Lesson 4: Fitting a line to data

Goals

- Compare and contrast (orally) values in a data set with predictions made using a given line.
- Comprehend that a model of data, such as a line of fit, can be used to predict values that are not given in the data.
- Identify (orally) obvious outliers on a scatter plot.

Learning Targets

- I can pick out outliers on a scatter plot.
- I can use a model to predict values for data.

Lesson Narrative

In the previous lesson, students focused primarily on the details of a scatter plot. In this lesson, their focus becomes more holistic and they begin to see a set of data points as a single thing that can be analysed, not just a bunch of disconnected points. For the first time students see that sometimes we can model the relationship between two variables with a line, although they continue to analyse the connections between the scatter plot and the line by comparing individual points. As they zoom out and see these connections, students begin to see the structure of the scatter plot and to use that structure to reason abstractly and quantitatively.

Addressing

- Construct and interpret scatter plots for bivariate measurement data to investigate patterns of association between two quantities. Describe patterns such as clustering, outliers, positive or negative association, linear association, and nonlinear association.
- Know that straight lines are widely used to model relationships between two quantitative variables. For scatter plots that suggest a linear association, informally fit a straight line, and informally assess the model fit by judging the closeness of the data points to the line.

Instructional Routines

- Discussion Supports
- Notice and Wonder
- Think Pair Share

Student Learning Goals

Let’s look at the scatter plots as a whole.

4.1 Predict This

Warm Up: 5 minutes

In previous lessons, students were asked to interpret individual points in a scatter plot or compare two points with each other. In this warm-up, students are asked think about a point in the context of the entire data set. They compare two different possible predictions for the dependent variable, given a value for the independent variable. This sets them up to understand and interpret a linear model for the relationship between independent and dependent variables. They use the potential location of a point not included in the scatter plot to answer a question about the context, based on the visual structure of the scatter plot.

Instructional Routines

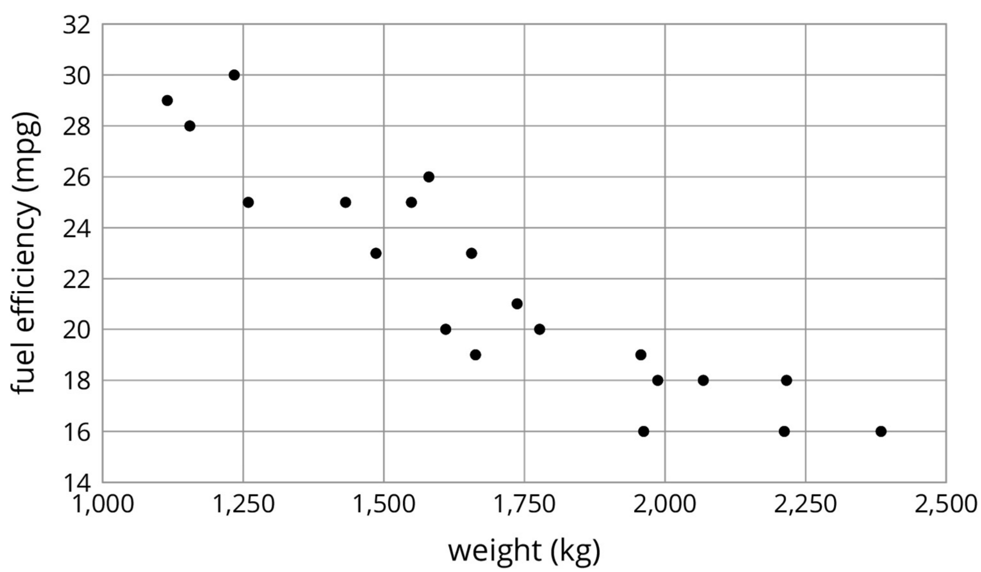
- Think Pair Share

Launch

Arrange students in groups of 2. Give 1 minute of quiet work time followed by 1 minute to check their solution with their partner. Follow with a whole-class discussion.

Student Task Statement

Here is a scatter plot that shows weights and fuel efficiencies of 20 different types of cars.



If a car weighs 1 750 kg, would you expect its fuel efficiency to be closer to 22 mpg or to 28 mpg? Explain your reasoning.

