

Least Squares Regression – ID: 9825

By Michele Patrick

Time required
60 minutes

Activity Overview

This activity introduces students to the least squares regression line, which minimizes the sum of the squares of differences between actual and predicted values (residuals). They will first adjust a line on a scatter plot to where they think it best shows the trend of the data. Then, they will measure the lengths of vertical segments representing the residuals and find the sum of the squares of these lengths. Trying to minimize this sum, they will adjust their lines, followed by calculating the regression equation and understanding how it minimizes the sum of the squares.

Concepts

- Scatter plots
- Residuals
- Minimizing the sum of the squares

Teacher Preparation

This activity is designed to be used in an Algebra 2 classroom. It can also be used in an introductory Statistics or advanced Algebra 1 classroom.

- *Students will be asked to compare the sum of the squares of the least squares regression line to the correlation coefficient, r . If students are not familiar with r , you can delete these questions from the worksheet without affecting the flow of the activity.*
- *The last part of Problem 2 asks students to find regression lines based on the sum of the absolute values of the residuals and the sum of the shortest distances between the predicted and actual values, and to compare those equations to the equation found using the sum of the squares of the residuals. This can be skipped if time does not allow.*
- *The screenshots on pages 2–5 (top) demonstrate expected student results. Refer to the screenshots on pages 5 (bottom) and 6 for a preview of the student TI-Nspire document (.tns file).*
- **To download the student .tns file and student worksheet, go to education.ti.com/exchange and enter “9825” in the quick search box.**

Classroom Management



- *This activity is intended to be mainly **teacher-led**, with breaks for individual student work. Use the following pages to present the material to the class and encourage discussion. Students will follow along using their handhelds.*
- *The student worksheet Alg2Act08_LeastSquares_worksheet_EN helps guide students through the activity and provides a place for students to record their answers.*
- *The TI-Nspire solution document Alg2Act08_LeastSquares_Soln_EN.tns shows the expected results of working through the activity.*
- *Information for an optional extension is provided at the end of this activity.*

TI-Nspire™ Applications

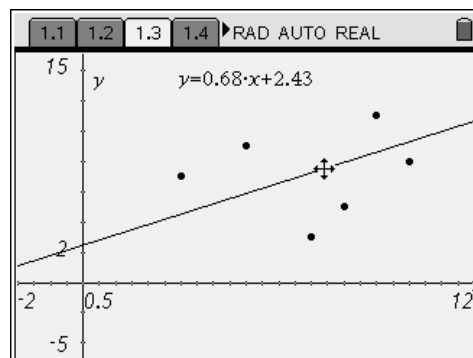
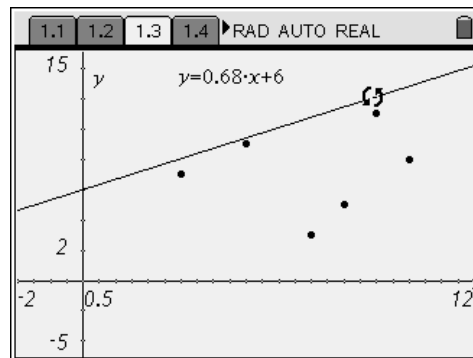
Graphs & Geometry, Lists & Spreadsheet, Notes

Problem 1 – A more scattered scatter plot

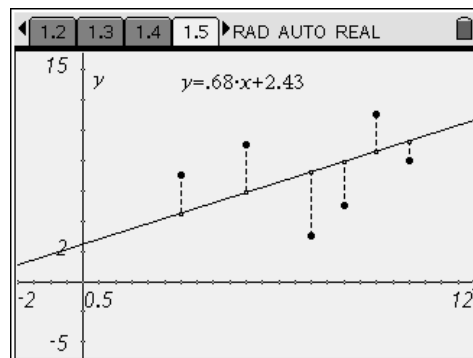
On page 1.3, students will see a scatter plot containing 6 data points and a movable line. (The data values for these points are listed in the spreadsheet on page 1.7.) Students are to grab and drag the line until they think it is a good fit for the data.

At this point, they should just be “eyeballing” a line that fits the general trend of the data. They can rotate the line when the cursor looks like this:  and translate the line when the cursor looks like this: .

The displayed equation of the line will update as they change the line. When students are satisfied with the location of their line, they should record the equation on their worksheets.

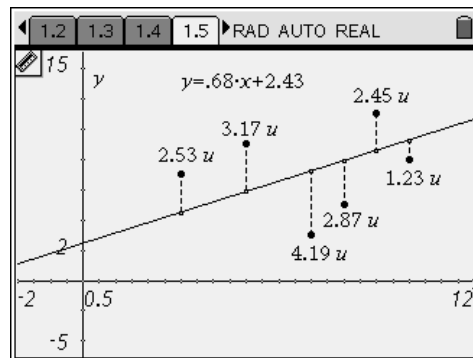


The scatter plot shown on page 1.5 is the same as the one on page 1.3, but this time, vertical segments connecting the data points to the line have been constructed. Students are to edit the text box with the equation to match the equation they found on page 1.3. (The equations students use here will vary.)



