



# PROJECT-SET

Statistics Education for Teachers

## What Fits?

### Overview of lesson

This activity has students explore what a line of best fit is and where a line of best fit should be placed in a scatterplot of data with a linear association. This is done through students' completion of an activity investigating the relationship between the drop height of a golf ball and the height of its first bounce, as well as analysis of other students' work on the same activity. Students will learn what the line of best fit means and appropriate criteria for its placement.

### GAISE Components

This activity follows all of the four components of statistical problem solving put forth in the Guidelines for Assessment and Instruction in Statistics Education (GAISE) Report. This is a Level B activity.

### Common Core State Standards for Mathematical Practice

3. Construct viable arguments and critique the reasoning of others.
4. Model with mathematics.
7. Look for and make use of structure.

### Common Core State Standards

#### Grade 8.SP: Investigate patterns of association in bivariate data

1. 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.
2. 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.
3. Use the equation of a linear model to solve problems in the context of bivariate measurement data, interpreting the slope and intercept.

High School: S-ID: Summarize, represent, and interpret data on two categorical and quantitative variables

6. Represent data on two quantitative variables on a scatter plot, and describe how the variables are related.

- a. Fit a function to the data; use functions fitted to data to solve problems in the context of the data. Use given functions or choose a function suggested by the context. Emphasize linear, quadratic, and exponential models.
- b. Informally assess the fit of a function by plotting and analyzing residuals.
- c. Fit a linear function for a scatterplot that suggests a linear association.

High School: S-ID: Interpret linear models.

7. Interpret the slope (rate of change) and the intercept (constant term) of a linear model in the context of the data.

**NCTM Principles and Standards for School Mathematics**

**Data Analysis and Probability Standards for Grades 6-8**

**Formulate questions that can be addressed with data and collect, organize, and display relevant data to answer them**

- formulate questions, design studies, and collect data about a characteristic shared by two populations or different characteristics within one population
- select, create, and use appropriate graphical representations of data, including histograms, box plots, and scatterplots

**Select and use appropriate statistical methods to analyze data**

- find, use, and interpret measures of center and spread, including mean and interquartile range
- discuss and understand the correspondence between data sets and their graphical representations, especially histograms, stem-and-leaf plots, box plots, and scatterplots

**Develop and evaluate inferences and predictions that are based on data**

- make conjectures about possible relationships between two characteristics of a sample on the basis of scatterplots of the data and approximate lines of fit

**Data Analysis and Probability Standards for Grades 9-12**

**Formulate questions that can be addressed with data and collect, organize, and display relevant data to answer them**

- understand histograms, parallel box plots, and scatterplots and use them to display data

**Select and use appropriate statistical methods to analyze data**

- identify trends in bivariate data and find functions that model the data or transform the data so that they can be modeled

### **Prerequisites**

Prior to completing this activity, students should :

- Be familiar with displaying bivariate quantitative data in scatterplots
- Understand a linear function has a constant rate of change
- Be able to find the y-coordinate of a point on a line given the line's equation and the point's x-coordinate
- Be able to find the y-coordinate of a point on a line given the line's graph and the point's x-coordinate
- Determine a line's slope and y-intercept given its graph and/or equation

### **Learning Targets**

Students will be able to:

- interpret scatter plots for bivariate measurement data to investigate patterns of association between two quantities.
- Describe patterns in scatterplots such as clustering, positive or negative association, and linear association.
- Explain why straight lines are used to model relationships between two quantitative variables with a linear association.
- Informally fit a straight line
- Informally assess the model's fit by judging the closeness of the data points to the line
- Analyze criteria for lines of best fit
- Explain what the line of best fit is
- Use the equation of a line of best fit to make a prediction and judge the reasonableness of the prediction
- Interpret the slope and y-intercept of the equation of a line of best fit in context and judge the reasonableness of those interpretations

### **Time Required**

The time required for this activity is roughly 90 minutes.

### **Materials required**

For this activity, the students will need copies of the activity sheets, short pieces of stiff wire (available at hobby or craft stores) or spaghetti, and pencils.

## **Instructional Lesson Plan**

### **The GAISE Statistical Problem-Solving Procedure**

#### **I. Formulate Question(s)**

This activity will be driven by the teacher-directed question: **“What is the relationship between the drop height of a golf ball and the height of its first bounce?”** Teachers should conduct a discussion using the following questions:

- What is our population of interest?
- Do you expect there to be a strong association between these two variables? Why?
- What form do you expect the relationship to follow (i.e., what function family would be used to model this relationship)? Why?
- What direction do you expect the association to take? Why?
- Do you expect a golf ball to bounce back up to its drop height? Lower than its drop height? Higher than its drop height?

An extensive discussion should take place surrounding these questions. A description of correct responses to the above questions is located in the answer key on page 15.

#### **II. Collect Data**

Students are provided with the data in the form of a scatterplot.

#### **III. Analyze the Data By Measures and Graphs**

Students will analyze a scatterplot and place a line of best fit on the graph.

To do this, the teacher will distribute the first page of the activity where students will read the introduction and interpret the scatterplot, comparing the actual graph to one(s) anticipated in phase I. Once students have determined that a linear relationship is most appropriate, then have students individually do as directed in the Part One directions:

Determine where you think the line of best fit should be placed on the scatterplot above by manipulating a piece of spaghetti or wire. Draw your line on the scatterplot once you have determined its best placement.

#### **IV. Interpret the Results**

Students will examine other students' work in order to critique the types of reasoning performed by other students while placing the line of best fit.

1) Lead a class discussion of the line of best fit as determined by the students individually. Include in your discussion:

- What things did students consider when determining the line's placement?
- What were the student-created criteria for determining where to place the line?
- Does the line match the relationship students anticipated in step I: Formulate a question?

- What are some reasons why the data are not perfectly linear?

2) Distribute remaining pages of the activity. Help students understand what they are asked to do in Part Two, as described at the top of page 2:

In the class that collected the data, the teacher gave the students a copy of the scatterplot and asked them each to find the line of best fit for the data. The teacher then asked each student to explain why they put the line where they did, defining the criterion for the line of best fit.