Student Response

We expect the fuel efficiency to be closer to 22 mpg. There are several cars close to that weight, and their fuel efficiency is between 18 and 22 mpg. The cars that have a fuel efficiency close to 28 mpg have a weight less than 1 250 kg, although there is one car that has a weight greater than 1 500 kg and a fuel efficiency of 26 mpg. But that one has a higher fuel efficiency than any other car with a weight between 1 500 and 1 750 kg.

Activity Synthesis

Display the graph for all to see. Poll the class to see if they think the fuel efficiency is closer to 22 mpg or 28 mpg. If they are all in agreement that the answer is closer to 22 mpg, ask a few students to share their reasoning. If there is disagreement, ask students to share their reasoning and come to an agreement. If it does not come up in the discussion, ask students to look at cars whose fuel efficiency is close to 28 mpg and note that their weights are quite a bit less. Then look at cars with a weight close to 1 750 kg, and note that their fuel efficiency is between 18 and 22 mpg. As a whole class, decide where to plot both potential points, and point out that one is close to the other nearby values and one is very far away.

4.2 Shine Bright

15 minutes (there is a digital version of this activity)

In this lesson, the meaning of the words *model* and *modelling* are explained in terms of a linear model. Students are not expected to define the words, but should be comfortable understanding and using them. A model is used to predict prices for diamonds not included in the data as well as compare existing data points to the model.

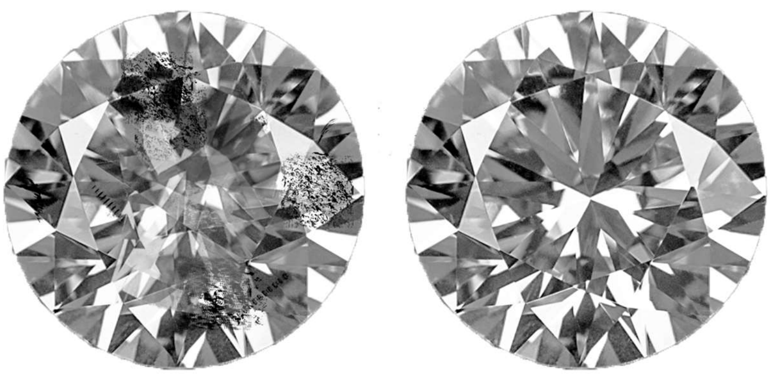
Instructional Routines

- Discussion Supports
- Think Pair Share

Launch

Keep students in groups of 2. Allow students 5 minutes quiet work time followed by partner and whole-class discussion.

Show the first image for all to see. Ask students, “Which diamond do you think has the highest price? Which diamond do you think has the lowest price?” Most likely, they will guess that the largest has the highest price and the smallest has the lowest price. Validate this intuition and let them know that the size of a diamond is *one* thing that its price is based on. Tell students that the size of a diamond is commonly measured in *carats*, which is a measure of weight. (One carat is equivalent to 200 milligrams.)



Show the second image for all to see. Ask students, “Which diamond do you think has the higher price?” Poll the class, and ask a few students to share their reasoning. Tell students that these diamonds are the same size, but they are not the same price. The price of a diamond is not just based on size, but also on how many flaws are inside the stone. Since the diamond on the right has fewer flaws (which makes it more sparkly), it has a higher price than the diamond on the left.

Tell students, “You will look at some data for the prices of diamonds of different sizes. You have used mathematics to analyse real-world situations, identifying variables in a situation and describing their relationships mathematically. This process is called *modelling,* and the mathematical description is called a *model*. Sometimes you made assumptions about the situation or ignored some features so that the model would be simpler.”

Action and Expression: Develop Expression and Communication. Invite students to talk about their ideas with a partner before writing them down. Display sentence frames to support students when they explain their ideas. For example, “It looks like…” and “We are trying

