Building Concepts: Modeling Linear Relationships

TEACHER NOTES

Lesson Overview

In this TI-Nspire lesson, students investigate patterns of association in bivariate data looking for direction, strength, and form of the relationship between the two variables.

Linear equations can be used to model the relationship between two variables whose scatter plot shows a linear pattern.

Learning Goals

1. Identify an association between two quantitative variables as positive, negative, strong, weak or none and interpret this association in terms of a given context;
2. find and interpret linear equations to model relationships between two quantitative variables;
3. understand and be able to calculate residuals using a model for the relationship between two quantitative variables;
4. assess the model fit by using the sum of the absolute value of the residuals;
5. analyze a residual plot to determine whether a model is actually a good fit for the relationship between two quantitative variables.

Prerequisite Knowledge

Modeling Linear Relationships is the twenty-third lesson in a series of lessons that explore the concepts of statistics and probability. In this lesson students investigate patterns of association in bivariate data. This lesson builds on the concepts of the previous lessons. Prior to working on this lesson students should have completed *Scatter Plots*. Students should understand how to:

- interpret points in a scatter plot;
- find the equation of a line;
- find and interpret rate of change.

Vocabulary

- **association:** a relationship between groups
- **linear:** straight or direct
- **regression:** the line modeling the relationship between two variables
- **rate of change:** describes how one thing changes relative to another
- **residual:** the difference between the observed value for a given input and the value predicted by the line.

Lesson Pacing

This lesson may take three or four days of class time to complete with students, though you may choose to extend, as needed.

Lesson Materials

- Compatible TI Technologies:

  TI-Nspire CX Handhelds, TI-Nspire Apps for iPad®, TI-Nspire Software
- Modeling Linear Relationships_Student.pdf
- Modeling Linear Relationships_Student.doc
- Modeling Linear Relationships.tns
- Modeling Linear Relationships_Teacher Notes
- To download the TI-Nspire activity (TNS file) and Student Activity sheet, go to http://education.ti.com/go/buildingconcepts.

Class Instruction Key

The following question types are included throughout the lesson to assist you in guiding students in their exploration of the concept:

Class Discussion: Use these questions to help students communicate their understanding of the lesson. Encourage students to refer to the TNS activity as they explain their reasoning. Have students listen to your instructions. Look for student answers to reflect an understanding of the concept. Listen for opportunities to address understanding or misconceptions in student answers.

Student Activity: Have students break into small groups and work together to find answers to the student activity questions. Observe students as they work and guide them in addressing the learning goals of each lesson. Have students record their answers on their student activity sheet. Once students have finished, have groups discuss and/or present their findings. The student activity sheet can also be completed as a larger group activity, depending on the technology available in the classroom.

Deeper Dive: These questions are provided for additional student practice and to facilitate a deeper understanding and exploration of the content. Encourage students to explain what they are doing and to share their reasoning.

Mathematical Background

In studying ratios, students have become familiar with proportional relationships, graphing collections of equivalent ratios and noting that the collection lies on a line through the origin. The equation of the line for a proportion is of the form $y = kx$, where k is the rate of change (and tied to a unit rate) as well as the constant of proportionality. In Lesson 22, *Scatter Plots*, students construct a scatter plot, describe the graph in terms of clusters, gaps, and unusual data points (much as in the univariate situation). In this lesson, students investigate patterns of association in bivariate data looking for an overall positive or negative relationship in the points, a linear or nonlinear (curved) pattern, and strong or weak association between the two variables. When scatter plots suggest a linear association or pattern, a line can be drawn through the “center” of the cloud of points to capture the essential nature of the relationship between the variables. How well the line “fits” the cloud of points is judged by how closely the points are packed around the line, considering that one or more outliers might have tremendous influence on the location of the line. This is analogous to finding the mean and MAD for a set of univariate data.

Building from the concept of rate of change developed in their work on ratios, students find the rate of change of a line modeling a linear relationship between two variables, interpret it in terms of the context and write an equation for the line. The line modeling the relationship is called a regression line. A residual or “error” is the difference between the observed value for a given input and the value predicted by the line. The fit of a regression line can be determined by two methods. In one method, the total “error” is summarized by using the sum of the absolute deviations (SAD), making connections to earlier work with univariate data. The SAD can be minimized by making adjustments to slope and intercept, thereby making a line of good fit. A second method to assess fit is a sum of the squared errors (residuals) in using the line to predict a response or dependent value for the given independent values or explanatory values. When the sum of the squared residuals is minimized, the line is called a least squares regression line. This lesson does not use a least squares regression line, leaving its development to later grades. Understanding the least squares regression line calls for an understanding of quadratics, which has not been developed in most typical courses at this point.

A second consideration in determining if a regression line is a good fit is to determine whether a plot of the errors (residuals) vs the corresponding independent values, called a *residual plot*, has a pattern or shows anything predictable. If the spread around the horizontal axis in a residual plot is random, using the line to predict outcomes will not have a systematic error. No pattern will exist in making predictions (e.g., always over predict when the independent variable is small), and the line is said to be a good model for the relationship between the variables.

Rate of change, used here instead of slope, describes how one thing changes relative to another. Rate of change has important practical interpretations for most statistical investigations where the units are typically not the same for the two variables, and students should be able to interpret the rate of change in different contexts. Students should be familiar with different forms of equations of lines and how to find the rate of change given two points or a graph.

Correlation is tied to association between the variables and in this lesson is only referred to as positive or negative to indicate the direction of change, and as strong, weak or none. A key take-away should be that strong association or correlation **never** indicates any cause and effect relationship between two quantitative variables unless they are part of a carefully designed experimental study. Causal attribution can only come when randomization of treatments to subjects is used in the design of the study, which is beyond the focus of this lesson.

The data sets in this lesson are very rich and can take a long time to investigate. Each one makes a different point that is important for students to consider. The first one, Store Prices, connects back to proportional relationships with lines only though the origin and introduces the sum of the absolute deviations as a way to measure a “good” model; the second extends to a general linear equation involving the relationship between mass and bite force; the third, Monopoly, introduces residuals as a way to assess how well a model “fits” the data; and the others introduce negative associations, no associations, and patterns that are not linear.

Resources

Erickson et al., (2012). PLoS ON http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0031781

http://listverse.com/2012/11/05/top-10-animal-bites-that-will-completely-destroy-you/

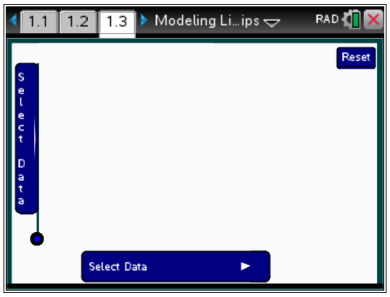
http://news.nationalgeographic.com/news/2012/03/120315-crocodiles-bite-force-erickson-science-plos-one-strongest/

http://en.wikipedia.org/wiki/List_of_U.S._states_by_poverty_rate

http://nces.ed.gov/ccd/tables/ACGR_2010-11_to_2012-13.asp#f2

Part 1, Page 1.3

Focus: Some scatter plots represent relationships between two quantitative variables that can be described by a proportional relationship $(y = kx)$; the "error" in using the equation is the difference between an observed value and the value predicted by the equation.



On page 1.3 a scatter plot for item prices at 2 stores can be created.

***x*-axis** or ***y*-axis** are used to choose a data set for the axes.

Draw Line shows a ray through the origin.

Selecting a point or tabbing will show/hide the item and ordered pair representing the prices at each store.

Move Line activates the arrow keys on the screen and the right/left arrows on the keyboard to move the line

Select a point to show/hide the vertical distance from the point to the line.

All Segments displays the vertical distances for each point.

