

Parabolic Path to a *Best Best-Fit Line*:

Finding the Least Squares Regression Line By Exploring the Relationship between Slope and Residuals

Objective: How does one determine a *best* best-fit line for a set of data? “Eyeballing it” may be a good place to start, but there is a more exact way. In this activity, you will not only find a *best* best-fit line for a data set, you will discover why it must be the best.

What to Do:

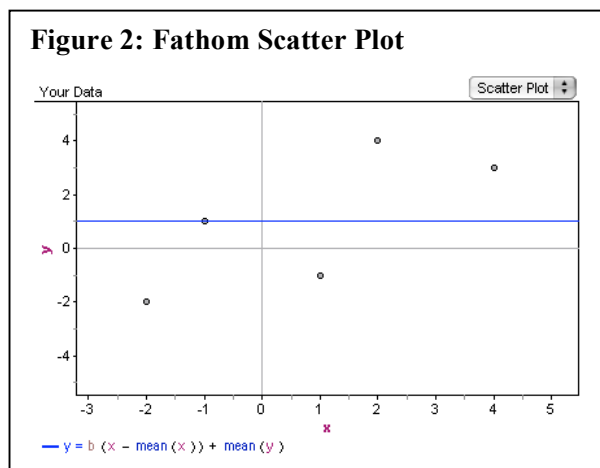
1. Open the *Fathom* file **Parabolic_Path_LSRL.ftm**.
2. In the Case Table (upper left corner of screen), input the data set as given in Figure 1.

Figure 1: Fathom Case Table

Your Data

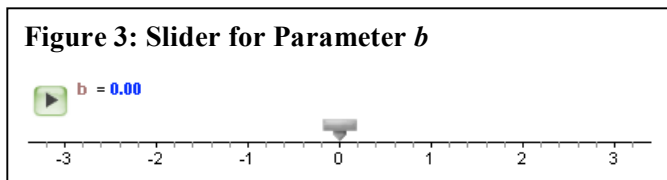
	x	y
1	-2	-2
2	1	-1
3	4	3
4	2	4
5	-1	1

3. Notice the data points are displayed in a scatter plot (upper right of screen) as in Figure 2. Also on the scatter plot, a blue horizontal line has appeared that we can use to find a best-fit line to model the data.



- a. The *general* equation for this line can be found just below the scatter plot. What is it? Write the equation using \bar{x} for “mean(x)” and \bar{y} for “mean(y)”.
- b. What specific slope does the initial horizontal (blue) line have?
- c. Based on that slope, simplify the general linear equation for the specific line graphed.

4. Below the Case Table (Figure 1) on the Fathom screen is a slider (Figure 3) that changes the values of parameter **b** in the linear equation. Try moving the slider to the left and right.



- What effect does changing the value of **b** have on the blue line? Why?
 - By changing **b**, can the line be vertically or horizontally *translated*?
 - By changing **b**, what type of *transformation* does occur?
5. In order to find an equation for a line, at least one point on the line must be known. For a best-fit line, a reasonable point to begin with is (\bar{x}, \bar{y}) , also called the *center of gravity* for the data set.
- Why would the center of gravity be a good point to include on a best-fit line?
 - Around what point does the line on the scatter plot appear to rotate when the slope, **b**, is changed with the slider?
 - In the Fathom window, a Summary Table as in Figure 4 displays statistics on the data, including the slope (**b**), \bar{x} , \bar{y} , the sum of the residuals, and the sum of squared residuals. Based on the data's statistics, what are the specific coordinates of the point of rotation?
 - Looking at the line graphed on the scatter plot, do these coordinates appear to be correct?

Figure 4: Fathom Summary Table

Summary Table

Your Data

	Residual
	0
	0.8
	1
	0
	26

S1 = **b**
 S2 = $\text{mean}(x)$
 S3 = $\text{mean}(y)$
 S4 = $\text{sum}(\text{Residual})$
 S5 = $\text{sum}(\text{Residual}^2)$

- e. Now look closely at the Case Table as in Figure 5 that contains the data points. It also includes other useful information: **YFitted** values and **Residual** values. How are the **YFitted** values – also known as *predicted values* – determined?

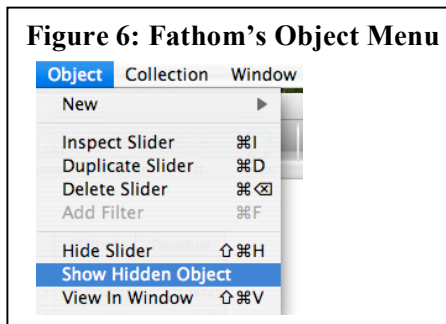
Figure 5: Fathom Case Table

Your Data

	x	y	YFitted	Residual
1	-2	-2	1	-3
2	1	-1	1	-2
3	4	3	1	2
4	2	4	1	3
5	-1	1	1	0

- f. How are the **Residuals** determined? What is a *residual*?
6. Knowing only one point (in this case, the center of gravity) is not sufficient information to determine any line, much less a best-fit line. We also need to find the best slope to fit the data.
- Adjust the slope of the line by using the slider to change the value of **b**. Try to find a line that is a good fit to the data. Record the value of the slope (**b**) for this initial estimate.
 - How did you decide what the slope of the best-fit line should be?
 - Do you think someone else would choose the same slope as you? Why or why not?
7. Is there a more accurate method for determining the best slope? Go to the **Object** menu and select **Show Hidden Objects** as in Figure 6. A graph of a function appears.