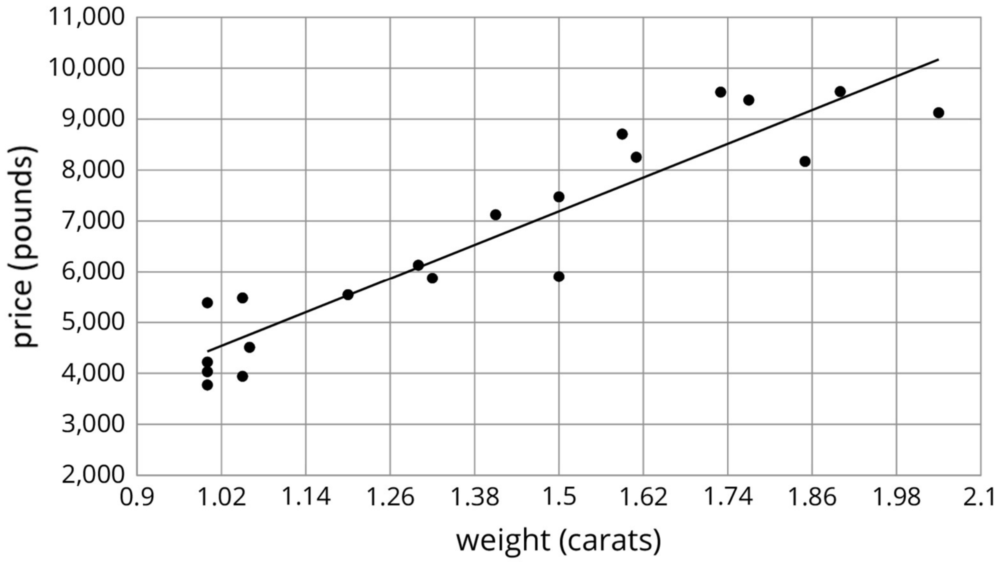
to…”
Supports accessibility for: Language; Organisation Speaking: Discussion Supports. Provide sentence frames to support students as they work with their partner and explain their reasoning. For example, “I think ____ because _____ .” or “I (agree/disagree) because ______ .”
Design Principle(s): Support sense-making; Optimise output for (explanation)

Student Task Statement

Here is a table that shows weights and prices of 20 different diamonds.

weight (carats)	actual price (pounds)	predicted price (pounds)
1	3 772	4 429
1	4 221	4 429
1	4 032	4 429
1	5 385	4 429
1.05	3 942	4 705
1.05	4 480	4 705
1.06	4 511	4 760
1.2	5 544	5 533
1.3	6 131	6 085
1.32	5 872	6 195
1.41	7 122	6 692
1.5	7 474	7 189
1.5	5 904	7 189
1.59	8 706	7 686
1.61	8 252	7 796
1.73	9 530	8 459
1.77	9 374	8 679
1.85	8 169	9 121
1.9	9 541	9 397
2.04	9 125	10 170

The scatter plot shows the prices and weights of the 20 diamonds together with the graph of $y = 5\,520x - 1\,091$.



The function described by the equation $y = 5\,520x - 1\,091$ is a *model* of the relationship between a diamond's weight and its price.

This model *predicts* the price of a diamond from its weight. These predicted prices are shown in the third column of the table.

1. Two diamonds that both weigh 1.5 carats have different prices. What are their prices? How can you see this in the table? How can you see this in the graph?

2. The model predicts that when the weight is 1.5 carats, the price will be £7 189. How can you see this in the graph? How can you see this using the equation?

3. One of the diamonds weighs 1.9 carats. What does the model predict for its price? How does that compare to the actual price?

4. Find a diamond for which the model makes a very good prediction of the actual price. How can you see this in the table? In the graph?

5. Find a diamond for which the model's prediction is not very close to the actual price. How can you see this in the table? In the graph?

Student Response

1. £7 474 and £5 904. You can see this in the two rows in the table that show 1.5 in the first column. You can see this in the graph by finding the two points with $x = 1.5$.

2. The point $(1.5, 7\,189)$ is on the graph of the line. We can substitute $x = 1.5$ into the equation to get $y = 5\,520 \times 1.5 - 1\,091 = 7\,189$.

3. The model predicts that the price will be £9 397, because $y = 5\,520 \times 1.9 - 1\,091 = 9\,397$. The actual price is £9 541. The actual price is £144 greater than the prediction, because $9\,541 - 9\,397 = 144$.

4. The best prediction is made for the diamond with a weight of 1.2 carats—the prediction is only off by £11. You can see this in the table in the row with 1.2 in the first column. You can see that the corresponding point in the scatter plot is very close to the line.

5. Answers vary. Sample response: The prediction for one of the diamonds that weighs 1.5 carats is off by over £1 000. You can see this in the second row with a 1.5 in the first column. You can see that the corresponding point in the scatter plot is far below the line.

Activity Synthesis

The goal of this discussion is to help students understand the relationship between the data and a linear model of the data.

