Unit 3, Lesson 7

The Correlation Coefficient

Goals

- Match the correlation coefficient to its appropriate scatter plot and linear model.
- Use the correlation coefficient to determine the goodness of fit for a linear model.

Learning Targets

- I can describe the goodness of fit of a linear model using the correlation coefficient.
- I can match the correlation coefficient with a scatter plot and linear model.

Lesson Narrative

The mathematical purpose of this lesson is for students to interpret the correlation coefficient and to use it to understand the strength of a linear relationship. The term **correlation coefficient** is introduced and is defined as a number that can be used to determine how well a line models the data.

When students sort scatter plots, they are given the opportunity to analyze representations, statements, and structures closely and make connections (MP2, MP7). When students take turns with a partner to match the correlation coefficients with scatter plots, students trade roles explaining their thinking and listening, which provides opportunities for them to explain their reasoning and critique the reasoning of others (MP3).

Standards

Addressing	**HSS-ID.B.6, HSS-ID.C.8**
Building Toward	**HSS-ID.C.8**

Instructional Routines

- Card Sort
- MLR8: Discussion Supports
- Take Turns
- Which Three Go Together?

Required Materials

Materials to Gather

- Math Community Chart: Activity 2

Materials to Copy

- Scatter Plot Fit Cards (1 copy for every 2 students): Activity 2

Student Facing Learning Goals

Let's see how good a linear model is for some data.

7.1 Which Three Go Together: Linear Models

5 min

Warm-up

Activity Narrative

This *Warm-up* prompts students to compare four scatter plots displaying data with linear and nonlinear trends. It gives students a reason to use language precisely (MP6). It gives the teacher an opportunity to hear how students use terminology and talk about characteristics of the items in comparison to one another.

Standards

Addressing	**HSS-ID.B.6**
Building Toward	**HSS-ID.C.8**

Instructional Routines

- Which Three Go Together?

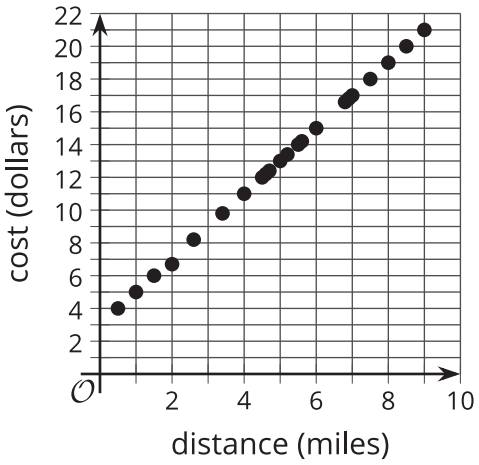
Launch

Arrange students in groups of 2–4. Display the scatter plots for all to see. Give students 1 minute of quiet think time and then time to share their thinking with their small group. In their small groups, tell each student to share their response with their group and then together find as many sets of three as they can.

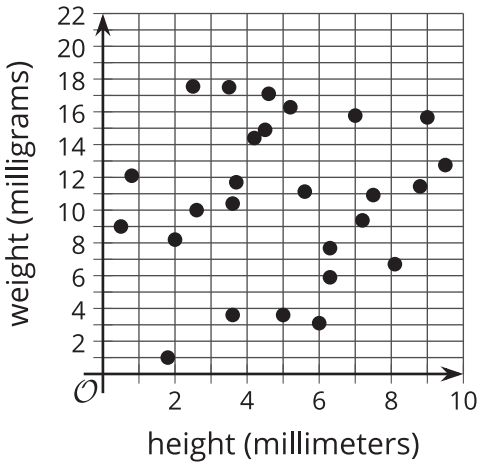
Student Task Statement

Which three go together? Why do they go together?

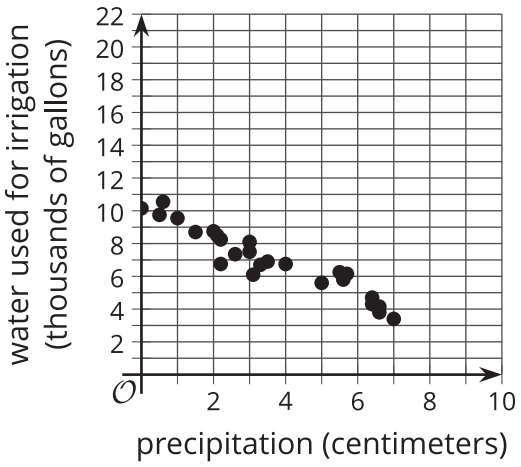
A



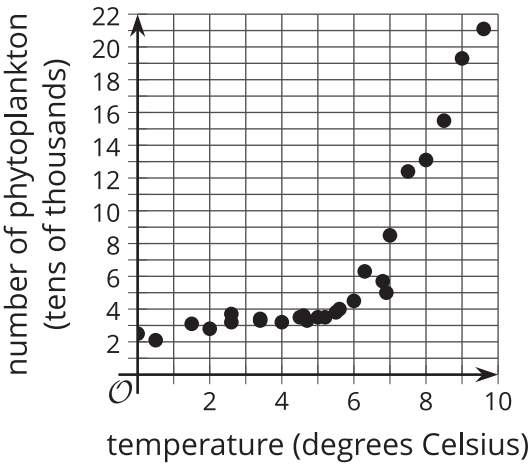
B



C



D



Student Response

Sample responses:

A, B, and C go together because the x-axes are all measured in units of length.

A, B, and D go together because the trends in the data appear to be increasing.

A, C, and D go together because it seems they will be fit well by a model function.

B, C, and D go together because they are not fit perfectly by a linear model.

