would this conjecture ever be true? If so, describe the relationship you would see in the data. If not, explain why this will never be true for any least squares line.

Initiating Lesson 3.3.1: Using Residuals to Determine If a Line Is a Good Fit

Estimated number of 50-minute class sessions: 1–2

This lesson is intended for between one and two 50-minute class sessions. Should this lesson require more than 50 minutes, the subsequent lesson can probably be covered in less than 50 minutes.

Learning Goals and Concept Flow

Statistics Learning Goals addressed by this module:

S.2. Distributional Thinking Goal: Students will demonstrate the use of distributional thinking to reason about the data in order to describe and summarize distributions of data, identify trends and patterns, judge the fit of a model to a distribution, and describe similarities and differences in comparing distributions.

Mathematics Learning Goals addressed by this module:

M.4: Functions and Modeling Goal: Students will understand functions as a way of modeling a correspondence between two variables. Students will be able to represent functions in various ways: verbally, algebraically, and graphically.

The series of tasks in this introductory lesson reinforces the concept of a residual. Specifically, the tasks highlight both a residual's relevance in determining a least-squares regression (LSR) line for bivariate data and its interpretive value in the context of an LSR line as a predictive model. Residual analysis methods are then introduced that can be used to assess the usefulness of an LSR line as a predictive model.

Students will understand that

- each datapoint can be viewed as composed of two parts: the part explained by the model and the error, called the *residual*. The residual can be due to chance variation or due to variables that are not measured.
- a residual is a measure of the error between a value predicted by the regression line and the actual value of the response variable.
- positive residuals indicate that the prediction underestimates the actual response (i.e., the actual response value is higher than the response value predicted by the model), while negative residuals indicate that the regression line overestimates the actual response (i.e., the actual response value is lower than the response value predicted by the model).
- the proportion of variability in y that can be explained by the regression model is given by r^2. Alternatively, the proportion of variability *not* explained by the regression model can be computed as the ratio of two quantities: SSE and SST. SSE is the sum of the squared residuals based on the least-squares line, while SST is the sum of the squared residuals based on the line $y = \bar{y}$ [i.e., a line that assumes a response variable's value (y) is not related to its explanatory variable's value (x), and thus a knowledge of x is of no help in terms of predicting y]. Since you want to minimize SSE, this means that a low SSE/SST ratio is desired. This also means that r^2 can be computed as $r^2 = 1 - \frac{\text{SSE}}{\text{SST}}$, and values closer to 1 (100%) are desired.

Initiating Lesson 3.3.1: Using Residuals to Determine If a Line Is a Good Fit

Students will be able to

- given the regression line and an observation from the dataset, compute the residual for that observation.
- interpret (in context) the value of a residual.
- interpret (in context) the value of the standard deviation of the residuals as a typical prediction error.
- interpret (in context) the value of r^2.
- explain why it is desirable to have a small value of s_e and a large value of r^2 in a regression context.

The rich tasks in this lesson are presented in two distinct sets. In the first set, the concept of a residual (introduced in a previous lesson) is reinforced. More formal language and interpretation regarding residuals is also discussed; specifically, students are asked to consider what information the size (absolute value) and sign of a residual communicates *in context* while relating visual and geometric characteristics of residuals to the size (absolute value) and sign of the residual. This work meets specific learning outcomes germane to residual analysis while laying the foundation for the second set of tasks. (**Note:** These first tasks include slightly more scaffolding than usual for an initial task.)

In the second set of tasks, students investigate how scatterplots that show corresponding regression models and residuals can help to broadly assess how effective a model is in predicting y from x. In that series of tasks, students are not formally developing the measures of s_e and r^2 at first. Rather, they develop observations about residual qualities and speculate as to how these qualities relate to the usefulness of a specific LSR line as a predictive model. To start, students compare two datasets that have the same regression line and the same number of observations but different scatter around the line. They are asked to decide which model is more likely to produce "good" predictions and why. (In Part II, s_e is introduced as a way of quantifying this, and then "we want a small value of s_e" is motivated.) Students then compare two datasets with the same regression line and the same SSE but a different number of observations. As before, students are tasked with deciding which model is more likely to produce good predictions and why.

Students again compare two datasets where s_e is small, but in one case, the regression line is not really useful for predicting y (i.e., a case where the line $y = \bar{y}$ also provides very good predictions such as a nearly horizontal regression line with small scatter). Here, students are asked to consider why in this situation the regression line is not useful even though s_e is small. Students then contrast the two datasets' regression lines as compared to the corresponding $y = \bar{y}$ lines to assess how much improvement (reduction) in the sum of the squared prediction errors is achieved when the LSR line is used to make predictions instead of the $y = \bar{y}$ line. (In Part II, r^2 is then introduced as a way of quantifying the improvement in squared prediction errors when the regression line is used to make predictions instead of the $y = \bar{y}$ line.) The tasks reinforce that a small s_e and large r^2 are both desired; however, the two may not be achieved simultaneously, and it is not enough to look at only one of these measurements since each communicates different aspects about the usefulness of the linear model.

Initiating Lesson 3.3.1: Using Residuals to Determine If a Line Is a Good Fit

Part I: Rich Tasks [Student Handout]

Recall that a residual (or error) is the difference between the actual value of the response variable and the value predicted by the regression line. As a formula, residual = observed y – predicted $y = y - \hat{y}$.

Analyzing residuals can help you assess the effectiveness of a least-squares regression (LSR) model for predicting values of the response variable.

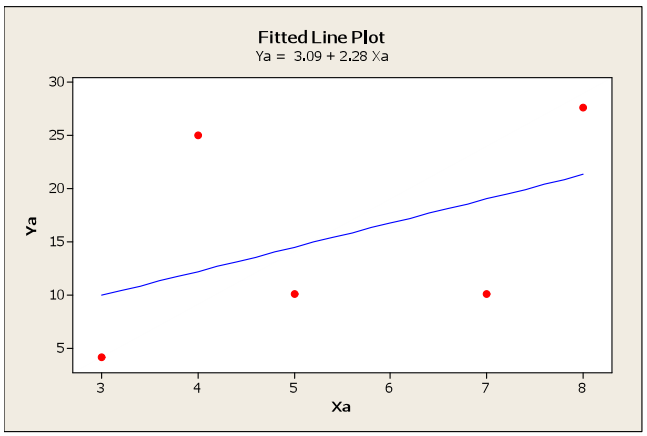
Note: The terms *residual* and *error* are used interchangeably in this lesson.

Task 1: More About the Size and Sign of Residuals [Student Handout]

Consider the scatterplot and its LSR line shown below.

Dataset A

x	y
3	4.08
4	25.08
5	10.08
7	10.08
8	27.7



(1) The equation of the regression line is $\hat{y} = 3.09 + 2.28x$. Compute the predicted value of y for each x-value and fill in the following table. For each observation, locate on the regression line a point with coordinates $(x, \hat{y})$.

Dataset A

x	y	$\hat{y} = 3.09 + 2.28x$
3	4.08	
4	25.08	
5	10.08	
7	10.08	
8	27.7	

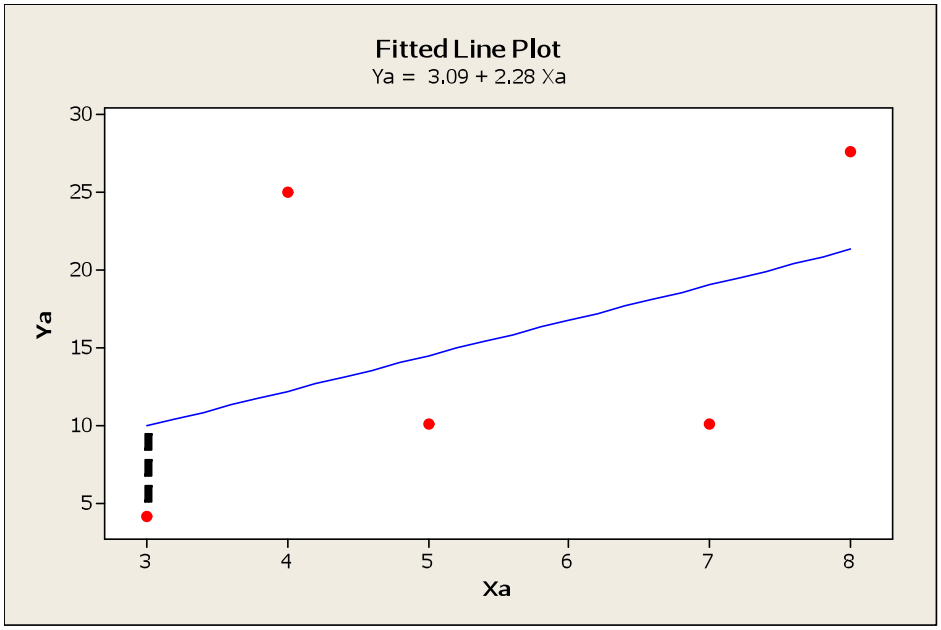
Initiating Lesson 3.3.1: Using Residuals to Determine If a Line Is a Good Fit

(2) Based on your predicted $\hat{y}$-values and the observed y-values in the original dataset, compute the residual (error) for each observation. Fill in the following table. (First, fill in $\hat{y}$-values from Question 1.)

Dataset A

x	y	$\hat{y} = 3.09 + 2.28x$	Residual $(y - \hat{y})$
3	4.08		
4	25.08		
5	10.08		
7	10.08		
8	27.7		

(3) On the scatterplot below, draw a vertical dashed segment between each datapoint and the LSR line. These segments represent the residuals for the data points. (**Note:** The first residual segment is already drawn.)



(4) How is the sign of each residual (positive or negative) represented in this diagram?

(5) What does the length of each vertical dashed segment tell you about the corresponding residual?

Initiating Lesson 3.3.1: Using Residuals to Determine If a Line Is a Good Fit

(6) Suppose an LSR model is created that predicts a subway fare based on miles traveled. Suppose an observation that represents the actual subway fare a person pays based on the miles traveled has a *positive* residual. On a scatterplot, does the point representing this observation appear above or below the LSR line? Is the actual fare the person paid more or less than the fare predicted by the model?

(7) Suppose you have a scatterplot that shows sale price and acreage for 60 homes in a particular county, and an LSR model is created that predicts a home's sale price based on the home's acreage. One particular home is represented by a datapoint that is *below* the regression line. Is the sale price of this home greater than the price predicted by the model or less than that price? What is the sign of this datapoint's residual? Another home has a sale price exactly equal to the price produced by the model. Is the datapoint for that home above the regression line, below the line, or on the line?

The LSR line is the line that minimizes the *sum of the squared residuals*. The acronym for sum of the squared residuals is SSE because *residuals* are also called *errors* (and the acronym SSR has another meaning in certain statistical analyses). As a formula, sum of the squared residuals = SSE = $\Sigma(y - \hat{y})^2$.

(8) Compute the SSE for Dataset A by completing the final column of the following table. Square the residual values you computed earlier and add up the squared residual values. (First, fill in $\hat{y}$-values and residual values from Question 2.)

Dataset A

x	y	$\hat{y}$ = 3.09 + 2.28x	Residual ($y - \hat{y}$)	Squared Residual
3	4.08			
4	25.08			
5	10.08			
7	10.08			
8	27.7			
			Total:	**= SSE**

Initiating Lesson 3.3.1: Using Residuals to Determine If a Line Is a Good Fit

Task 2: Using the Sum of the Squared Residuals to Assess a Model's Effectiveness in Predicting *y* from *x* [Student Handout]

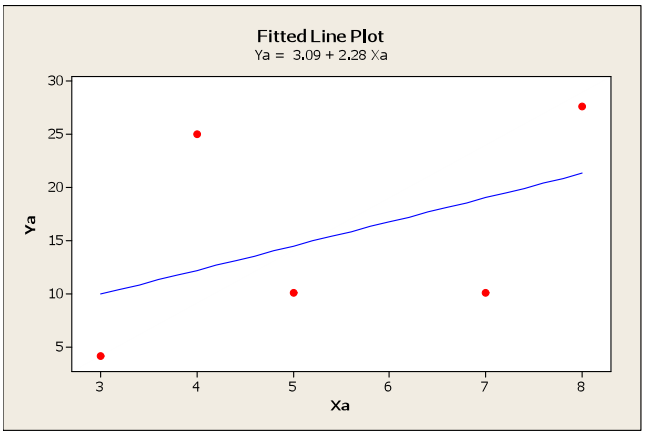
Although the LSR line is often referred to as a "line of best fit," it is important to assess how useful the LSR line is as a prediction model. The remaining tasks in this lesson address this question.

Part A

(9) Two datasets with their corresponding scatterplots and LSR lines are shown below. Both plots have approximately *the same regression line* but *different scatter* around the line. Based on visual assessment, which regression line do you think is more likely to produce better predictions, the one for Dataset A or the one for Dataset B? Or would they be equally good prediction models? Explain your reasoning, and carefully describe any visual characteristics of the scatterplots that lead to your decision.

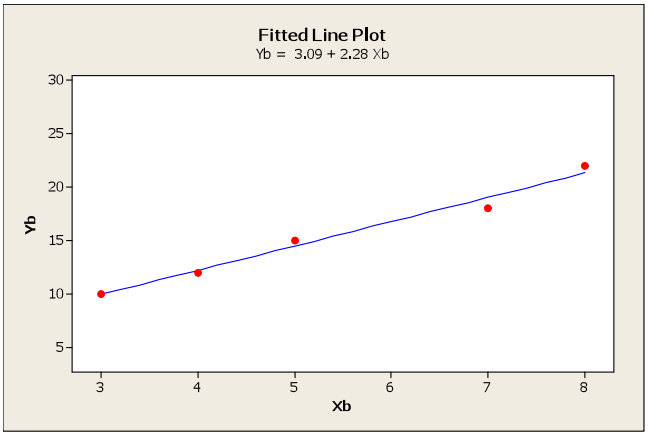
Dataset A

x	y
3	4.08
4	25.08
5	10.08
7	10.08
8	27.7



Dataset B

x	y
3	10
4	12
5	15
7	18
8	22



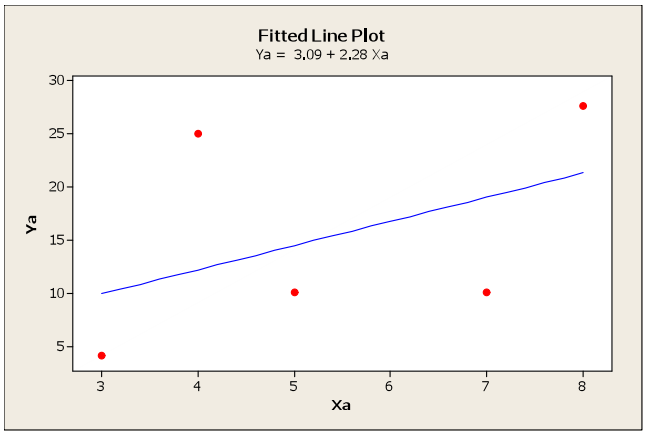
(10) Which regression model from Question 9 do you think has the lower SSE? Why?

Initiating Lesson 3.3.1: Using Residuals to Determine If a Line Is a Good Fit

(11) Two datasets with their corresponding scatterplots and LSR lines are shown below. Both plots have approximately the *same regression line* and approximately the *same SSE value* but a *different number of observations* in each case. Based on visual assessment, which regression line do you think is more likely to produce better predictions, the one for Dataset A or the one for Dataset C? Or would they be equally good prediction models? Explain your reasoning, and carefully describe any visual characteristics of the scatterplots that led to your decision.