Below are some sample student responses to this task. For each student's response,

- 1) Describe what the student's line of best fit tells them about the relationship between the bounce height of a ball and the height it was dropped from.
- 2) Interpret the slope and y-intercept of the student's line in the context of this data set. Do their interpretations seem reasonable in this context?
- 3) Use the student's line to predict the height the golf ball will bounce to if it dropped from 60 centimeters. Is the prediction reasonable?
- 4) Analyze the student's criterion. Will the criterion always work to produce a line which accurately models any data set? If it will, explain why. If it won't, draw at least one example of a scatterplot with the line placed using that criterion and explain why the criterion produces a bad line of best fit.

Give students time to complete Part Two in groups.

At this point in the lesson, a teacher may help students begin to develop a way to distinguish a "good" trend line from a "bad" one. For example, looking at the sum of the absolute values of the residuals would be one way to begin to do that.

- 3) Have students revisit their work in Part One and make corrections if necessary.
- 4) Have students answer two summary questions, listed at the end of the activity sheet:
  - 1) What is the best criterion to use when determining where the line of best fit is for any data set? Justify your response.
  - 2) Another student asks you "What is the line of best fit?". Write your response to this student.

### **Possible Extensions**

Students can start to explore data sets that have:

- a) Multiple linear trends to them, so are best fit through multiple lines of best fit
- b) Curvature that can be straightened out through linearization techniques (like exponential, which can be transformed to be linear by plotting  $\log(y)$  vs.  $x$ ) after which linear models can be found

## References

Casey, S., & Kaplan, J. (under review). *Statistical knowledge for teaching: The case of informal line of best fit*.

Casey, S., & Wilson, D. (2012). *Student conceptions of the line of best fit*. Research proposal.

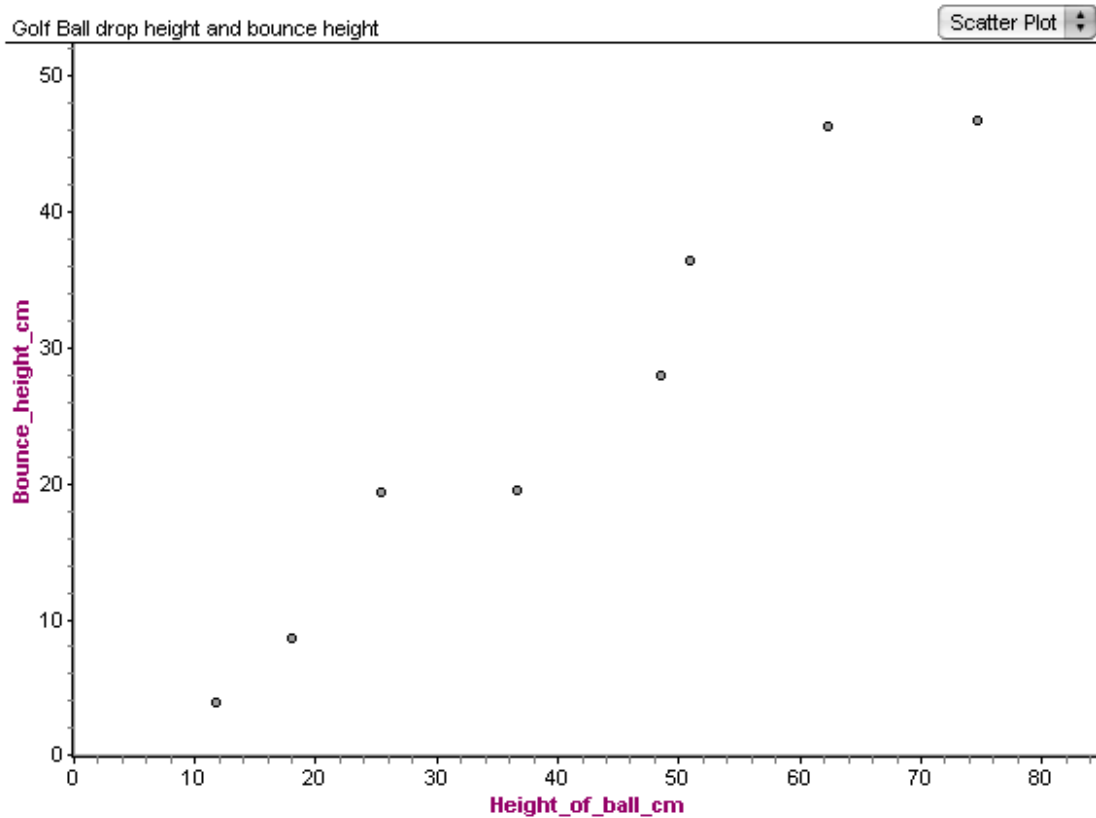
Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2007). *Guidelines for assessment and instruction in statistics education (GAISE) report: A preK–12 curriculum framework*. Alexandria, VA: American Statistical Association.

## Activity

### What Fits?

#### Introduction

An eighth grade class was investigating how high a golf ball bounces when it is dropped from different heights. Students were given eight set heights to drop the ball from. Then they dropped a golf ball from each of those heights and measured how high the ball bounced back up. Below is the scatterplot they made of their data.



#### Part One Instructions

Determine where you think the line of best fit should be placed on the scatterplot above by manipulating a piece of spaghetti or wire. Draw your line on the scatterplot once you have determined its best placement.

Use your line to predict the height the golf ball will bounce to if it is dropped from 60 centimeters. Is the prediction reasonable?