Activity Synthesis

Invite each group to share one reason why a particular set of three goes together. Record and display the responses for all to see. After each response, ask the class if they agree or disagree. Since there is no single correct answer to the question of which three go together, attend to students' explanations and ensure the reasons given are correct.

During the discussion, ask students to explain the meaning of any terminology they use, such as "linear," "nonlinear," and "random." Also, press students on unsubstantiated claims.

7.2 Card Sort: Scatter Plot Fit

20 min

Activity Narrative

In this activity, students are given cards displaying scatter plots of data that can be fit by linear models with varying accuracy. Cards show data that is random, poorly fit by a linear model, well fit by a linear model, and better fit by another type of function, such as quadratic or exponential. Students should begin to recognize these differences and the connection to the correlation coefficient.

Students sort different scatter plots during this activity. A sorting task gives students opportunities to analyze representations, statements, and structures closely and make connections (MP2, MP7).

Monitor for different ways groups choose to categorize the scatter plots, but especially for categories that distinguish

between plots that would be modeled well with a linear function and those that would not.

Standards

Addressing	**HSS-ID.B.6**
Building Toward	**HSS-ID.C.8**

Instructional Routines

- Card Sort
- MLR8: Discussion Supports
- Take Turns

Launch

Math Community

Display the Math Community Chart for all to see. Give students a brief quiet think time to read the norms, or invite a student to read them out loud. Tell students that during this activity they are going to practice looking for their classmates putting the norms into action. At the end of the activity, students can share what norms they saw and how the norm supported the mathematical community during the activity.

Arrange students in groups of 2, and distribute pre-cut cards. Tell them that in this activity, they will sort some cards into categories of their choosing. When they sort the scatter plots, they should work with their partner to come up with categories.

Access for English Language Learners

MLR8 Discussion Supports. Students should take turns finding a match and explaining their reasoning to their partner. Display the following sentence frames for all to see: "I noticed _____, so I matched . . . ." Encourage students to challenge each other when they disagree.
Advances: Conversing

Student Task Statement

Your teacher will give you a set of cards that show scatter plots.

1. Sort the cards into categories of your choosing. Be prepared to describe your categories.
   Pause for a whole-class discussion.
2. Sort the cards into new categories in a different way. Be prepared to describe your new categories.

Student Response

Sample responses:

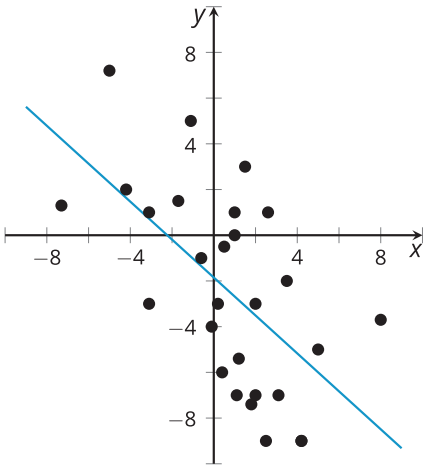
- Scatter plots that look linear: C, D, F, G. Scatter plots that are not very linear: A, B, E, H, I, J.
- Generally increasing: E, F, G, I. Generally decreasing: A, C, D, J. Not really increasing or decreasing: B, H.

Activity Synthesis

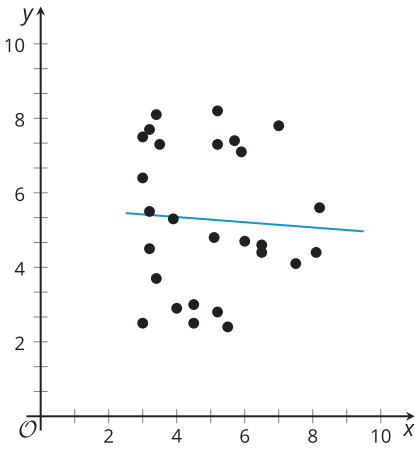
Select groups of students to share their categories and how they sorted their scatter plots. Discuss as many different types of categories as time allows, but ensure that one set of categories distinguishes between plots that would be modeled well with a linear function and those that would not. Attend to the language that students use to describe their categories and scatter plots, giving them opportunities to describe their scatter plots more precisely. Highlight the use of terms like "linear model," "fit," and "nonlinear."

Display the scatter plots with the best-fit lines and r-values.