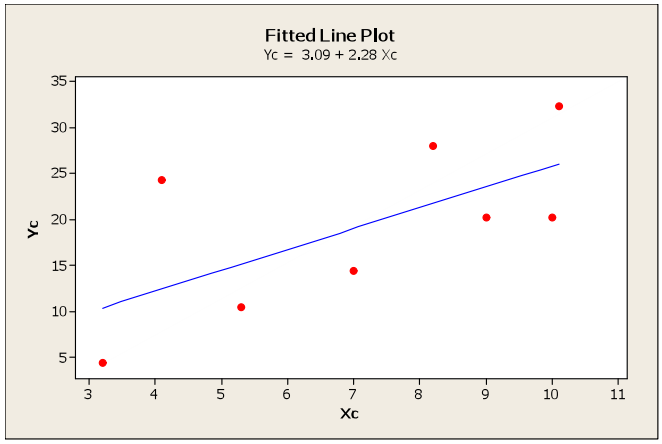
Dataset A

x	y
3	4.08
4	25.08
5	10.08
7	10.08
8	27.7



Dataset C

x	y
3.2	4.34
4.1	24.26
5.3	10.5
7	14.36
8.2	27.96
9	20.26
10	20.26
10.1	32.37



(12) Summarize your observations:

Given the regression lines for two datasets, the one with the (*circle one*: higher/lower) SSE is probably the better predictor. If the two datasets have the same SSE, the regression line for the one with (*circle one*: more/fewer) datapoints is probably the better predictor.

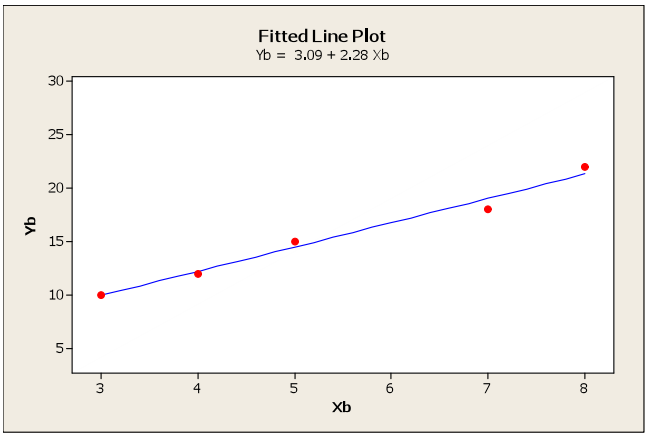
Initiating Lesson 3.3.1: Using Residuals to Determine If a Line Is a Good Fit

Part B

Two scatterplots with their corresponding LSR lines are shown below. Both datasets have the *same number of observations* and approximately the *same SSE value*.

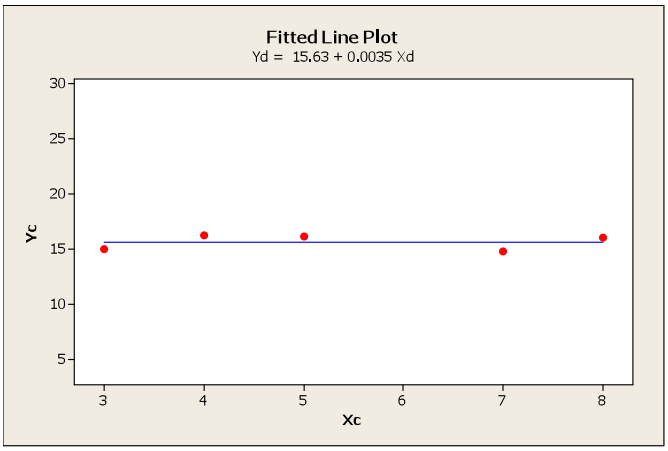
Dataset B

x	y
3	10
4	12
5	15
7	18
8	22



Dataset D

x	y
3	15
4	16.3
5	16.12
7	14.83
8	16



(13) For each dataset, compute the mean of the response variable (y) values.

Dataset B: $\bar{y}$ = _____ Dataset D: $\bar{y}$ = _____

(14) For the Dataset B scatterplot, add the corresponding horizontal line $y = \bar{y}$ to the scatterplot. Do the same for Dataset D.

The $y = \bar{y}$ line is very important in helping you assess the usefulness of an LSR line as a predictive model. The $y = \bar{y}$ line serves as an appropriate model if a response variable's value (y) is not related to its explanatory variable's value (x). In other words, if y does not appear to increase or decrease as x increases, the $y = \bar{y}$ line is an *appropriate model* for predicting y. This $y = \bar{y}$ line can be used as a baseline.

Initiating Lesson 3.3.1: Using Residuals to Determine If a Line Is a Good Fit

The following is a common method for assessing how much of the variability in y is accounted for by an LSR model of y on x:

1. Pretend that the $y = \bar{y}$ line is the line of best fit.
2. Compute the residuals based on this assumption. Since the predicted value of y is always $\bar{y}$ in such a case, the residual is $y - \bar{y}$ for each observation.
3. Compute the sum of these squared residuals (i.e., the sum of the squares of $y - \bar{y}$). Call this sum the *sum of squares total* (SST). As a formula, sum of squares total = SST = $\Sigma(y - \bar{y})^2$.
4. Now compute the real LSR line and compute the real SSE based on that LSR line.
5. Calculate the ratio SSE/SST.

The lower the value of SSE/SST, the better the job the LSR model is doing in terms of accounting for the variability in your response variable (y).

(15) Calculate the SSE/SST ratio for Datasets A–D.

(16) If the value of SSE/SST is close to 1, how are the SSE and SST values related? In that case, since the SSE value comes from the real LSR line and the SST value comes from the $y = \bar{y}$ line, do you think the equations of the real LSR line and $y = \bar{y}$ line would be similar or different? Of Datasets B and D, which one best fits that description of an SSE/SST ratio close to 1? In that case, is the LSR line much of an improvement over the $y = \bar{y}$ line in terms of predicting y? [**Hint:** Notice that the SSE and SST formulas are identical when the predicted value of y ($\hat{y}$) is equal to the mean of the y-values ($\bar{y}$), or in other words, when your best prediction of y is the mean of the y-values ($\bar{y}$).]

(17) If the value of SSE/SST is close to 0, how are the SSE and SST values related? In that case, since the SSE value comes from the real LSR line and *since SSE is something you want to minimize*, do you think that the LSR line model is doing a good or bad job in predicting y? Of Datasets B and D, which one best fits that description of an SSE/SST ratio close to 0? In that case, is the LSR line an improvement over the $y = \bar{y}$ line in terms of predicting y?

(18) If y does not appear to increase or decrease as x increases, the $y = \bar{y}$ line serves as an appropriate model for predicting y. Of Datasets B and D, which one best fits that description? Do you think that the *size* (absolute value) of the correlation coefficient (r) for that dataset is a small or large value?

(19) Consequently, if y increases or decreases as x increases, the $y = \bar{y}$ line is most likely *not* an appropriate model for predicting y. Of Datasets B and D, which one best shows y increasing or decreasing as x increases? Do you think that the *size* (absolute value) of the correlation coefficient (r) for that dataset is a small or large value?

(20) Rank the least-squares models of Datasets A–D from best to worst in terms of how good of a job the model seems to be doing in predicting y from x. For each model, think about the scatter of the residuals around a given line and the visual distinction between the real LSR line and the $y = \bar{y}$ line as well as other characteristics of the models and scatterplots. Comment on what characteristics of the models and the scatterplots led to your rankings. How much of a factor was the scatter of the residuals around a given line? What about the SSE/SST ratio? Was the visual distinction between the real LSR line and the $y = \bar{y}$ line much of a factor? What else was important in your ranking?

	Dataset (A, B, C, or D)	**Comments**
Best Model		
Next Best Model		
Third Best Model		
Worst Model of the 4		

Part I Wrap-Up/Transition to Part II

As has been the case in previous lessons, the tasks in Part I should provide students the opportunity to struggle with the important ideas—in this case, which visual characteristics of the scatterplot, the regression line drawn on the scatterplot, and the residual values as seen on the scatterplot help in assessing the usefulness of a linear regression model? Students should realize that lower values of SSE (and the visual concept of less scatter around the regression line) are desired, but they should also notice that this measurement alone does not tell the whole story regarding a model's usefulness. In Part II, you begin to address the formal computation of s_e and r^2 to demonstrate that SSE is an important component of these measurements, to show via formula that SSE alone does not tell the whole story (i.e., it needs to be considered relative to the number of observations available, and it needs to be compared to the SST value), and to address any errors in student reasoning and computation.

Consider polling the class on their responses to Question 20. Also encourage students to discuss as a group the characteristics of the datasets, scatterplots, and regression lines that influenced their decisions in Question 20.

Have students discuss their answers to Questions 6 and 7 either as a class or in small groups. It may be prudent to take time here to verify that students have mastered the idea that a positive residual occurs when a point is above the regression line, that a 0 residual occurs when a point is exactly on the regression line, and that a negative residual occurs when a point is below the regression line. Also reinforce that the context of the story behind the data is very important. For example, since a negative residual means that an observed value is less than the expected value, this may be great when you are buying a soda ("I am paying less than I was expecting.") but not so great when you are predicting your income ("I am earning less than I was expecting.").

Part II: Computing Statistical Measures That Use the SSE to Assess a Model's Effectiveness in Predicting *y* from *x*

Introduction

Tell students that they are now going to develop some measurements for more carefully quantifying some of the previous concepts. Reinforce that previous visual observations and conjectures will prove useful and that SSE is an important component in both measurements to be discussed. As the SSE value for the regression model in Dataset A is presented in the table for Question 21, ask students to verify that this value is what they computed back in Question 8.

Have students work in groups (or check with one another periodically) to see if correct/similar values of s_e and r^2 are calculated throughout the tasks. Also encourage students to share their observations regarding how the calculations fit (or do not fit) previous conjectures.

Questions and Tasks [Student Handout]

One Measurement: s_e

The standard error of the regression (s_e) is a formal way of measuring the typical amount that an observation deviates from the least-squares line. It is a representation of the size of the average vertical distance that observations fall from the LSR line. As such, the measurement units of s_e are the same as the measurement units of the response variable (y). (**Note:** The use of s in this term is directly related to the *standard deviation* work from previous lessons in that a concept of *measuring spread* is employed here.)

A smaller s_e value for your regression implies that your model will do a better job of predicting the response variable since s_e is measuring the size of a "typical prediction error," and you want that quantity to be small.

As you may have suspected, the value of s_e is related to SSE. It is also related to the number of observations used to develop the regression equation.

As a formula, $s_e = \sqrt{\frac{\text{SSE}}{n-2}}$.

(21) Using the table below, compute the s_e-values for each of the four cases you previously examined. Which model has the highest s_e? Which one has the lowest s_e? Did the model that you thought was doing the best job (from your previous rankings) have the lowest s_e? Did any model that you thought was doing a poor job also have a small s_e?

Dataset	SSE	Number of Observations	s_e
A	340.34		
B	1.86		
C	340.28		
D	1.86		

Even though you want a small s_e value, that alone does not fully assess the usefulness of the LSR line as a prediction model. As shown in Dataset D, a small s_e can still occur even when the LSR line is not much better than the $y = \bar{y}$ line.

Another Measurement: Coefficient of Determination

A formal measurement of the percentage of variability in *y* that is accounted for by an LSR model of *y* on *x* is called the *coefficient of determination.* This measurement quantifies the improvement in SSE (specifically, the reduction in SSE) when the LSR line is used to make predictions instead of the $y = \bar{y}$ line.

This coefficient of determination value is closely related to the SSE/SST ratio discussed earlier. As a formula, coefficient of determination = 1 – (SSE/SST).

Since the SSE must be a value greater than or equal to 0, *the highest value that a coefficient of determination can have is 1* (or 100%), and that occurs when SSE = 0. Since the SSE value can only be as great as the SST value (which happens when $\hat{y}$-values equal the $\bar{y}$-value), *the lowest value that a coefficient of determination can have is 0* (or 0%), and that occurs when SSE = SST.

(22) Based on this formula, if you have a dataset with a very small SSE (relative to SST), do you have a high coefficient of determination value (closer to 1) or a low coefficient of determination value (closer to 0)? Given that the objective in LSR is to minimize the SSE, do you want a high coefficient of determination value (closer to 1) or a low coefficient of determination value (closer to 0) for your LSR?

(23) Compute the coefficient of determination for the LSR models for each dataset in the table below. Which one has the highest coefficient of determination? Which one has the lowest coefficient of determination? Did the model that you thought was doing the best job have the highest coefficient of variation? Did any model that you thought was doing a poor job still have a large coefficient of variation?

Dataset	SSE	SST	Coefficient of Determination
A	340.34	429.74	
B	1.86	91.2	
C	340.28	597.91	
D	1.86	1.86	

In Questions 18 and 19, you considered what the size of the correlation coefficient (r) might be for cases where the $y = \bar{y}$ line most likely serves as an appropriate model for predicting y and for cases where the $y = \bar{y}$ line most likely does *not* serve as an appropriate model for predicting y. Recall the characteristics of the datasets and scatterplots that fit each case.

There is a specific mathematical relationship between the correlation coefficient (r) between two variables (x and y) and the coefficient of determination for the least-squares model that predicts the response variable's value based on a single explanatory variable (x). For those cases, coefficient of determination = r^2 = (correlation coefficient)^2. For this reason, the coefficient of determination is often called r^2.

(24) Since the coefficient of determination has the same value as the square of the correlation coefficient (r) for two variable cases such as those discussed in this lesson, do you think that datasets with strong correlation values (values of r that are closer to –1 or 1) yield LSR models that have a high r^2-value (close to 1) or a low r^2-value (close to 0)? Check your previous work and the previous scatterplots if needed. Do you think that datasets with strong correlation values (values of r that are closer to –1 or 1) yield LSR models that explain a great deal of the variability in y or not much of the variability in y? Why does that make sense visually in terms of residuals?

Wrap-Up Questions/Direct Instruction About Statistical Concepts

Have students to share how their calculations from Questions 21 and 23 fit (or did not fit) their previous conjectures and rankings (from Question 20) of how well a model seemed to be doing in predicting y from x.

Via discussion or lecture, highlight the following:

- Positive residuals indicate that the prediction underestimates the actual response (i.e., the actual response value is higher than the response value predicted by the model), while negative residuals indicate that the regression line overestimates the actual response (i.e., the actual response value is lower than the response value predicted by the model).
- The context of the story behind the data is very important. For example, since a negative residual means that an observed value is less than the expected value, this may be great when you are buying a soda ("I am paying less than I was expecting.") but not so great when you are predicting your income ("I am earning less than I was expecting.").
- When bivariate data have no correlation (i.e., when y does not vary at all with x), a horizontal line model of $y = \bar{y}$ is the best-fitting line that you can use to predict y. (The horizontal line $y = \bar{y}$ in that case is actually the LSR line that minimizes the sum of the squared residuals.) However, if the data have any correlation, a specific nonhorizontal LSR line (i.e., one with nonzero slope) is the line of best fit.
- While a low value of the SSE is desired (remember that is what you are minimizing in a *least-squares* regression model), when evaluating and comparing regression models in terms of their ability to predict y from x, you need to examine the SSE in relation to
  - the number of observations used to generate the LSR model, and
  - the SST of the LSR.

Initiating Lesson 3.3.1: Using Residuals to Determine If a Line Is a Good Fit

- When assessing a linear regression model, you want a small s_e-value and a large r^2-value. However, it is not enough to look at only one of these measurements. Even though they are both based on the regression's SSE value, both measurements communicate very different things about the usefulness of the linear model. A desirable value in one does not necessarily guarantee a good value in the other.
    - A model with a low s_e and a low r^2 means that although the predictions have little error, the model is not much better than $y = \bar{y}$ (which is an equation that implies that your knowledge of x is of no help in predicting y). See Dataset D.
    - A model with a high r^2 and a high s_e means that the regression model of y on x accounts for much of the variability in y but that a typical prediction error may be large (addressed in a Homework example).

Note: While the analysis of s_e-values and r^2-values has only been presented here for cases of *linear* regression models with one response variable (y) and one explanatory variable (x), analysis of these measures is also used for assessing other regression models such as nonlinear and multiple regression models not covered in the Statway course.