## Part Two Instructions

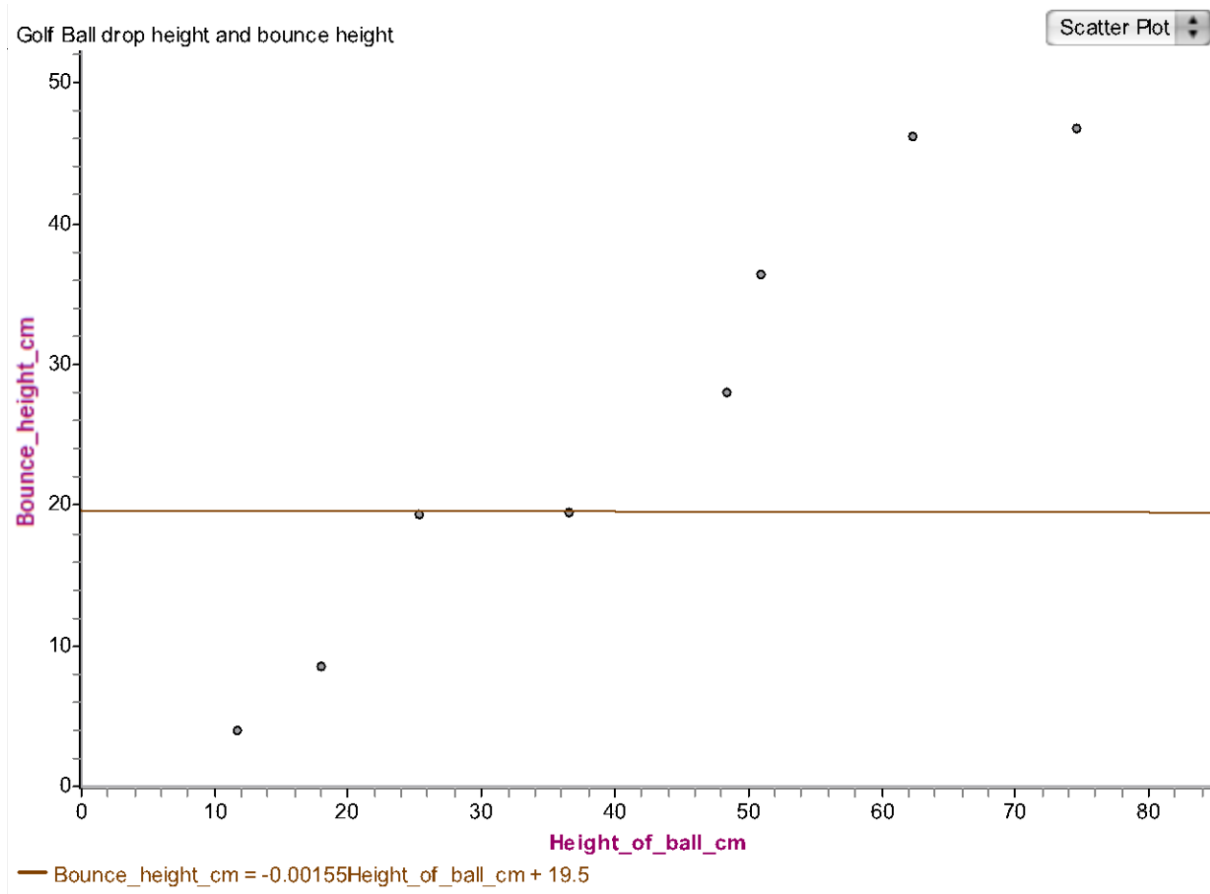
Students were given a copy of the scatterplot and asked to find the line of best fit for the data. Then, students had to explain why they put the line where they did by defining the criterion for the line of best fit.

Below are the responses of some of the students to this task. For each student's response,

- 1) Describe what the student's line of best fit tells them about the relationship between the bounce height of a ball and the height it was dropped from.
- 2) Interpret the slope and y-intercept of the student's line in the context of this data set. Do their interpretations seem reasonable in this context?
- 3) Use the student's line to predict the height the golf ball will bounce to if it dropped from 60 centimeters. Is the prediction reasonable?
- 4) Analyze the student's criterion. Will the criterion always work to produce a line that accurately models any data set? If it will, explain why. If it won't, draw at least one example of a scatterplot with the line placed using that criterion and explain why the criterion produces a bad line of best fit.

