AJUSTE POR MÍNIMOS CUADRADOS. RECTA DE REGRESIÓN.

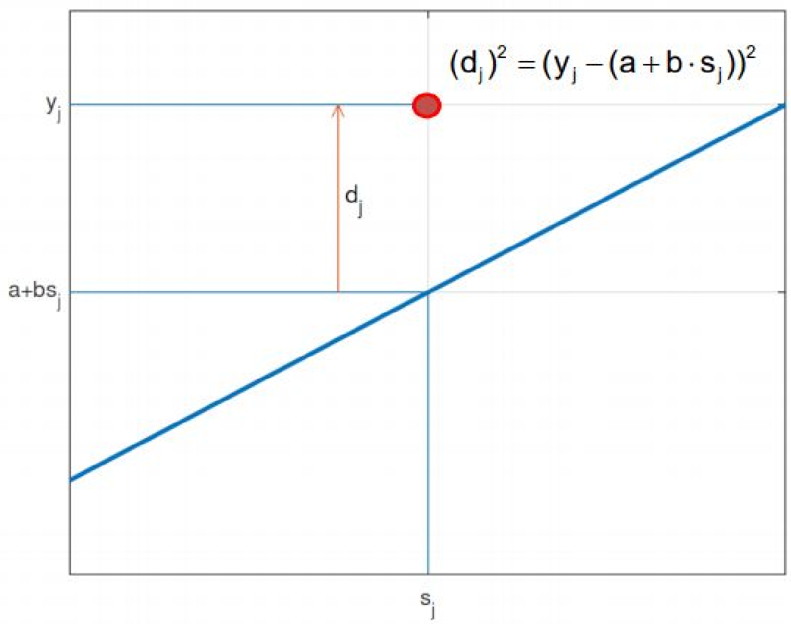
¿QUÉ ES?

El ajuste por mínimos cuadrados es un método de análisis numérico que se trata de un proceso de optimización. Dados un conjunto de valores (x,y), se intenta encontrar la función continua que mejor se ajuste a los datos proporcionados de acuerdo con el criterio de mínimo error cuadrático, expuesto por el matemático Gauss. La función continua a la que queremos aproximar los datos puede ser de distintos tipos, no necesariamente una recta. Así pues, podemos ajustar una función de tipo parábola (o polinomios de grados mayores), exponencial, logarítmica, trigonométrica, etc.

CASO LINEAL. RECTA DE REGRESIÓN.

En este documento, nos vamos a centrar en el caso lineal, el más sencillo. Como se muestra en la imagen a continuación, buscamos obtener una recta r(x) que se ajuste mejor a los datos proporcionados (y_j). Por ello, se va a intentar minimizar la distancia entre los puntos (d_j). Así, esta distancia va a ser la diferencia de la imagen del punto en esa coordenada (y_j) menos el valor de de la recta en esa coordenada.

Elevando esta expresión al cuadrado, se minimiza la distancia. La explicación de este proceso es compleja y fue demostrada por Gauss. A nosotros no nos queda otro remedio que tomarlo como axioma.



R(x)=a+b.x

PROCEDIMIENTO.

En esta explicación, vamos a realizar el proceso necesario para determinar los valores a y b de la recta de regresión:

1. **Emplear sumatorios**. La expresión que acabamos de ver en el apartado anterior es aplicada a un valor en el eje X concreto. No obstante, no tenemos sólo un valor, sino una nube de puntos, por lo que es necesario emplear sumatorios:

$$\sum_{j=1}^{n} (dj)^2 = \sum_{j=1}^{n} (y_j - (a + b \cdot s_j))^2$$

Como depende de dos variables, se trata de una función:

$$f(a,b) = \sum_{j=1}^{n} (d_j)^2 = \sum_{j=1}^{n} (y_j - (a + b \cdot s_j))^2$$

2. **Derivar parcialmente.** Debido a que el objetivo es minimizar la distancia de los puntos a la recta de regresión, derivamos la función y la igualamos a 0 con el fin de encontrar el mínimo . Sin embargo, la función depende de dos parámetros, a y b, luego tendremos que derivar f respecto de a y de b. Derivadas parciales:

$$\frac{\partial f}{\partial a} = 2 \sum_{j=1}^{n} [(y_j - (a + b \cdot s_j)) \cdot (-1)]$$

$$\frac{\partial f}{\partial b} = 2 \sum_{j=1}^{n} [(y_j) - (a + b \cdot s_j)) \cdot (-s_j)]$$

3. **Simplificar la expresión**. Una vez que hemos derivado parcialmente, seguimos operando. Nos interesa dejar a la izquierda lo que depende de a y b, y a la derecha el término independiente:

$$\left\{ \begin{array}{c} \sum_{j=1}^{n} [(y_j - (a + b \cdot s_j)) \cdot (-1)] = 0 \\ \sum_{j=1}^{n} [(y_j - (a + b \cdot s_j)) \cdot (-s_j)] = 0] \end{array} \right\} \Rightarrow \left\{ \begin{array}{c} \sum_{j=1}^{n} a + b \sum_{j=1}^{n} s_j = \sum_{j=1}^{n} y_j \\ \sum_{j=1}^{n} (a \cdot s_j) + b \sum_{j=1}^{n} (s_j)^2 = \sum_{j=1}^{n} (y_j \cdot s_j) \end{array} \right\}$$

$$\Rightarrow \left\{ \begin{array}{c} na + b \sum_{j=1}^{n} s_j = \sum_{j=1}^{n} y_j \\ a \sum_{j=1}^{n} s_j + b \sum_{j=1}^{n} (s_j)^2 = \sum_{j=1}^{n} (y_j \cdot s_j) \end{array} \right\}$$