Initiating Lesson 3.3.1: Using Residuals to Determine If a Line Is a Good Fit

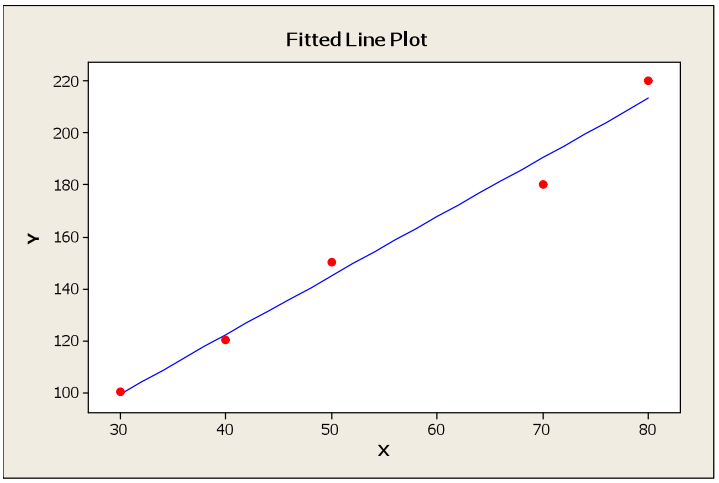
Homework [Student Handout]

(1) Verify that the coefficient of determination value (r^2) that you computed for the LSR line in Dataset A (Question 23 in Part II) is equal to the square of the correlation coefficient (r) for that dataset.

(2) A new dataset and its corresponding scatterplot with the LSR line for predicting y from x are shown below.

New Dataset

x	y
30	100
40	120
50	150
70	180
80	220



(a) From visual inspection, compared to the four datasets previously discussed, where do you think this regression model ranks in terms of its usefulness in predicting y from x for its dataset? (See Question 20 in Part I.) Do you put it near the top of the list, near the bottom, or somewhere in the middle? Why? Do you think that this model has a particularly good r^2-value? A particularly good s_e-value? Explain your reasoning.

(b) Compute the LSR line for this dataset. What do you predict for y when $x = 45$? Compute the r^2-value and s_e-value for the regression model. Do these measurements support your ranking for this model in Question 2a?

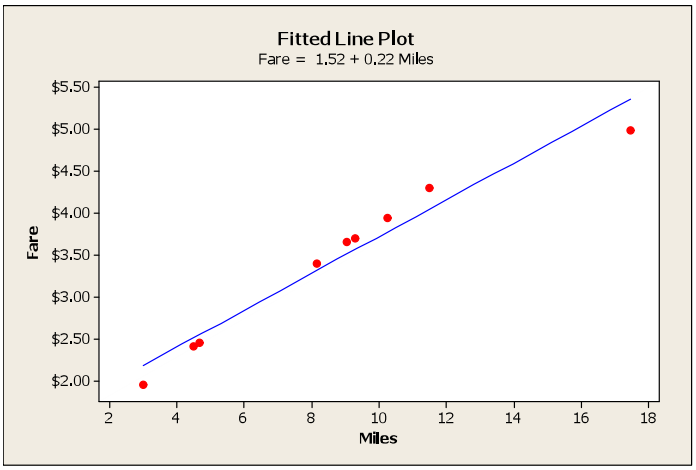
(c) Based on your visual inspection in Question 2a and your computations in Question 2b, what is good about this model that makes it useful for predicting y from x? What cause for concern do you have (if any) regarding the model's usefulness in predicting y from x?

(3) The Metro is the subway rail service used for Washington, D.C., and its immediate suburbs. When using this service, a rider must pay a fare that is relative to the distance traveled from Starting Point A to Ending Point B. For example, if a passenger boarded a subway train at a given station and traveled 10 miles, he or she pays a greater fare than if he or she had only traveled 7 miles. The further you travel from a given starting point, the more you generally have to pay.

Initiating Lesson 3.3.1: Using Residuals to Determine If a Line Is a Good Fit

The following data show the miles traveled and the standard nonpeak (reduced) fare amount needed for travel from the Metro Center station stop to nine other Metro stations.[1]

Station	Miles	Fare
Pentagon	2.98	$1.95
Virginia Square-GMU	4.47	$2.40
Congress Heights	4.66	$2.45
Medical Center	8.15	$3.40
Branch Ave	9.02	$3.65
West Falls Church-VT/UVA	9.29	$3.70
New Carrollton	10.23	$3.95
Greenbelt	11.49	$4.30
Shady Grove	17.44	$5.00



Predicting fare (y) based on miles traveled (x), the LSR model is $\hat{y} = 1.52 + 0.22x$.

(a) In the context of the data presented, what does the slope value of 0.22 estimate? Use words such as *dollars*, *cents*, *miles*, and *fare* in your description.

(b) What is the residual for the point that represents the fare to the Greenbelt station? Does this residual value mean that a rider pays more or less than the model predicts for that trip?

(c) The value of s_e for the regression model is 0.220715. (**Note:** It is only coincidental that this statistic's value is close to the value of the slope in the LSR equation.) How do you interpret this s_e-value in the context of this equation? What are the units of s_e? Explain what this s_e-value implies in terms of the quality of your predictions using this model.

(d) The r^2-value for the regression model is 95.6%. How do you specifically interpret that value in the context of this equation? Explain what this r^2-value implies in terms of the quality of your predictions using this model.

[1] Retrieved from www.wmata.com/rail/station_detail.cfm?station_id=1.

Initiating Lesson 3.3.1: Using Residuals to Determine If a Line Is a Good Fit

Part I

Recall that a residual (or error) is the difference between the actual value of the response variable and the value predicted by the regression line. As a formula, residual = observed y – predicted $y = y - \hat{y}$.

Analyzing residuals can help you assess the effectiveness of a least-squares regression (LSR) model for predicting values of the response variable.

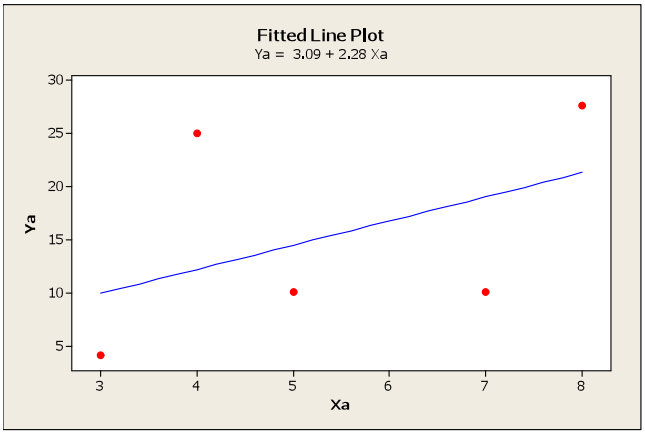
Note: The terms *residual* and *error* are used interchangeably in this lesson.

Task 1: More About the Size and Sign of Residuals

Consider the scatterplot and its LSR line shown below.

Dataset A

x	y
3	4.08
4	25.08
5	10.08
7	10.08
8	27.7



(1) The equation of the regression line is $\hat{y} = 3.09 + 2.28x$. Compute the predicted value of y for each x-value and fill in the following table. For each observation, locate on the regression line a point with coordinates $(x, \hat{y})$.

Dataset A

x	y	$\hat{y} = 3.09 + 2.28x$
3	4.08	
4	25.08	
5	10.08	
7	10.08	
8	27.7	

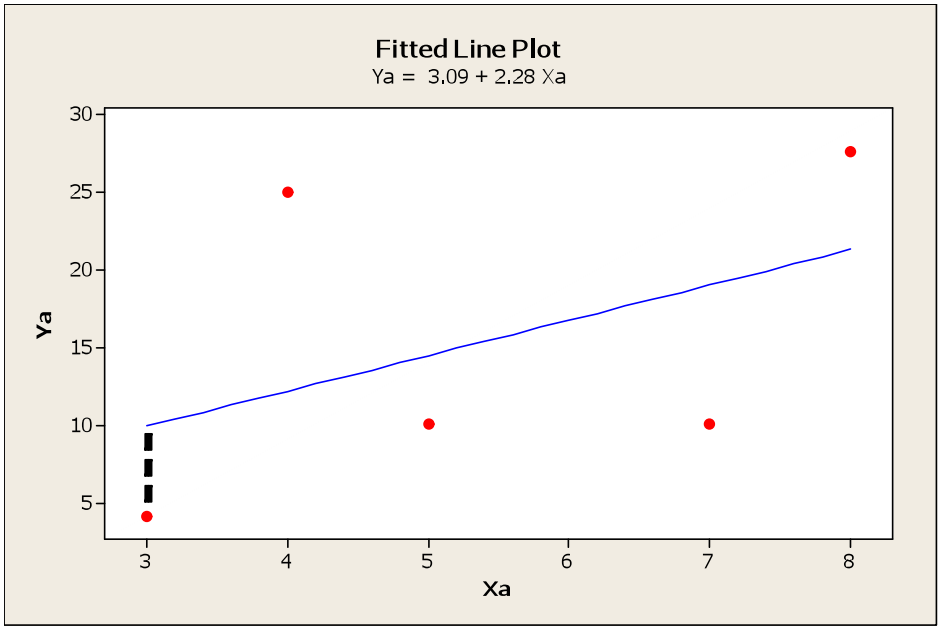
Initiating Lesson 3.3.1: Using Residuals to Determine If a Line Is a Good Fit

(2) Based on your predicted $\hat{y}$-values and the observed y-values in the original dataset, compute the residual (error) for each observation. Fill in the following table. (First, fill in $\hat{y}$-values from Question 1.)

Dataset A

x	y	$\hat{y} = 3.09 + 2.28x$	Residual ($y - \hat{y}$)
3	4.08		
4	25.08		
5	10.08		
7	10.08		
8	27.7		

(3) On the scatterplot below, draw a vertical dashed segment between each datapoint and the LSR line. These segments represent the residuals for the data points. (**Note:** The first residual segment is already drawn.)



(4) How is the sign of each residual (positive or negative) represented in this diagram?

(5) What does the length of each vertical dashed segment tell you about the corresponding residual?

Initiating Lesson 3.3.1: Using Residuals to Determine If a Line Is a Good Fit

(6) Suppose an LSR model is created that predicts a subway fare based on miles traveled. Suppose an observation that represents the actual subway fare a person pays based on the miles traveled has a *positive* residual. On a scatterplot, does the point representing this observation appear above or below the LSR line? Is the actual fare the person paid more or less than the fare predicted by the model?

(7) Suppose you have a scatterplot that shows sale price and acreage for 60 homes in a particular county, and an LSR model is created that predicts a home's sale price based on the home's acreage. One particular home is represented by a datapoint that is *below* the regression line. Is the sale price of this home greater than the price predicted by the model or less than that price? What is the sign of this datapoint's residual? Another home has a sale price exactly equal to the price produced by the model. Is the datapoint for that home above the regression line, below the line, or on the line?

The LSR line is the line that minimizes the *sum of the squared residuals*. The acronym for sum of the squared residuals is SSE because *residuals* are also called *errors* (and the acronym SSR has another meaning in certain statistical analyses). As a formula, sum of the squared residuals = SSE = $\Sigma(y - \hat{y})^2$.

(8) Compute the SSE for Dataset A by completing the final column of the following table. Square the residual values you computed earlier and add up the squared residual values. (First, fill in $\hat{y}$-values and residual values from Question 2.)

Dataset A

x	y	$\hat{y}$ = 3.09 + 2.28x	Residual ($y - \hat{y}$)	Squared Residual
3	4.08			
4	25.08			
5	10.08			
7	10.08			
8	27.7			
			Total:	**= SSE**

Initiating Lesson 3.3.1: Using Residuals to Determine If a Line Is a Good Fit

Task 2: Using the Sum of the Squared Residuals to Assess a Model's Effectiveness in Predicting *y* from *x*

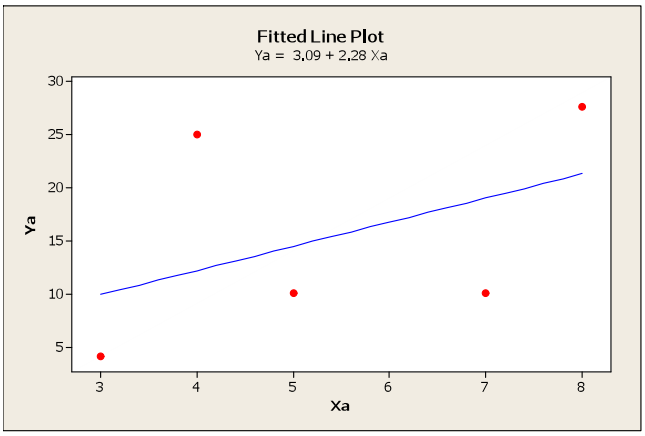
Although the LSR line is often referred to as a "line of best fit," it is important to assess how useful the LSR line is as a prediction model. The remaining tasks in this lesson address this question.

Part A

(9) Two datasets with their corresponding scatterplots and LSR lines are shown below. Both plots have approximately *the same regression line* but *different scatter* around the line. Based on visual assessment, which regression line do you think is more likely to produce better predictions, the one for Dataset A or the one for Dataset B? Or would they be equally good prediction models? Explain your reasoning, and carefully describe any visual characteristics of the scatterplots that led to your decision.

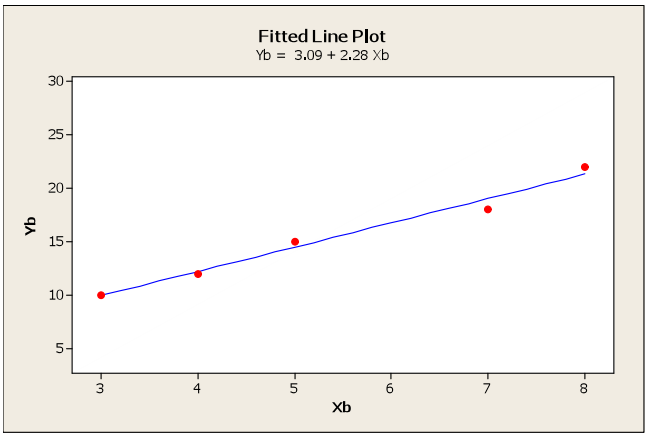
Dataset A

x	y
3	4.08
4	25.08
5	10.08
7	10.08
8	27.7



Dataset B

x	y
3	10
4	12
5	15
7	18
8	22



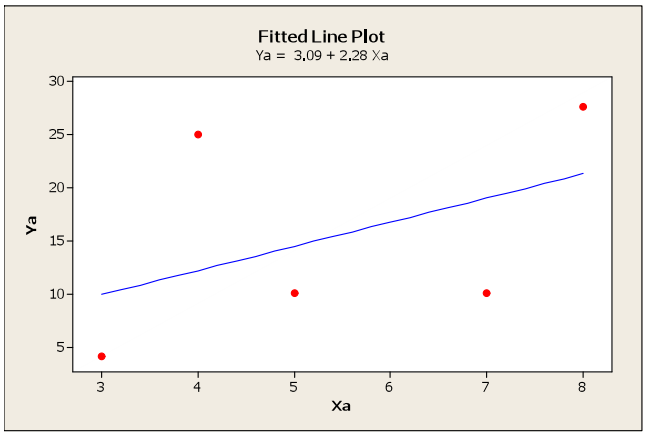
Initiating Lesson 3.3.1: Using Residuals to Determine If a Line Is a Good Fit

(10) Which regression model from Question 9 do you think has the lower SSE? Why?

(11) Two datasets with their corresponding scatterplots and LSR lines are shown below. Both plots have approximately the *same regression line* and approximately the *same SSE value* but a *different number of observations* in each case. Based on visual assessment, which regression line do you think is more likely to produce better predictions, the one for Dataset A or the one for Dataset C? Or would they be equally good prediction models? Explain your reasoning, and carefully describe any visual characteristics of the scatterplots that led to your decision.