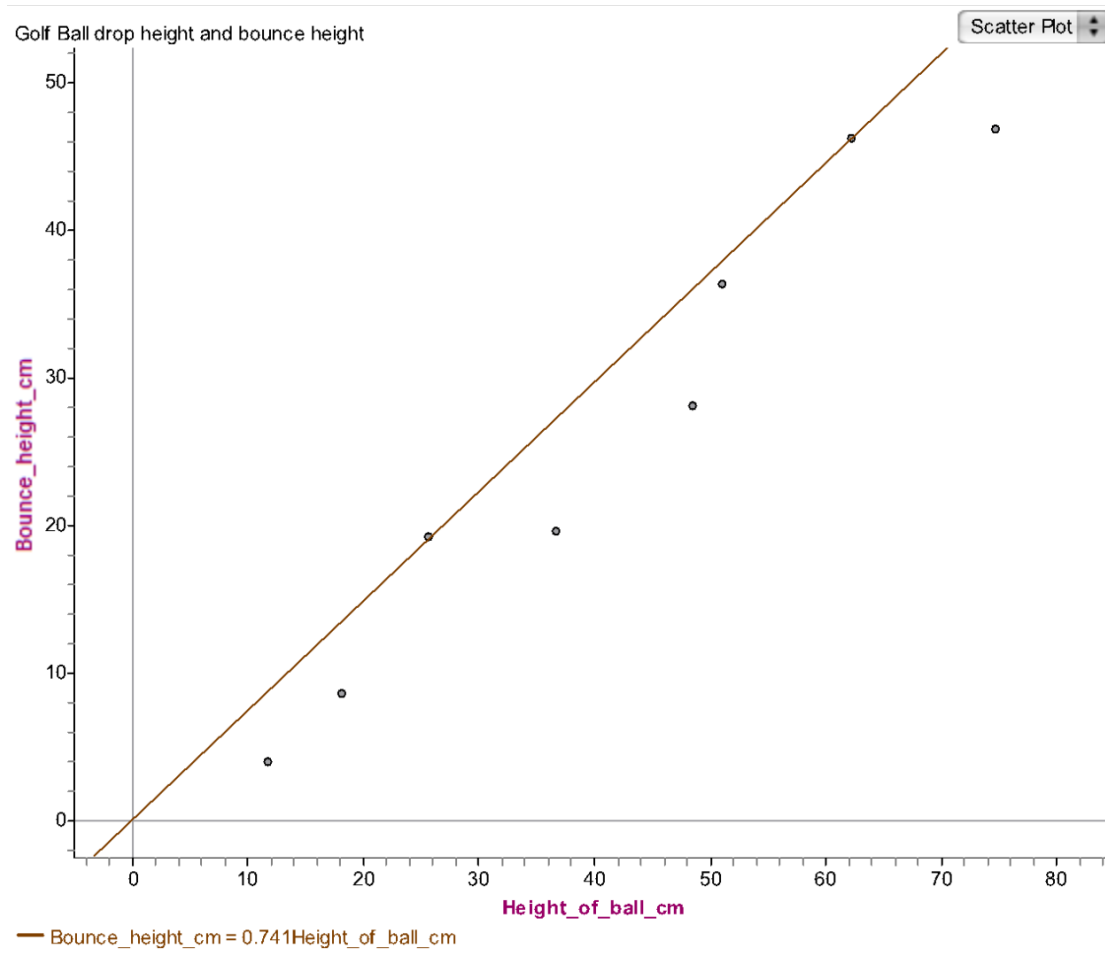


Christina's response



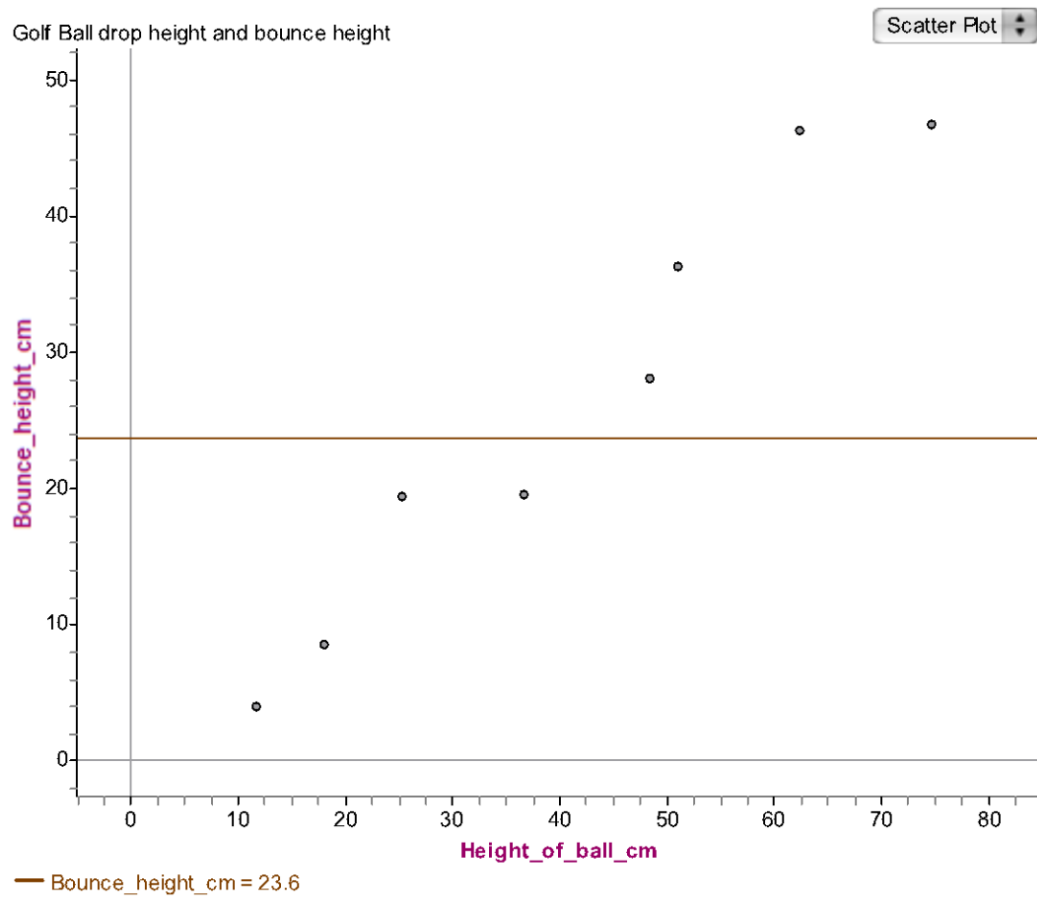
I thought of the line of best fit like the mode, so I put the line through two points with the same y-coordinate because they occur most often.

Edward's response



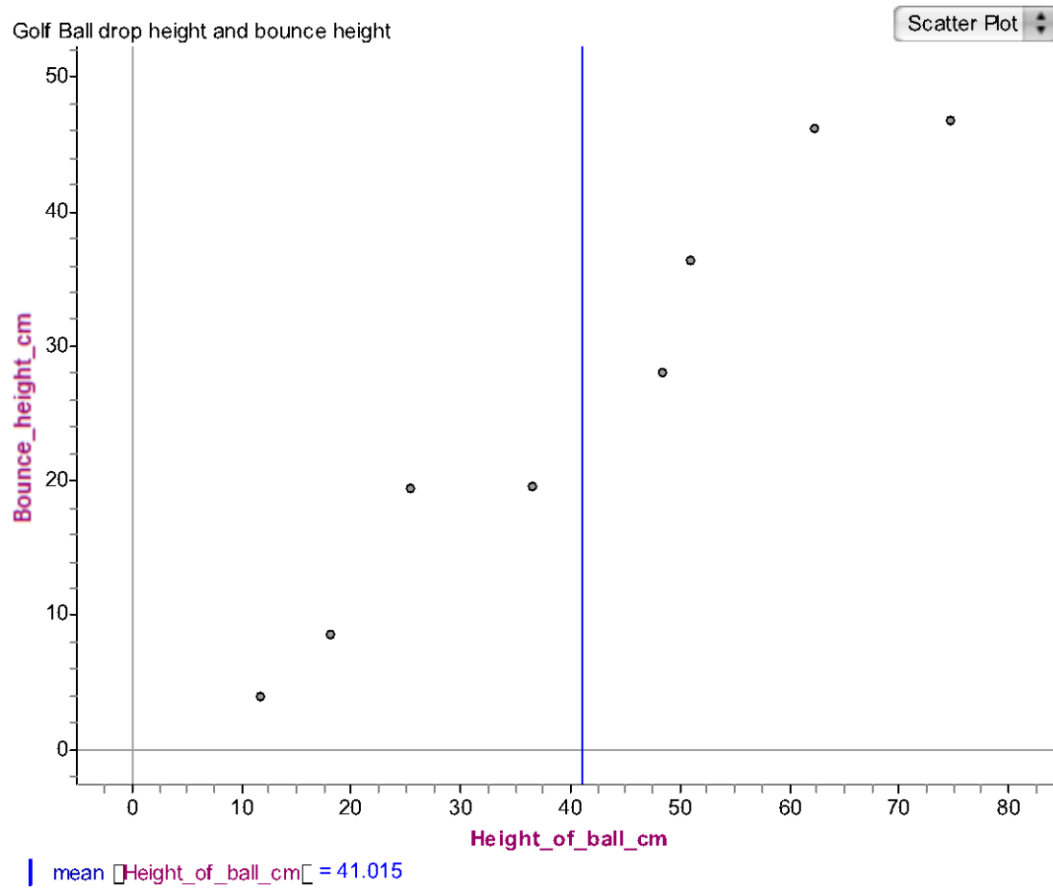
The line needs to start at (0,0) then go through the most dots. I got my line to go through two of the dots so I put it there.