Show SAD displays the sum of the absolute deviations (vertical distances).

Add Point adds a point at the origin, which can be moved by dragging or using the arrow keys on the screen or keypad.

Tab Key changes whether tab cycles through points on the graph or cycles through the buttons on screen.

Remove point allows a point to be selected and removed.

TI-Nspire Technology Tips

menu accesses page options.

tab cycles through points on the graph or the buttons.

enter selects a highlighted pointed or activates a button.

esc releases selected points and quits adding or removing points.

ctrl del resets the page.

Class Discussion

In these questions, students describe the strength, direction, and form of the relationship between the variables in the scatter plot and use a line through the origin to model the pattern.

Have students…	Look for/Listen for…
Select Store A for the horizontal axis and Store B for the vertical axis. The scatter plot displays the prices for the same brands and sizes of items purchased at two different grocery stores.	
• ***Use tab or the cursor to select a point. What does the point represent?***	Answers will vary. The point in the bottom left corner shows the prices of the least expensive item, frozen peas, which cost \$0.88 at Store A and \$1.00 at Store B.

Class Discussion (continued)

- ***Describe the scatter plot.***

  Answers may vary. The pattern in the plot seems to be linear; the points go in a relatively straight line from the lower left to the upper right. The most expensive item at both stores is paper towels, the point in the upper right, which cost $5.97 at Store A and $7.69 at Store B. A bunch of items cost from $2.50 to $3.00 at Store A and around $3.00 at Store B.

- ***Which of the following sentences describes the relationship between the prices at the two stores?***

  i. ***As the prices in Store A increase, the prices in Store B decrease.***

  ii. ***As the prices in Store A increase, the prices in Store B increase.***

  iii. ***As the prices in Store A decrease, the prices in Store B increase*.**

  Answer: ii

- ***Would you describe the association between the prices at the two stores as positive, negative, strong, weak or no association? Explain your reasoning.***

  Answer: Strong and positive because as the prices at Store A increase, the prices for the corresponding item increase as well and there doesn't appear to be a lot of variation from that pattern.

Select **Draw Line**.

- ***What is the equation of the line and what does a point on the line represent in terms of the prices?***

  Answer: The equation of the line is $B = 1A$. A point on the line would represent an item that cost the same at both stores.

- ***Interpret the rate of change of the line in context of prices of items.***

  Answer: The rate of change of the line is 1, which would mean that for each $1 increase in the cost of an item at Store A, the item would also increase by $1 at Store B.

- ***What does a point above the line represent? Give an example that supports your reasoning.***

  Answers may vary. Points above the line represent items that were less expensive at Store A than Store B. The point for cereal is above the line, and cereal costs $3.52 at Store A and $4.79 at Store B.

Class Discussion (continued)

- ***Will the point for an item that costs $4.00 at Store A and $3.50 at Store B lie above or below the line? Select* menu> Add Point*. Move the point to represent the new item. Does the location agree with your answer?***

  Answer: The point for the item will lie below the line.

- ***Select* menu> Remove Point**. ***Overall, which store seems to have the cheaper prices? Explain your reasoning.***

  Answer: All but one of the points representing the original items lie above $B = 1A$ so the prices at Store B are more expensive than prices for the same item at Store A.

- ***Sami suggests moving the line to have the equation $B = 1.18A + 0.25$. Do you think Sami's line would be a good line to model the prices? Why or why not?***

  Answer: Typically the line would go through the point (0, 0), like a proportion where the 1.18 is the constant of proportionality. For Sami's line, if an item was free at Store A, it would cost $0.25 at Store B, which could happen, but most stores don't give items away for free.

- ***Move the line to find a model you think reflects the pattern in the prices between the two stores. Interpret the rate of change for your line in terms of the store prices.***

  Answers will vary. For the equation $B = \$1.25A$ for every dollar something costs at Store A, it will cost approximately $1.25 at Store B.

- ***Remember studying ratios and proportions. How does this equation relate to a proportional relationship?***

  Answer: A proportional relationship can be graphed as a collection of equivalent ratios. For example, for the line above, the points on the line in the scatter plot belong to all of the ratios equivalent to 1.25:1.58. The rate of change is the constant of proportionality.

In the following questions, students make predictions about the outcome for a given input using a line they think models the linear relationship and compare the predictions from their line to the actual outcomes and confront the need to deal with negative differences.

Use your line from the question above to answer the following.

- ***Predict how much an item that cost $6.00 at Store A will cost at Store B. Explain how you made your prediction.***

  Answers will vary: If the equation of the line was $B = \$1.25A$; when $A = 6$, $B = 7.50$, so the item would cost $7.50.

- ***Use the scatter plot to find how much a box of cereal would cost at Store B, if the box costs $3.52 at Store A.***

  Answers will vary: $4.79

Class Discussion (continued)

- ***Find the difference between the actual price you found in the question above for the cereal at Store B and the price your line predicts for the cost of box of cereal at Store B.***

  Answers will vary: The line predicts $4.40; the difference is $0.39.

Student Activity Questions—Activity 1

1. a. Refer to the graph you created on page 1.3. Select the point representing a box of cereal, then Enter. Explain what the number and vertical segment tell you.

Answer: The number is the difference I found between the actual and predicted cost for the cereal. The vertical segment shows the difference on the graph—the vertical distance from the point to the line.

b. Select the point you think will have the greatest error (difference between the actual and predicted cost). Find the difference and explain what it means.

Answers will vary. Ketchup is the farthest from the line $B = 1.25A$. The difference in cost is about $\$2.19 - \$3.40 = -\$1.21$. The price of Ketchup predicted by the line at Store B is $1.21 cheaper than the actual price at Store B.

c. Fill in the table with the missing values using your line.

Item	**Store A Price $**	**Store B Price $**	**Store B price predicted by line*($)**	**Actual price minus predicted price ($)**
Whole Grain Cereal (10 oz)	3.52	4.79	4.40	$\$4.79 - \$4.40 = \$0.39$
Raisins	1.94	2.99	2.42	$\$2.99 - \$2.42 = \$0.57$
Peanuts	4.98	5.89	6.23	$\$5.89 - \$6.23 = \$0.34$
Ketchup (24 oz)	2.72	2.19	3.40	$\$2.19 - \$3.40 = -\$1.21$

Answers may be slightly different due to rounding.

d. For the items in the table, what is the total of the differences in the actual prices for the items minus the prices predicted by your line? Explain what your answer means in terms of the prices.

Answers will vary. The total cost of the items in Store B is $15.86. The cost predicted by my line is $16.45. The difference is $\$15.86 - \$16.45 = \$0.59$, which says that the total cost of the items predicted by my line is $0.59 more than the actual cost. $\$0.96 - \$1.55 = \$0.59$

Student Activity Questions—Activity 1 (continued)

2. **For an article they were writing, the class wanted to describe how far off the predicted prices were from the actual prices for the items at the two stores.**

   a. **Marj claimed the sum of the differences in actual price minus predicted price for the items in the table was -0.59, or about $\frac{-60}{4}=-15$ cents per item. Petra disagreed and said if you found the mean the way Marj did, you could have a sum of the differences in actual minus predicted prices to be 0, when there really were a lot of "differences". Whose reasoning makes the most sense and why? Give an example to support your thinking.**

   Answer: Petra is correct because if one item was 50 cents more at one store than the other and another item was 50 cents less, the sum of the differences would be 0 $(-0.5+0.5)$. This would make it seem like all of the prices were really the same. You have to worry about the positive and negative signs.

   b. **Hilary suggested finding the sum of the differences by taking the absolute value of the differences and then adding. What do you think about Hilary's reasoning?**

   Answer: Hilary's reasoning makes sense; it is like finding the mean absolute deviation. It accounts for the positive and negative signs and she will get a total amount of deviation.

   c. **Anita said she would find the mean difference by squaring all the differences and finding the sum of the squared differences. What do you think about Anita's reasoning?**

   Answer: Anita's reasoning is a good way to make sure you have taken care of the negative values, but she will have a very large number for the mean because she squared everything (she has squared $ as a unit). She might take the square root of her answer and then it would at least be in the right unit, $.

   d. **Use one of the methods described in b) and c) to find the sum of the differences in prices. Compare your answer with others. What might explain differences?**

   Answers will vary. Those who squared will have a much larger difference than those who used the absolute value.