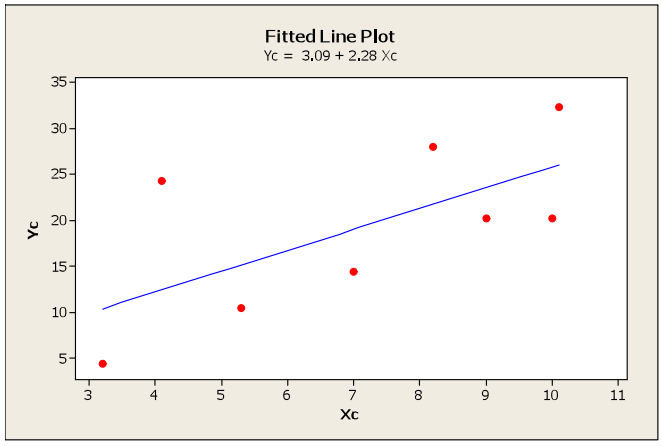
Dataset A

x	y
3	4.08
4	25.08
5	10.08
7	10.08
8	27.7



Dataset C

x	y
3.2	4.34
4.1	24.26
5.3	10.5
7	14.36
8.2	27.96
9	20.26
10	20.26
10.1	32.37



Initiating Lesson 3.3.1: Using Residuals to Determine If a Line Is a Good Fit

(12) Summarize your observations:

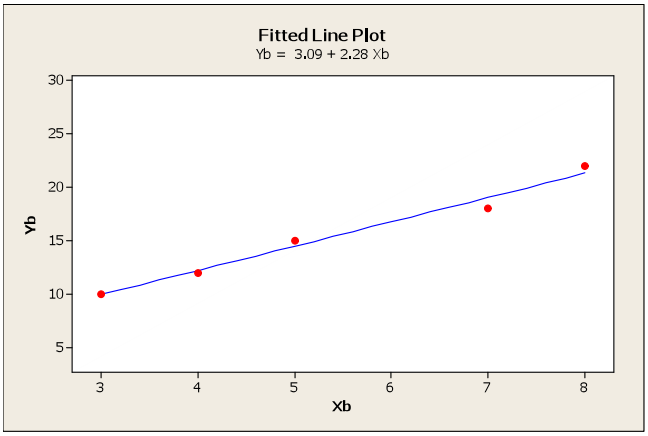
Given the regression lines for two datasets, the one with the (*circle one*: higher/lower) SSE is probably the better predictor. If the two datasets have the same SSE, the regression line for the one with (*circle one*: more/fewer) datapoints is probably the better predictor.

Part B

Two scatterplots with their corresponding LSR lines are shown below. Both datasets have the *same number of observations* and approximately the *same SSE value*.

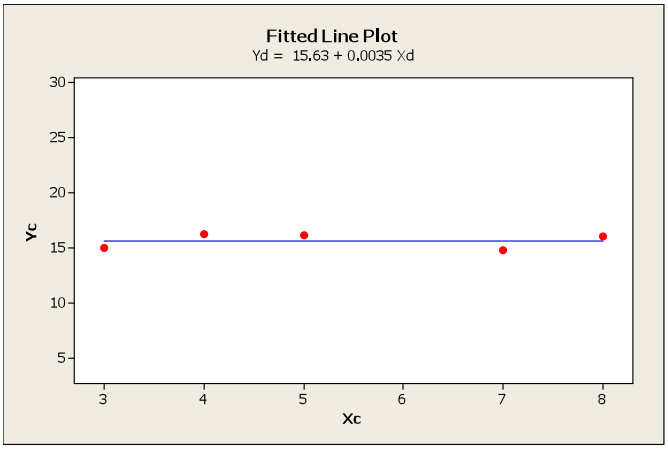
Dataset B

x	y
3	10
4	12
5	15
7	18
8	22



Dataset D

x	y
3	15
4	16.3
5	16.12
7	14.83
8	16



(13) For each dataset, compute the mean of the response variable (y) values.

Dataset B: $\bar{y}$ = _____ Dataset D: $\bar{y}$ = _____

Initiating Lesson 3.3.1: Using Residuals to Determine If a Line Is a Good Fit

(14) For the Dataset B scatterplot, add the corresponding horizontal line $y = \bar{y}$ to the scatterplot. Do the same for Dataset D.

The $y = \bar{y}$ line is very important in helping you assess the usefulness of an LSR line as a predictive model. The $y = \bar{y}$ line serves as an appropriate model if a response variable's value (y) is not related to its explanatory variable's value (x). In other words, if y does not appear to increase or decrease as x increases, the $y = \bar{y}$ line is an *appropriate model* for predicting y. This $y = \bar{y}$ line can be used as a baseline.

The following is a common method for assessing how much of the variability in y is accounted for by an LSR model of y on x:

1. Pretend that the $y = \bar{y}$ line is the line of best fit.
2. Compute the residuals based on this assumption. Since the predicted value of y is always $\bar{y}$ in such a case, the residual is $y - \bar{y}$ for each observation.
3. Compute the sum of these squared residuals (i.e., the sum of the squares of $y - \bar{y}$). Call this sum the *sum of squares total* (SST). As a formula, sum of squares total = SST = $\Sigma(y - \bar{y})^2$.
4. Now compute the real LSR line and compute the real SSE based on that LSR line.
5. Calculate the ratio SSE/SST.

The lower the value of SSE/SST, the better the job the LSR model is doing in terms of accounting for the variability in your response variable (y).

(15) Calculate the SSE/SST ratio for Datasets A–D.

(16) If the value of SSE/SST is close to 1, how are the SSE and SST values related? In that case, since the SSE value comes from the real LSR line and the SST value comes from the $y = \bar{y}$ line, do you think the equations of the real LSR line and $y = \bar{y}$ line would be similar or different? Of Datasets B and D, which one best fits that description of an SSE/SST ratio close to 1? In that case, is the LSR line much of an improvement over the $y = \bar{y}$ line in terms of predicting y? [**Hint:** Notice that the SSE and SST formulas are identical when the predicted value of y ($\hat{y}$) is equal to the mean of the y-values ($\bar{y}$), or in other words, when your best prediction of y is the mean of the y-values ($\bar{y}$).]

Initiating Lesson 3.3.1: Using Residuals to Determine If a Line Is a Good Fit

(17) If the value of SSE/SST is close to 0, how are the SSE and SST values related? In that case, since the SSE value comes from the real LSR line and *since SSE is something you want to minimize*, do you think that the LSR line model is doing a good or bad job in predicting y? Of Datasets B and D, which one best fits that description of an SSE/SST ratio close to 0? In that case, is the LSR line an improvement over the $y = \bar{y}$ line in terms of predicting y?

(18) If y does not appear to increase or decrease as x increases, the $y = \bar{y}$ line serves as an appropriate model for predicting y. Of Datasets B and D, which one best fits that description? Do you think that the *size* (absolute value) of the correlation coefficient (r) for that dataset is a small or large value?

(19) Consequently, if y increases or decreases as x increases, the $y = \bar{y}$ line is most likely *not* an appropriate model for predicting y. Of Datasets B and D, which one best shows y increasing or decreasing as x increases? Do you think that the *size* (absolute value) of the correlation coefficient (r) for that dataset is a small or large value?

Initiating Lesson 3.3.1: Using Residuals to Determine If a Line Is a Good Fit

(20) Rank the least-squares models of Datasets A–D from best to worst in terms of how good of a job the model seems to be doing in predicting y from x. For each model, think about the scatter of the residuals around a given line and the visual distinction between the real LSR line and the $y = \bar{y}$ line as well as other characteristics of the models and scatterplots. Comment on what characteristics of the models and the scatterplots led to your rankings. How much of a factor was the scatter of the residuals around a given line? What about the SSE/SST ratio? Was the visual distinction between the real LSR line and the $y = \bar{y}$ line much of a factor? What else was important in your ranking?

	Dataset (A, B, C, or D)	**Comments**
Best Model		
Next Best Model		
Third Best Model		
Worst Model of the 4		

Initiating Lesson 3.3.1: Using Residuals to Determine If a Line Is a Good Fit

Part II: Computing Statistical Measures That Use the SSE to Assess a Model's Effectiveness in Predicting *y* from *x*

One Measurement: s_e

The standard error of the regression (s_e) is a formal way of measuring the typical amount that an observation deviates from the least-squares line. It is a representation of the size of the average vertical distance that observations fall from the LSR line. As such, the measurement units of s_e are the same as the measurement units of the response variable (*y*). (**Note:** The use of *s* in this term is directly related to the *standard deviation* work from previous lessons in that a concept of *measuring spread* is employed here.)

A smaller s_e value for your regression implies that your model will do a better job of predicting the response variable since s_e is measuring the size of a "typical prediction error," and you want that quantity to be small.

As you may have suspected, the value of s_e is related to SSE. It is also related to the number of observations used to develop the regression equation.

As a formula, $s_e = \sqrt{\frac{SSE}{n-2}}$.

(21) Using the table below, compute the s_e-values for each of the four cases you previously examined. Which model has the highest s_e? Which one has the lowest s_e? Did the model that you thought was doing the best job (from your previous rankings) have the lowest s_e? Did any model that you thought was doing a poor job also have a small s_e?

Dataset	SSE	Number of Observations	s_e
A	340.34		
B	1.86		
C	340.28		
D	1.86		

Even though you want a small s_e value, that alone does not fully assess the usefulness of the LSR line as a prediction model. As shown in Dataset D, a small s_e can still occur even when the LSR line is not much better than the $y = \bar{y}$ line.

Initiating Lesson 3.3.1: Using Residuals to Determine If a Line Is a Good Fit

Another Measurement: Coefficient of Determination

A formal measurement of the percentage of variability in *y* that is accounted for by an LSR model of *y* on *x* is called the *coefficient of determination.* This measurement quantifies the improvement in SSE (specifically, the reduction in SSE) when the LSR line is used to make predictions instead of the $y = \bar{y}$ line.

This coefficient of determination value is closely related to the SSE/SST ratio discussed earlier. As a formula, coefficient of determination = 1 – (SSE/SST).

Since the SSE must be a value greater than or equal to 0, *the highest value that a coefficient of determination can have is 1* (or 100%), and that occurs when SSE = 0. Since the SSE value can only be as great as the SST value (which happens when $\hat{y}$-values equal the $\bar{y}$-value), *the lowest value that a coefficient of determination can have is 0* (or 0%), and that occurs when SSE = SST.

(22) Based on this formula, if you have a dataset with a very small SSE (relative to SST), do you have a high coefficient of determination value (closer to 1) or a low coefficient of determination value (closer to 0)? Given that the objective in LSR is to minimize the SSE, do you want a high coefficient of determination value (closer to 1) or a low coefficient of determination value (closer to 0) for your LSR?

(23) Compute the coefficient of determination for the LSR models for each dataset in the table below. Which one has the highest coefficient of determination? Which one has the lowest coefficient of determination? Did the model that you thought was doing the best job have the highest coefficient of variation? Did any model that you thought was doing a poor job still have a large coefficient of variation?

Dataset	SSE	SST	Coefficient of Determination
A	340.34	429.74	
B	1.86	91.2	
C	340.28	597.91	
D	1.86	1.86	

Initiating Lesson 3.3.1: Using Residuals to Determine If a Line Is a Good Fit

In Questions 18 and 19, you considered what the size of the correlation coefficient (r) might be for cases where the $y = \bar{y}$ line most likely serves as an appropriate model for predicting y and for cases where the $y = \bar{y}$ line most likely does *not* serve as an appropriate model for predicting y. Recall the characteristics of the datasets and scatterplots that fit each case.

There is a specific mathematical relationship between the correlation coefficient (r) between two variables (x and y) and the coefficient of determination for the least-squares model that predicts the response variable's value based on a single explanatory variable (x). For those cases, coefficient of determination = r^2 = (correlation coefficient)^2. For this reason, the coefficient of determination is often called r^2.

(24) Since the coefficient of determination has the same value as the square of the correlation coefficient (r) for two variable cases such as those discussed in this lesson, do you think that datasets with strong correlation values (values of r that are closer to –1 or 1) yield LSR models that have a high r^2-value (close to 1) or a low r^2-value (close to 0)? Check your previous work and the previous scatterplots if needed. Do you think that datasets with strong correlation values (values of r that are closer to –1 or 1) yield LSR models that explain a great deal of the variability in y or not much of the variability in y? Why does that make sense visually in terms of residuals?

Initiating Lesson 3.3.1: Using Residuals to Determine If a Line Is a Good Fit

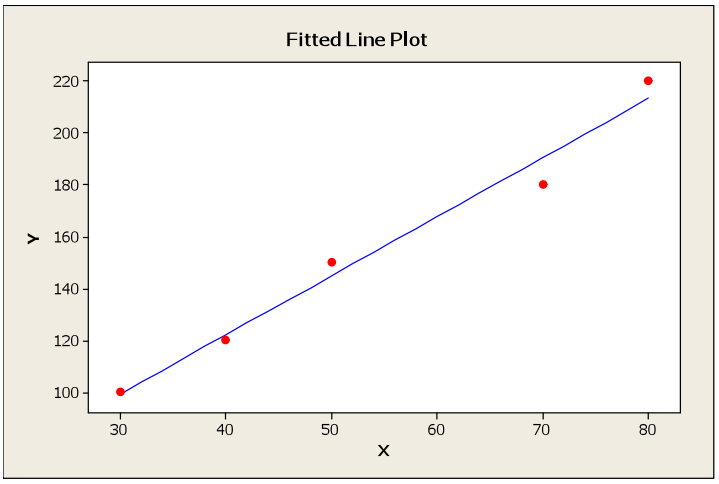
Homework

(1) Verify that the coefficient of determination value (r^2) that you computed for the LSR line in Dataset A (Question 23 in Part II) is equal to the square of the correlation coefficient (r) for that dataset.

(2) A new dataset and its corresponding scatterplot with the LSR line for predicting y from x are shown below.

New Dataset

x	y
30	100
40	120
50	150
70	180
80	220



(a) From visual inspection, compared to the four datasets previously discussed, where do you think this regression model ranks in terms of its usefulness in predicting y from x for its dataset? (See Question 20 in Part I.) Do you put it near the top of the list, near the bottom, or somewhere in the middle? Why? Do you think that this model has a particularly good r^2-value? A particularly good s_e-value? Explain your reasoning.

(b) Compute the LSR line for this dataset. What do you predict for y when $x = 45$? Compute the r^2-value and s_e-value for the regression model. Do these measurements support your ranking for this model in Question 2a?

(c) Based on your visual inspection in Question 2a and your computations in Question 2b, what is good about this model that makes it useful for predicting y from x? What cause for concern do you have (if any) regarding the model's usefulness in predicting y from x?

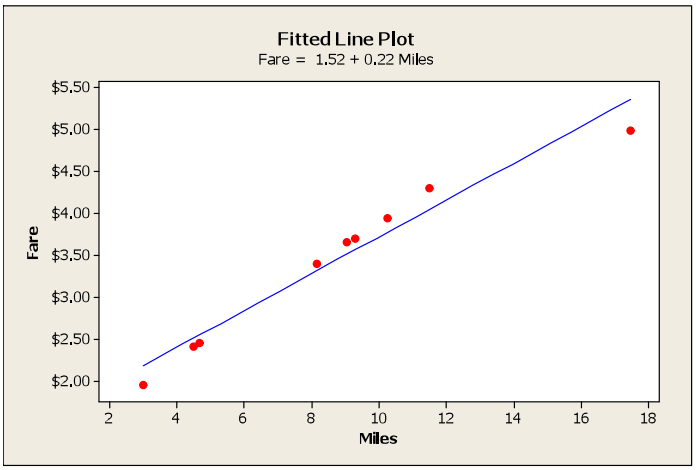
(3) The Metro is the subway rail service used for Washington, D.C., and its immediate suburbs. When using this service, a rider must pay a fare that is relative to the distance traveled from Starting Point A to Ending Point B. For example, if a passenger boarded a subway train at a given station and traveled 10 miles, he or she pays a greater fare than if he or she had only traveled 7 miles. The further you travel from a given starting point, the more you generally have to pay.