Molly's response



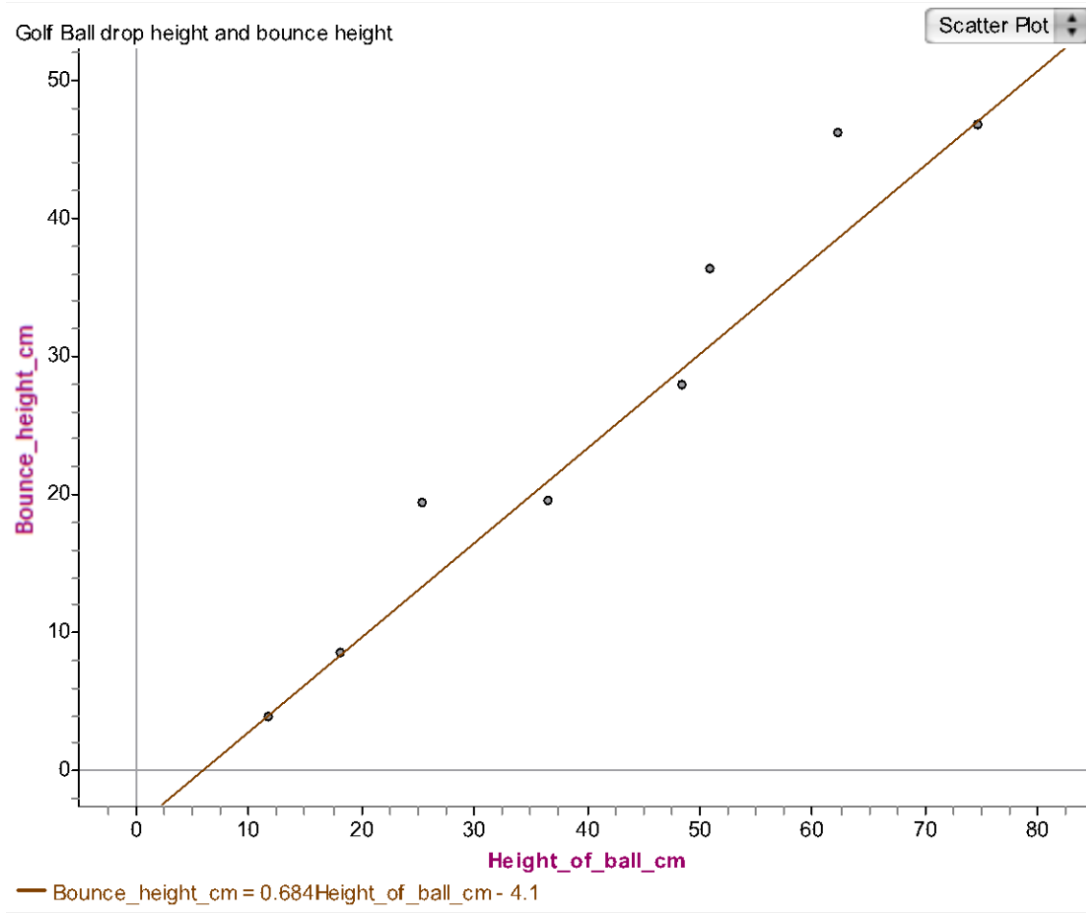
The line should be in the middle of the highest and lowest points because that's like the average.

*Petra's response*



I know the line should go in the middle of the data. I put it here so it would be in the middle, four points on each side.

*Talia's response*



I tried to make my line go through the most dots.

### Summary Questions

- 1) What is the best criterion to use when determining where the line of best fit is for any data set? Justify your response.
- 2) Another student asks you “What *is* the line of best fit?” Write your response to this student.

## Answers to Part I Discussion Questions

This activity will be driven by the teacher-directed question: **“What is the relationship between the drop height of a golf ball and the height of its first bounce?”** Teachers should conduct a discussion using the following questions:

- What is our population of interest?

The population of interest is all golf balls of this type.

- Do you expect there to be a strong association between these two variables? Why?

A strong association is expected as long as the data is collected accurately because there should be a relationship between the height the ball is dropped from and how high it bounces.

- What form do you expect the relationship to follow (i.e., what function family would be used to model this relationship)? Why?

A linear model would make sense because I expect the bounce height to increase at a constant rate.

- What direction do you expect the association to take? Why?

It should have a positive association since it will bounce higher when it is dropped from higher heights.

- Do you expect a golf ball to bounce back up to its drop height? Lower than its drop height? Higher than its drop height?