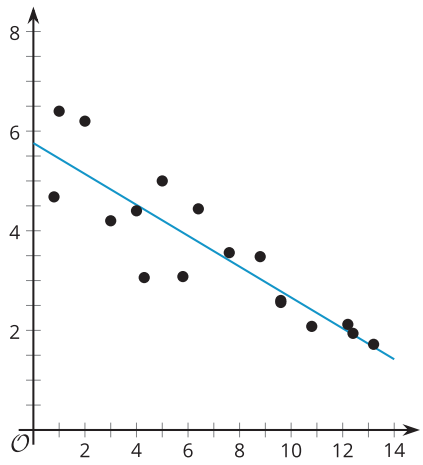
A: $y = -0.83x - 1.85, r = -0.61$



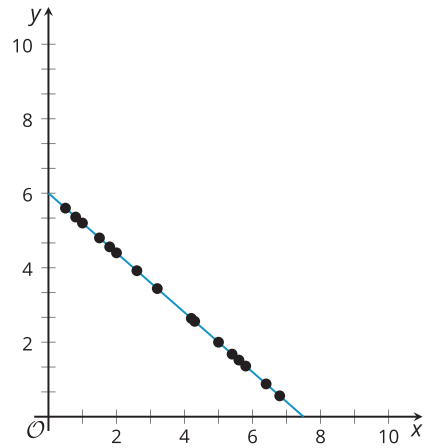
B: $y = -0.07x + 5.63, r = -0.06$



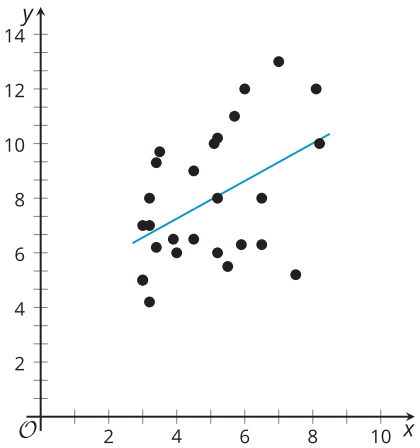
C: $y = -0.31x + 5.76, r = -0.88$



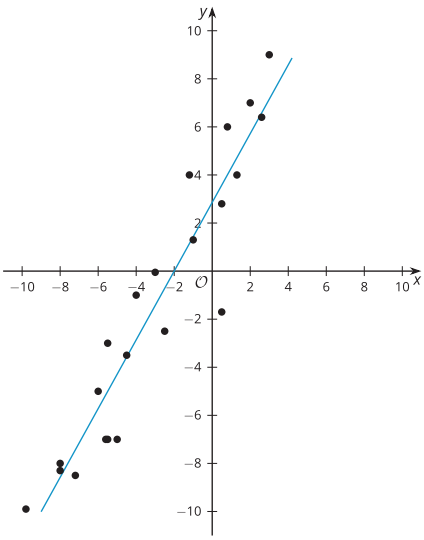
D: $y = -0.8x + 6, r = -1$



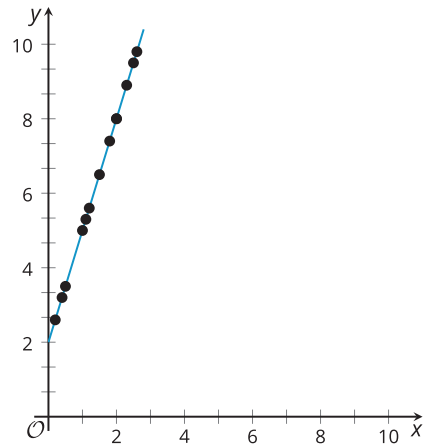
E: $y = 0.69x + 4.49, r = 0.46$



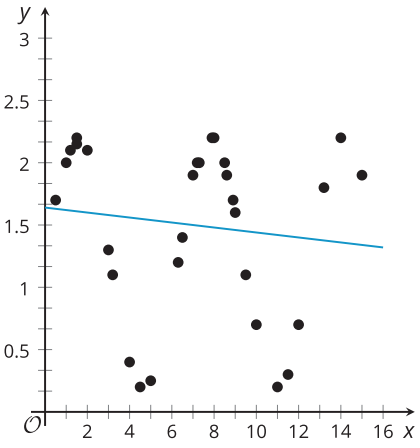
F: $y = 1.43x + 2.86, r = 0.94$



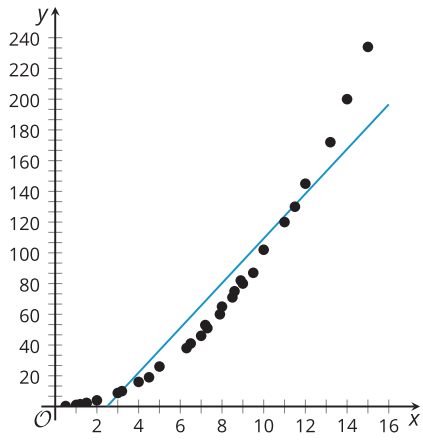
G: $y = 3x + 2, r = 1$



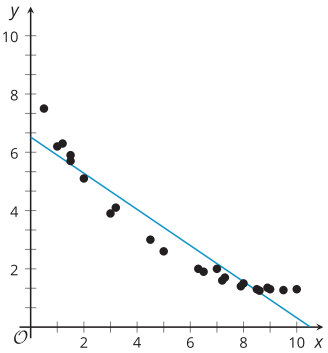
H: $y = -0.02x + 1.64, r = -0.13$



I: $y = 14.55x - 36.18, r = 0.95$



J: $y = -0.62x + 6.52, r = -0.96$



Give students 1 minute of quiet think time and then 1 minute to discuss with their partner the things they notice, then follow with a whole-class discussion.

Among things students should notice are:

- The sign of r is the same as the sign of the slope of the best-fit line.
- The values for r seem to go from -1 to 1.
- The closer r is to 1 or -1, the stronger the linear relationship between the variables.
- The closer r is to 0, the weaker the linear relationship between the variables.

Note that the sign of the correlation coefficient matches the sign of the slope of the best-fit line, but the value for r is not otherwise related to the slope. If $r = 0.8$, the best-fit line will have a positive slope, but whether the slope is 0.2 or 2,000 is not clear without examining the data.

Math Community

Conclude the discussion by inviting 2–3 students to share a norm they identified in action. Provide this sentence frame to help students organize their thoughts in a clear, precise way:

- "I noticed our norm '_____' in action today and it really helped me/my group because _____."

Access for Students with Disabilities

Representation: Develop Language and Symbols. Create a display of important terms and vocabulary. Invite students to suggest language or diagrams to include that will support their understanding of the following concepts: fit, linear model, nonlinear, increasing, decreasing, pattern, random, and correlation coefficient. *Supports accessibility for: Memory, Language*

7.3 Matching Correlation Coefficients

10 min

Activity Narrative

In this activity, students gain a better understanding of correlation coefficients by taking turns with a partner to match scatter plots and correlation coefficients. Students trade roles, explaining their thinking and listening, which provides opportunities for them to explain their reasoning and critique the reasoning of others (MP3).