The following data show the miles traveled and the standard nonpeak (reduced) fare amount needed for travel from the Metro Center station stop to nine other Metro stations.[1]

Station	Miles	Fare
Pentagon	2.98	$1.95
Virginia Square-GMU	4.47	$2.40
Congress Heights	4.66	$2.45
Medical Center	8.15	$3.40
Branch Ave	9.02	$3.65
West Falls Church-VT/UVA	9.29	$3.70
New Carrollton	10.23	$3.95
Greenbelt	11.49	$4.30
Shady Grove	17.44	$5.00

[1]Retrieved from www.wmata.com/rail/station_detail.cfm?station_id=1.

Initiating Lesson 3.3.1: Using Residuals to Determine If a Line Is a Good Fit



Predicting fare (y) based on miles traveled (x), the LSR model is $\hat{y} = 1.52 + 0.22x$.

(a) In the context of the data presented, what does the slope value of 0.22 estimate? Use words such as *dollars*, *cents*, *miles*, and *fare* in your description.

(b) What is the residual for the point that represents the fare to the Greenbelt station? Does this residual value mean that a rider pays more or less than the model predicts for that trip?

Initiating Lesson 3.3.1: Using Residuals to Determine If a Line Is a Good Fit

(c) The value of s_e for the regression model is 0.220715. (**Note:** It is only coincidental that this statistic's value is close to the value of the slope in the LSR equation.) How do you interpret this s_e-value in the context of this equation? What are the units of s_e? Explain what this s_e-value implies in terms of the quality of your predictions using this model.

(d) The r^2-value for the regression model is 95.6%. How do you specifically interpret that value in the context of this equation? Explain what this r^2-value implies in terms of the quality of your predictions using this model.

Supporting Lesson 3.3.2: Using Residuals to Determine If a Line Is an Appropriate Model

Estimated number of 50-minute class sessions: 1

Learning Goals

Students will understand that

- residual plots can determine
    - if the model is appropriate for the data. (The residual plot has a seemingly random scatter.)
    - if another model is necessary. (A pattern exists in the residual plot. This pattern can indicate a better choice for model.)
    - if an observation exerts a high amount of "leverage" or influence on the slope and intercept estimates.
- no matter how high the r^2-value is or how low the s_e-value is, the values of r^2 and s_e alone are not sufficient for assessing the fit of a model. If the residual plot implies that the least-squares regression model developed is *not* an appropriate choice for the given dataset (as presented), another type of model should be considered and/or observations with strong influence may need to be removed (if justified).

Students will be able to construct a residual plot and use it to assess the appropriateness of employing a linear model for describing the relationship between the response variable and the predictor variable.

Introduction

Students are provided with three types of bivariate data cases:

- where it is fairly obvious from the scatterplot that a linear model is appropriate,
- where there may be some concern from the scatterplot that a linear model is *not* appropriate, and
- where it appears from the scatterplot at first glance that a linear model is appropriate, but in fact it is not appropriate (as will be determined later by further investigation via residual plots).

Students are first asked to assess if a linear model is appropriate based on the scatterplot (y versus x) alone.

Students develop one residual plot via a step-by-step approach with specific calculated values. Following this task, prepared residual plots are presented (i.e., students do not have to make them) for the regression models that students have seen previously, and students' previous conjectures are examined and discussed. The intent is for students to understand that a residual plot is an important and necessary way of assessing the appropriateness of a modeling approach (i.e., viewing the scatterplot alone may not suffice) and that a pattern in a residual plot is an indicator that the least-squares regression model developed is *not* an appropriate choice for the given dataset (as presented), no matter how high the r^2-value is or how low the s_e value is. Analysis of the residual plot can help determine if another type of model should be considered and/or if observations with strong influence should be removed (if justified).

Supporting Lesson 3.3.2: Using Residuals to Determine If a Line Is an Appropriate Model

The Homework activities include a complete task where students are provided with data from a previous lesson in which a linear model was employed (with a high r^2-value and low s_e-value), but in this case, they are asked to assess the appropriateness of that linear model via a residual plot analysis.

Tasks [Student Handout]

In Lesson 3.3.1, you developed measurements to assess how useful a least-squares regression (LSR) line was as a prediction model for a given bivariate dataset. However, in addition to assessing how effective a linear model is in predicting *y* from *x*, you should also examine if the use of a linear model is even a good idea in the first place. In fact, determining if a linear model is *appropriate* is actually more important than assessing its *usefulness*. Several characteristics of a bivariate dataset can make a general linear model (such as the ones you used in previous lessons) an inappropriate choice of model. A few of these problematic characteristics are as follows:

- The overall trend (or form) of the data is not linear (e.g., a curved pattern is present between *x* and *y*, or a *distinctly* cyclical pattern is shown where *y* systematically goes up and down as *x* increases or decreases).
- The distance between the observations and the linear regression line (i.e., the size of the residuals) systematically increases or decreases as *x* increases or decreases.
- An observation has an unusually large positive or negative residual value and/or exerts an unusually large amount of influence on the slope and *y*-intercept calculations of the regression line.

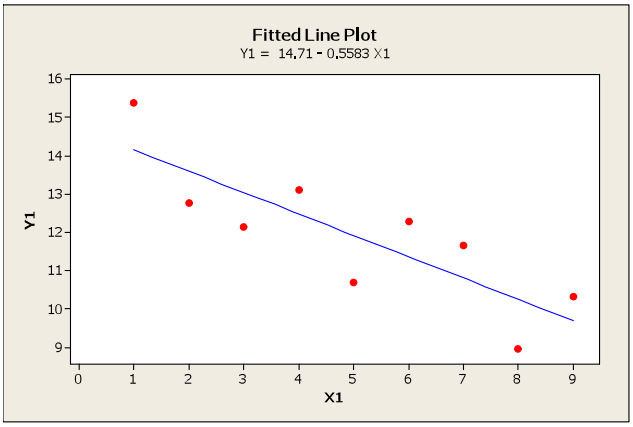
When characteristics such as these are present, they are often noticeable in a scatterplot of the response variable (*y*) and the explanatory variable (*x*).

Five datasets with their corresponding scatterplots and the LSR line for predicting *y* from *x* are provided.

(1) Based on the scatterplots for each of the following five models, do you think that the LSR line (included on each scatterplot) is an *appropriate* model for predicting *y* from *x*? For each scatterplot, record your decision and comment on what characteristics of the scatterplot led to your decision.

Dataset #1

X1	Y1
1	15.38
2	12.76
3	12.14
4	13.12
5	10.7
6	12.28
7	11.66
8	8.94
9	10.32

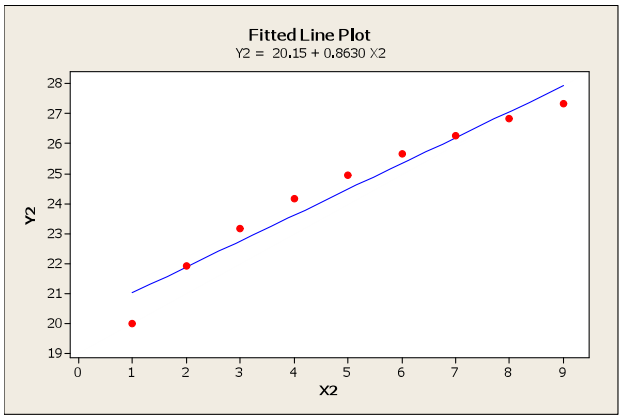


(a) For Dataset 1, it appears from visual inspection of the scatterplot that a linear model is (*circle one*: appropriate/inappropriate) because…

Supporting Lesson 3.3.2: Using Residuals to Determine If a Line Is an Appropriate Model

Dataset #2

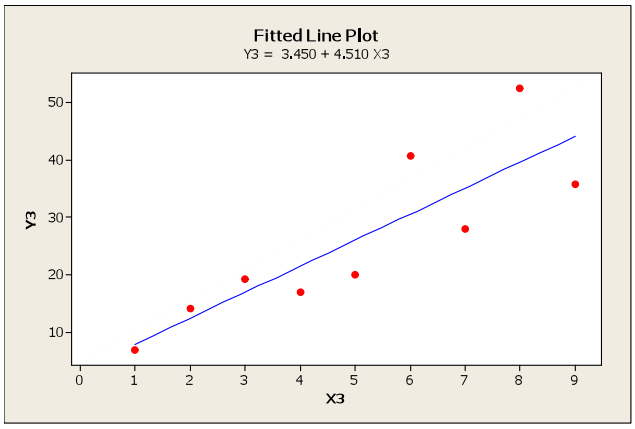
X2	Y2
1	20.00
2	21.89
3	23.16
4	24.14
5	24.95
6	25.65
7	26.27
8	26.82
9	27.32



(b) For Dataset 2, it appears from visual inspection of the scatterplot that a linear model is (*circle one*: appropriate/inappropriate) because…

Dataset #3

X3	Y3
1	7.0
2	14.1
3	19.3
4	16.9
5	20.0
6	40.7
7	28.0
8	52.3
9	35.7

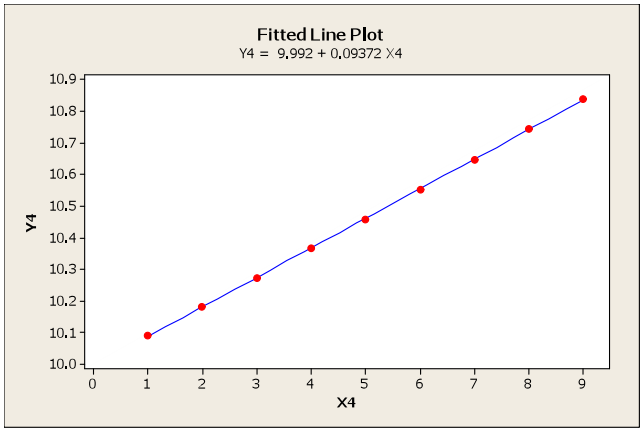


(c) For Dataset 3, it appears from visual inspection of the scatterplot that a linear model is (*circle one*: appropriate/inappropriate) because…

Supporting Lesson 3.3.2: Using Residuals to Determine If a Line Is an Appropriate Model

Dataset #4

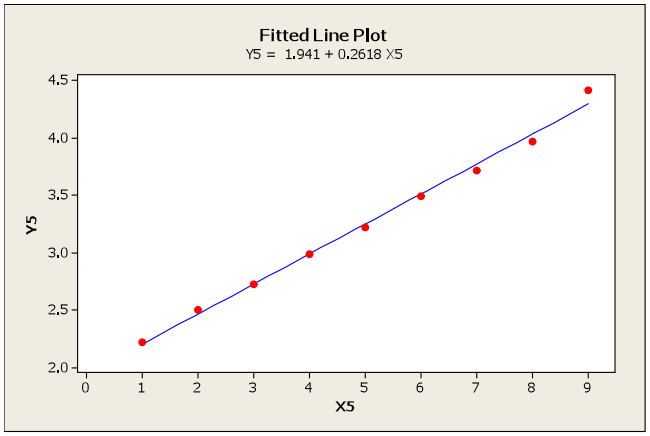
X4	Y4
1	10.09
2	10.18
3	10.27
4	10.36
5	10.46
6	10.55
7	10.65
8	10.74
9	10.84



(d) For Dataset 4, it appears from visual inspection of the scatterplot that a linear model is (*circle one*: appropriate/inappropriate) because…

Dataset #5

X5	Y5
1	2.22
2	2.50
3	2.72
4	2.99
5	3.22
6	3.49
7	3.72
8	3.97
9	4.42



(e) For Dataset 5, it appears from visual inspection of the scatterplot that a linear model is (*circle one*: appropriate/inappropriate) because…

Wrap-Up/Instructor Note

At this point, consider polling the class as to its conjectures regarding the appropriateness of a linear model for each dataset. In addition to the count of *appropriate* and *inappropriate* for each dataset, discuss and record the students' observations as to why they came to these decisions. See if any students wish to change their decisions based on the shared discussion.

Supporting Lesson 3.3.2: Using Residuals to Determine If a Line Is an Appropriate Model

Residual Plots [Student Handout]

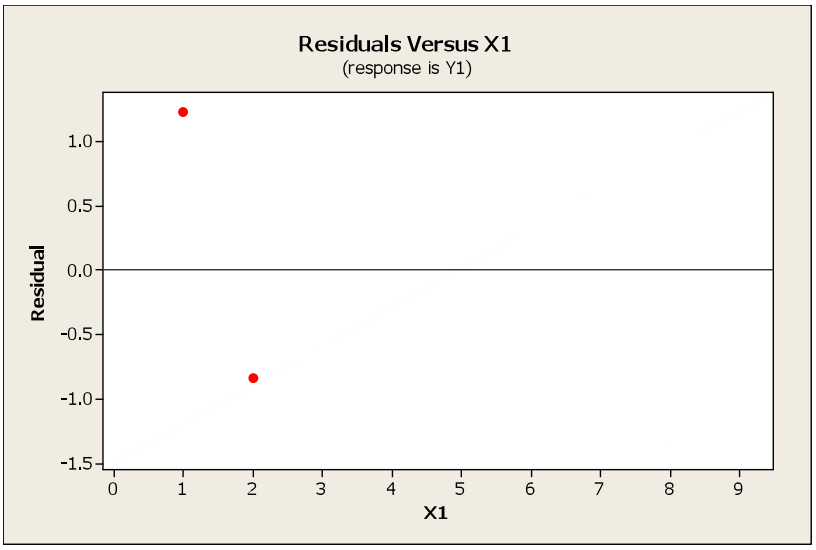
A residual plot is a special scatterplot that displays the relationship between a residual's value and the *x*-value of that residual's corresponding observation. The plot is a highly useful tool for determining if an LSR line model is appropriate for a bivariate dataset.

(2) Compute the residual value for each observation in Dataset 1 (see Question 1) based on the LSR line model developed for the dataset.

Dataset 1

X1	Y1	$\hat{y} = 14.71 - 0.5583x$	Residual $(y - \hat{y})$
1	15.38	14.1517	1.23
2	12.76	13.5934	–0.83
3	12.14	13.0351	
4	13.12	12.4768	
5	10.70	11.9185	
6	12.28	11.3602	
7	11.66	10.8019	
8	8.94	10.2436	
9	10.32	9.6853	

(3) Using the scatterplot template, plot each residual value with its corresponding coordinate's *x*-value. The first two residual values have already been included on the plot.



Supporting Lesson 3.3.2: Using Residuals to Determine If a Line Is an Appropriate Model

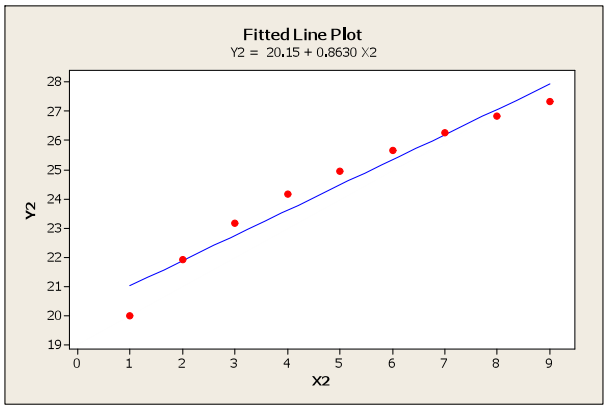
(4) What is similar between the residual plot (residuals versus *x*) for Dataset 1 and the original scatterplot (*y* versus *x*) for Dataset 1? What is different? Does the residual plot visually correspond to the scatterplot? (i.e., Is this a reasonable looking residual plot based on the original scatterplot?) Does there appear to be a pattern in the residual plot, and/or are there any unusual residual values of concern?