These questions introduce the sum of the absolute differences (SAD) and provide a reflection about the work. The end task is to find a model (equation) that will make the overall sum of the absolute errors as small as possible. Note that in doing some of the computation to find the error or fill out the table, students might want to use the calculator scratch pad.

3. **SAD is the sum of the absolute value of all of the differences. Select All Segments then Show SAD.**

   a. **How does the value for SAD relate to the scatter plot and line?**

   Answer: Each segment represents the difference between the actual cost of an item at Store B and the cost predicted by my line for a particular item at Store A. The SAD for my line is 7.28 (measured in dollars), which is the sum of the absolute values of the actual costs minus predicted costs for the items.

Student Activity Questions—Activity 1 (continued)

b. **The predicted cost of all of the items at Store B using the equation $B = 1.25A$ is $63.38, and the actual cost was $59.58. How does the difference, actual total cost minus predicted total cost compare to the SAD? Explain any differences.**

Answer: The difference between the two total costs was $-\$3.80$. It is smaller than the SAD, $7.28 because the SAD uses the absolute value of the differences.

c. **Move the line. Describe what happens to the SAD.**

Answer: As the line gets farther from the pattern in the data, lengths of the segments increase and the SAD increases. The closer the line is to modeling the pattern, the shorter the segments and the smaller the SAD.

d. **Find a new line that has a smaller SAD than your original line. Explain what your SAD represents.**

Answers will vary. One line, $B = 1.18A$, had a SAD of 6.33. The sum of the absolute value of the differences, actual price minus predicted price, for all of the items was $6.33.

4. **Identify the following statements as true or false. Be ready to explain your thinking.**

a. **The line $y = x$ represents a relationship where all of the ordered pairs are of the form (*x*, *x*).**

Answer: True. The values for *x* and *y* are equal.

b. **If the equation for a linear *relationship* is $y = 2x$, the *x*-values are twice as large as the *y*-values.**

Answer: False. The *y*-values are twice as large as the *x*-values (or the *x*-values are half as large as the y-values).

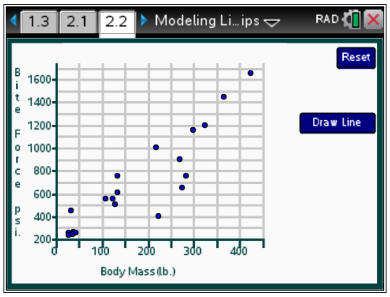
c. **Suppose the equation that models a linear relationship is $B = A$, where A and B are greater than or equal to 0. If the point (*A*, *B*) is below the line, then B will be greater than A.**

Answer: False. B will never be greater than A because below the line $B = A$, B is always less than A.

d. **If the difference, actual price of an item minus the price predicted by a line, is negative, the point representing that item lies below the line.**

Answer: True because to have a negative difference, the vertical coordinate of the actual price has to be smaller (below) the vertical coordinate of the predicted price.

Part 2, Page 2.2



Focus: In general relationships between two variables that have a fairly strong linear association can be modeled by a linear equation, and the fit of the equation can be assessed by using the sum of the absolute values of the residuals.

The commands and behavior of the buttons and menu on page 2.2 are similar to those on page 1.3.

Class Discussion

Teacher Tip: The focus of these questions is investigating the relationship between two variables (weight, bite force), where scientists were trying to predict the bite force of crocodiles and alligators knowing their weight. Students interpret the scatter plot, find a model to represent the linear pattern, and use their model to predict bite force, including extrapolating to a heavier alligator that was not part of the original study. Note that equations that model the relationship will not contain the point (0, 0) but have a *y*-intercept that is not 0. Many real contexts do not make sense outside of a certain domain, and in these cases, the *y*-intercept may not make sense in the context. (And note that the axes are often truncated; in this case, the horizontal axis begins at 0 but the vertical axis starts at 200, a point raised in these questions.)

The scatter plot on page 2.2 represents the bite force in pounds per square inch and the body mass in pounds of 27 crocodiles and alligators from around the world.

- ***Which of these reptiles has the largest body mass and bite force?***

  Answer: Croc E at 465 pounds of body mass and 1650 pounds per square inch (psi) of bite force.

- ***Which of the reptiles seems to be the farthest from the pattern?***

  Answers may vary. The Indian Ghrial (225, 400) seems to have a smaller bite force for its body mass than the others.

- ***Describe the relationship between body mass and bite force you can see from the scatter plot.***

  Answer: The larger the body mass, the larger the bite force for these crocodiles and alligators.

- ***Is the linear relationship strong, mild or weak? Explain your thinking.***

  Answers will vary. The scatter plot looks relatively linear, but the points representing Indian Gharial, Nile and Croc D are a bit off of the main trend line.

Class Discussion (continued)

Select* Draw Line*. The equation of the line drawn to reflect the linear relationship is shown above the graph.

- ***The independent variable is body mass and the dependent variable is pounds per square inch. Explain what this means about predicting bite force.***	Answer: Figuring out the pounds per square inch of bite force for an alligator or crocodile depends on knowing the body mass; you use the body mass to predict the pounds per square inch.
- ***Interpret the rate of change in terms of the crocodiles and alligators.***	Answer: The rate of change is 3.42 or $\frac{3.42}{1}$, which indicates that for each one pound gain in body mass, the bite force increases by 3.42 psi.
- ***Tina says the graph goes through the origin. What would you say to Tina?***	Answer: She is wrong because the point at the lower left where the two axes intersect is not at the origin (0, 0) but at (0, 200).
- ***Suppose a new crocodile was found with a body mass of 175 pounds. Predict the bite force for this crocodile and explain how you found your answer.***	Answers may vary. Some might read the prediction off the graph, while others might use the equation, replacing M with 175, to get a predicted bite force of 798.5 pounds.

Student Activity Questions—Activity 2

1. **Determine the error in using the line to predict the bite force for the Indian Gharial.**

   a. **What does the value tell you?**

   Answer: The –569.5 indicates that the predicted bite force by the line is 569.5 pounds more than the actual bite force for that crocodile.

   b. **Thom says that the length of the segment from this point to the line is larger than all of the other segments together. How could you decide if he is correct?**

   Answer: Determine the SAD and see if it is more than half of the SAD. The SAD is 2858, so the sum of the absolute differences between the predicted and the observed bite force is 2,858 pounds. 569.5 is only about $\frac{1}{4}$ of the overall total, so Thom is incorrect.

   c. **Move the line to minimize the total sum of the absolute distances. What is the equation of your line? Compare your line to others in your class. Who has the smallest SAD?**

   Answers will vary. One equation might be $F = 3.42M + 124$ with $\text{SAD} = 2286.08$ psi.

   d. **What is the mean absolute deviation of the difference in the actual values for a given weight and the values predicted using your model?**

   Answers will vary. There were 27 crocodiles and alligators represented in the plot, so using the equation above, the MAD would be $\frac{2286.08}{27}$, which is about 84.67 psi.