Standards

Addressing **HSS-ID.C.8**

Instructional Routines

- MLR8: Discussion Supports
- Take Turns

Launch

Tell students that the r-value is called a correlation coefficient. A **correlation coefficient** is one way to measure the strength of a linear relationship. Tell students that:

- The sign of r is the same as the sign of the slope of the best-fit line.
- The values for r go from -1 to 1 (inclusive).
- The closer r is to 1 or -1, the stronger the linear relationship between the variables.
- The closer r is to 0, the weaker the linear relationship between the variables.

Arrange students in groups of 2. Tell students that for each scatter plot, one partner finds the associated correlation coefficient and explains why they think it goes with that scatter plot. The other partner's job is to listen and make sure they agree. If they don't agree, the partners discuss until they come to an agreement. For the next scatter plot, the students swap roles. If necessary, demonstrate this protocol before students start working.

Access for Students with Disabilities

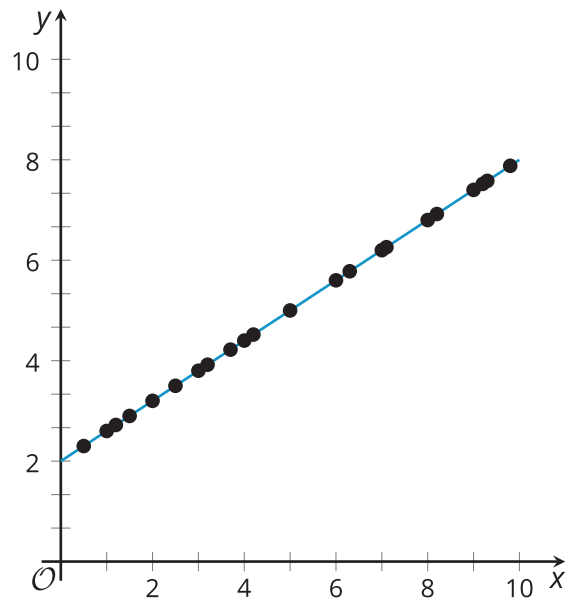
Representation: Develop Language and Symbols. Provide students with access to charts with symbols and meanings. Create a chart of the r-value ranging from -1 to 1. Label with the corresponding features of the various domains of the r-value. Select students to label the chart with the corresponding descriptors for each domain. Small sketches or printouts of example scatter plots can be added to the appropriate areas of the chart. *Supports accessibility for: Conceptual Processing, Memory*

Student Task Statement

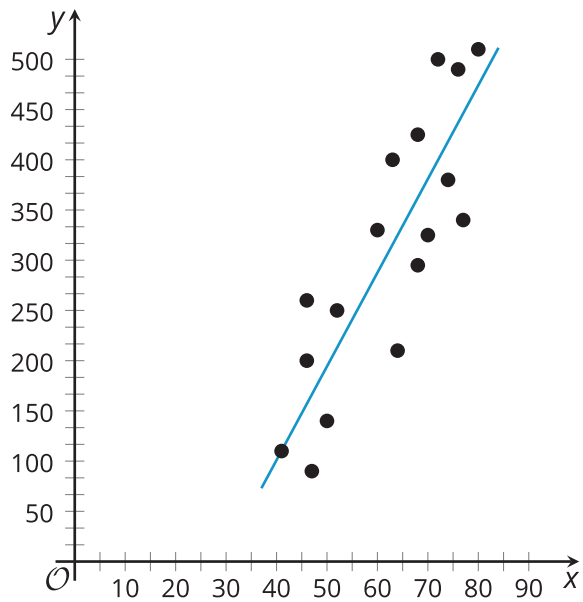
1. Take turns with your partner to match a scatter plot with a **correlation coefficient**.
2. For each match you find, explain to your partner how you know it's a match.
3. For each match your partner finds, listen carefully to their explanation. If you disagree, discuss your thinking and work to reach an agreement.

1. $r = -1$
2. $r = -0.95$
3. $r = -0.74$
4. $r = -0.06$
5. $r = 0.48$
6. $r = 0.65$
7. $r = 0.9$
8. $r = 1$

