Teacher Notes

Parabolic Path to a *Best* Best-Fit Line: Finding the Least Squares Regression Line by Exploring the Relationship between Slope and Residuals

Authors: Jeff Reinhardt and Joe Simons

What Is It? In this activity, students use dynamic statistical software to explore important characteristics of a least squares regression line, especially the quadratic relationship between its slope and the squared residuals.

Grade Level: High school

Class Time: 1-2 class periods

Mathematics Strands:

- Statistics: scatter plot, residuals, mean ($\bar{x}$ notation), center of gravity ($\bar{x}$, $\bar{y}$), least squares regression line
- Functions: linear, quadratic: graphical properties, point-slope form, slope-intercept form, standard form, algebraic manipulation
- Geometry: transformations (translations and rotations), area

Materials:

- Software: Fathom 2 : Dynamic Data Software, Key Curriculum Press; PC/MAC
- Fathom file: Parabolic_Path_LSRL.ftm
- Documents: Parabolic_Path_Activity.doc
  Parabolic_Path_Teacher_Notes.doc

Sample Solutions:

3. **a.** $y = b(x - \bar{x}) + \bar{y}$
b. Slope, *b*, is 0.
c. $y = \bar{y}$

4. **a.** It changes the slope, since *b* is the slope in the general equation.
b. No
c. Rotation about a fixed point.

5. **a.** Answers vary. The line would be based on the averages of both the *x*–values and the *y*–values.
b. The center of gravity, $(\bar{x}, \bar{y})$.
c. $(0.8,\ 1.0)$
d. Yes.

e. Each "YFitted" value, or "predicted *y*- value," is the output given by the blue line's equation for the corresponding *x*- value.

f. Each "Residual" value is found by subtracting the predicted *y*-value from the actual *y*-value for a given *x*-value. That is, "each residual = actual *y* – predicted *y*." A residual is a vertical deviation from the best-fit line, and as such, it is the error in prediction.

6. **a.** Answers vary.
b. Answers vary; "tried to center it," "equal errors above and below the line," "eye-balled it," etc.
c. No; reasons vary.

7. **a.** Explanatory variable is the slope of the blue (best-fit) line.
b. Response variable is the sum of the squared residuals.
c. Quadratic function
d. Parabola

8. **a.** The blue line fits the data better as the point moves closer to the vertex.
b. ≈ 0.79
c. Answers vary. The parabola's vertex determines which slope, *b*, results in the smallest sum of the squared residuals (or "total error") for a line of best fit.

9. **a.** $y - 1 = 0.79(x - 0.8)$
b. $y = 0.79x + 0.368$
c. Answers vary. The LSRL is the best-fit line that minimizes (that is, creates the *least*) sum of the *squared residuals*.

10. **a.** Fathom's LSRL: $y = 0.789x + 0.37$; they are nearly identical equations.
b. Answers vary. The five squares represent the squared residuals. The total sum of the area of the squares equals the sum of the squared residuals.
c. Sum of squared residuals $= 11.79$; on the vertex.
d. The LSRL minimizes the sum of the areas of the green squares, and so, the LSRL minimizes the sum of the squared residuals).

11. The Least Squares Regression Line (LSRL) must:
1) Contain the center of gravity, ($\bar{x}$, $\bar{y}$). This point represents the average of the explanatory values and the response values. Including this point effectively "centers" or "balances" the best-fit line.
2) Have a slope that minimizes the sum of the squared residuals of the line with the data. Doing so will result in a best-fit line that has the smallest total of (squared) vertical deviations from the actual values, thereby minimizing error.

Extension 1 (only)

a.

x	y	$\hat{y} = b(x - 0.8) + 1.0$	**Residual =** $y - \hat{y}$
-2	-2	-2.8b + 1	-2.8b - 3
1	-1	0.2b + 1	0.2b - 2
4	3	3.2b + 1	3.2b + 2
2	4	1.2b + 1	1.2b + 3
-1	1	-1.8b + 1	-1.8b

b.

Squared Residuals = $(y - \hat{y})^2$
$9 + 16.8b + 7.84b^2$
$4 - 0.8b + 0.04b^2$
$4 + 12.8b + 10.24b^2$
$9 + 7.2b + 1.44b^2$
$3.24b^2$

Sum of Squared Residuals = $\sum_{i=1}^{n} (residuals)^2 = \sum_{i=1}^{n} (y_i - \hat{y})^2 = 26 + 36b + 22.8b^2$

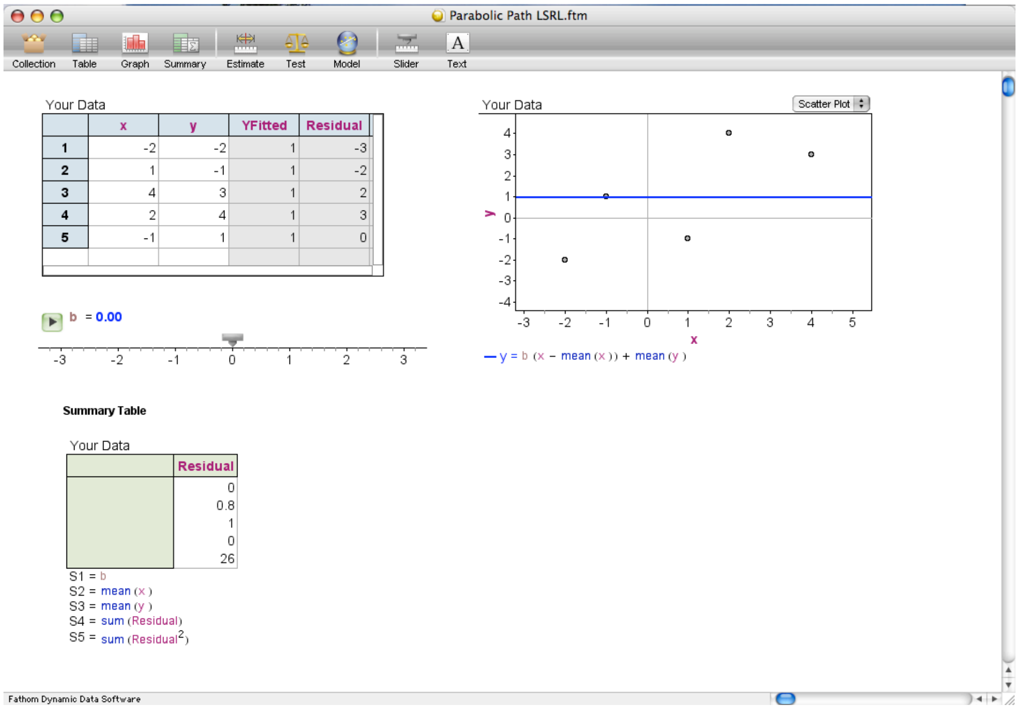
c. A quadratic function describes the relationship between the slope, b, and the sum of the squared residuals for a line that includes the center of gravity.

d. Show that $\sum_{i=1}^{n} (y_i - \overline{y})^2 = 26$, $-2\sum_{i=1}^{n} [y_i(x_i - \overline{x})] = 36$, and $\sum_{i=1}^{n} (x_i - \overline{x})^2 = 22.8$.
A calculator or computer with a statistical package would be quite useful for this.

e. Begin with $\sum_{i=1}^{n} (y_i - \hat{y})^2 = \sum_{i=1}^{n} [y_i - [b(x_i - \overline{x}) + \overline{y}]]^2$. Then expand and simplify.

Selected Screenshots

Sample Screenshot 1: Step 3 of Student Activity



Sample Screenshot 2: Step 10 of Student Activity