Student Activity Questions—Activity 2 (continued)

2. **Another crocodile, Croc G, with a body mass of 350 pounds had a bite force of 1100 psi.**

   a. **How do you think the SAD will change if you add Croc G to the plot?**

   Answers will vary: Using the equation $F = 3.42M + 124$, the SAD should increase about 221 psi because that is the distance from the psi predicted by the line and the actual psi for Croc G.

   b. **What do you think will happen to the original SAD of 2288.08 psi if a crocodile that weighs 200 pounds and has a bite force of 800 psi is added to the plot?**

   Answer: For the equation $F = 3.42M + 124$, the SAD should remain almost the same because the observed and the predicted values are 800 and 808 respectively.

   c. **If you add a point representing a crocodile that weighs 200 pounds with a bite force of 1000 psi, do you think the SAD will increase, decrease or remain the same?**

   Answer: The SAD should increase as the difference between the absolute value of the observed and predicted is always positive.

3. **In 2012, an American alligator named Hercules was weighed and his bite force measured.**

   a. **Use your equation to predict the bite force given that Hercules weighed 665 pounds.**

   Answers will vary. Using the equation in 1c), the predicted bite force would be about 2,399 pounds per square inch.

   b. **Use the MAD you found in 1d) to find an interval that would typically contain the bite force for Hercules.**

   Answers will vary. Using the equation and work above, the interval would be the predicted bite force +/- the MAD, which would be about 2315 to 2484 psi.

   c. **Hercules actually had a bite force of 2125 psi. How well did your model predict his bite force?**

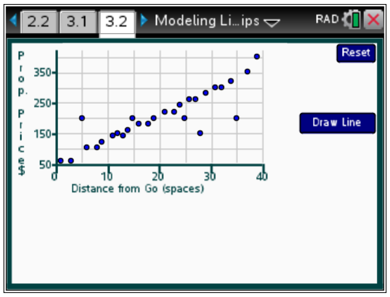
   The model predicted that typically, the bite force would be quite a bit higher; but it is not too unusual to have values outside of the mean +/- MAD when working with a set of data.

   d. **An adult male lion a weighs about 420 pounds and has a bite force of about 600 psi. Do you think it would be reasonable to use your model to estimate the bite force of other animals? Explain why or why not.**

   Answers will vary. The model in the example above would predict 1550 psi for the lion, which is way out of line for the number 600 psi given in the research. It would seem that a model constructed based on information about alligators and crocodiles might not be adequate for other animals with a different body and jaw construction.

Part 3, Page 3.2

Focus: The fit of a model to a pattern in a scatter plot can be informed by investigating the residual plot for any patterns or predictability.



Page 3.2 displays a scatterplot of the value of a property from the board game Monopoly and the number of spaces from Go.

Draw Line displays a residual plot as well as a line. Selecting a point in the plot highlights the corresponding point in the residual plot.

Residuals will show the vertical segments and the residuals or the SAD. The other commands behave like those on pages 1.3 and 2.2.

Class Discussion

The following questions use the game of Monopoly to build on the notion of "error" in predicting from a linear model to formally introduce the idea of residuals (the distance from the actual to the value predicted by a model for a given input). Evidence of patterns or predictability in the residuals provides clues for whether the model is actually appropriate to express the relationship between the two variables. The prices for two types of properties (utilities and railroads) are constant, and students will investigate what happens to the model and to the SAD when those values are removed from the plot. Students may find the Scratchpad on the handheld useful in doing some of the calculations.

Property	Spaces from GO	Cost
Mediterranean Avenue	1	60
Baltic Avenue	3	60
Reading Railroad	5	200
Oriental Avenue	6	100
Vermont Avenue	8	100
Connecticut Avenue	9	120
St. Charles Place	11	140
Electric Company	12	150
States Avenue	13	140
Virginia Avenue	14	160
Penn Railroad	15	200
St. James Place	16	180
Tennessee Avenue	18	180
New York Avenue	19	200

Property	Spaces from GO	Cost
Kentucky Avenue	21	220
Indiana Avenue	23	220
Illinois Avenue	24	240
B & O Railroad	25	200
Atlantic Avenue	26	260
Ventnor Avenue	27	260
Water Works	28	150
Marvin Gardens	29	280
Pacific Avenue	31	300
North Carolina Avenue	32	300
Pennsylvania Avenue	34	320
Short Line Railroad	35	200
Park Place	37	350
Boardwalk	39	400

Class Discussion (continued)

The scatter plot shows the relationship between the distance in spaces from Go and the price of lots in the board game Monopoly.

- ***Look at the scatter plot. Do you notice any patterns or anything interesting?***

  Answer: The overall trend seems to be linear but several points, (5, 200), (25, 200), (28, 250) and (35, 200), do not seem to be part of the linear trend

- ***What is the independent variable? The dependent or response variable? Explain what these mean in the context of the game Monopoly.***

  Answer: The independent (explanatory) variable is the distance from Go, and the dependent (response) variable is the price of the property. Being able to predict the cost of a property depends on knowing the distance of the property from Go.

- ***If a property is located 20 spaces from Go, use the plot to estimate about how much that property would cost.***

  Answer: The cost would probably be between $200 and $220.

Select* Draw Line. *Recall that, the error or the difference between the actual cost and the predicted cost is called the residual. The plot at the bottom of the screen is called a residual plot. A residual plot shows a plot of the residual (error) for each corresponding independent value. The horizontal axis in the plot represents a residual value of zero.

- ***Interpret the slope on the scatterplot in terms of the context. Could the y-intercept have any meaning in the context?***

  Answer: The slope of $\frac{9.25}{1}$ would mean that for every space further from Go, the cost of a property increases by $9.25. The *y* intercept is $50 wouldn't have much meaning in this context.

- ***Select the point (28, 150), then Enter. An ordered pair shows on each plot. What –159 in the ordered pair in the residual plot (28, –159) represent?***

  Answer: -159 is the difference between the actual cost of the property 28 spaces from Go, $150, and the cost predicted by the line, $309. $150-309=-159$. It is negative because the point lies below the line.

- ***If you were to use the line to predict the cost of the property 34 spaces from Go, would your predicted cost be over or under the actual cost? Explain your reasoning.***

  Answer: The predicted cost would be above the actual cost because the actual cost is $320 and the line would predict a cost of $364.50.

Class Discussion (continued)

- ***If you predicted the cost of a property for any of the spaces from Go, would your prediction ever be lower than the actual cost?***

  Answer: Only two –maybe three- of the points are above the line, so for those, your prediction from the line would be lower than the actual cost.

Look at the scatter plot and the residual plot again and in particular at the points that have large residuals.

- ***Do you notice anything unusual about the points?***

  Answers may vary. Four of the points are in a horizontal line at $C = 200$. Then at the top right, one point is above the line and in the middle one point pretty far below the line.

- ***In Monopoly, the costs of the railroads are always the same price. Find the railroads and indicate how many spaces from Go they are.***

  Answer: The railroads are 5, 15, 25 and 35 spaces from Go.

- ***The purpose of finding a model to relate two variables like distance and price is to be able to predict the second (dependent) such as price from a given value for the first (independent), such as distance. If another railroad were 32 spaces from Go, how much would the property cost?***

  Answer: $200

- ***Write an equation for the price of a railroad property located at any distance from Go.***

  Answer: $P = 200$.

Two utilities are located on the board, Water Works and Electric Company, at 12 and 28 spaces from Go. The price to buy a utility is always the same.

- ***What is the price to buy a utility?***

  Answer: $150

- ***What is the price of a utility predicted by the line?***

  Answers will vary depending on the line, Using one example, $P = 8.15(12) + 50 = \$147.80$, and $P = 8.15(28) + 50 = \$278.20$.

- ***What is worrisome about your answers to the question above?***

  Answer: The predicted prices are different and they should be the same.

Ceci claimed that because you always knew the price of the railroads and utilities, you should create a linear model using just the other properties without the utilities.