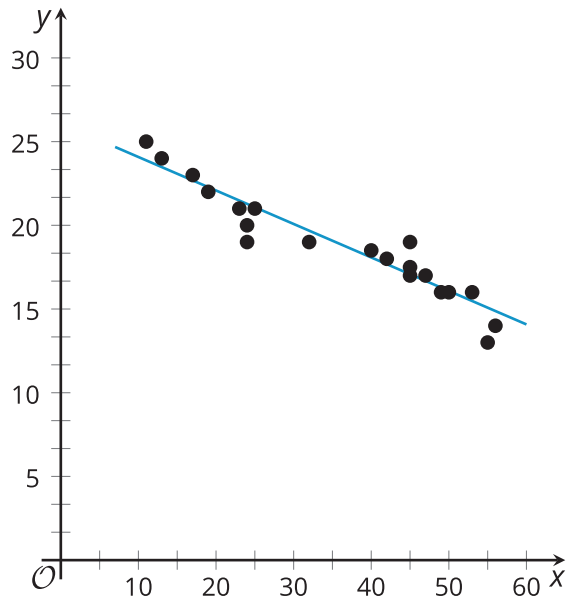
A


y
10
8
6
4
2
O
2
4
6
8
10
x

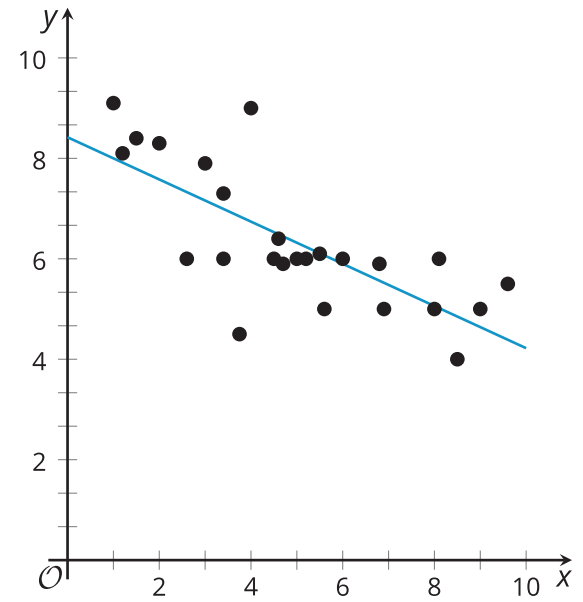
B


y
500
450
400
350
300
250
200
150
100
50
O
10
20
30
40
50
60
70
80
90
x

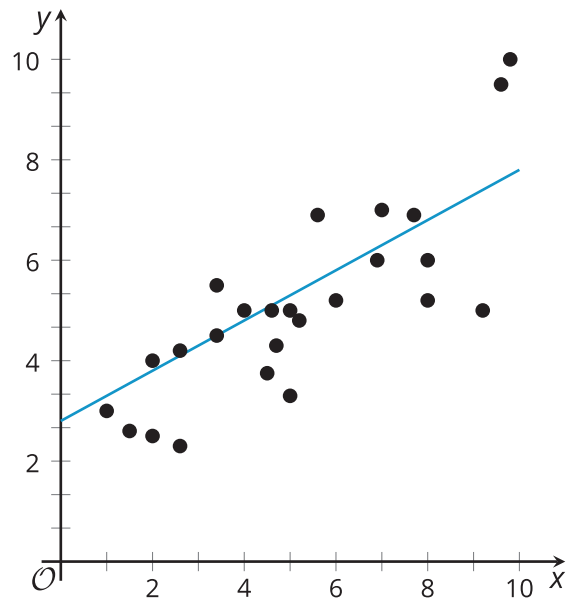
C


y
30
25
20
15
10
5
O
10
20
30
40
50
60
x

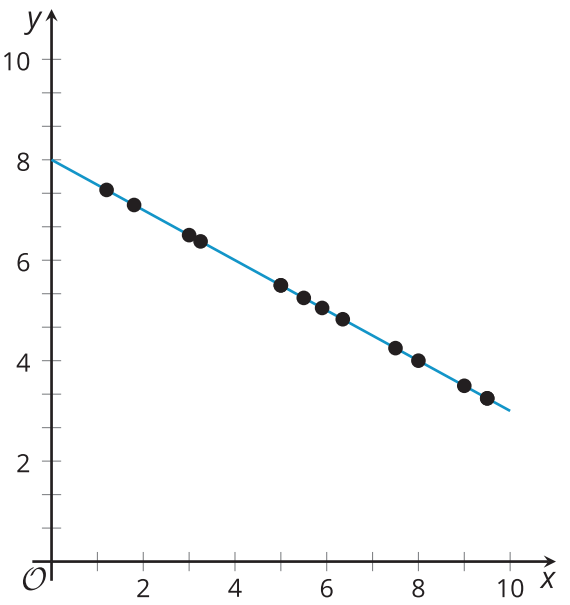
D


y
10
8
6
4
2
O
2
4
6
8
10
x

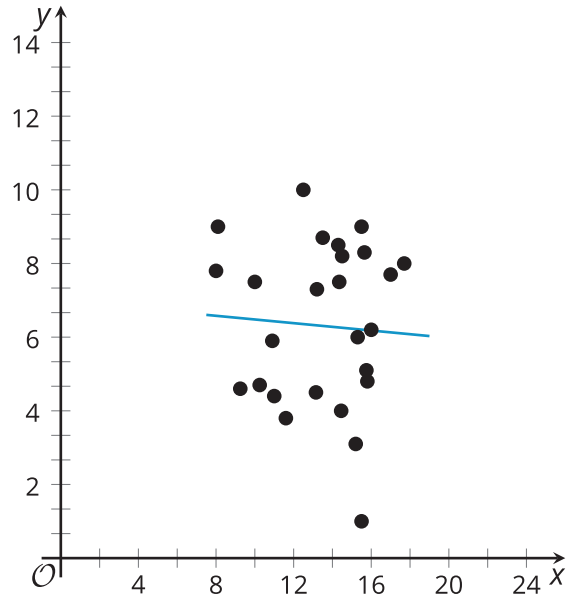
E



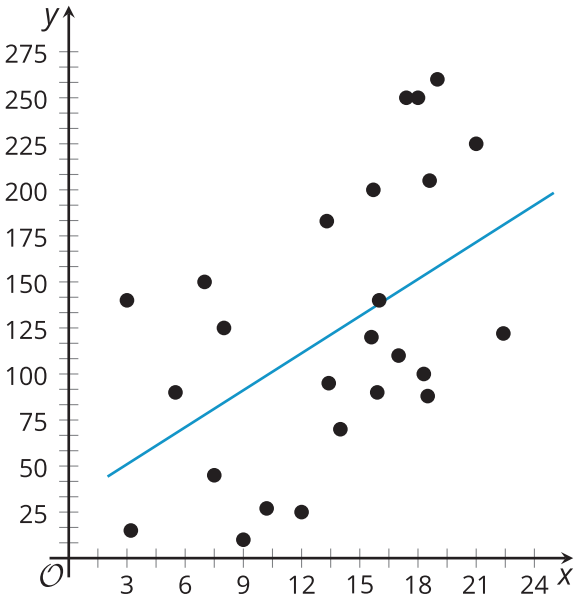
F



G



H



Student Response

1. F
2. C
3. D
4. G
5. H
6. E
7. B

8. A

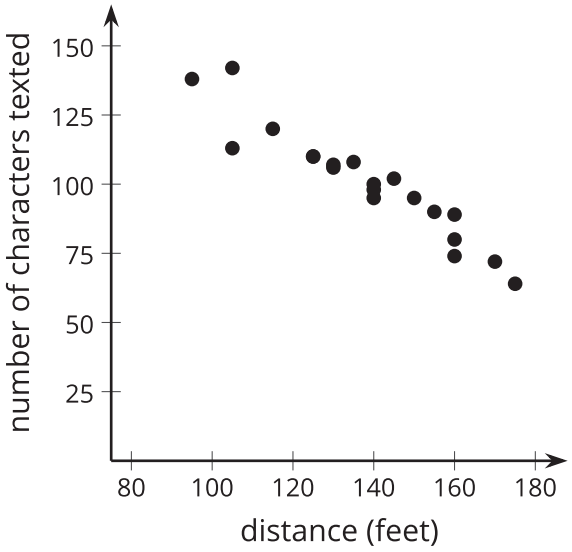
Building on Student Thinking

Students may struggle with starting to match the scatter plots with a correlation coefficient. Guide students by asking them about the sign of the correlation coefficients. Ask them to sort the cards into groups that make sense and use those to make a connection to the correlation coefficient values. Ask them: "How does the sign of the correlation coefficient relate to the linear model?"

Are You Ready for More?

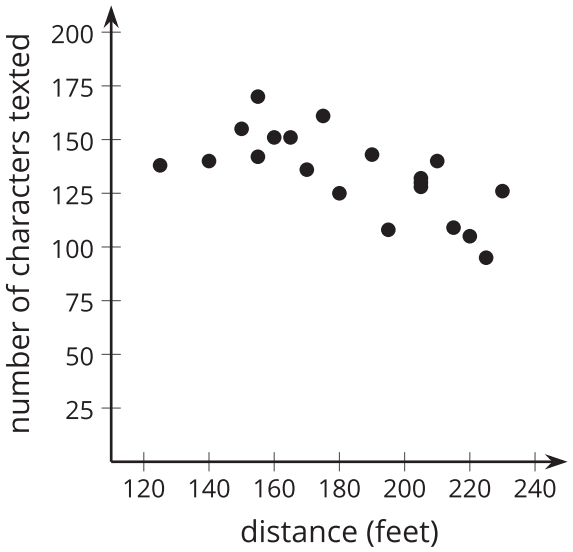
Jada wants to know if the speed that people walk is correlated with their texting speed. To investigate this, she measured the distance, in feet, that 5 of her friends walked in 30 seconds and the number of characters they texted during that time. Each of the 5 friends took 4 walks for a total of 20 walks. Here are the results of the first 20 walks.

distance (feet)	number of characters texted	distance (feet)	number of characters texted
105	142	95	138
125	110	125	110
115	120	160	80
140	98	175	64
145	102	130	106
160	89	140	95
170	72	150	95
140	100	155	90
130	107	160	74
105	113	135	108



Over the next few days, the same 5 friends practiced walking and texting to see if they could walk faster and text more characters. They did not record any more data while practicing. After practicing, each of the 5 friends took another 4 walks. Here are the results of the final 20 walks.

distance (feet)	number of characters texted	distance (feet)	number of characters texted
140	140	165	151
150	155	170	136
160	151	190	143
155	170	205	132
180	125	205	128
205	130	210	140
225	95	215	109
175	161	220	105
195	108	230	126
155	142	225	138



1. What do you notice about the two scatter plots?
2. Jada noticed that her friends walked further and texted faster during the last 20 walks than they did during the first 20 walks. Since both were faster, she predicts that the correlation coefficient of the line of best fit for the last 20 walks will be closer to -1 than the correlation coefficient of the line of best fit for the first 20 walks will be. Do you agree with Jada? Explain your reasoning.