To highlight their relationship, ask:

- "What does a point in the scatter plot represent?" (The *actual* weight and price of a diamond.)
- "What does the line represent?" (The *predicted* price of a diamond based on its weight.)
- "What does it mean when a point is close to the line? When it is far away from the line?" (When it is close, the model predicts the price well. The further away in the vertical direction, the worse the prediction.)
- "How can you use the graph to predict the price of a diamond that weighs 1.1 carats? How can you use the equation?" (Find the point on the line that lines up with $x = 1.1$ or substitute $x = 1.1$ into the model equation.)

4.3 The Agony of the Feet

15 minutes (there is a digital version of this activity)

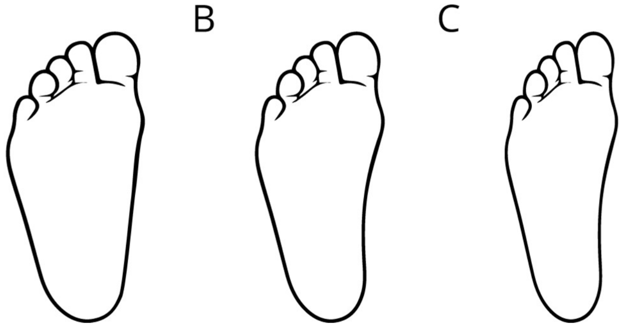
A scatter plot is shown and the points interpreted in context. Later, a linear model is graphed with the scatter plot to help students see an obvious outlier. In the discussion, the term outlier is introduced.

Instructional Routines

- Discussion Supports
- Notice and Wonder

Launch

A B C



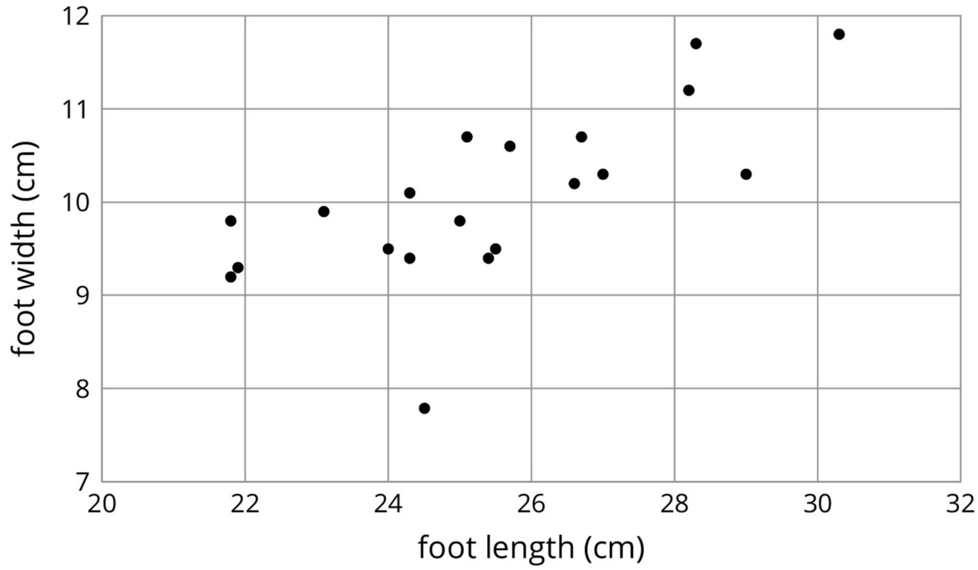
Ask students, "What do you notice? What do you wonder?" Ideally, they will notice that these are pictures of a foot of three different people. All three feet are approximately the same length but different widths. Tell students that all of these feet are a size 8. However, they wouldn't all necessarily find the same shoe equally comfortably, because of the varying widths. (Besides the numerical size, some shoes also come in different widths.) Students should understand that human feet can vary in both length and width.

If using the digital activity, after this discussion, clarify for students that a scatter plot has been created for them to analyse feet dimensions, using technology. Students can then work to complete the digital task which includes using a line of best fit to identify outliers.