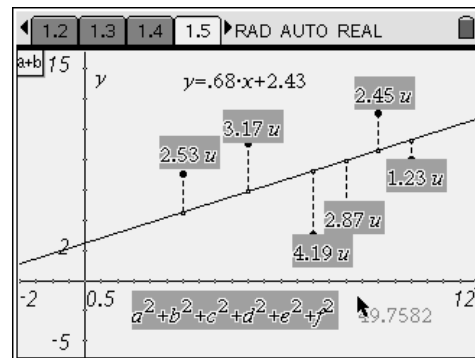
Discuss the *residuals*—the difference between the actual points and the predicted points (the ones on the line). Students should use the **Length** tool (**MENU > Measurement > Length**) to find the values of their six residuals.

Be sure to explain that a residual is defined as “actual value minus predicted value,” and thus the residuals for data points *below* the line are actually negative. However, because students will be *squaring* the residuals, which always results in a nonnegative value, it is okay to use the lengths as they are.



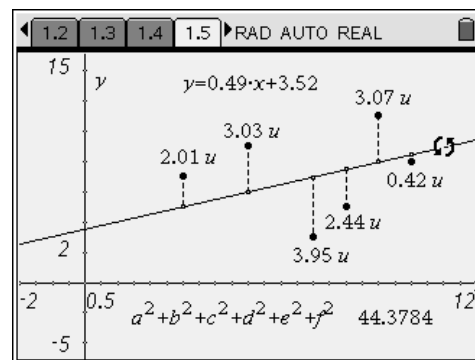
Next, students are to calculate the sum of the squares of the residuals. To do this, they need to use the **Text** tool (**MENU > Actions > Text**) to display the expression $a^2 + b^2 + c^2 + d^2 + e^2 + f^2$ and then use the **Calculate** tool (**MENU > Actions > Calculate**) to evaluate the expression for the six values just measured.

Since students may be using different equations, there will most likely be many different sums. Ask some of the students for their sums. Discuss that those with lower sums have a line that better fits the data than those with higher sums. The goal is to find the line that gives the lowest sum possible.



Challenge students to adjust their line to make the lowest sum they can find. Remind them that they can change both the slope and the y -intercept.

Have them record the equation and sum on their worksheets.



On page 1.7, students should select **MENU > Statistics > Stat Calculations > Linear Regression (mx+b)**. Enter **a[]** for X List, **b[]** for Y List, and **c[]** for 1st Column Result. Then press enter .

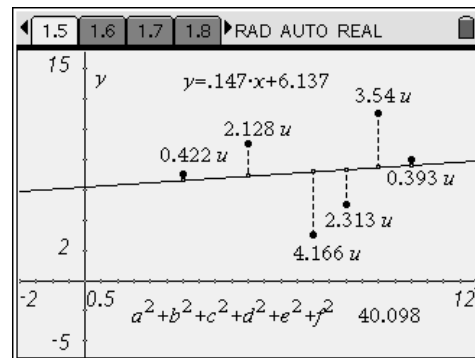
The slope (m) and y -intercept (b) of the “true” least squares regression line will appear in cells D3 and D4. Have students copy these values as well as the value of r (to at least three decimal places) on their worksheets, for future reference.

	A	B	C	D	E	F
	list1	list2				
				=LinRegM		
3	8	5	m	.147059		
4	5	9	b	6.13725		
5	9	11	r ²	.018007		
D3				=.14705882352941		

Perform a **Linear Regression (mx+b)**.

Returning to page 1.5, students should edit the text box to show the equation for the regression line they just found. After students press enter , ask what happened to the sum of the squares (it went down to 40.098). See which students had sums close to this minimal value.

Students may still drag the line to see that this sum is truly the least possible sum that can be obtained.



You can also have students go back to the spreadsheet and tab down to see the value of the residuals for the least squares regression line. Point out the negative residuals, reminding them of the discussion earlier.

Slight differences between these values and those shown on page 1.5 will be due to rounding the values of m and b in the equation.