4. **Adoptar una nomenclatura y expresarlo en forma matricial.** Nos ha quedado un sistema de ecuaciones con dos incógnitas. Simplificamos la notación para que sea más fácil operar:

$$SX \equiv \sum_{j=1}^{n} s_j\,; \quad SY \equiv \sum_{j=1}^{n} y_j; \quad SX2 \equiv \sum_{j=1}^{n} (s_j)^2; \quad SSY \equiv \sum_{j=1}^{n} (y_j \cdot s_j)$$

A continuación, escribimos el sistema en forma matricial:

$$\left\{\begin{matrix} na + bSX = SY \\ aSX + bSX2 = SSY \end{matrix}\right\} \Leftrightarrow A \cdot z = F \Rightarrow \begin{pmatrix} n & SX \\ SX & SX2 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} SY \\ SSY \end{pmatrix}$$

5. **Resolver el sistema matricial por Cramer.** Por último, una vez que tenemos el sistema matricial, sólo queda resolverlo. Para ello, se pueden aplicar distintos métodos, pero el más sencillo será aplicando Cramer:

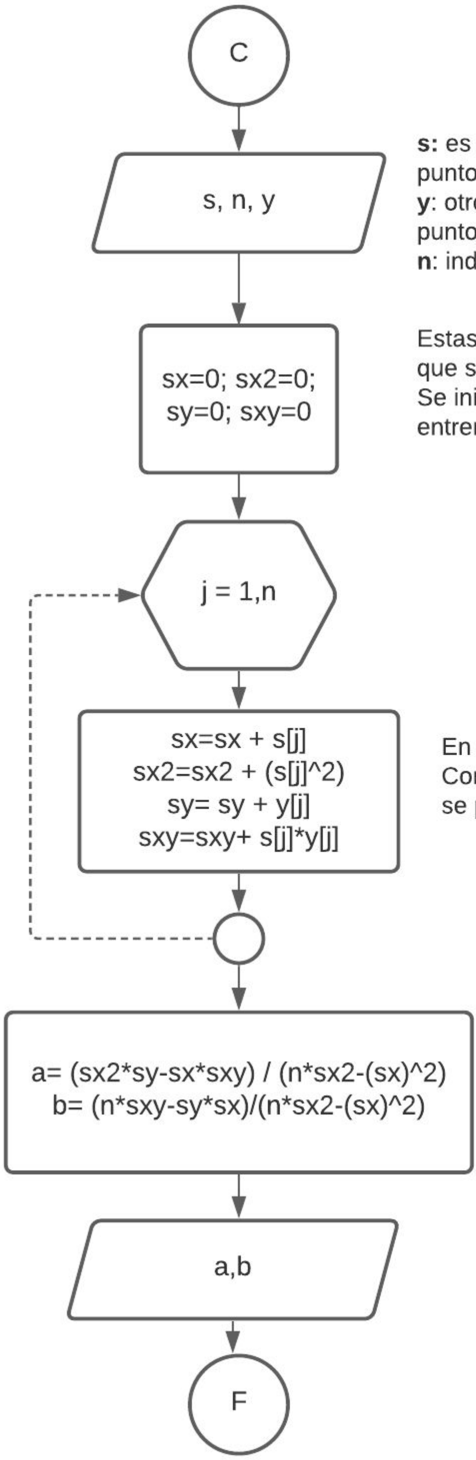
$$a = \frac{\begin{vmatrix} SY & SX \\ SSY & SX2 \end{vmatrix}}{\begin{vmatrix} n & SX \\ SX & SX2 \end{vmatrix}} = \frac{(SY \cdot SX2) - (SX \cdot SSY)}{(n \cdot SX2) - (SX)^2}$$

$$b = \frac{\begin{vmatrix} n & SY \\ SX & SSY \end{vmatrix}}{\begin{vmatrix} n & SX \\ SX & SX2 \end{vmatrix}} = \frac{(n \cdot SSY) - (SX \cdot SY)}{(n \cdot SX2) - (SX)^2}$$

Por tanto, la recta de regresión, R(X)= a + bx

ORGANIGRAMA.

A continuación se expone y se explica el algoritmo que determina los valores a y b de la recta de regresión aplicando el ajuste por mínimos cuadrados.



Se introducen los tres datos:

s: es un vector que contiene los valores de las x de todos los puntos

y: otro vector que contiene los valores de las y en todos los puntos

n: indica el número de puntos

Estas variables equivalen a los sumatorios que se necesitan para determinar a y b.
Se inicializan a cero, para que una vez que entren en el bucle, valgan 0.

En esta operación se realizan los sumatorios.
Como los cuatro sumatorios van de 1 hasta n, se pueden en el mismo bucle.

Se resuelve el sistema de ecuaciones por Cramer, como se ha visto anteriormente