Class Discussion (continued)

- ***Do you agree? Why or why not?***

  Answer: Because the prices are constant, the utilities and railroads have their own models for the price. It makes sense to create a model for the other properties because that is where the variability or noise is. Using the utilities in your model makes the SAD larger.

- ***Predict what will happen to the SAD for your model if you remove the points representing the utilities and railroads. Then select* menu> Remove Point *and delete those points to check your conjecture*.**

  Answer: I think the SAD will decrease. When I removed the points for the model $P = 8.15D + 50$, the SAD had gone down from $631 to $175.

- ***Describe the residual plot after the utilities and railroads have been removed.***

  Answer: All but three or four of the residuals are on or below the line so your prediction would almost always be an overestimate.

- ***Timor noted that the price for Boardwalk also seems to be an "outlier"- different from the pattern in the rest of the scatter plot. He suggests removing Boardwalk as well. What would you say to Timor?***

  Answers may vary. Removing Boardwalk would not be the same as removing the railroads and utilities because they were always the same price no matter where they were located and so were not the same kind of properties. Even though the price of Boardwalk is high compared to the rest and to the model, it should stay in the data set.

- ***See if you can improve your line by considering both the residuals and the SAD, then use your model to answer the following three questions.***

  Answer: $C = 8.05D + 45$ with a SAD of $137. The residuals look kind of strange but don't really have a predictable pattern.

- ***What is the mean absolute deviation for your model? Interpret your answer in terms of the prices of properties in Monopoly.***

  Answers will vary. For the example above, the MAD is $\frac{137}{22}$, which is about $6.23. The typical price for a property will be about $6.23 from the price predicted by the linear model.

- ***Suppose it was possible to purchase a Community Chest located 17 spaces from Go. Use your equation to predict what price. Give an interval estimate using the MAD.***

  Answers will vary. The equation above gives $\$181.85 +/- \$6.23$, so anywhere from $188.08 to $175.62

- **Suppose the board were a regular pentagon instead of a square. Predict the price of a property located 45 spaces from Go.**

  Answers will vary. Using the equation above, the price would be $407.25.

Student Activity Questions—Activity 3

1. Many different lines can be drawn to fit a linear relationship. One method to find a good fit is to keep the sum of all the absolute values of the differences between the predicted and actual values (SAD) as small as possible. Another method is to see if there is any pattern or predictability in the residuals (errors). If you can describe a pattern in the error in predicting the dependent variable, you might be able to find a better model. The "scatter" around a line you are using to model a linear relationship (called a regression line) should appear to be random.

a. Select Show SAD. What is the SAD for the line on the screen? What does that number represent?

Answer: SAD = 1131, the sum of the absolute value of the differences between the actual and predicted costs for all of the spaces that can be sold totals $1131.

b. Trey moved the line and the equation for his new line is $P = 9.44D + 25$. Create Trey's line. Did his new line reduce the SAD?

Answer: Yes, since the SAD was $823, which was smaller than $1131.

c. Do the points on the residual plot seem to be randomly scattered? Explain your reasoning.

Answer: Not really. The points for all the properties less than 18 spaces from Go are above the line, and for all properties greater than 18 spaces from Go the residuals are either on the line or below it. That would mean your predictions would always be an underestimate for properties below 18 spaces from Go and overestimates or exact for properties above 18 spaces from Go. The "noise" has a pattern and is not quite random.

d. Move the line until the residuals seem to be random. What is the equation for your line? The SAD?

Answers will vary. One might be $C = 8.15D + 50$ with SAD = $631

e. Compare your line with others. How well does each seem to satisfy the criteria of a small SAD and no patterns in the residuals?

Answers will vary. Some might have small SADs but patterns in the residuals and others might have larger SADS. Answers will vary. Some might have small SADs but patterns in the residuals and others might have larger SADS.

Student Activity Questions—Activity 3 (continued)

2. **Match the sentence starters with an ending that makes a true sentence. Not all parts will have matches.**

Sentence Starters

a. If a scatter plot displays a linear pattern between two variables

b. A model is a good fit for the linear pattern in the data

c. A residual is

d. An equation that models the relationship between two variables

e. If a scatter plot shows a clear pattern

Sentence Enders

f. the distance between a point and a line used to model the relationship between the variables.

g. a straight line can be used to model the relationship.

h. if the residual plot is randomly distributed around the horizontal axis.

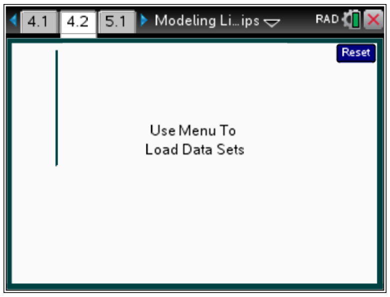
i. can be used to predict one value given the other.

j. the difference between the actual outcome and the predicted outcome for a given input.

k. if the sum of the absolute value of the residuals is small.

Answers: a and g; b and h; c and j or f; d and i; e and k do not have a match.

Part 4, Page 4.2



Use Menu To
Load Data Sets

Focus: Modeling linear relationships includes considering correlation, minimizing error in prediction, and interpretation of the model.

On page 4.2, students select data to investigate.

Load Data can be used to choose from four different data sets: poverty and graduation rates by state, first and last name lengths, alligator lengths and weights, maximum recorded speeds and life spans for selected animal types.

The page options are similar to those on page 3.2.

Class Discussion

Teacher Tip: The tasks in this part can be done as a whole class, by individual students or by students working in small groups or pairs. The key ideas have been developed in the first three parts; looking for linear relationships in a scatter plot, fitting a linear model to the trend, judging the fit of the linear model by using the SAD and examining residuals for possible patterns, interpreting the rate of change in a model in terms of the context and using the model to make predictions. Each of the data sets in this part adds an important idea, but a complete set of questions is not given. Rather each data set has two or three guiding questions to focus students on what is new about fitting a linear model in this particular context.

Several of the data sets relate to the concept of correlation, which is an actual measure of the strength of the linear association in paired data. The activity does not employ the formula for correlation but uses general terms such as strong, weak or none, and in particular distinguishes between a clear pattern (which may not be linear) and a linear pattern. You might want to refer back to: the Store prices where the correlation was very strong; (weight, bite force) for the alligators and crocodiles, which had fairly strong correlation; and the relationship between number of spaces from Go and price of properties in Monopoly. Students should clearly recognize that the existence of a strong correlation does not imply causation; for example, shoe sizes and reading vocabulary are strongly associated for younger children - but having large feet does not cause an increase in the number of words children know. Both large feet and number of words are a function of age: as children grow older, both their feet and their grasp of words increase.

Name Length

These data are the number of letters in the first and last names of a class of students. For example, the name Susanna Dei would be represented by the ordered pair (3, 7). The data show no association (correlation) as there is no linear pattern. Note that the line $F = 0L + 5$ or $F = 5$ is a horizontal line that divides the data so six points are above the line and seven points below, with a SAD of 19. Moving the line can produce a SAD of 18 but not much lower.

Students might investigate the following questions:

- ***Describe the pattern you see in the data and in the residuals. Include a description of the correlation between the number of letters in the first names and number of letters in the last names.***
- ***Move the line to obtain a small SAD.***
- ***Suppose you knew someone had eight letters in their last name. What would you predict for the number of letters in their first name? Include a description of the variability you might expect in your prediction.***
- ***What is your conclusion about fitting a model to these data?***

Class Discussion (continued)

Alligators

These data are the length and weight of alligators in Florida that were captured and tagged. The data show a clear pattern but it is not linear, which makes sense because length is one-dimensional (linear measurement) and weight is three-dimensional (like volume of a cube). Note that a predictable pattern will be present in the residual plot of a linear model fit to these data; the line will underestimate the weight of small alligators and of large alligators and overestimate the weight of those in between (or in extreme cases, over or under estimate everything).