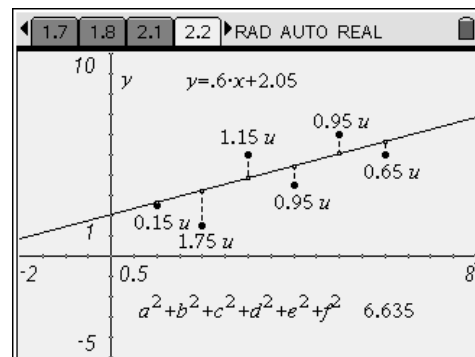
A	list1	B	list2	C	D	E
					=LinRegM	
5	9	11	r^2		.018007	
6	10	8	r		.134191	
7			Resid		{.42156...}	

Perform a **Linear Regression (mx+b)**.

Problem 2 – A less scattered scatter plot

On page 2.2, students will find another scatter plot with 6 data points. (This data may be found on page 2.4.) A movable line and six vertical segments representing the residuals are already drawn.

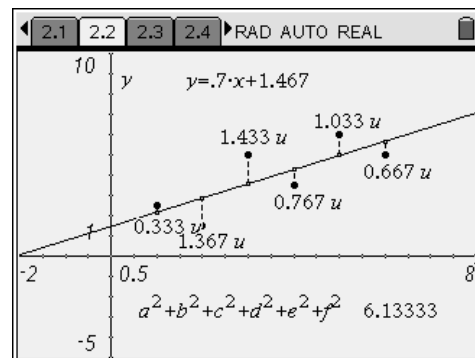
Have them proceed as in Problem 1 to find sum of the squares of the residuals, recording the equation and sum on their worksheets. To spice things up, you can have a friendly competition to see who can obtain the lowest sum—without using the spreadsheet of course!



Using the spreadsheet on page 2.4, to calculate the regression line, students should record the appropriate values and return to page 2.2 and adjust the equation in the text box to match the regression equation just obtained, recording again the minimized sum.

See which student or students came close. You may wish to reward them with a small prize.

If students are familiar with the correlation coefficient, r , they can compare the values of r and the sum of the squares of the residuals for the two data sets and make a conjecture. In Problem 1, $r \approx 0.134$ and the sum is about 40. In Problem 2, $r \approx 0.764$ and the sum is about 6. Ask them for the value of the sum of the residuals if r was -1 or 1 (zero; all of the points would lie on the line).

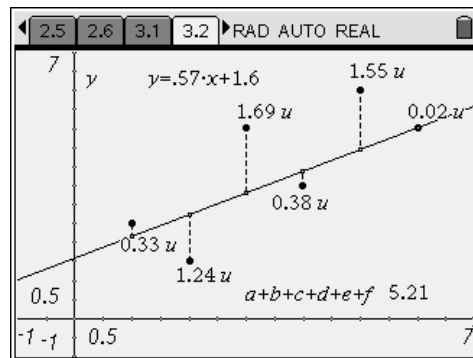


Correlation and regression are strongly related and students will study this in more detail in more advanced courses. For now, it is enough to know that weaker correlations result in greater values for the sum of the squares and stronger correlations result in lower values.

The scatter plots on pages 3.2 and 3.4 both show the data in the spreadsheet on page 2.4. You can choose to end the activity at page 2.6 or continue with pages 3.1 through 3.4, which explore finding a regression line using methods other than the least squares method.

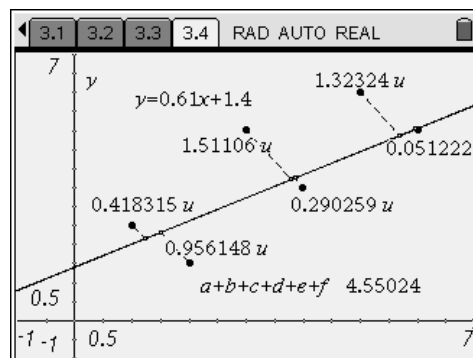
Extension

On page 3.2, students first need to replace the equation with the equation of the least squares regression line (calculated on the spreadsheet in Problem 2). Next, they should calculate the sum of the *absolute values* of the residuals (remind students that some residuals are positive and some are negative, depending on their relative location to the regression line). Because length is always nonnegative, students can simply find the sum of the lengths.



Students should then adjust the line until this sum is minimized. Ask how the equation of this regression line compares to the equation of the line using the least squares method.

On page 3.4, students should again replace the equation shown with the equation of the least squares regression line. Then, they should find the sum of the *shortest distances* between the actual and predicted points. (Because they are not vertical distances, they are not residuals. Also, because these are distances, they all are nonnegative, regardless of each point's relation to the line.)



Students will again adjust the line until this sum is minimized. Ask how the equation of this regression line compares to the two previous methods.

Discuss with students that these two methods of regression are rarely used because mathematicians have shown that the least squares regression line may be generalized. In some cases, one approach may give a line that appears to have most of the points "closer" to the line (i.e., with the least "error,") but for a different set data, another approach will have the least error. The least squares regression line will always give the smallest sum of squared residuals. Additionally, the method using absolute values can produce more than one line that minimizes the sum of the residuals. Therefore, unless it is otherwise noted, a linear regression line is always the least squares regression line.

