

## Some Notes on Least Squares Fit

Carl Lee

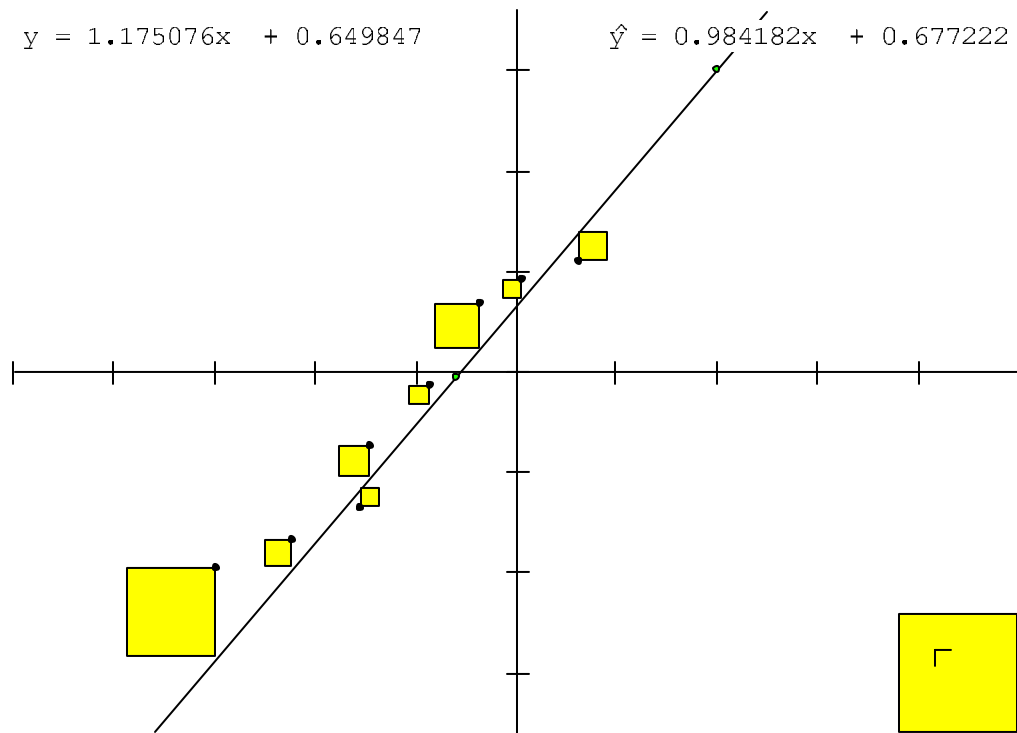
ACE Problem 9 of Investigation 1 of *Samples and Populations* introduces the students to the idea of approximating a set of data points with a straight line. What is a “best” line?

One way to define this is to consider your set of points  $(x_1, y_1), \dots, (x_n, y_n)$  and the equation of a proposed line  $y = mx + b$ . The problem is to find “best” choices of the slope  $m$  and the intercept  $b$ . You could substitute the various  $x_i$  values into the equation and see how far apart  $mx_i + b$  is from the desired  $y_i$  in each case. The error is  $e_i = mx_i + b - y_i$ , which may be positive or negative.

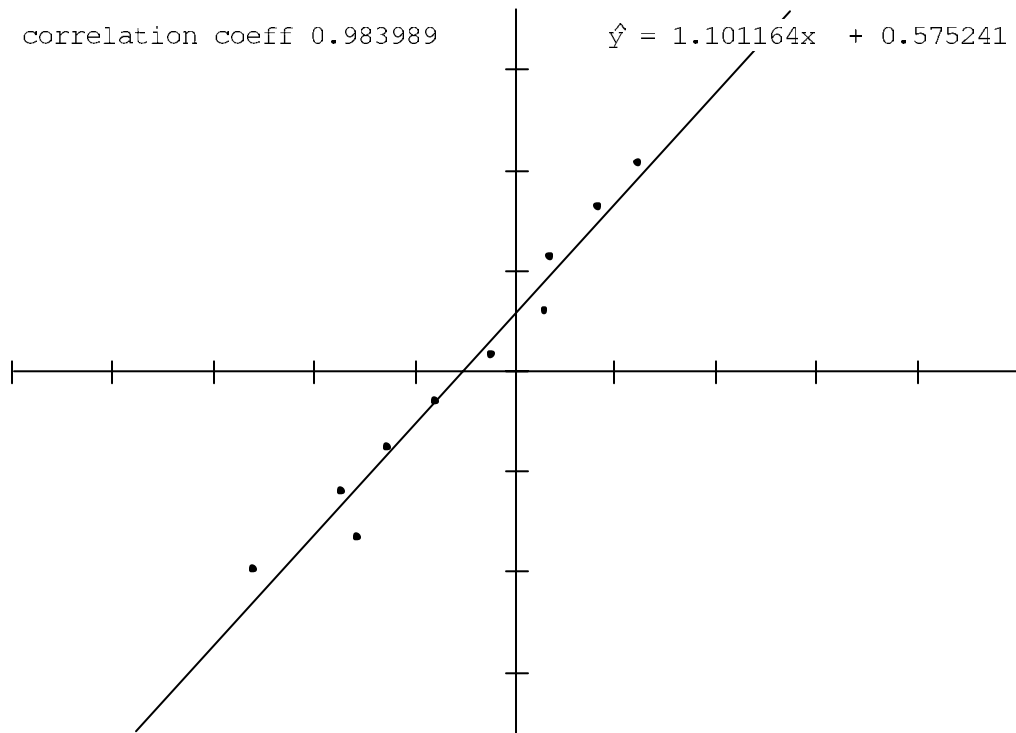
For a best “least squares” fit, we try to make the sum of the squares of the errors as small as possible. That is to say, we try to choose  $m$  and  $b$  to minimize

$$\sum_{i=1}^n e_i^2.$$

For a geometric demonstration of this, start the program “Winstats”. Use “Window→Demos→Least squares...” to open the demo. Create points by clicking on the coordinate system with the right mouse button. The squares of the various errors will be represented by yellow squares, and the total squared error will be represented by a square in the lower right-hand corner. Use the left mouse button to click and drag the small green points around to change the line, and see if you can visually find a “best line”. The least error found so far will be indicated by a small “angle” within the lower right-hand square. You can use “Misc→Show→Regression equation” to see the equation of the actual best line. Use “Misc→Show→Variable equation” to show the equation of *your* line at any given time—this will change as you drag the line around.



Here is another Winstats demo. Start the program “Winstats” and use “Window→Demos→Correlation...” to open another window. Create points by clicking with the right mouse button. Then use “Misc→Show→Regression Line” to display the line with the best least squares fit, and “Misc→Show→Regression Equation” to show its equation.



Your TI-73 calculator can also compute the equation of this regression line. The command is “LinReg(ax+b)” under “2nd STAT CALC”. Instructions are given in Chapter 7 of the manual.

Calculus or linear algebra can be used to get the formulas for the desired  $m$  and  $b$ . They turn out to be:

$$m = \frac{n \sum x_i y_i - (\sum x_i)(\sum y_i)}{n \sum x_i^2 - (\sum x_i)^2}$$

$$b = \frac{(\sum x_i^2)(\sum y_i) - (\sum x_i)(\sum x_i y_i)}{n \sum x_i^2 - (\sum x_i)^2}$$

For a derivation, see, for example, the article on “Least Squares Fitting” in [math-world.wolfram.com](http://math-world.wolfram.com).