Students might investigate the following questions:

- ***How would you expect the length and weight of alligators to be associated?***
- ***Move the line to see if you can get a good fit with no pattern in the residuals and a small SAD.***
- ***What is your conclusion about fitting a model to these data?***

Poverty and Graduation Rates

State	**Poverty (%)**	**Graduate (%)**
Alabama	16.7	80
Alaska	10.0	72
Arizona	15.2	75
Arkansas	15.9	85
California	13.2	80
Colorado	11.4	77
Connecticut	9.7	86
Delaware	9.2	80
District of Columbia	20.7	62
Florida	11.1	76
Georgia	14.4	72
Hawaii	8.6	82
Idaho	9.9	*
Illinois	11.5	83
Indiana	12.6	87
Iowa	11.3	90
Kansas	12.5	86
Kentucky	14.8	86
Louisiana	18.3	74
Maine	12.6	86
Maryland	9.7	85
Massachusetts	10.1	85
Michigan	12.0	77
Minnesota	8.1	80
Mississippi	20.1	76
Missouri	11.6	86
Montana	13.8	84
Nebraska	9.5	88
Nevada	10.6	71
New Hampshire	5.6	87
New Jersey	6.8	88
New Mexico	17.9	70
New York	14.5	77
North Carolina	13.1	83
North Dakota	11.2	88
Ohio	12.3	82
Oklahoma	15.6	85
Oregon	12.0	69
Pennsylvania	11.2	86
Rhode Island	12.1	80
South Carolina	15.0	78
South Dakota	11.8	83
Tennessee	15.0	86
Texas	16.2	88
Utah	9.2	83
Vermont	7.6	87
Virginia	9.2	84
Washington	10.2	76
West Virginia	15.4	81
Wisconsin	10.2	88
Wyoming	10.6	77

* data not available

Class Discussion (continued)

The data are the percentage of the population living in poverty and the percentage rate of high school graduation by state in the United States. The association (correlation) is negative and fairly strong. This is the first example of a model that has a negative rate of change.

Students might investigate the following questions:

- ***Describe the pattern in the scatter plot. Use words like increasing and decreasing.***
- ***Identify the following states: highest poverty rate and lowest graduation rate; least poverty rate and highest graduation rate.***
- ***How does your state compare to the others with respect to the two variables, poverty rate and high school graduation rate?***
- ***Fit a line to the data using both the residuals and the SAD. Interpret both the slope and the SAD in terms of the data.***
- ***Puerto Rico, a US territory, has a poverty rate of about 45%. What would you predict for the graduation rate> Give an interval estimate.***

Animals

The data are the maximum recorded speeds and maximum life spans of types of animals. The association (correlation) is negative and mildly linear.

Students might investigate the following questions:

- ***Describe the pattern in the plot.***
- ***Fit a model that seems to model the relationship between the maximum speeds and life spans of the animal types.***
- ***For which animals would your model produce the largest residual (error) in predicting speeds for given life spans? Explain how you know.***
- ***Four animal types are clustered in the lower left and do not seem to be part of the pattern. Do these animals have anything in common? Which animals seem to have a much higher speed for their given life span than other types with about the same life span?***
- ***These data do not have a clear independent, dependent relationship. Plot the data (maximum speed, maximum life span). Describe the plot and how your analysis will be different from that looking at the plot of (maximum life span, maximum speed).***

Deeper Dive — Page 1.3

Suppose the original scatter plot on page 1.3 was (Store B prices, Store A prices). How would that change each of the following?

- ***the description of the association?***

  Answer: The association is still strong in the positive direction and linear.

- ***the equation used to model the relationship?***

  Answer: One equation might be $A = 0.84B$ compared to $B = 1.12A$

- ***the interpretation of the rate of change?***

  Answer: 0.84 would indicate that for every $1 spent at Store B, you would spend $0.84 at Store A, while 1.12 would indicate that for every $1 spent at Store A you would spend $1.12 at Store B. In both cases, Store A is cheaper by about 12 to 16 cents.

- ***the smallest SAD?***

  Answers will vary. An answer for the equation for (B, A) is $5.38, while an equation for (A, B) might have a SAD of $6.60.

- ***Bre argued that because the same data are used for both the x-variable and the y-variable, the equation for (x, y) should be the same as the equation for (y, x). Do you agree or disagree with Bre?***

  Answers may vary. Bre is wrong because the arrangements of the points are different, the slopes would have to be reciprocals of each other; and because the relationship among the points is slightly different in each plot, the slopes are not likely to be reciprocals.

The goal of finding a line to summarize an association between two variables is finding a model that allows you to predict with some sense of accuracy an outcome for a given independent (explanatory) variable. Decide if each sentence ending is correct and describe an example from the TNS activity that supports your reasoning.

A good model for a relationship between two quantities

a. will go through as many actual data points as possible.

b. will always include the first and last data point.

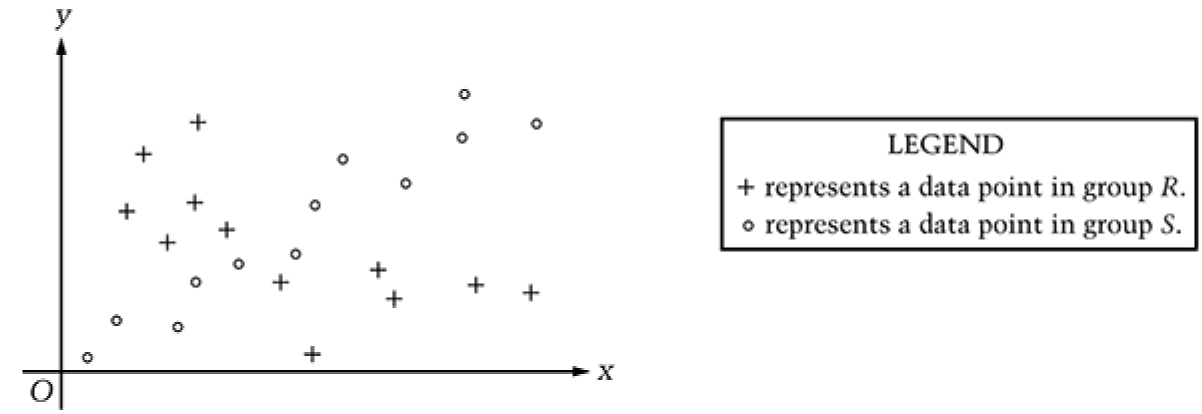
c. will have the same number of data points above the line as below the line.

d. has a SAD that is as small as possible.

Answers will vary. a) Having all of the data close to the line could be a better model than one that has five points right on the line and the others far from the line; b) the first and last points might not be representative of the relationship and would produce a line that does not really generalize the pattern; c) a horizontal line can be drawn that will have the same number of points above as below but will not represent the pattern in the data; d) is correct because this will give small differences between the observed and predicted.

Sample Assessment Items

After completing the lesson, students should be able to answer the following types of questions. If students understand the concepts involved in the lesson, they should be able to answer the following questions without using the TNS activity.



1. The scatterplot above shows data for groups *R* and *S*. Which of the following statements is true about the correlation between the *x* and *y* values of group *R* and the correlation between the *x* and *y* values of group *S*?

   a. The *x* and *y* values appear to be negatively correlated in both *R* and *S*.

   b. The *x* and *y* values appear to be positively correlated in both *R* and *S*.

   c. The *x* and *y* values appear to be negatively correlated in *R*, but positively correlated in *S*.

