Lines of Best Fit

Suppose we are given n data points $(x_1, y_1), \ldots (x_n, y_n)$. There might not be a line that passes through all the points, but we may want to find the "closest" line in the following sense. Let $\ell(x) = mx + b$ be a linear function. Define the ith *error* to be $|y_i - \ell(x_i)|$. We seek m and b that minimizes the sum of the squared errors. In other words, we want to choose m and b that makes

$$\sum_{i=1}^{n} |y_i - \ell(x_i)|^2$$

as small as possible.

One way to say that there is no line through all of these points is to say that $A\mathbf{x} = \mathbf{b}$ is inconsistent where

$$A = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} m \\ b \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

In such a case, we could look for a vector $\hat{\mathbf{x}} \in \mathbf{R}^n$ that is "as close as possible to a solution" in the sense that $A\hat{\mathbf{x}}$ is as close to $\mathbf{b}$ as possible. This leads us to the following definition.

Least Squares Solution

A **least squares solution** to $A\mathbf{x} = \mathbf{y}$ is a vector $\hat{\mathbf{x}}$ such that for all $\mathbf{x}$,

$$||A\hat{\mathbf{x}} - \mathbf{y}|| \leq ||A\mathbf{x} - \mathbf{y}||.$$

The term "least squares" is because we are minimizing the square root of a sum of squares, and that amounts to minimizing a sum of squares.

Least Squares Solution:
A vector $\hat{\mathbf{x}}$ is a least squares solution to $A\mathbf{x} = \mathbf{y}$ if and only if $A^T A\hat{\mathbf{x}} = A^T\mathbf{y}$. In particular, if $A^T A$ is invertible, then there is a unique least squares solution, given by
$$\hat{\mathbf{x}} = (A^T A)^{-1} A^T \mathbf{y}.$$

Example 1. Find a least squares solution to the equation

$$\begin{bmatrix} 1 & 1 \\ 2 & 1 \\ 3 & 1 \end{bmatrix} \mathbf{x} = \begin{bmatrix} 1 \\ 2 \\ 4 \end{bmatrix}$$

Calculating $\hat{\mathbf{x}}$ using the formula above, we get $\hat{\mathbf{x}} = \langle 3/2, -2/3 \rangle$. The corresponding line of best fit is $y = \frac{3}{2}x - \frac{2}{3}$. $\square$

Using the formula, we can derive general formulas for m and b. We have

$$A^T A = \begin{bmatrix} \sum x_i^2 & \sum x_i \\ \sum x_i & n \end{bmatrix}$$

where all the sums go $i = 1, 2, ..., n$. We need to solve the system associated with the augmented matrix

$$[A^T A | A^T \mathbf{y}] = \begin{bmatrix} \sum x_i^2 & \sum x_i & \sum x_i y_i \\ \sum x_i & n & \sum y_i \end{bmatrix}$$

You can do this by high school algebra, but it's not hard by row reduction if we introduce some notation. Define the **mean**s of the x and y values as

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad \overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i.$$

Then the augmented matrix is

$$\begin{bmatrix} \sum x_i^2 & n\overline{x} & \sum x_i y_i \\ n\overline{x} & n & n\overline{y} \end{bmatrix}$$

Divide the 2nd row by n. Then replace the 1st row by $n\overline{x}$ times the resulting second row:

$$\begin{bmatrix} \sum x_i^2 - n\overline{x}^2 & 0 & \sum x_i y_i - n\overline{xy} \\ \overline{x} & 1 & \overline{y} \end{bmatrix}$$

Now define

$$S = \sum x_i^2 - n\overline{x}^2, \quad T = \sum x_i y_i - n\overline{x}\overline{y}.$$

Divide row 1 by S and subtract the new first row $\overline{x}$ times from row 2:

$$\begin{bmatrix} 1 & 0 & T/S \\ 0 & 1 & \overline{y} - (T/S)\overline{x} \end{bmatrix}$$

We now have formulas for m and b, but not the way statisticians usually write them. It is a nice exercise with manipulation of sums to prove

$$\begin{aligned} \sum x_i^2 - n\overline{x}^2 &= \sum (x_i - \overline{x})^2 \\ \sum x_i y_i - n\overline{x}\overline{y} &= \sum (x_i - \overline{x})(y_i - \overline{y}). \end{aligned}$$

The quantities on the right are of interest to mathematicians, because they are simply n times what they call the **variance** of x and the **covariance** of x and y, which are defined by

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2, \quad \sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}).$$

Thus

$$m = \frac{T}{S} = \frac{n\sigma_{xy}}{n\sigma_x^2} = \frac{\sigma_{xy}}{\sigma_x^2},$$

and we have finally proved the theorem in the form statisticians know it:

Theorem 1. The best-fitting line to the data points (x_1, y_1),..., (x_n, y_n) is

$$y = \frac{\sigma_{xy}}{\sigma_x^2}x + (\overline{y} - \frac{\sigma_{xy}}{\sigma_x^2}\overline{x}).$$

We can rearrange this to

$$y - \overline{y} = \frac{\sigma_{xy}}{\sigma_x^2}(x - \overline{x}).$$

In particular, the best line is the line that goes through the point $(\overline{x}, \overline{y})$ and has slope the ratio of the covariance to the variance of x. This is the easiest way to remember the best fit formula.

July 1, 2025