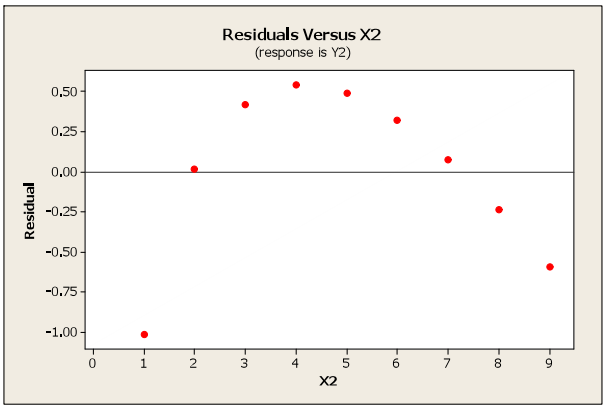
Residual Plot: "Pattern = Problem" [Student Handout]

As seen in the previous example, when a regression model is an appropriate model for a bivariate dataset, the residual plot should display a seemingly pattern-free and random scatter of the residuals. In other words, when a regression model is an appropriate model, there is no pattern in the residual plot.

When a pattern is observed in a residual plot for general linear regression models, it may imply that a problematic characteristic is present in the original dataset that renders a linear model inappropriate. In general terms, if there is a pattern in a residual plot, it means that there is some systematic behavior in your prediction errors (residuals) indicating that another model and/or the removal of an unusual observation may be warranted.

The scatterplot (*y* versus *x*) and the residual plot (residual versus *x*) are shown below for Dataset 2.

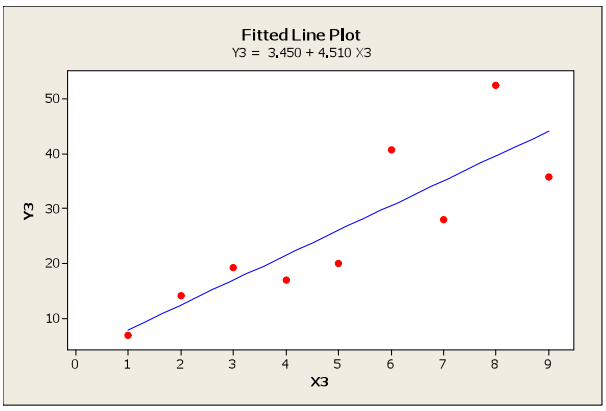


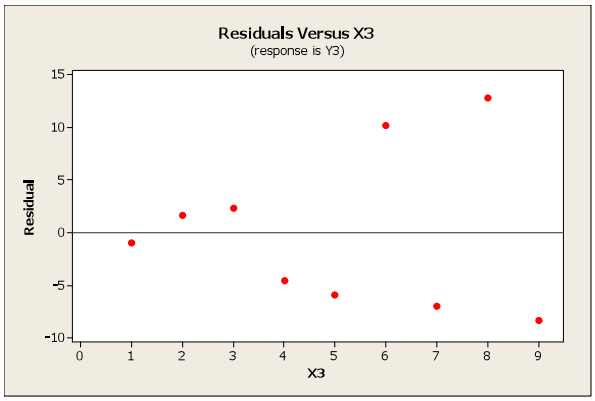


(5) For Dataset 2, does the residual plot visually correspond to the scatterplot? (i.e., Is this a reasonable looking residual plot based on the original scatterplot?) Does there appear to be a pattern in the residual plot, and/or are there any unusual residual values of concern? Does your analysis of the residual plot support your earlier conjecture as to whether a linear model is appropriate?

Supporting Lesson 3.3.2: Using Residuals to Determine If a Line Is an Appropriate Model

The scatterplot (*y* versus *x*) and the residual plot (residual versus *x*) are shown below for Dataset 3.



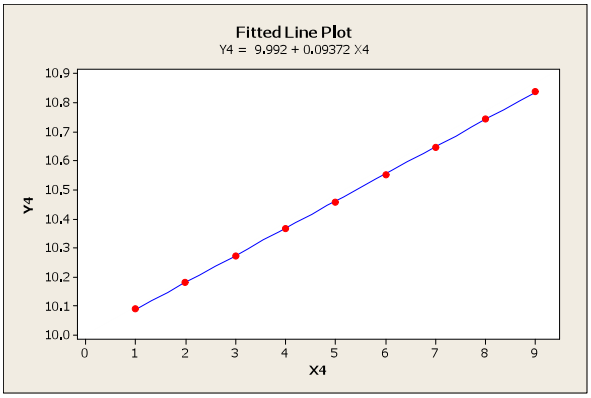


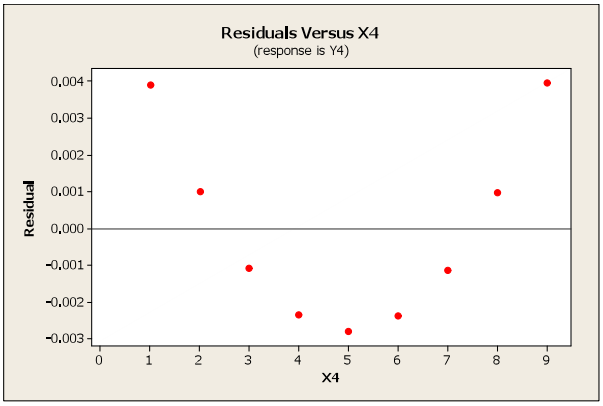
(6) For Dataset 3, does the residual plot visually correspond to the scatterplot? (i.e., Is this a reasonable looking residual plot based on the original scatterplot?) Does there appear to be a pattern in the residual plot, and/or are there any unusual residual values of concern? Does your analysis of the residual plot support your earlier conjecture as to whether a linear model is appropriate?

Perhaps based on the scatterplots for Datasets 2 and 3, you previously suspected that a linear model was *not* appropriate. If so, you may wonder "why bother" with residual plots—as in some cases, with careful examination, the original scatterplot for a bivariate dataset can be somewhat effective in terms of assessing if there is a problematic characteristic.

However, in some cases, a problematic characteristic that renders a standard linear model inappropriate for a bivariate data set can be difficult to see in an original scatterplot and is *much easier* to see in a residual plot.

The scatterplot (*y* versus *x*) and the residual plot (residual versus *x*) are shown below for Dataset 4.

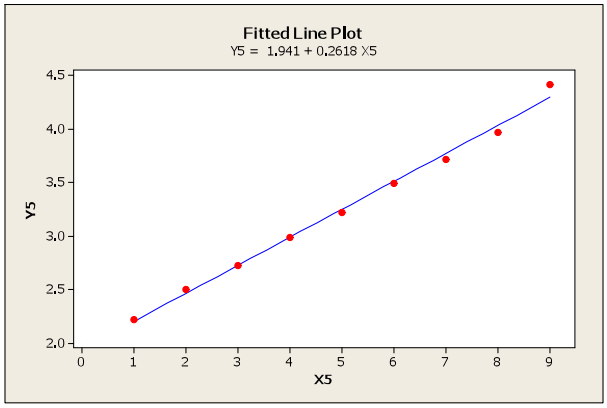


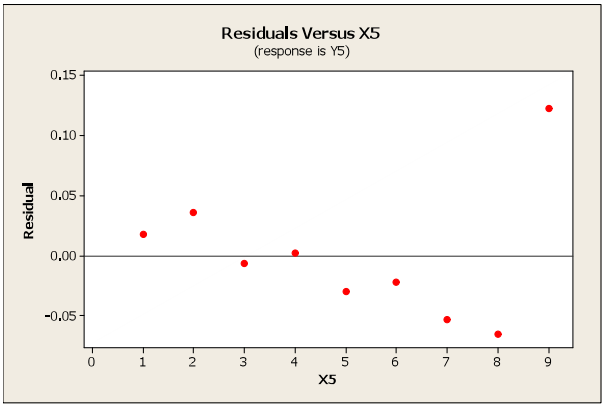


Supporting Lesson 3.3.2: Using Residuals to Determine If a Line Is an Appropriate Model

(7) For Dataset 4, does the residual plot visually correspond to the scatterplot? (i.e., Is this a reasonable looking residual plot based on the original scatterplot?) Does there appear to be a pattern in the residual plot, and/or are there any unusual residual values of concern? Does the analysis of the residual plot support your earlier conjecture as to whether a linear model is appropriate?

The scatterplot (*y* versus *x*) and the residual plot (residual versus *x*) are shown below for Dataset 5.





(8) For Dataset 5, does the residual plot visually correspond to the scatterplot? (i.e., Is this a reasonable looking residual plot based on the original scatterplot?) Does there appear to be a pattern in the residual plot, and/or are there any unusual residual values of concern? Does the analysis of the residual plot support your earlier conjecture as to whether a linear model is appropriate?.

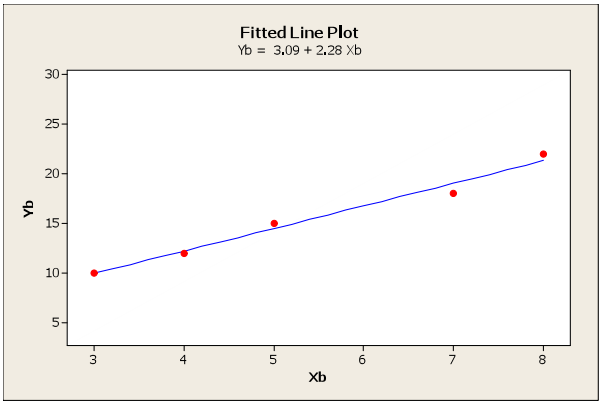
The values of r^2 and s_e alone (covered in the previous lesson) are not sufficient for assessing the fit of the linear model. A pattern in a residual plot implies that the regression model developed is *not* an appropriate choice for the given dataset (as presented), no matter how high the r^2-value is or how low the s_e-value is.

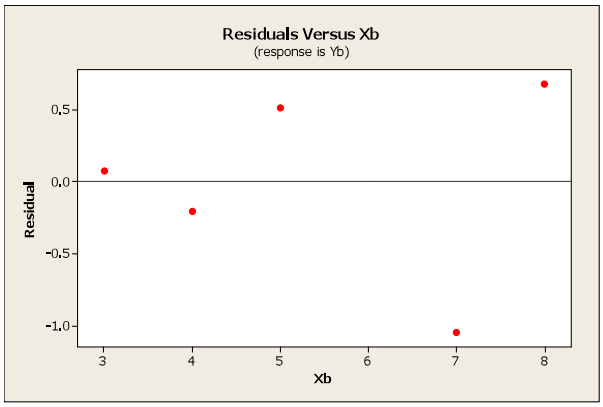
Important Note [Student Handout]

When residual plots are developed from a regression based on very few datapoints, the idea of discerning a pattern in the plot can be difficult or even subjective.

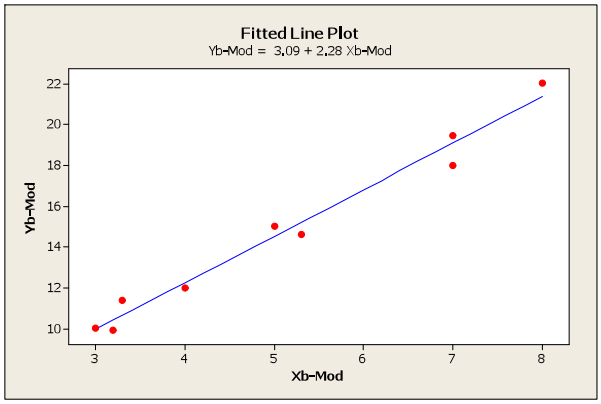
For example, in the previous lesson, the regression line for Dataset B was considered to be a good model for predicting *y* from *x* due to its high r^2-value and low s_e-value. However, the lesson did not ask you to develop or examine a residual plot. In the following residual plot for Dataset B, you could argue that the plot shows a wavy or up-and-down pattern to the residuals and that the residuals' sizes are increasing as *x* increases—two potential problematic characteristics.

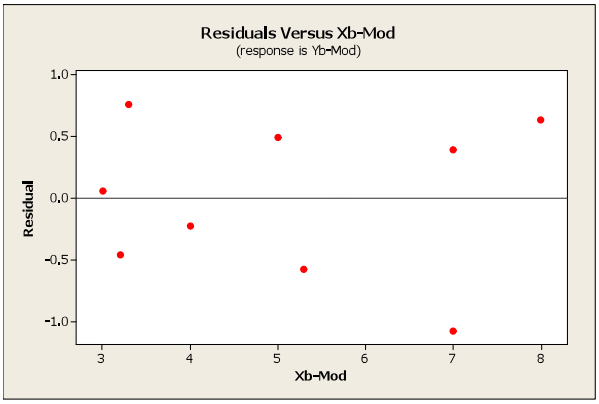
Supporting Lesson 3.3.2: Using Residuals to Determine If a Line Is an Appropriate Model





On the other hand, with a few more datapoints, residuals from the same regression model do not seem to imply any patterns or problems.





The moral: When you use a very small number of observations, the visual evidence of a pattern (or lack of pattern) in the residual plot may not be as definitive as it can be with larger datasets. Generally, when you are trying to determine if a linear regression model is appropriate based on a residual plot that is developed from a small number of observations, use caution and carefully communicate this caution in your analysis.

Wrap-Up Questions/Direct Instruction About Statistical Concepts

Via discussion or lecture, highlight the following:

- A residual plot (a scatterplot of residuals versus *x*) can indicate when a model does not fit the data well.
  - If the model is appropriate for the data, the residual plot has a seemingly random scatter.
  - When the model is *not* appropriate for the data as presented, the residual plot generally displays some sort of pattern. A pattern in a residual plot is a bad thing!

Supporting Lesson 3.3.2: Using Residuals to Determine If a Line Is an Appropriate Model

- Several characteristics of a bivariate dataset can make a general linear model an inappropriate choice of model. A few of these problematic characteristics are as follows:
    - The overall trend (or form) of the data is not linear (e.g., a curved pattern is present between *x* and *y*, or a *distinctly* cyclical pattern is shown where *y* systematically goes up and down as *x* increases or decreases).
    - The distance between the observations and the linear regression line (i.e., the size of the residuals) systematically increases or decreases as *x* increases or decreases.
    - An observation has an unusually large positive or negative residual value and/or exerts an unusually large amount of influence on the slope and *y*-intercept calculations of the regression line.
- If residuals based on a linear model show a pattern/system in a residual plot, another type of model is needed and/or you need to carefully examine if an individual observation is strongly influencing your least-squares line estimates of the slope and *y*-intercept.
- In some cases, a problematic characteristic that makes a standard linear model inappropriate for a bivariate data set is *much easier* to see in a residual plot than in a scatterplot of *y* on *x*.
- The values of r^2 and s_e alone (covered in Lesson 3.3.1) are not sufficient for assessing the *appropriateness* of the linear model. A pattern in a residual plot implies that an LSR line model is *not* an appropriate choice for the given dataset (as presented), no matter how high the r^2-value is or how low the s_e-value is.

Note: While the analysis of residual plots has only been presented here for cases of one response variable (*y*) and one explanatory variable (*x*), analysis of residual plots is also used for assessing other regression models such as nonlinear and multiple regression models not covered in the Statway course. Note also that in some cases, software generates residual plots comparing residual values to $\hat{y}$-values rather than to *x*-values.