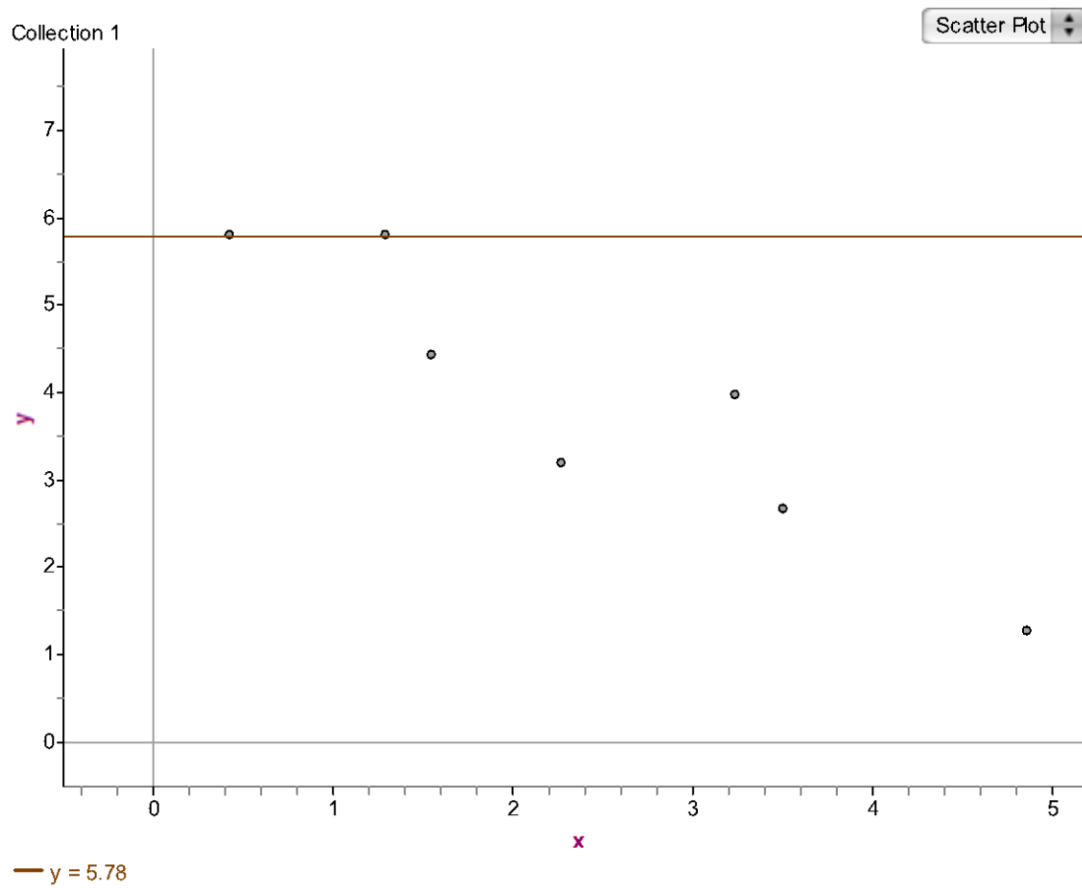
The golf ball will likely bounce up to lower than its drop height.

## Answers to Activity Part Two

Below are possible correct responses.

### Christina

- 5) The ball bounces to the same height no matter how high the ball is dropped from.
- 6) The bounce height of the ball goes down by 0.00155 cm on average when the drop height is increased by 1 cm. This doesn't make sense because it should get higher as you drop the ball from higher heights.  
The ball would bounce to 19.5 cm if it wasn't dropped. This doesn't make sense because it should not bounce if it's not dropped, so the bounce height should be zero.
- 7) 19.4 cm. This is not reasonable because the bounce height for 60 should be between 38 cm and 48 cm, the bounce heights for drop heights of 52 and 65 cm.
- 8) No, this method won't usually work. Below is a scatterplot where the first two points have the same y-value of 5.8. Putting the line through those points produces a line that is horizontal, but the points overall have a clearly negative trend so the line doesn't fit the data well.

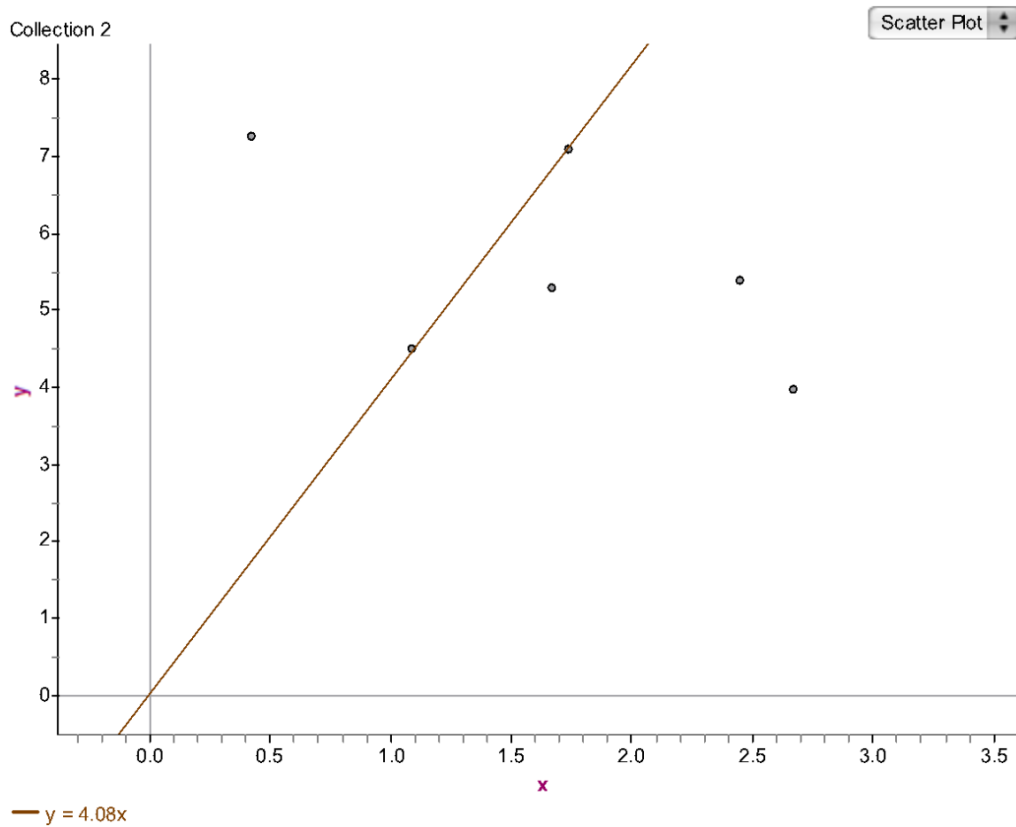


### Edward

- 1) The line shows the bounce height of the ball increases as the drop height increases.
- 2) The bounce height of the ball goes up by 0.741 cm on average when the drop height is increased by 1 cm. This is reasonable because the bounce height increases when the drop height increases, and I would expect the bounce height to increase by less than 1 cm if you increase the drop height by 1 cm.  
The ball would bounce to 0 cm if you dropped it from 0 cm. This makes sense because the ball shouldn't bounce if you don't drop it!
- 3) 44.46 cm. That is reasonable because it's near where the data points for drop heights of 52 and 65 cm are.
- 4) It makes sense to have the line go through (0,0) here because of the context about ball bounce and drop heights, but it won't make sense in every context. Also, going through the most dots won't always work because it's possible the dots that line up aren't in the middle of the cloud of points.