Speaking: Discussion Supports. Use this routine to clarify the meaning of "width" and "length" in the context of the task. Connect the terms width and length multi-modally by utilising different types of sensory inputs, such as demonstrating the width and length on the images of each foot using gestures or inviting students to do so. Invite students to chorally repeat phrases that include these words in context. This will help students produce and make sense of the language needed to understand the context of the task.
Design Principle(s): Support sense-making

Student Task Statement

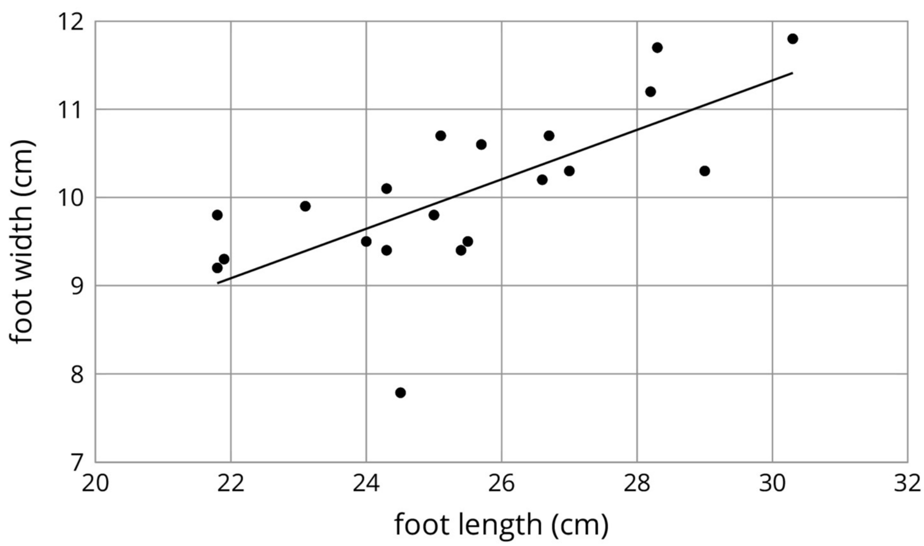
Here is a scatter plot that shows lengths and widths of 20 different left feet.



1. Estimate the widths of the longest foot and the shortest foot.

2. Estimate the lengths of the widest foot and the narrowest foot.

3. Here is the same scatter plot together with the graph of a model for the relationship between foot length and width.



Circle the data point that seems weird when compared to the model. What length and width does that point represent?

Student Response

1. Longest has a width of about 11.9 cm, shortest has a width of about 9.3 cm

2. Widest has a length of about 30.4 cm, narrowest has a length of about 24.5 cm

3. The point that "seems weird" represents length approximately 24.5 cm and width about 7.8 cm

Activity Synthesis

Introduce the term **outlier**. An outlier is a point that is separated from the rest of the data. Sometimes data sets have outliers. Sometimes that's because there really is a data point that is very different than the others. Sometimes it is because there was an error in collecting the data. Sometimes it is because there was an error in entering the data. When there are outliers, one has to make a judgement about whether to include it in the analysis or not.

Lesson Synthesis

To help students see the connection between a scatter plot and a linear model for a data set, ask:

- "What kind of *model* for a data set did we investigate today?" (A linear model.)
- "What does this kind of model help us do?" (See the trend in the data more clearly and make predictions.)
- "What does it mean when a data point is closer in the vertical direction to the line that represents a linear model? What does it mean when a data point is farther from the line in the vertical direction?" (When the data point is closer in the vertical direction to the line, it represents data that fits the prediction well. When the point is farther from the line, the data does not fit the prediction well.)
- "In your own words, how can you identify an outlier from a scatter plot?" (A point that is far from the other points in the scatter plot represents an outlier.)