3. Use technology to find an equation of the line of best fit and the correlation coefficient for each data set. Was your answer to the previous question correct?
4. Why do you think the correlation coefficients for the two data sets are so different? Explain your reasoning.

Extension Student Response

Sample responses:

1. I noticed that in the second scatter plot, the students traveled longer distances. In addition, the students seemed to type more characters for any given distance in the second scatter plot than in the first scatter plot.
2. I do not agree with Jada. The correlation coefficient deals with how closely the line of best fit models the data. The comparison between the walking distance and the number of characters texted is about the slope of the line—not the fit of the line.
3. The correlation coefficient for the first line of best fit is -0.95, and the correlation coefficient for the second line of best fit is -0.68. My answer to the previous question is correct, and I noticed that the slopes are different and so are the vertical intercepts.
4. I think the correlation coefficients are different because the data in the second scatter plot is more scattered than the data in the first scatter plot. On average, the data in the first scatter plot is much closer to the line of best fit than the data in the second scatter plot is.

Activity Synthesis

The purpose of this discussion is for students to understand that the correlation coefficient is a formal way to quantify the strength of a linear relationship between variables and that the sign of the correlation coefficient tells us whether or not the variables show a positive or negative association.

Here are some questions for discussion.

- "What does the sign of the correlation coefficient tell you about the data?" (If it is negative, then y tends to decrease as x increases. If it is positive, then y tends to increase as x increases.)
- "What does it mean to have a correlation coefficient of 1 or -1?" (It means that the data is perfectly linear and is fit exactly by a linear function.)

Access for English Language Learners

MLR8 Discussion Supports. Revoice student ideas to demonstrate and amplify mathematical language use. For example, revoice the student statement "The graph goes up" as "y increases as x increases."
Advances: Speaking

Lesson Synthesis

Here are some questions for discussion.