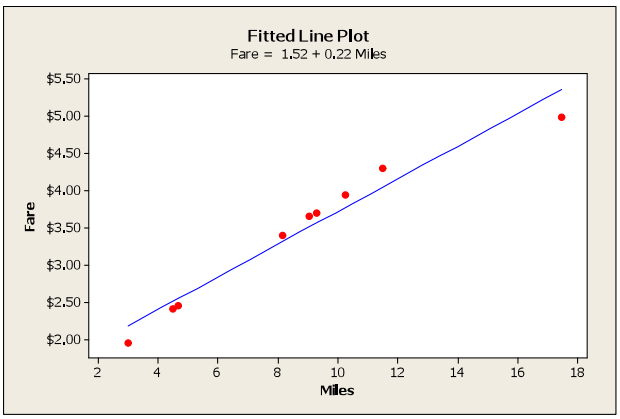
Supporting Lesson 3.3.2: Using Residuals to Determine If a Line Is an Appropriate Model

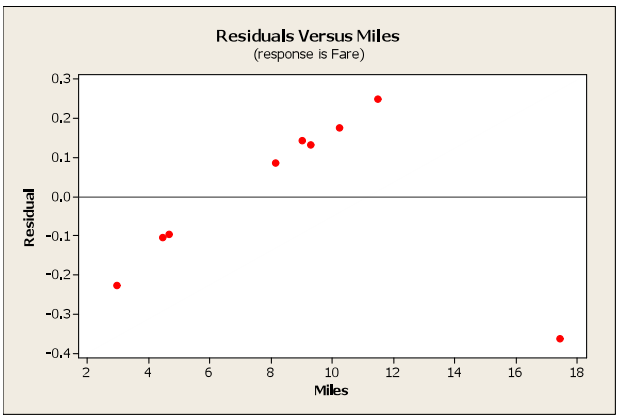
Homework [Student Handout]

In Lesson 3.3.1, the regression line for predicting the Washington, D.C., Metro reduced-rate fare (y) for travel from the Metro Center station stop based on miles traveled (x) was considered to be a good model due to its high r^2-value and relatively low s_e-value. However, the lesson did not ask you to develop or examine a residual plot.

Below are the original dataset and the original scatterplot (with the LSR line included) for the Metro fare dataset. A copy of the residual plot is now also included.

Station	Miles	Fare
Pentagon	2.98	$1.95
Virginia Square-GMU	4.47	$2.40
Congress Heights	4.66	$2.45
Medical Center	8.15	$3.40
Branch Ave	9.02	$3.65
West Falls Church-VT/UVA	9.29	$3.70
New Carrollton	10.23	$3.95
Greenbelt	11.49	$4.30
Shady Grove	17.44	$5.00





(1) What information does the residual plot provide regarding the appropriateness of the LSR line model previously developed?

(2) The Metro system charges a maximum reduced-rate fare of $5 for any trip from Metro Center regardless of distance traveled. This implies that some long-distance trips with a fare of $5 may not adhere to the same linear model as trips of a shorter distance. In addition, the Metro system charges a minimum reduced rate fare of $1.95 for any trip from Metro Center regardless of distance traveled. This implies that certain shorter trips with a fare of $1.95 may not adhere to the same linear model as trips of a slightly longer distance. Knowing this information, it is appropriate to

Supporting Lesson 3.3.2: Using Residuals to Determine If a Line Is an Appropriate Model

remove \$5 trips with particularly long distances and as well as \$1.95 trips with particularly short distances and then to compute another LSR model on the remaining data. Considering this strategy, based on the data and graphs, which station's (or stations') observation(s) do you consider removing from the original dataset?

(3) Compared to the previously developed prediction model (from Lesson 3.3.1), what do you think will happen upon generating a new regression model based on the modification (observation removal) described in Question 2? Do you think the r^2-value and s_e-values will change? If so, how might they change? Do you think the residual plot will improve? Explain your reasoning.

(4) Examine your conjectures from Question 3 by removing any observations listed in your answer to Question 2 and then recalculating the LSR model. What happened to the slope and *y*-intercept estimates, the values of r^2 and s_e, and the residual plot?

Supporting Lesson 3.3.2: Using Residuals to Determine If a Line Is an Appropriate Model

In Lesson 3.3.1, you developed measurements to assess how useful a least-squares regression (LSR) line was as a prediction model for a given bivariate dataset. However, in addition to assessing how effective a linear model is in predicting *y* from *x*, you should also examine if the use of a linear model is even a good idea in the first place. In fact, determining if a linear model is *appropriate* is actually more important than assessing its *usefulness*. Several characteristics of a bivariate dataset can make a general linear model (such as the ones you used in previous lessons) an inappropriate choice of model. A few of these problematic characteristics are as follows:

- The overall trend (or form) of the data is not linear (e.g., a curved pattern is present between *x* and *y*, or a *distinctly* cyclical pattern is shown where *y* systematically goes up and down as *x* increases or decreases).
- The distance between the observations and the linear regression line (i.e., the size of the residuals) systematically increases or decreases as *x* increases or decreases.
- An observation has an unusually large positive or negative residual value and/or exerts an unusually large amount of influence on the slope and *y*-intercept calculations of the regression line.

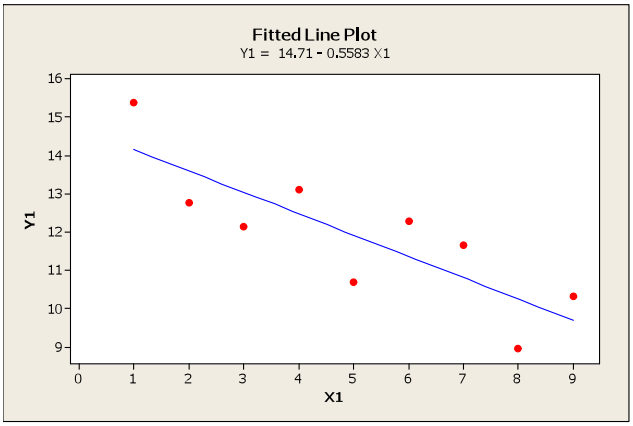
When characteristics such as these are present, they are often noticeable in a scatterplot of the response variable (*y*) and the explanatory variable (*x*).

Five datasets with their corresponding scatterplots and the LSR line for predicting *y* from *x* are provided.

(1) Based on the scatterplots for each of the following five models, do you think that the LSR line (included on each scatterplot) is an *appropriate* model for predicting *y* from *x*? For each scatterplot, record your decision and comment on what characteristics of the scatterplot led to your decision.

Dataset #1

X1	Y1
1	15.38
2	12.76
3	12.14
4	13.12
5	10.7
6	12.28
7	11.66
8	8.94
9	10.32

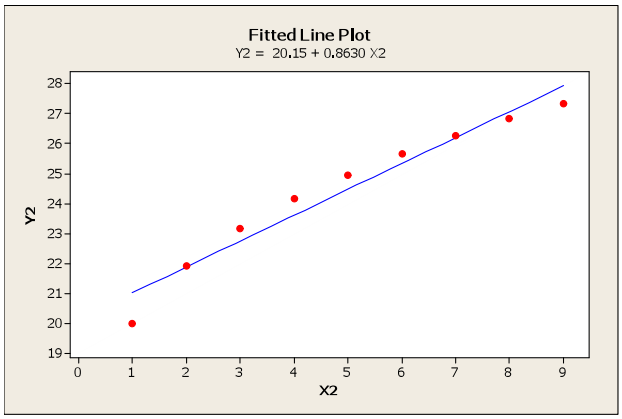


(a) For Dataset 1, it appears from visual inspection of the scatterplot that a linear model is (*circle one*: appropriate/inappropriate) because…

Supporting Lesson 3.3.2: Using Residuals to Determine If a Line Is an Appropriate Model

Dataset #2

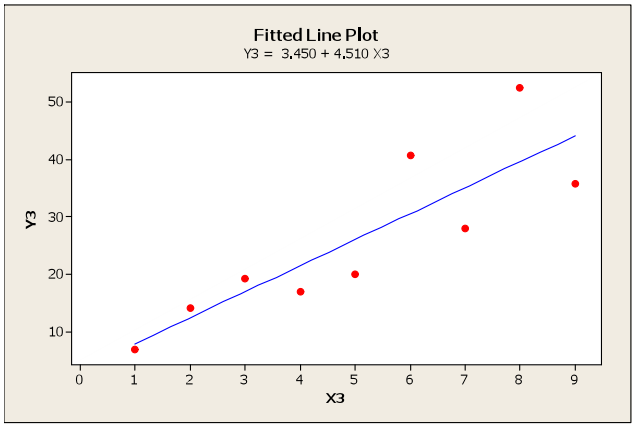
X2	Y2
1	20.00
2	21.89
3	23.16
4	24.14
5	24.95
6	25.65
7	26.27
8	26.82
9	27.32



(b) For Dataset 2, it appears from visual inspection of the scatterplot that a linear model is (*circle one*: appropriate/inappropriate) because…

Dataset #3

X3	Y3
1	7.0
2	14.1
3	19.3
4	16.9
5	20.0
6	40.7
7	28.0
8	52.3
9	35.7

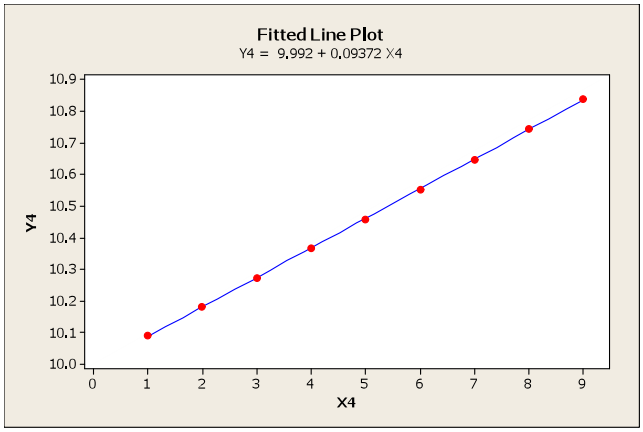


(c) For Dataset 3, it appears from visual inspection of the scatterplot that a linear model is (*circle one*: appropriate/inappropriate) because…

Supporting Lesson 3.3.2: Using Residuals to Determine If a Line Is an Appropriate Model

Dataset #4

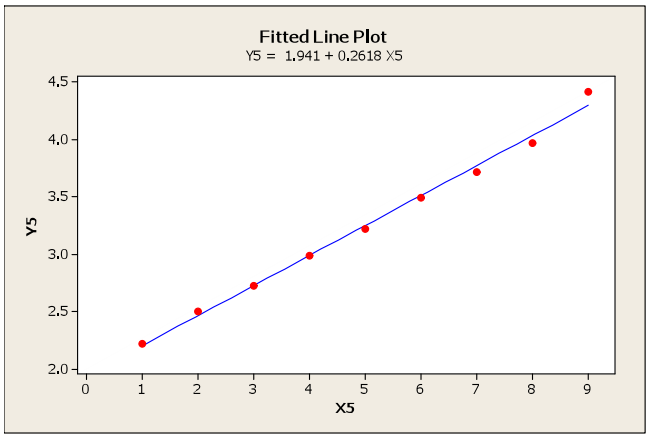
X4	Y4
1	10.09
2	10.18
3	10.27
4	10.36
5	10.46
6	10.55
7	10.65
8	10.74
9	10.84



(d) For Dataset 4, it appears from visual inspection of the scatterplot that a linear model is (*circle one*: appropriate/inappropriate) because…

Dataset #5

X5	Y5
1	2.22
2	2.50
3	2.72
4	2.99
5	3.22
6	3.49
7	3.72
8	3.97
9	4.42



(e) For Dataset 5, it appears from visual inspection of the scatterplot that a linear model is (*circle one*: appropriate/inappropriate) because…

Supporting Lesson 3.3.2: Using Residuals to Determine If a Line Is an Appropriate Model

Residual Plots

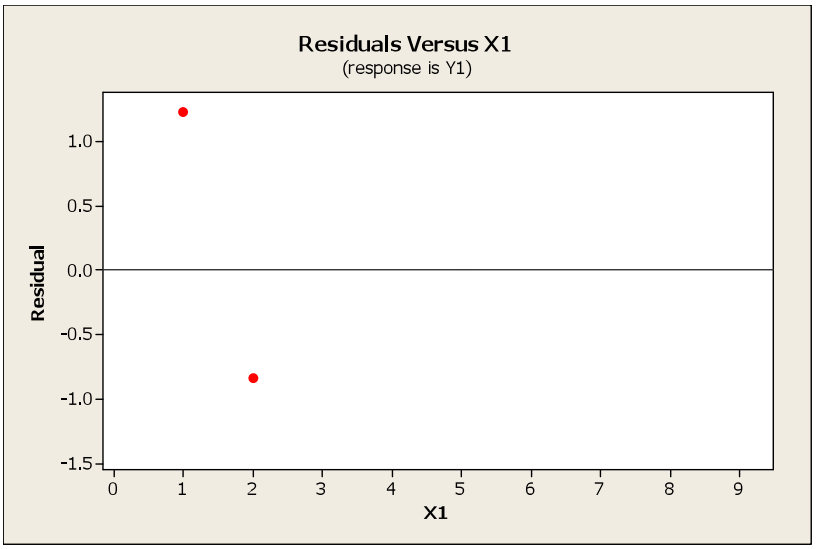
A residual plot is a special scatterplot that displays the relationship between a residual's value and the *x*-value of that residual's corresponding observation. The plot is a highly useful tool for determining if an LSR line model is appropriate for a bivariate dataset.

(2) Compute the residual value for each observation in Dataset 1 (see Question 1) based on the LSR line model developed for the dataset.

Dataset 1

X1	Y1	$\hat{y} = 14.71 - 0.5583x$	Residual $(y - \hat{y})$
1	15.38	14.1517	1.23
2	12.76	13.5934	–0.83
3	12.14	13.0351	
4	13.12	12.4768	
5	10.70	11.9185	
6	12.28	11.3602	
7	11.66	10.8019	
8	8.94	10.2436	
9	10.32	9.6853	

(3) Using the scatterplot template, plot each residual value with its corresponding coordinate's *x*-value. The first two residual values have already been included on the plot.



Supporting Lesson 3.3.2: Using Residuals to Determine If a Line Is an Appropriate Model

(4) What is similar between the residual plot (residuals versus *x*) for Dataset 1 and the original scatterplot (*y* versus *x*) for Dataset 1? What is different? Does the residual plot visually correspond to the scatterplot? (i.e., Is this a reasonable looking residual plot based on the original scatterplot?) Does there appear to be a pattern in the residual plot, and/or are there any unusual residual values of concern?

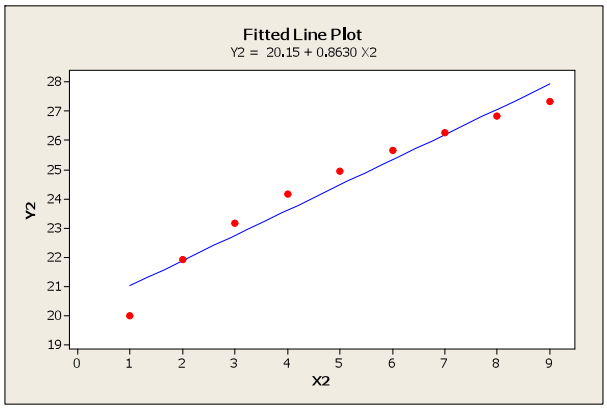
Supporting Lesson 3.3.2: Using Residuals to Determine If a Line Is an Appropriate Model

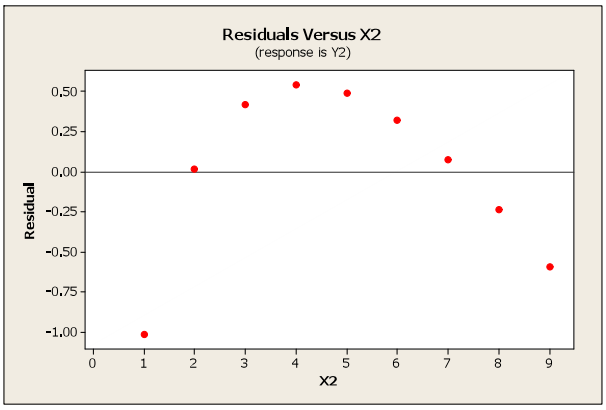
Residual Plot: "Pattern = Problem"

As seen in the previous example, when a regression model is an appropriate model for a bivariate dataset, the residual plot should display a seemingly pattern-free and random scatter of the residuals. In other words, when a regression model is an appropriate model, there is no pattern in the residual plot.

When a pattern is observed in a residual plot for general linear regression models, it may imply that a problematic characteristic is present in the original dataset that renders a linear model inappropriate. In general terms, if there is a pattern in a residual plot, it means that there is some systematic behavior in your prediction errors (residuals) indicating that another model and/or the removal of an unusual observation may be warranted.

