General solution of linear regression problem

Problem: Let $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ be a finite set of points in the plane. Find a formula for the linear function $f(x) = ax + b$ which minimizes the sum of squares of errors.

Sigma notation

If $a_1, \ldots, a_m$ is a sequence of numbers, then

$$\sum_{i=1}^{m} a_i := a_1 + \cdots + a_m$$

For example, if $a_1 = 3$, $a_2 = 20$, $a_3 = 12$, the $\sum_{i=1}^{3} a_i = 3 + 20 + 12 = 35$; while $\sum_{i=2}^{3} a_i = 20 + 12 = 32$.

Sometimes, the inidices are omitted. So that we write $\sum a$ for $\sum_{i=1}^{n} a_i$.

Formula for sum of squares of errors

$$
\begin{aligned}
S(a,b) &= (f(x_1)-y_1)^2+\cdots+(f(x_n)-y_n)^2 \\
&= \sum_{i=1}^{n}(f(x_i)-y_i)^2 \\
&= \sum(ax_i+b-y_i)^2 \\
&= \sum(a^2x_i^2+b^2+y_i^2+2abx_i-2ax_iy_i-2by_i) \\
&= a^2(\sum x^2)+(\sum_{i=1}^{n} b^2)+(\sum y^2) \\
&\quad +2ab(\sum x)-2a(\sum xy)-2b(\sum y) \\
&= a^2(\sum x^2)+nb^2+2ab(\sum x) \\
&\quad -2a(\sum xy)-2b(\sum y)+(\sum y^2)
\end{aligned}
$$

Derivatives of the sum of squares of errors

$$
\begin{aligned}
\frac{\partial S}{\partial a} &= 2(\sum x^2)a+2(\sum x)b-2(\sum xy) && (1)\\
\frac{\partial S}{\partial b} &= 2(\sum x)a+2nb-2(\sum y) && (2)
\end{aligned}
$$

Solving for a and b

Setting $\frac{\partial S}{\partial b} = 0$ we find

$$b = \frac{(\sum y) - a(\sum x)}{n}$$

Substituting in the equation $\frac{\partial S}{\partial a} = 0$ and clearing denominators, we find

$$0 = n(\sum x^2)a + (\sum x)(\sum y) - (\sum x)^2 a - n(\sum xy)$$

which yields

$$a = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

Verifying minimization

$$\frac{\partial S}{\partial a} = 2(\sum x^2)a + 2(\sum x)b - 2(\sum xy) \quad (3)$$

$$\frac{\partial S}{\partial b} = 2(\sum x)a + 2nb - 2(\sum y) \quad (4)$$

$$\frac{\partial^2 S}{\partial a^2} = 2(\sum x^2) \quad (5)$$

$$\frac{\partial^2 S}{\partial b^2} = 2n \quad (6)$$

$$\frac{\partial^2 S}{\partial a \partial b} = 2(\sum x) \quad (7)$$

$$D_S = 4n(\sum x^2) - 4(\sum x)^2 \quad (8)$$

Verification, continued

As a general rule, $\sum_{i=1}^{n} x_i^2 \geq (\sum_{i=1}^{n} x_i)^2$. Thus, $D_S > 0$ (except when $n = 1$!). As $\frac{\partial^2 S}{\partial b^2} = 2n > 0$, the point we found is a minimum.

Example

Find the line which best fits $\{(0, 2), (-3, 8), (5, 2), (2, 1)\}$.

Solution

$$
\begin{aligned}
n &= 4 && (9)\\
\sum x &= 4 && (10)\\
\sum y &= 13 && (11)\\
\sum xy &= -12 && (12)\\
\sum x^2 &= 38 && (13)
\end{aligned}
$$

Solution, continued

$$
\begin{aligned}
a &= \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}\\
&= \frac{4(-12) - (4)(13)}{4(38) - (4)^2}\\
&= \frac{-100}{136}\\
&= \frac{-25}{34}\\
&\approx -0.74
\end{aligned}
$$

Solution, continued

So

$$\begin{aligned} b &= \frac{(\sum y) - a(\sum x)}{n} \\ &= \frac{13 - (\frac{-25}{34})4}{4} \\ &= \frac{123}{34} \\ &\approx 3.62 \end{aligned}$$