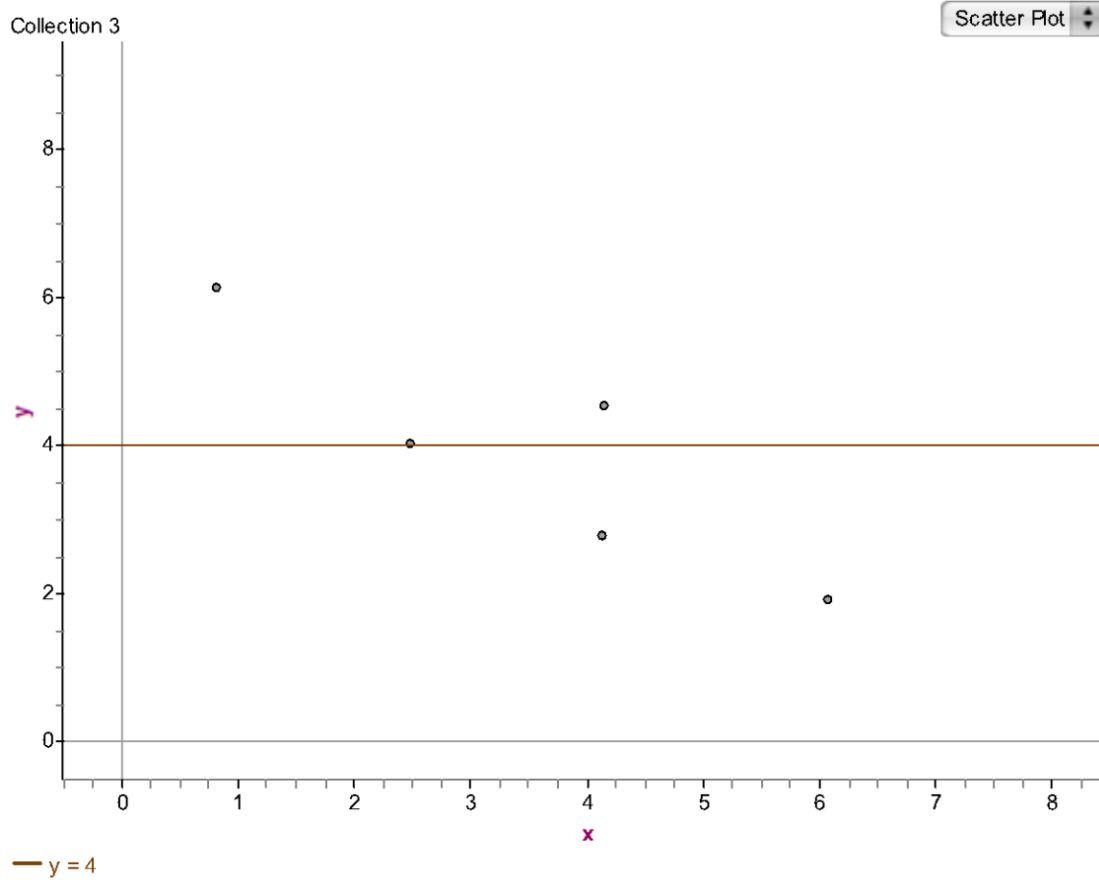
Here's an example of a scatterplot where Edward's criteria doesn't work well. Since the data have a negative association, it doesn't make sense to make the line start at (0,0). Also the line through the most points is going up, when it needs to be going down.



### Molly

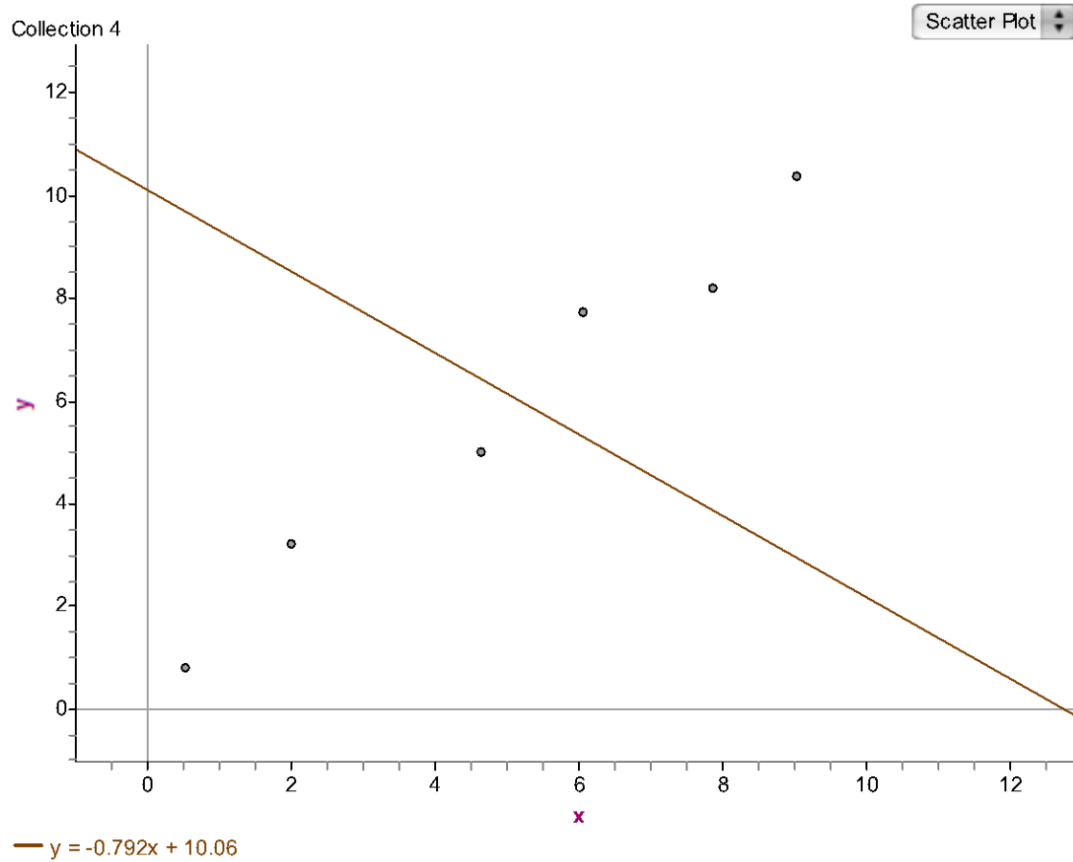
- 1) The ball bounces to the same height no matter how high the ball is dropped from.
- 2) This line has a slope of zero, which means that if the ball is bounced from higher heights the bounce height doesn't change. This doesn't make sense because the ball should bounce higher if it's dropped from higher heights.  
The ball would bounce to 23.6 cm if it wasn't dropped. This doesn't make sense because it should not bounce if it's not dropped, so the bounce height should be zero.
- 3) 23.6 cm. That's not reasonable because it's less than the bounce height of a ball dropped from 50 cm, and it should be greater.
- 4) The student's idea that the line should be in the middle is fine, but the line needs to be in the middle of all the points not just the highest and lowest. It also needs to have the same angle as the data, not horizontal every time.

