MAC-CPTM Situations Project

Situation 33: Least Squares Regression

Prepared at Pennsylvania State University
Mid-Atlantic Center for Mathematics Teaching and Learning
14 July 2005 – Sue, Evan, Donna

Edited at Pennsylvania State University
16, 21 January 2009 – Maureen, Heather, Lana
9 September 2009 – M. Kathleen Heid

Prompt

During a discussion of lines of best fit, a student asks why the sum of the squared differences between predicted and actual values is used. Why use squared differences to find the line of best fit? Why use differences rather than some other measure to find the line of best fit?

Commentary

The set of foci provide reasons why the sum of the squared residuals, differences between the predicted and actual values, is used when determining lines of best fit. The first focus highlights why summing the residuals is not sufficient for determining a line of best fit. The remaining foci highlight the computational advantages of the squared residuals over other alternatives. The second focus examines why one would sum the squared residuals as opposed to summing the absolute value of the residuals. The third focus deals with why one would sum the squared residuals as opposed to summing the squared perpendicular distances.

Mathematical Foci

Mathematical Focus 1:

The line of best fit is not the only line having residuals sum to zero.

The sum of the residuals for the line of best fit is zero. However, the line of best fit is not unique in producing this sum. For any bivariate set of data, the sum of the lengths of the vertical segments from each y_i to any line passing through the mean of the x's and the mean of the y's will also equal zero.

Consider the following set of data: (5, 20), (10, 30), (15, 33), (20, 39), (25, 48)

The line of best fit is $y = 1.3x+14.5$. The sum of the residuals is zero, and the sum of the squared residuals is 11.5.

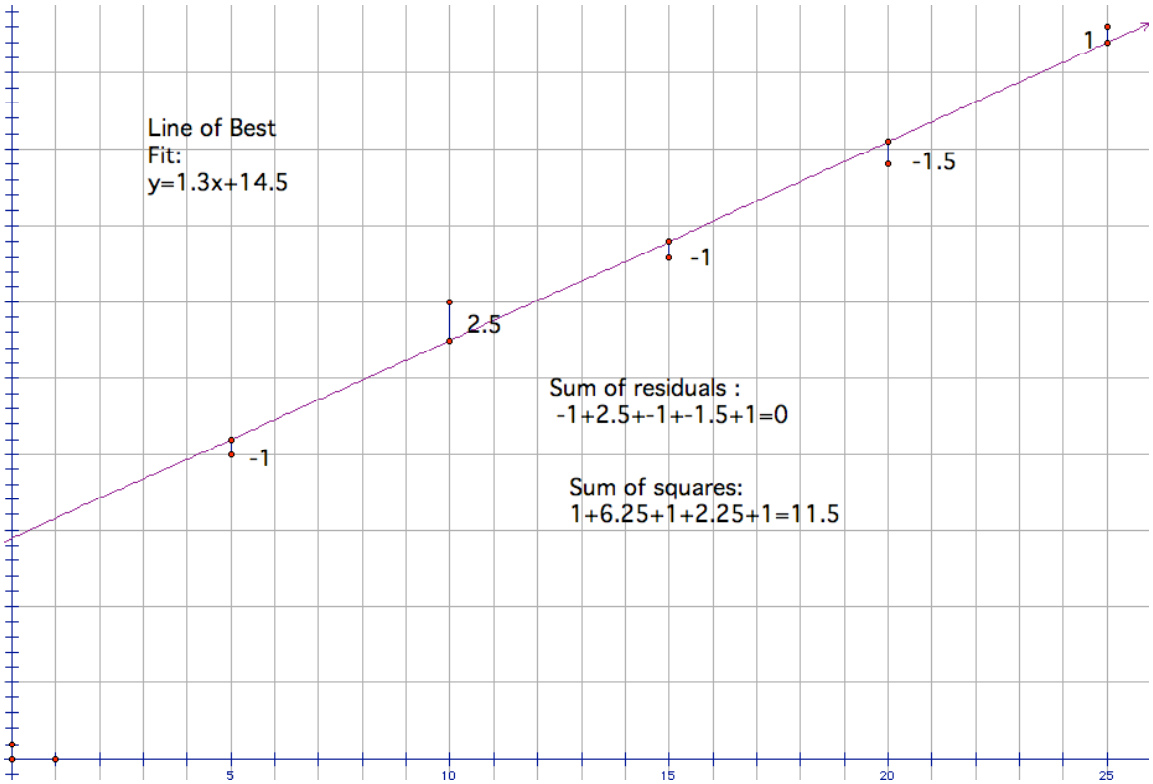

Figure 1 The line of best fit, given by y=1.3x+14.5

For this set of data, the line *y=-x+49* also yields residuals totaling zero, but the sum of the squared residuals is now 1334. The line given by y=-x+49 has slope equal to -1, and is not a good fit for the data (see Figure 2).