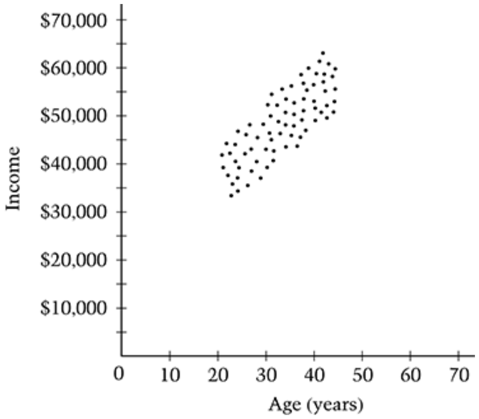
   d. The *x* and *y* values appear to be positively correlated in *R*, but negatively correlated in *S*.

   e. The *x* and *y* values appear to be more highly correlated in *R* than in *S*.

   Answer: c. The x and y values appear to be negatively correlated in R, but positively correlated in S.

NAEP grade 8 2009

Questions 2 and 3 refer to the following scatterplot.



2. The plot below shows the distribution of the median number of hours per week spent doing homework A random sample of graduates from a particular college program reported their ages and incomes in response to a survey. Each point on the scatterplot above represents the age and income of a different graduate. Which of the following equations best fits the data above?

   a. $y = -1{,}000x + 15{,}000$

   b. $y = 1{,}000x$

   c. $y = 1{,}000x + 15{,}000$

   d. $y = 10{,}000x$

   e. $y = 10{,}000x + 15{,}000$

   Answer: c. $y = 1{,}000x + 15{,}000$

NAEP 2009 grade 8

3. Based on the data in the scatterplot, predictions can be made about the income of a 35 year old and the income of a 55 year old. For which age is the prediction more likely to be accurate?

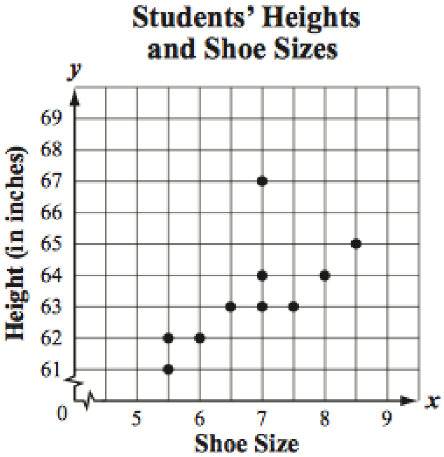
   ○ 35 year old ○ 55 year old

   Justify your answer.

   Answer: 35 year old as the plot does not show any data for those over 50 years old.

NAEP, 2009 grade 8

4. The scatterplot below shows the relationship between the height, in inches, and the shoe size of each of 10 students in a class.

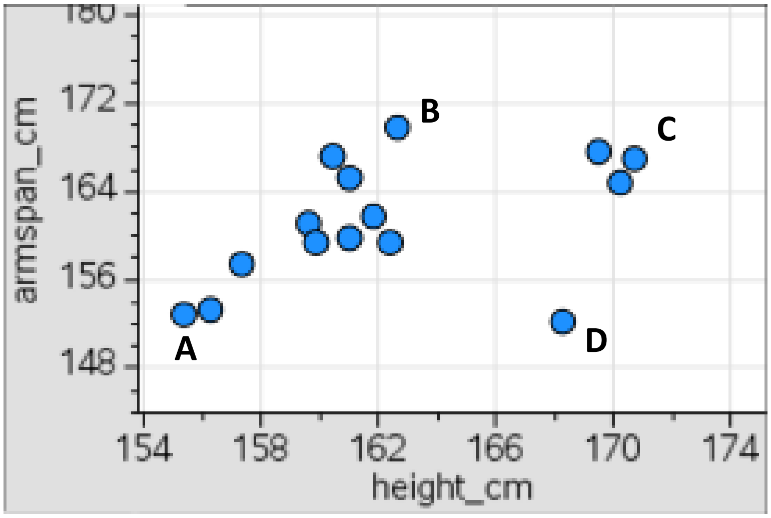


   Based on the scatterplot, what ordered pair represents the outlier in the data?

   Answer: (7, 67)

   (MA 2015 grade 8 release items, http://www.doe.mass.edu/mcas/2015/release/Gr8-Math.pdf)

5. The scatter plot below shows the relationship between height, in centimeters, and the arm span in centimeters of 15 students in a class.



Based on the scatter plot, determine which ordered pair would be farthest from the best fit line.

a) A
b) B
c) C
d) D

Answer: D

Student Activity Solutions

In these activities you will investigate patterns of association in bivariate data. After completing the activities, discuss and/or present your findings to the rest of the class.

Activity 1 [Page 1.3]

1. a. Refer to the graph you created on page 1.3. Select the point representing a box of cereal, then **Enter**. Explain what the number and vertical segment tell you.

Answer: The number is the difference I found between the actual and predicted cost for the cereal. The vertical segment shows the difference on the graph—the vertical distance from the point to the line.

b. Select the point you think will have the greatest error (difference between the actual and predicted cost). Find the difference and explain what it means.

Answers will vary. Ketchup is the farthest from the line B=1.25A. The difference in cost is about $\$2.19-\$3.40=-\$1.21$*. The price of Ketchup predicted by the line at Store B is $1.21 cheaper than the actual price at Store B.*

c. Fill in the table with the missing values using your line.

Item	Store A Price $	Store B Price $	Store B price predicted by line*($)	Actual price minus predicted price ($)
Whole Grain Cereal (10 oz)	3.52	4.79	4.40	$\$4.79-\$4.40=\$0.39$
Raisins	1.94	2.99	2.42	$\$2.99-\$2.42=\$0.57$
Peanuts	4.98	5.89	6.23	$\$5.89-\$6.23=\$0.34$
Ketchup (24 oz)	2.72	2.19	3.40	$\$2.19-\$3.40=-\$1.21$

Answers may be slightly different due to rounding.

d. For the items in the table, what is the total of the differences in the actual prices for the items minus the prices predicted by your line? Explain what your answer means in terms of the prices.

Answers will vary. The total cost of the items in Store B is $15.86. The cost predicted by my line is $16.45. The difference is $\$15.86-\$16.45=\$0.59$*, which says that the total cost of the items predicted by my line is $0.59 more than the actual cost.* $\$0.96-\$1.55=\$0.59$

2. For an article they were writing, the class wanted to describe how far off the predicted prices were from the actual prices for the items at the two stores.

   a. Marj claimed the sum of the differences in actual price minus predicted price for the items in the table was -0.59, or about $\frac{-60}{4}=-15$ cents per item. Petra disagreed and said if you found the mean the way Marj did, you could have a sum of the differences in actual minus predicted prices to be 0, when there really were a lot of "differences". Whose reasoning makes the most sense and why? Give an example to support your thinking.

   Answer: Petra is correct because if one item was 50 cents more at one store than the other and another item was 50 cents less, the sum of the differences would be 0 (-0.5+0.5). This would make it seem like all of the prices were really the same. You have to worry about the positive and negative signs.

   b. Hilary suggested finding the sum of the differences by taking the absolute value of the differences and then adding. What do you think about Hilary's reasoning?

   Answer: Hilary's reasoning makes sense; it is like finding the mean absolute deviation. It accounts for the positive and negative signs and she will get a total amount of deviation.

   c. Anita said she would find the mean difference by squaring all the differences and finding the sum of the squared differences. What do you think about Anita's reasoning?

   Answer: Anita's reasoning is a good way to make sure you have taken care of the negative values, but she will have a very large number for the mean because she squared everything (she has squared $ as a unit). She might take the square root of her answer and then it would at least be in the right unit, $.

   d. Use one of the methods described in b) and c) to find the sum of the differences in prices. Compare your answer with others. What might explain differences?