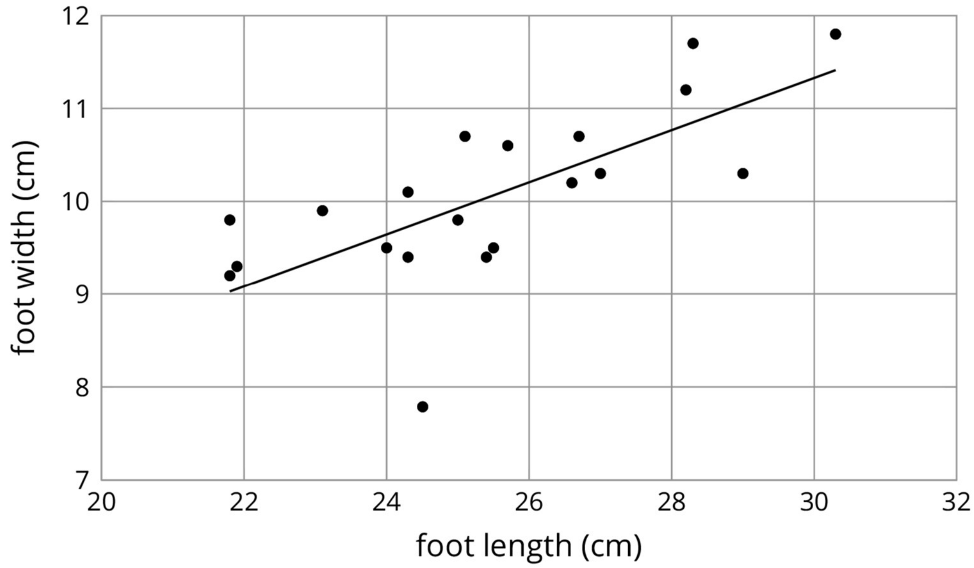
4.4 A 1 Foot Foot

Cool Down: 5 minutes

Given a scatter plot, students find the point that is closest to a given value and then compare this point to the predicted value based on a linear model.

Student Task Statement

Here is a scatter plot that shows lengths and widths of 20 left feet, together with the graph of a model of the relationship between foot length and width.



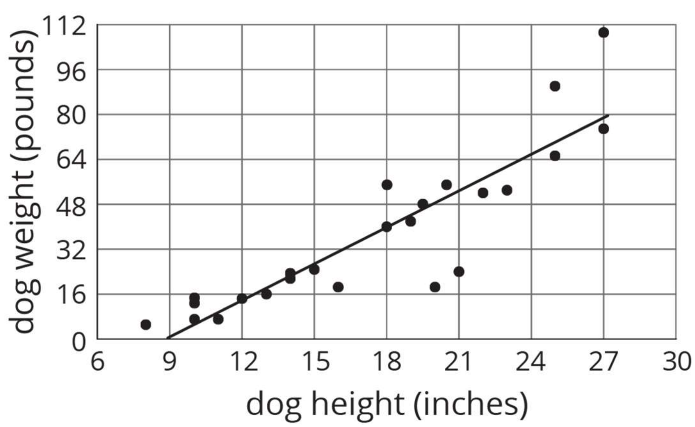
1. Draw a box around the point that represents the foot with length closest to 29 cm.
2. What is the approximate width of this foot?
3. What width does the model predict for a foot with length 29 cm?

Student Response

1. A box is drawn around the point at approximately $(29.1, 10.4)$.
2. About 10.4 cm
3. About 11 cm (or a little more than 11 cm)

Student Lesson Summary

Sometimes, we can use a linear function as a model of the relationship between two variables. For example, here is a scatter plot that shows heights and weights of 25 dogs together with the graph of a linear function which is a model for the relationship between a dog's height and its weight.



We can see that the model does a good job of predicting the weight given the height for some dogs. These correspond to points on or near the line. The model doesn't do a very good job of predicting the weight given the height for the dogs whose points are far from the line.

For example, there is a dog that is about 20 inches tall and weighs a little more than 16 pounds. The model predicts that the weight would be about 48 pounds. We say that the model *overpredicts* the weight of this dog. There is also a dog that is 27 inches tall and weighs about 110 pounds. The model predicts that its weight will be a little less than 80 pounds. We say the model *underpredicts* the weight of this dog.

Sometimes a data point is far away from the other points or doesn't fit a trend that all the other points fit. We call these **outliers**.