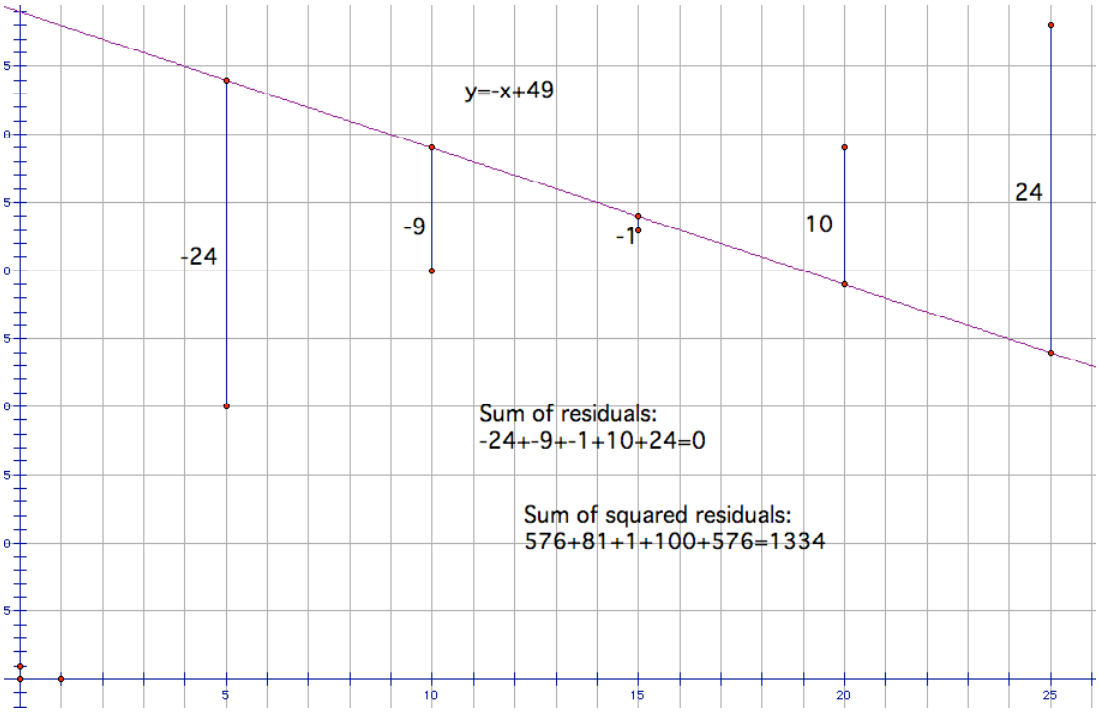

Figure 2 The line given by $y=-x+49$

Mathematical Focus 2:

Although summing the absolute value residuals seems reasonable, it presents more calculation challenges than summing the squared residuals..

The least squares regression line results from minimizing the sum of squared residuals. To examine the difference between summing the squared residuals versus summing the absolute value residuals, one needs to understand that the sum of quadratic expressions is quadratic, whereas the sum of absolute value expressions cannot be written as a single absolute value expression. Further, in order to minimize the sum by using calculus (specifically the derivative), the calculations are straightforward in the case of a quadratic function but quite cumbersome for a function defined as a sum of absolute values. This argument is not as compelling if powerful technology is available to perform the calculations.

Mathematical Focus 3:

Sums of the squared perpendicular distances from a point to the line of best fit provide less helpful information regarding the goodness-of-fit of predictions than do the sums of the squared residuals.

The main purpose for using least-squares regression is to make predictions for the response variable based on a given value for the explanatory variable. Thus, statisticians are interested in determining the goodness-of-fit of their predictions,

i.e., they are interested in finding their prediction error, which is the residual, and minimizing this error. Additionally, the calculations for minimizing the sum of squared residuals for ordinary least squares regression are less cumbersome (minimizing $\sum_{i=1}^{n}\left[y_i-(a+bx_i)\right]^2$) than the calculations for minimizing the sum of squared perpendicular distances (minimizing $\sum_{i=1}^{n}\frac{\left[y_i-(a+bx_i)\right]^2}{1+b^2}$).