This scatterplot shows another example of why this criterion is bad.



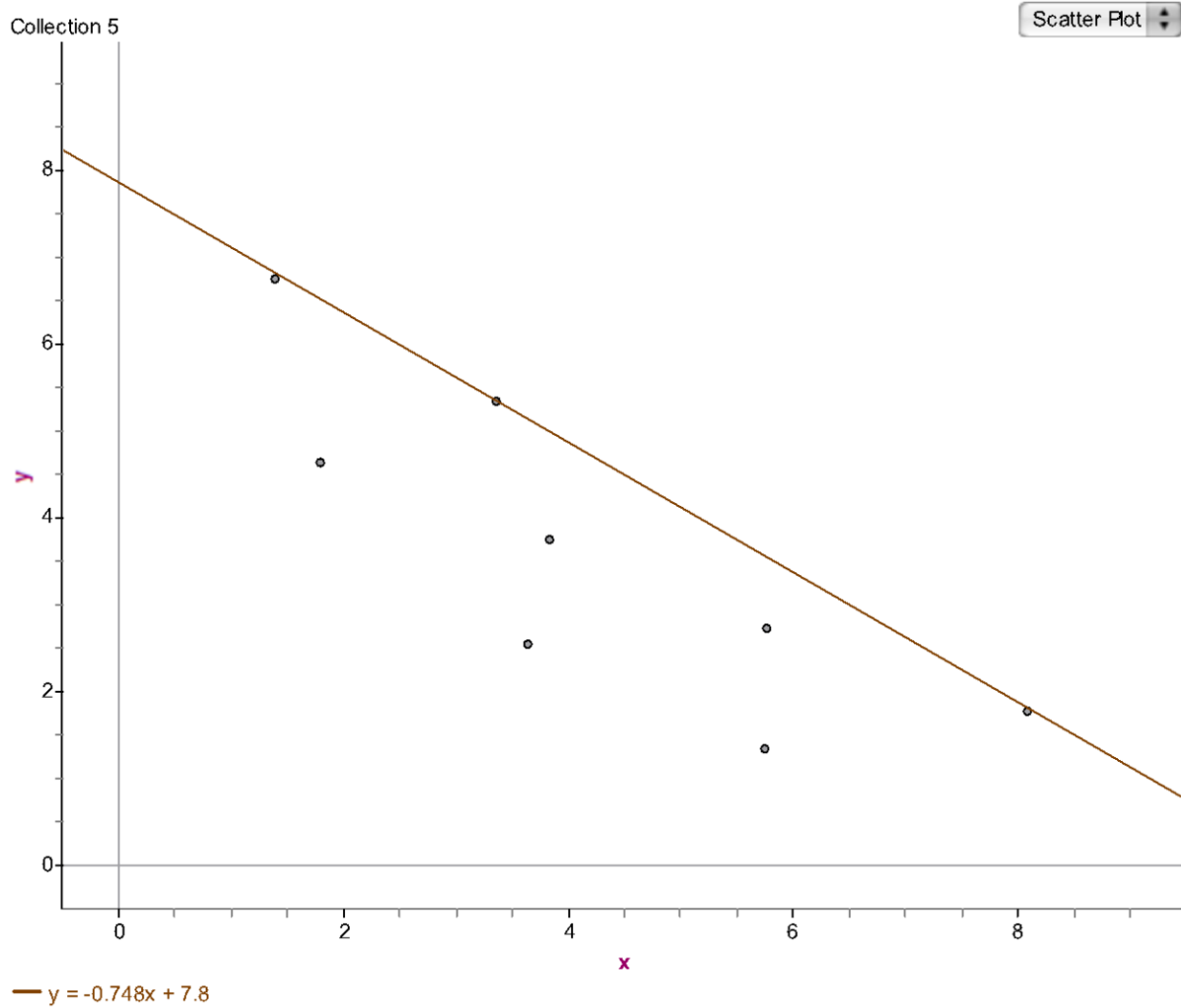
### Petra

- 1) The line says the drop height of the ball is always the same for every bounce height.
- 2) This line's slope is undefined, which means that the drop height of the ball is always 41.105 cm. This doesn't make sense because the ball was and can be dropped from a bunch of different heights, not just 41.105 cm.  
The line doesn't have a y-intercept.
- 3) The line can't be used to predict the bounce height.
- 4) Like Molly, Petra has the right idea that the line should go in the middle of the data, and even the idea that there's the same number of points on each side of the line can work sometimes. But this scatterplot shows that just having that for the criterion doesn't make a good line of best fit.



## Talia

- 1) The line shows the bounce height of the ball increases as the drop height increases.
- 2) The bounce height of the ball goes up by 0.648 cm on average when the drop height is increased by 1 cm. This is reasonable because the bounce height increases when the drop height increases, and I would expect the bounce height to increase by less than 1 cm if you increase the drop height by 1 cm.  
The ball should bounce -4.1 cm if dropped from height zero, which means the ball isn't dropped. This is not reasonable because a ball can't bounce a negative height.
- 3) 34.78 cm. I think this prediction is too low. That's about the height the ball bounced when dropped from 52 cm and it should bounce higher than that if it's dropped from 8 cm higher.
- 4) Going through the most points can make a good line of best fit but it won't always. The scatterplot below shows a case where it makes a bad line of best fit. The line is too high.



### Possible Answers to Summary Questions

- 3) I think the best criterion is the one that makes the total distances of the points above and below the line the same. That way it gets the line in the middle and follows the trend of the data set.
- 4) It's a line that shows your best guess as to where the data would be for those values that you don't have.