- "What might a scatter plot look like when its line of best fit has a correlation coefficient of 0.9? Sketch it." (It looks like points that follow a linear model very closely. The linear model has a positive slope.)
- "What does a scatter plot look like when its line of best fit has a correlation coefficient of -0.5? Sketch it." (It looks like a loosely scattered cloud of data that trends downward from left to right.)
- "One line of best fit has a correlation coefficient of 0.88, and the other line of best fit has a correlation coefficient of -0.88. Han claims that the one with a positive correlation coefficient fits its data better. Is Han correct? Explain your reasoning." (Han is probably not correct. The sign of the correlation coefficient tells us about the relationship between the variables—not the fit of the data. The positive correlation coefficient just means that as x increases, y also tends to increase. The residuals should also be examined in both cases to determine which data is better fit by a linear model.)
- "Why is it important to know the correlation coefficient for a linear model?" (The correlation coefficient is a way to measure the strength and direction of a linear relationship.)

7.4 What Is a Correlation Coefficient?

5 min

Cool-down

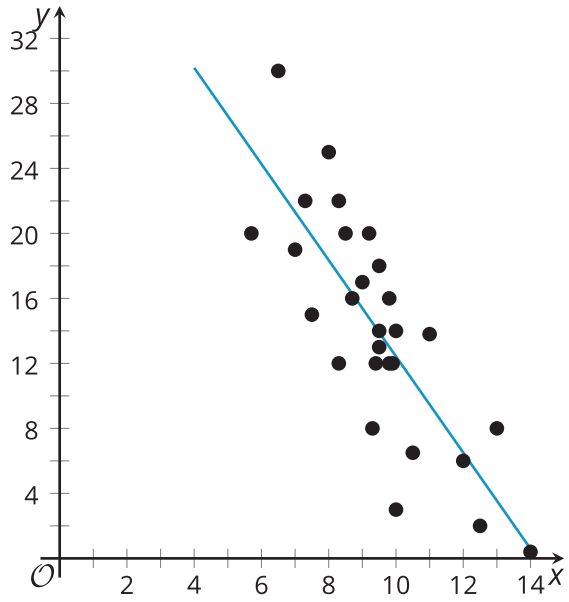
Standards

Addressing **HSS-ID.C.8**

Student Task Statement

1. What information does a correlation coefficient tell us about the data in a scatter plot?

2. Which value best estimates the value for the correlation coefficient of the scatter plot: -1, -0.8, -0.2, 0.2, 0.8, or 1? Explain your reasoning.



Student Response

Sample response:

1. The sign of the correlation coefficient matches the sign of the slope of the best-fit line. The closer the correlation coefficient value is to 0, the worse the fit of the best-fit line. The closer the correlation coefficient is to 1 or -1, the better the best-fit line fits the data.
2. -0.8, since the data appears to be decreasing and a line is an okay fit for the data, but not perfect

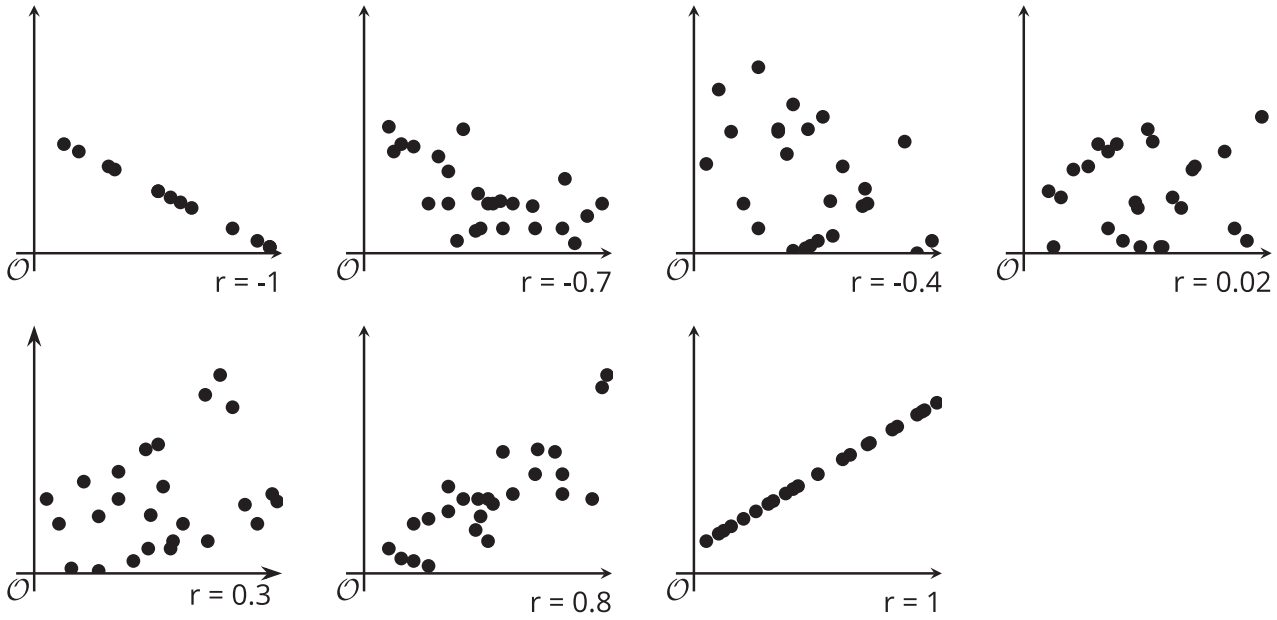
Responding to Student Thinking

More Chances
Students will have more opportunities to understand the mathematical ideas addressed here. There is no need to slow down or add additional work to the next lessons.

Lesson 7 Summary

While residuals can help pick the best-line to fit the data among all lines, we still need a way to determine the strength of a linear relationship. Scatter plots of data that are close to the best-fit line are better modeled by the line than are scatter plots of data that are farther from the line.