   Answers will vary. Those who squared will have a much larger difference than those who used the absolute value.

3. SAD is the sum of the absolute value of all of the differences. Select All Segments then Show SAD.

   a. How does the value for SAD relate to the scatter plot and line?

   Answer: Each segment represents the difference between the actual cost of an item at Store B and the cost predicted by my line for a particular item at Store A. The SAD for my line is 7.28 (measured in dollars), which is the sum of the absolute values of the actual costs minus predicted costs for the items.

   b. The predicted cost of all of the items at Store B using the equation B = 1.25 A is $63.38, and the actual cost was $59.58. How does the difference, actual total cost minus predicted total cost compare to the SAD? Explain any differences.

   Answer: The difference between the two total costs was -$3.80. It is smaller than the SAD, $7.28 because the SAD uses the absolute value of the differences.

c. Move the line. Describe what happens to the SAD.

Answer: As the line gets farther from the pattern in the data, lengths of the segments increase and the SAD increases. The closer the line is to modeling the pattern, the shorter the segments and the smaller the SAD.

d. Find a new line that has a smaller SAD than your original line. Explain what your SAD represents.

Answers will vary. One line, B = 1.18A, had a SAD of 6.33. The sum of the absolute value of the differences, actual price minus predicted price, for all of the items was $6.33.

4. Identify the following statements as true or false. Be ready to explain your thinking.

a. The line *y* = *x* represents a relationship where all of the ordered pairs are of the form (*x*, *x*).

Answer: True. The values for x and y are equal.

b. If the equation for a linear *relationship* is *y* = 2*x*, the *x*-values are twice as large as the *y*-values.

Answer: False. The y-values are twice as large as the x-values (or the x-values are half as large as the y-values).

c. Suppose the equation that models a linear relationship is B = A, where A and B are greater than or equal to 0. If the point (A, B) is below the line, then B will be greater than A.

Answer: False. B will never be greater than A because below the line B = A, B is always less than A.

d. If the difference, actual price of an item minus the price predicted by a line, is negative, the point representing that item lies below the line.

Answer: True because to have a negative difference, the vertical coordinate of the actual price has to be smaller (below) the vertical coordinate of the predicted price.

Activity 2 [Page 2.2]

1. Determine the error in using the line to predict the bite force for the Indian Gharial.

a. What does the value tell you?

Answer: The -569.5 indicates that the predicted bite force by the line is 569.5 pounds more than the actual bite force for that crocodile.

b. Thom says that the length of the segment from this point to the line is larger than all of the other segments together. How could you decide if he is correct?

Answer: Determine the SAD and see if it is more than half of the SAD. The SAD is 2858, so the sum of the absolute differences between the predicted and the observed bite force is 2,858 pounds. 569.5 is only about 1/4 of the overall total so Thom is incorrect.

c. Move the line to minimize the total sum of the absolute distances. What is the equation of your line? Compare your line to others in your class. Who has the smallest SAD?

Answers will vary. One equation might be $F = 3.42M + 124$ *with* $\text{SAD} = 2286.08$ *psi.*

d. What is the mean absolute deviation of the difference in the actual values for a given weight and the values predicted using your model?

Answers will vary. There were 27 crocodiles and alligators represented in the plot, so using the equation above, the MAD would be 2286.08/27, which is about 84.67 psi.

2. Another crocodile, Croc G, with a body mass of 350 pounds had a bite force of 1100 psi.

a. How do you think the SAD will change if you add Croc G to the plot?

Answers will vary: Using the equation $F = 3.42M + 124$*, the SAD should increase about 221 psi because that is the distance from the psi predicted by the line and the actual psi for Croc G.*

b. What do you think will happen to the original SAD of 2288.08 psi if a crocodile that weighs 200 pounds and has a bite force of 800 psi is added to the plot?

Answer: For the equation $F = 3.42M + 124$*, the SAD should remain almost the same because the observed and the predicted values are 800 and 808 respectively.*

c. If you add a point representing a crocodile that weighs 200 pounds with a bite force of 1000 psi, do you think the SAD will increase, decrease or remain the same?

Answer: The SAD should increase as the difference between the absolute value of the observed and predicted is always positive.

3. In 2012, an American alligator named Hercules was weighed and his bite force measured.

a. Use your equation to predict the bite force given that Hercules weighed 665 pounds.

Answers will vary. Using the equation in 1c), the predicted bite force would be about 2,399 pounds per square inch.

b. Use the MAD you found in 1d) to find an interval that would typically contain the bite force for Hercules.

Answers will vary. Using the equation and work above, the interval would be the predicted bite force +/- the MAD, which would be about 2315 to 2484 psi.

c. Hercules actually had a bite force of 2125 psi. How well did your model predict his bite force?

The model predicted that typically, the bite force would be quite a bit higher; but it is not too unusual to have values outside of the mean +/- MAD when working with a set of data.

d. An adult male lion a weighs about 420 pounds and has a bite force of about 600 psi. Do you think it would be reasonable to use your model to estimate the bite force of other animals? Explain why or why not.

Answers will vary. The model in the example above would predict 1550 psi for the lion, which is way out of line for the number 600 psi given in the research. It would seem that a model constructed based on information about alligators and crocodiles might not be adequate for other animals with a different body and jaw construction.

Activity 3 [Page 3.2]

1. Many different lines can be drawn to fit a linear relationship. One method to find a good fit is to keep the sum of all the absolute values of the differences between the predicted and actual values (SAD) as small as possible. Another method is to see if there is any pattern or predictability in the residuals (errors). If you can describe a pattern in the error in predicting the dependent variable, you might be able to find a better model. The "scatter" around a line you are using to model a linear relationship (called a regression line) should appear to be random.

   a. Select Show SAD. What is the SAD for the line on the screen? What does that number represent?

   Answer: SAD = 1131, the sum of the absolute value of the differences between the actual and predicted costs for all of the spaces that can be sold totals $1131.

   b. Trey moved the line and the equation for his new line is P = 9.44D+25. Create Trey's line. Did his new line reduce the SAD?

   Answer: Yes, since the SAD was $823, which was smaller than $1131.

   c. Do the points on the residual plot seem to be randomly scattered? Explain your reasoning.

   Answer: Not really. The points for all the properties less than 18 spaces from Go are above the line, and for all properties greater than 18 spaces from Go the residuals are either on the line or below it. That would mean your predictions would always be an underestimate for properties below 18 spaces from Go and overestimates or exact for properties above 18 spaces from Go. The "noise" has a pattern and is not quite random.

   d. Move the line until the residuals seem to be random. What is the equation for your line? The SAD?

   Answers will vary. One might be C=8.15D+50 with SAD = $631

   e. Compare your line with others. How well does each seem to satisfy the criteria of a small SAD and no patterns in the residuals?

   Answers will vary. Some might have small SADs but patterns in the residuals and others might have larger SADS. Answers will vary. Some might have small SADs but patterns in the residuals and others might have larger SADS.

2. Match the sentence starters with an ending that makes a true sentence. Not all parts will have matches.

Sentence Starters	**Sentence Enders**
a. If a scatter plot displays a linear pattern between two variables	f. the distance between a point and a line used to model the relationship between the variables.
b. A model is a good fit for the linear pattern in the data	g. a straight line can be used to model the relationship.
c. A residual is	h. if the residual plot is randomly distributed around the horizontal axis.
d. An equation that models the relationship between two variables	i. can be used to predict one value given the other.
e. If a scatter plot shows a clear pattern	j. the difference between the actual outcome and the predicted outcome for a given input.
	k. if the sum of the absolute value of the residuals is small.

Answers: a and g; b and h; c and j or f; d and i; e and k do not have a match.