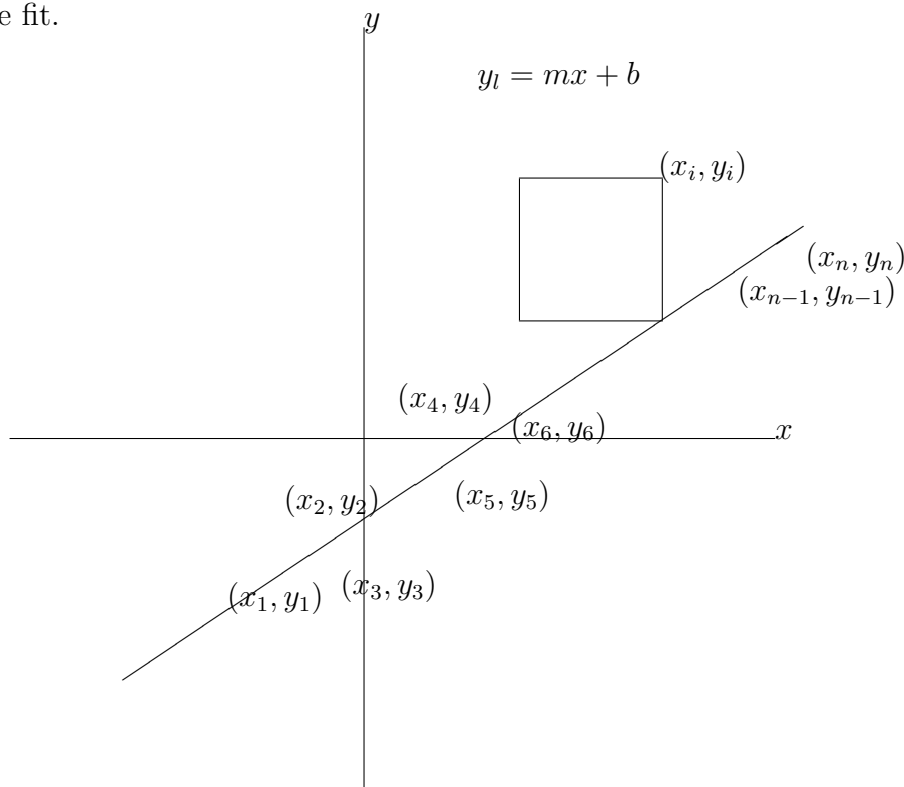


The Least Squares Regression Line An Exercise in Optimization

I. The Problem: Given data points (x_i, y_i) , find the least squares straight line fit.



II. The setup: The data points, (x_i, y_i) are given, and we will have the line if we can calculate m , the slope and b , the y -intercept. The plan is to do this by minimizing the total area of all the n squares. The i^{th} square's side is $y_i - y_l = (y_i - (mx_i + b))$, or

$$(y_i - mx_i - b)$$

Thus the area of the i^{th} square is

$$(y_i - mx_i - b)^2$$

and this expands to

$$y_i^2 - mx_i y_i - y_i b - mx_i y_i + x_i^2 m^2 + x_i b m - y_i b + x_i b m + b^2$$

and combining like terms, we get the Thing we want to minimize, which is a function of m and b .

$$Th(m, b) = \sum_{i=1}^n (y_i^2 - 2mx_iy_i - 2y_ib + 2x_ibm + x_i^2m^2 + b^2) \quad (1)$$

and we can distribute the summation sign and move the constants outside to get

$$= \sum_{i=1}^n y_i^2 - 2m \sum_{i=1}^n x_iy_i - 2b \sum_{i=1}^n y_i + 2bm \sum_{i=1}^n x_i + m^2 \sum_{i=1}^n x_i^2 + nb^2 \quad (2)$$

Now to save writing, and make this somewhat less of a mess, let's give the \sum -sign terms nicknames,

$$\sum_{i=1}^n y_i^2 = Y^2 \quad (3)$$

$$\sum_{i=1}^n x_i^2 = X^2 \quad (4)$$

$$\sum_{i=1}^n y_i = Y \quad (5)$$

$$\sum_{i=1}^n x_i = X \quad (6)$$

$$\sum_{i=1}^n x_iy_i = XY \quad (7)$$

So that

$$Th(m, b) = Y^2 - 2m(XY) - 2Yb + 2Xbm + X^2m^2 + nb^2 \quad (8)$$

III. Note that the independent variables are m and b , because God gave us the data, all we can do is move the line around. So the game is to minimize $Th(m, b)$ – but we know how to do that. We also know that any critical point is a minimum, since the form of $Th(m, b)$ is the square of something.

IV. So get thee up a mess of partials.

$$\frac{\partial Th}{\partial m} = -2XY + 2Xb + 2X^2m \quad \text{and} \quad (9)$$

$$\frac{\partial Th}{\partial b} = -2Y + 2Xm + 2nb \quad (10)$$

and set them to zero to get

$$2X^2m = 2XY - 2Xb \quad \text{and} \quad (11)$$

$$2nb = 2Y - 2Xm \quad (12)$$

This is a set of two equations in two unknowns, which looks promising.

V. So I am going to solve the second and stuff it into the first and see what appeareth.

$$b = \frac{2Y - 2Xm}{2n} \quad (13)$$

$$2X^2m = 2XY - 2X \left(\frac{2Y - 2Xm}{2n} \right) \quad (14)$$

$$2X^2m = 2XY - \left(\frac{4XY + 4XXm}{2n} \right) \quad (15)$$

$$4X^2nm = 4nXY - 4XY + 4XXm \quad (16)$$

$$nX^2m - XXm = nXY - XY \quad (17)$$

$$m = \frac{nXY - XY}{nX^2 - XX} \quad (18)$$

And if we translate the nicknames we get:

$$m = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \quad (19)$$

which looks rather amazingly like the stuff on page 195. These things are not carted off mountains engraved on stone.

Inspired by success, we lumber forward

$$2nb = 2Y - 2Xm \quad (\text{and we have } m) \quad (20)$$

$$2nb = 2Y - 2X \left(\frac{nXY - XY}{nX^2 - XX} \right) \quad (21)$$

$$nb = Y - \left(\frac{nXXY + XXY}{nX^2 - XX} \right) \quad (22)$$

$$nb = \frac{Y(nX^2 - XX) - nXXY + (XXY)}{nX^2 - XX} \quad (23)$$

$$b = \frac{1}{n} \left(\frac{nX^2 - (X)^2Y - n(X)(XY) + (X)^2Y}{nX^2 - (X)^2} \right) \quad (24)$$

$$(25)$$

and mercifully a few things combine to give us

$$b = \frac{x^2Y - X(XY)}{nx^2 - XX} \quad (26)$$

which is

$$b = \frac{\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i - \sum_{i=1}^n x_i \sum_{i=1}^n x_i y_i}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \quad (27)$$

Whew!