The **correlation coefficient** is a convenient number that can be used to describe the strength and direction of a linear relationship. Usually represented by the letter *r*, the correlation coefficient can take values from -1 to 1. The sign of the correlation coefficient is the same as the sign of the slope for the best-fit line. The closer the correlation coefficient is to 0, the weaker the linear relationship. The closer the correlation coefficient is to 1 or -1, the better a linear model fits the data.



While it is possible to try to fit a linear model to any data, we should always look at the scatter plot to see if there is a possible linear trend. The correlation coefficient and residuals can also help determine whether the linear model makes sense to use to estimate the situation. In some cases, another type of function might be a better fit for the data, or the two variables we are examining may be uncorrelated, and we should look for connections using other variables.

Glossary

- correlation coefficient

Lesson 7 Practice Problems

1 Student Task Statement

Select **all** the values for r that indicate a positive slope for the line of best fit.

A. 1

B. -1

C. 0.5

D. -0.5

E. 0

F. 0.8

G. -0.8

Solution

A, C, F

2 Student Task Statement

The correlation coefficient, r, is given for several different data sets. Which value for r indicates the strongest correlation?

A. 0.01

B. -0.34

C. -0.82

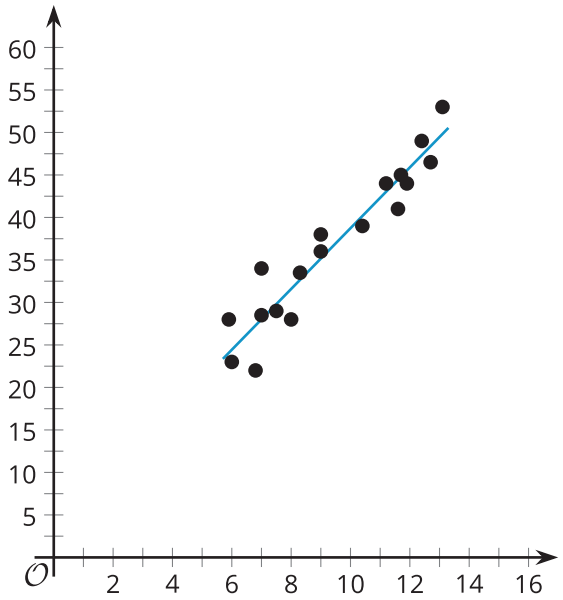
D. -0.95

Solution

D

3 Student Task Statement

Which of the values is the best estimate of the correlation coefficient for the line of best fit for the data represented by the scatter plot?



A. -0.9

B. -0.4

C. 0.4

D. 0.9

Solution

D

4 from Unit 3, Lesson 5

Student Task Statement

Technology required.

A study investigated the relationship between the amount of daily food waste measured in pounds and the number of people in a household. The data in the table displays the results of the study.

number of people in household, x	food waste (pounds), y
2	3.4
3	2.5
4	8.9
4	4.7
4	3.5
4	4
5	5.3
5	4.6
5	7.8
6	3.2
8	12

Use graphing technology to create the line of best fit for the data in the table.

a. What is the equation of the line of best fit for this data? Round numbers to two decimal places.

b. What is the slope of the line of best fit? What does it mean in this situation? Is this realistic?

c. What is the y-intercept of the line of best fit? What does it mean in this situation? Is this realistic?

Solution

a. $y = 1.22x - 0.09$

b. Sample response: The slope of the line is 1.22, which means that for every additional person in a household, an additional 1.22 pounds of food is wasted. This is realistic, as we expect that more people will increase the amount of food waste.

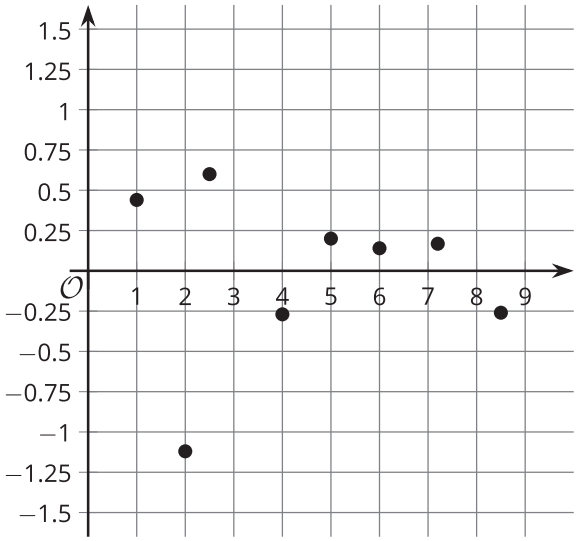
c. Sample response: The y-intercept of the line is $(0, -0.09)$, which means that if there are 0 people in a household, there is -0.09 pound of food wasted. This is not exactly realistic because if a household had no one, there would not be anyone to create food waste. On the other hand, -0.09 is very close to 0, so it is not that far off.

5 from Unit 3, Lesson 6

Student Task Statement

A table of values and the plot of the residuals for the line of best fit are shown.

x	y
1	10
2	8
2.5	9.5
4	8
5	8
6	7.5
7.2	7
8.5	6



a. Which point does the line estimate the best?

b. Which point does the line estimate the worst?

Solution

a. $(6, 7.5)$

b. $(2, 8)$

6

from Unit 3, Lesson 6

Student Task Statement

Tyler creates a scatter plot that displays the relationship between the grams of food a hamster eats, x, and the total number of rotations that the hamster's wheel makes, y. Tyler creates a line of best fit and finds that the residual for the point $(1.4, 1250)$ is -132. The point $(1.2, 1364)$ has a residual of 117. Interpret the meaning of 117 in the context of the problem.

Solution

Sample response: The point $(1.2, 1364)$ represents 117 rotations more than the total number of rotations estimated by the line of best fit.