Least Squares Regression – ID: 9825

(Student)TI-Nspire File: *Alg2Act08_LeastSquares_EN.tns*

1.1 1.2 1.3 1.4 RAD AUTO REAL

**LEAST SQUARES
REGRESSION**

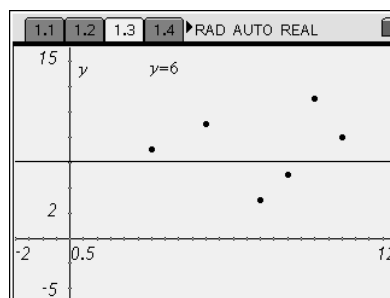
Algebra 2

Scatter plots, residuals,
and sums of squares

1.1 1.2 1.3 1.4 RAD AUTO REAL

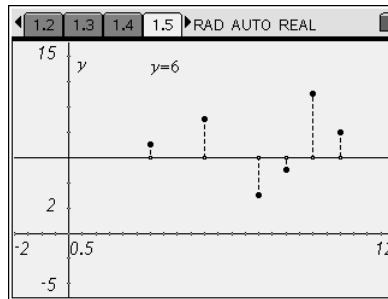
On page 1.3, a scatter plot of six data points is shown, as is the horizontal line $y = 6$.

Drag and rotate the line to better show the trend of the data. Record your equation on your worksheet.



1.1 1.2 1.3 1.4 ▸ RAD AUTO REAL

On the next page, edit the displayed equation to the one you found on page 1.3. Use the **Length** tool to find the values of the *residuals*—the vertical distance from each point to the line. Then use the **Text** and **Calculate** tools find the sum of the squares of the residuals.



1.3 1.4 1.5 1.6 ▸ RAD AUTO REAL

Question

Drag the line again. What is the lowest sum you can get? What is the equation that results in this sum?

Answer

1.4 1.5 1.6 1.7 ▸ RAD AUTO REAL

	A list1	B list2	C	D	E	F
1	3	7				
2	7	3				
3	8	5				

A7 |

Perform a **Linear Regression (mx+b)**.

1.5 1.6 1.7 1.8 ▸ RAD AUTO REAL

Question

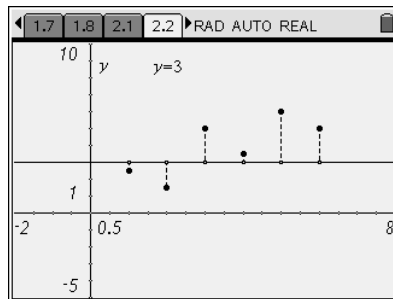
How does the equation you found on page 1.5 compare to the linear regression equation? How do the sums compare?

Answer

1.6 1.7 1.8 2.1 ▸ RAD AUTO REAL

On the next page, find the equation of the line that minimizes the sum of the squares of the residuals.

Use the **Length**, **Text**, and **Calculate** tools as you did in Problem 1.



1.8 2.1 2.2 2.3 ▸ RAD AUTO REAL

Question

What is the lowest sum you can get? What is the equation that results in this sum?

Answer

2.1 2.2 2.3 2.4 ▸ RAD AUTO REAL

	A list1	B list2	C	D	E	F
1	1	2.5				
2	2	1.5				
3	3	5				

A7 |

Perform a **Linear Regression (mx+b)**.

2.2 2.3 2.4 2.5 ▸ RAD AUTO REAL

Question

How does the equation you found on page 2.2 compare to the linear regression equation? How do the sums compare?

Answer

2.3 2.4 2.5 2.6 ▸ RAD AUTO REAL

Question

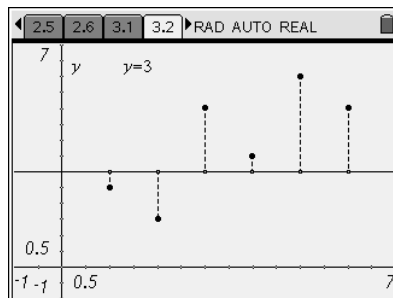
Compare the correlation coefficient, r , and the sums for the regression equations from Problems 1 and 2 and make a conjecture.

Answer

2.4 2.5 2.6 3.1 ▸ RAD AUTO REAL

On the next page, find the equation of the line that minimizes the sum of the *absolute values* of the residuals.

How does this compare to the method using the sum of the *squares* of the residuals?



2.6 3.1 3.2 3.3 ▸ RAD AUTO REAL

On the next page, find the equation of the line that minimizes the sum of the *shortest* distances (not the vertical distances) from each data point to the line.

How does this compare to the other two methods?