The scatterplot (*y* versus *x*) and the residual plot (residual versus *x*) are shown below for Dataset 2.

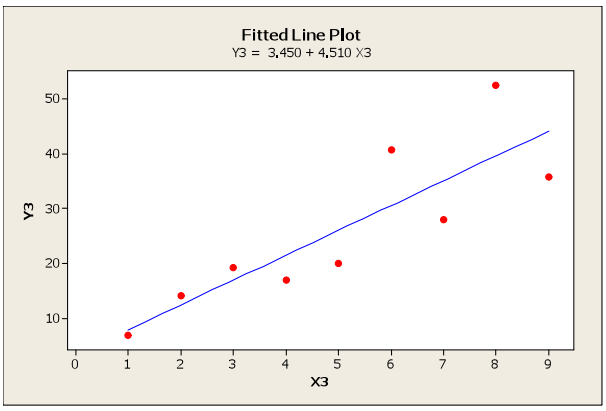


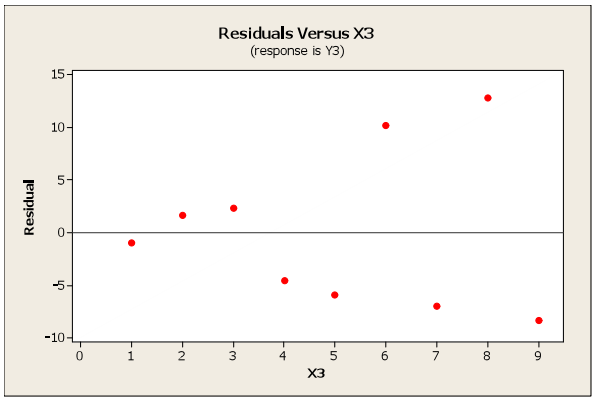


(5) For Dataset 2, does the residual plot visually correspond to the scatterplot? (i.e., Is this a reasonable looking residual plot based on the original scatterplot?) Does there appear to be a pattern in the residual plot, and/or are there any unusual residual values of concern? Does your analysis of the residual plot support your earlier conjecture as to whether a linear model is appropriate?

Supporting Lesson 3.3.2: Using Residuals to Determine If a Line Is an Appropriate Model

The scatterplot (*y* versus *x*) and the residual plot (residual versus *x*) are shown below for Dataset 3.





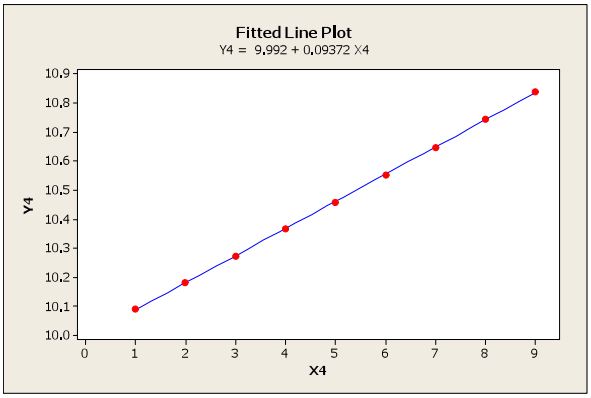
(6) For Dataset 3, does the residual plot visually correspond to the scatterplot? (i.e., Is this a reasonable looking residual plot based on the original scatterplot?) Does there appear to be a pattern in the residual plot, and/or are there any unusual residual values of concern? Does your analysis of the residual plot support your earlier conjecture as to whether a linear model is appropriate?

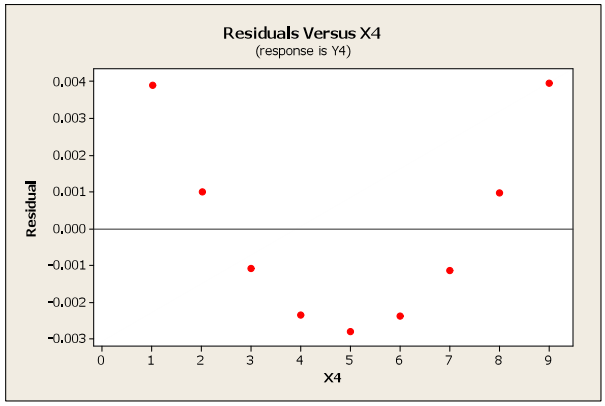
Perhaps based on the scatterplots for Datasets 2 and 3, you previously suspected that a linear model was *not* appropriate. If so, you may wonder "why bother" with residual plots—as in some cases, with careful examination, the original scatterplot for a bivariate dataset can be somewhat effective in terms of assessing if there is a problematic characteristic.

However, in some cases, a problematic characteristic that renders a standard linear model inappropriate for a bivariate data set can be difficult to see in an original scatterplot and is *much easier* to see in a residual plot.

Supporting Lesson 3.3.2: Using Residuals to Determine If a Line Is an Appropriate Model

The scatterplot (*y* versus *x*) and the residual plot (residual versus *x*) are shown below for Dataset 4.

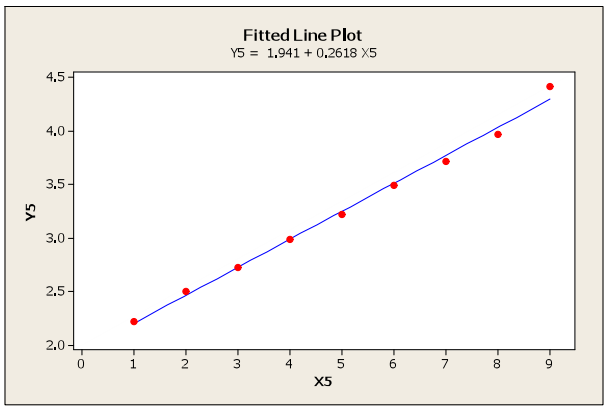


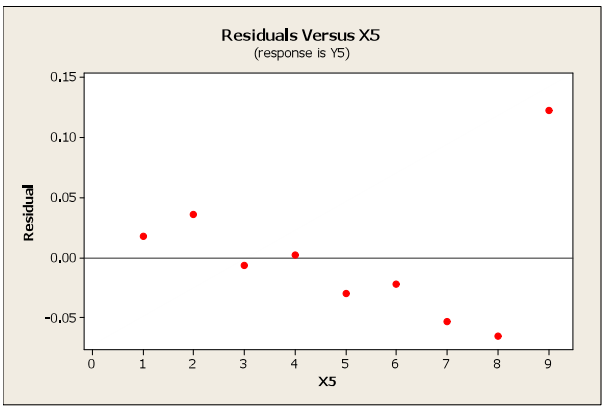


(7) For Dataset 4, does the residual plot visually correspond to the scatterplot? (i.e., Is this a reasonable looking residual plot based on the original scatterplot?) Does there appear to be a pattern in the residual plot, and/or are there any unusual residual values of concern? Does the analysis of the residual plot support your earlier conjecture as to whether a linear model is appropriate?

Supporting Lesson 3.3.2: Using Residuals to Determine If a Line Is an Appropriate Model

The scatterplot (*y* versus *x*) and the residual plot (residual versus *x*) are shown below for Dataset 5.





(8) For Dataset 5, does the residual plot visually correspond to the scatterplot? (i.e., Is this a reasonable looking residual plot based on the original scatterplot?) Does there appear to be a pattern in the residual plot, and/or are there any unusual residual values of concern? Does the analysis of the residual plot support your earlier conjecture as to whether a linear model is appropriate?.

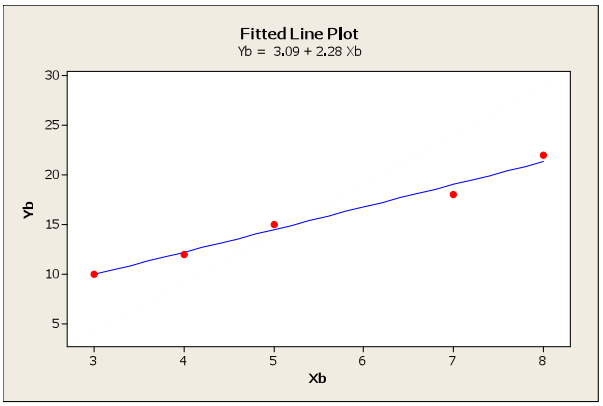
The values of r^2 and s_e alone (covered in the previous lesson) are not sufficient for assessing the fit of the linear model. A pattern in a residual plot implies that the regression model developed is *not* an appropriate choice for the given dataset (as presented), no matter how high the r^2-value is or how low the s_e-value is.

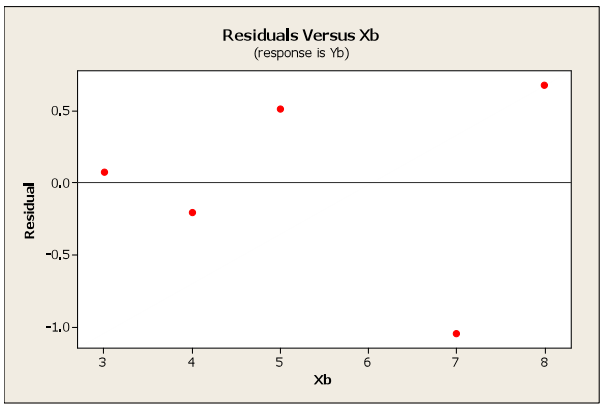
Supporting Lesson 3.3.2: Using Residuals to Determine If a Line Is an Appropriate Model

Important Note

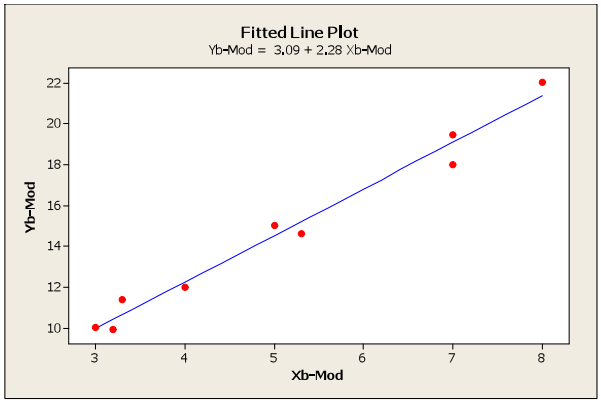
When residual plots are developed from a regression based on very few datapoints, the idea of discerning a pattern in the plot can be difficult or even subjective.

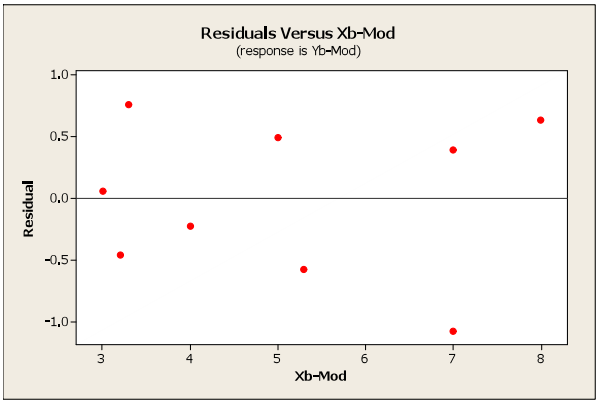
For example, in the previous lesson, the regression line for Dataset B was considered to be a good model for predicting *y* from *x* due to its high r^2-value and low s_e-value. However, the lesson did not ask you to develop or examine a residual plot. In the following residual plot for Dataset B, you could argue that the plot shows a wavy or up-and-down pattern to the residuals and that the residuals' sizes are increasing as *x* increases—two potential problematic characteristics.





On the other hand, with a few more datapoints, residuals from the same regression model do not seem to imply any patterns or problems.





The moral: When you use a very small number of observations, the visual evidence of a pattern (or lack of pattern) in the residual plot may not be as definitive as it can be with larger datasets. Generally, when you are trying to determine if a linear regression model is appropriate based on a residual plot that is developed from a small number of observations, use caution and carefully communicate this caution in your analysis.

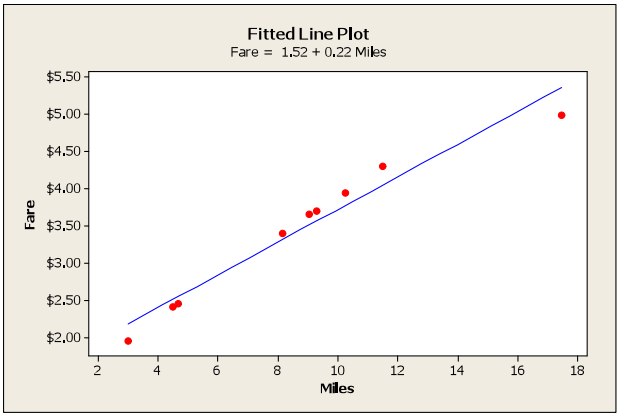
Supporting Lesson 3.3.2: Using Residuals to Determine If a Line Is an Appropriate Model

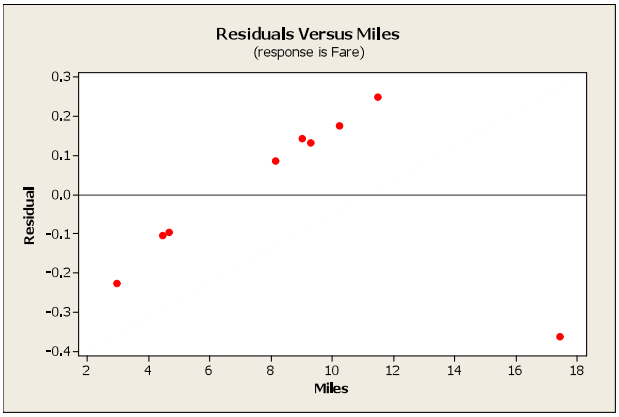
Homework

In Lesson 3.3.1, the regression line for predicting the Washington, D.C., Metro reduced-rate fare (y) for travel from the Metro Center station stop based on miles traveled (x) was considered to be a good model due to its high r^2-value and relatively low s_e-value. However, the lesson did not ask you to develop or examine a residual plot.

Below are the original dataset and the original scatterplot (with the LSR line included) for the Metro fare dataset. A copy of the residual plot is now also included.

Station	Miles	Fare
Pentagon	2.98	$1.95
Virginia Square-GMU	4.47	$2.40
Congress Heights	4.66	$2.45
Medical Center	8.15	$3.40
Branch Ave	9.02	$3.65
West Falls Church-VT/UVA	9.29	$3.70
New Carrollton	10.23	$3.95
Greenbelt	11.49	$4.30
Shady Grove	17.44	$5.00





(1) What information does the residual plot provide regarding the appropriateness of the LSR line model previously developed?

Supporting Lesson 3.3.2: Using Residuals to Determine If a Line Is an Appropriate Model

(2) The Metro system charges a maximum reduced-rate fare of $5 for any trip from Metro Center regardless of distance traveled. This implies that some long-distance trips with a fare of $5 may not adhere to the same linear model as trips of a shorter distance. In addition, the Metro system charges a minimum reduced rate fare of $1.95 for any trip from Metro Center regardless of distance traveled. This implies that certain shorter trips with a fare of $1.95 may not adhere to the same linear model as trips of a slightly longer distance. Knowing this information, it is appropriate to remove $5 trips with particularly long distances and as well as $1.95 trips with particularly short distances and then to compute another LSR model on the remaining data. Considering this strategy, based on the data and graphs, which station's (or stations') observation(s) do you consider removing from the original dataset?

(3) Compared to the previously developed prediction model (from Lesson 3.3.1), what do you think will happen upon generating a new regression model based on the modification (observation removal) described in Question 2? Do you think the r^2-value and s_e-values will change? If so, how might they change? Do you think the residual plot will improve? Explain your reasoning.

(4) Examine your conjectures from Question 3 by removing any observations listed in your answer to Question 2 and then recalculating the LSR model. What happened to the slope and *y*-intercept estimates, the values of r^2 and s_e, and the residual plot?