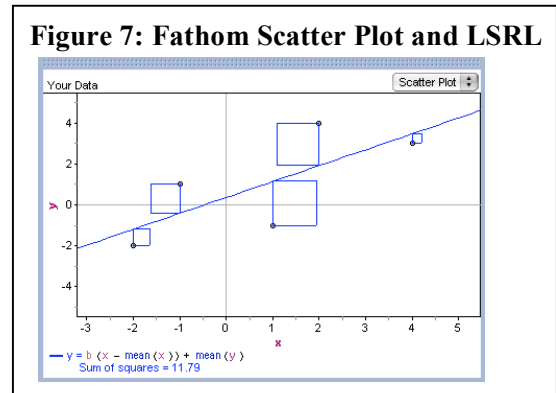
- What is the explanatory (independent) variable of this function? [Hint: How are the axes labeled?]
- What is the response (dependent) variable?



- c. What *type* of function models the relationship between these two variables?
 - d. What is the shape of this function's graph?
8. Try adjusting the slope (**b**) of the line once again. This time notice how the mysterious point on the parabola moves in response. This point and the slope of our best-fit line are connected.
- a. Adjust the slope to move the mysterious point closer to the vertex of the parabola. What effect does this have on how well the line fits the data?
 - b. Use the parabola and its vertex to determine the best slope for your linear model. Once you are satisfied with your choice, record the coordinate values of the mysterious point to the nearest 2 decimal places. (Refer to the Summary Table for values.)
 - c. At the vertex point, notice that the response variable has a minimum value. Considering the meaning of the response variable, why is the vertex helpful in determining the slope of our best-fit line?
9. Determine an equation for the best-fit line for the data set.
- a. Use the center of gravity and the slope found with the parabola to write an equation for the best-fit line in point-slope form, that is, $y - y_1 = b(x - x_1)$.
 - b. Rewrite your equation above in slope-intercept form, $y = bx + a$.
 - c. The type of "best-fit" line that found in this activity is known as the "Least Squares Regression Line," or LSRL, for short. Why is it referred to by this name?

10. Now that we have determined the LSRL, we can check it with the LSRL as computed by Fathom.

- a. Highlight (click on) the scatter plot. Then select **Least-Squares Line** from the **Graph** menu. Fathom graphs the LSRL (green line) on the scatter plot with ours (blue line). Are they close? What is Fathom's LSRL equation, and how does it compare to ours?
- b. Return to the **Graph** menu and select the **Show Squares** option. Notice that squares connected to the LSRL appear as in Figure 7. What do these squares represent? And what does the total sum of the areas of these squares represent?



- c. Below the equation for Fathom's LSRL is a value for its "Sum of squares." What is this value? Where can this value be found on the graph of the parabola?
 - d. Change the value of the slope (**b**) of the line and notice the effect on the size of the squares. To determine a *best* best-fit line for a data set, we found a special line through the center of gravity that also does what to the sum of the areas of the green squares?
11. What two primary characteristics must a Least Squares Regression Line (a best-fit line) have to model a set of data, and why is each characteristic significant?

Extensions

1. *Why* is there a *quadratic* relationship between the slope of a best-fit line and the sum of the squared residuals? Investigate this relationship in the following:
 - a. Use a line in the form $y = b(x - \bar{x}) + \bar{y}$ that contains the center of gravity, $\bar{x} = 0.8$ and $\bar{y} = 1.0$. Find the residual in terms of \mathbf{b} for each of the five data points in the activity. (It may help to organize the results in a table, and you may want to collaborate with others.)
 - b. Square each residual. Then find the sum of the squared residuals in terms of \mathbf{b} (the slope).
 - c. What type of function describes the relationship between the slope and the sum of squared residuals for a line that includes the center of gravity?
 - d. Verify that the general quadratic function (given in the parabola's window in Fathom and also below) works for the data set in the activity.

$$\sum_{i=1}^n (\text{residuals})^2 = \sum_{i=1}^n (y_i - \bar{y})^2 - 2b \cdot \sum_{i=1}^n y_i (x_i - \bar{x}) + b^2 \cdot \sum_{i=1}^n (x_i - \bar{x})^2$$

- e. Show that the above holds true for any general data set.
2. In the activity, we determined a LSRL line for a data set of five specific points. What if the data set changed?
 - a. How does changing a point in the data set affect the quadratic relationship between the slope and the sum of the squared residuals for a possible best-fit line? Choose any one of the five data points and drag it around to new locations. In what ways does this affect the parabola used to determine the LSRL? How can the effects on the parabola caused by changing a point in the data set be explained in terms of the parabola's equation?
 - b. How does changing a point in the data set affect the Least Squares Regression Line? Highlight (click on) the scatter plot. Then under the **Graph** menu, de-select **Show Squares**. Choose any one of the five data points and drag it around to new locations. In what ways does this affect the LSRL (the green line in the Fathom window)? Specifically, how does changing a point in the data set affect each of the two defining characteristics of the Least Squares Regression Line?