Glossary

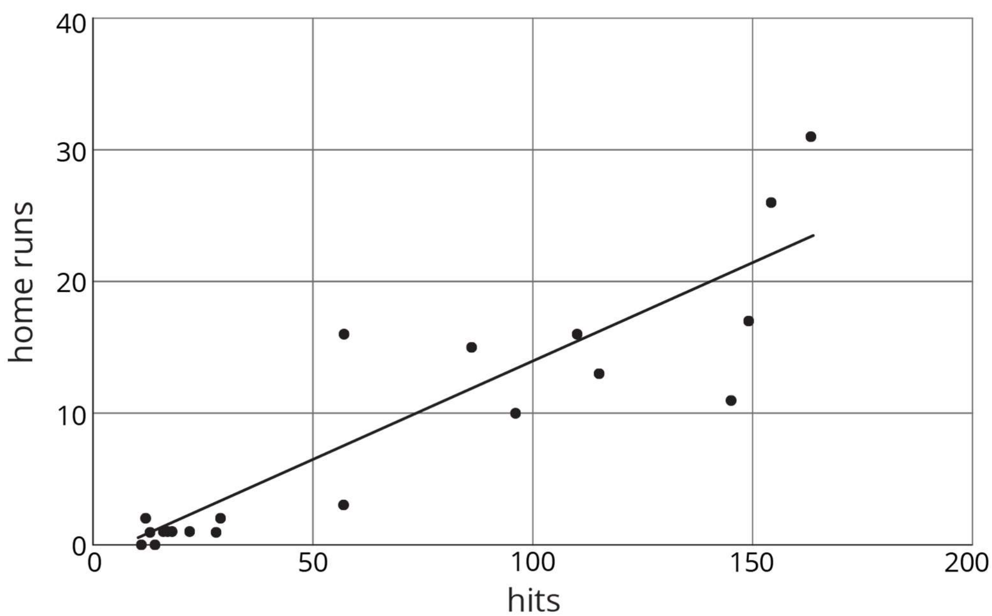
- outlier

Lesson 4 Practice Problems

Problem 1 Statement

The scatter plot shows the number of hits and home runs for 20 baseball players who had at least 10 hits last season. The table shows the values for 15 of those players.

The model, represented by $y = 0.15x - 1.5$, is graphed with a scatter plot.



Use the graph and the table to answer the questions.

a. Player A had 154 hits in 2015. How many home runs did he have? How many was he predicted to have?

b. Player B was the player who most outperformed the prediction. How many hits did player B have last season?

c. What would you expect to see in the graph for a player who hit many fewer home runs than the model predicted?

hits	home runs	predicted home runs
12	2	0.3
22	1	1.8
154	26	21.6
145	11	20.3
110	16	15
57	3	7.1
149	17	20.9
29	2	2.9
13	1	0.5
18	1	1.2
86	15	11.4
163	31	23
115	13	15.8
57	16	7.1
96	10	12.9

Solution

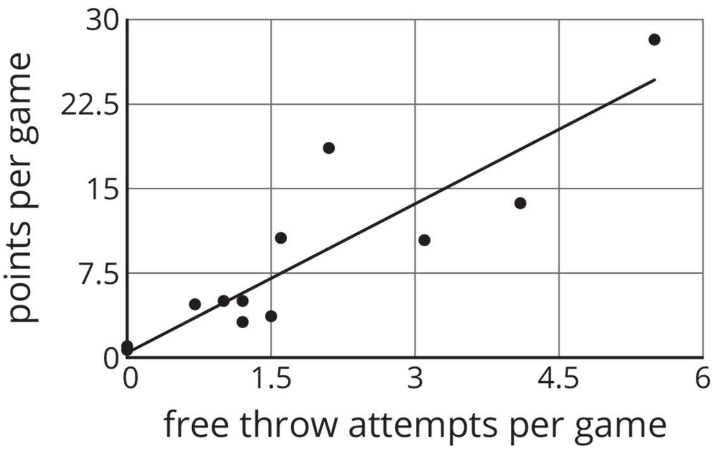
a. Home runs: 26. Predicted home runs: 21.6

b. 57

c. The point should be much lower on the graph than the line.

Problem 2 Statement

Here is a scatter plot that compares points per game to free throw attempts per game for basketball players in a tournament. The model, represented by $y = 4.413x + 0.377$, is graphed with the scatter plot. Here, x represents free throw attempts per game, and y represents points per game.



a. Circle any data points that appear to be outliers.

b. What does it mean for a point to be far above the line in this situation?

c. Based on the model, how many points per game would you expect a player who attempts 4.5 free throws per game to have? Round your answer to the nearest tenth of a point per game.

d. One of the players scored 13.7 points per game with 4.1 free throw attempts per game. How does this compare to what the model predicts for this player?

Solution

a. Circle the point at (2.1, 18.6).

b. A point above the line represents a player who scores more points per game than predicted by their number of free throw attempts.

c. 20.2 points per game, because $4.413(4.5) + 0.377$ is roughly equal to 20.2.

d. The model predicts that with 4.1 free throw attempts per game, the player should score $4.413(4.1) + 0.377$, or about 18.5 points per game. That means the player is scoring less than the model predicts they should.