MAC-CPTM Situations Project

Situation 33: Least Squares Regression

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

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?

Mathematical Foci

Mathematical Focus 1

One possible path is to examine the need to use a positive value rather than the residual value to calculate the line of best fit. One needs to understand that the sum of the residuals for the line of best fit is zero and that 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 segment from each y_i to the horizontal line at the mean of the y_i's will equal zero as well.

Mathematical Focus 2

Another possible path is to examine the need to square the residuals rather than use the absolute value for the 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.

Mathematical Focus 3

A third possible path is to examine why the residuals are used rather than the perpendicular distance between a point and the line of best fit. 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 $\sum_{i=1}^{n}\frac{\left[y_i-\left(a+bx_i\right)\right]^2}{1+b^2}$).