Mínimos Cuadrados (recta de regresión lineal)

La relación entre dos magnitudes x e y lo podemos relacionar a través de la ecuación lineal:

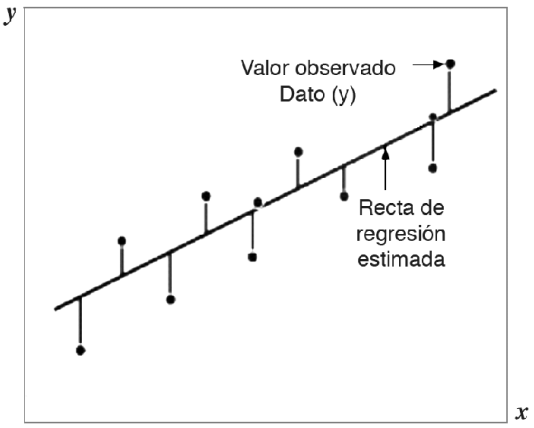
$$r(x)= y = ax + b$$

En donde la constante *a* es la pendiente de la recta y *b* es la ordenada en el origen. Todos hemos utilizado esta fórmula anteriormente, ya sea para problemas de matemáticas, física o química, y sabemos perfectamente cómo se despejan estas constantes en la ecuación, ya que con solamente dos puntos en el eje podemos ser capaces de sacar la ecuación lineal. Sin embargo, ¿qué pasaría si tuviéramos 100 datos? ¿Cómo sacaremos su recta de regresión?

Pues aquí, en Fundamentos de Programación aprenderemos a calcular tanto la constante *a*, como la constante *b* por uno de los métodos más efectivos para determinar estos parámetros: la técnica de mínimos cuadrados. De manera que no perdemos mucho tiempo calculando cada operación porque lo podemos programar todo.

Tal y como lo define Miprofe.com, un mínimo cuadrado es *"un procedimiento de análisis numérico en la que, dados un conjunto de datos se intenta determinar la función continua que mejor se aproxime a estos valores, por medio de recta de regresión, proporcionando una demostración visual de la relación entre los puntos de los mismos"*.

Dicho así, busca minimizar la suma de cuadrados de las diferentes ordenadas y buscando que se aproxime todo lo que pueda a todos los puntos que se encuentran en la gráfica.
Antes de hallar a y b necesitaremos saber algunos conceptos previamente, para que nos dé una mejor aproximación para un número *n* de puntos dados (s[j]).



Tomemos como ejemplo la imagen anterior, donde cada punto de la recta está situado en las coordenadas (*sj, yj*) siendo *j* el número de datos que tengamos. La distancia entre el punto y la recta la podemos definir como d[j], tal que:

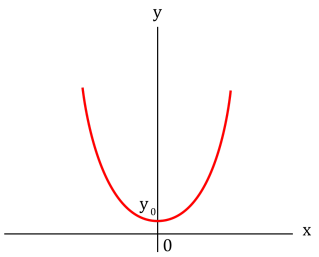
$$d[j]= y[j] - r(s[j]) \rightarrow d[j] = y[j] - (a + b*(s[j]))$$

Queremos encontrar una recta de la forma r(x)=a+bx que nos dé la mejor aproximación posible para un número *n* de puntos dados, *sj* - es decir, que pase por el mayor número de puntos posible.
Para ello, vamos a buscar una función que nos proporcione los valores de *a* y *b* que minimicen el valor *dj* (cuanto menor sea esta distancia, más precisa será la aproximación de la recta de regresión).
Planteamos una función *f* que sea igual a la suma de los cuadrados de dichas distancias, y buscamos el valor para el cual sea mínima.
(**Recuerda:** Para buscar el mínimo de una función, se buscan las soluciones de su derivada).

$$f(a,b) = \sum_{j=1}^{n} dj^2 \rightarrow \sum_{j=1}^{n} (yj - (a + b \cdot sj)^2$$

Para realizar este cálculo se usan los cuadrados de las distancias, porque como $d[j]^2$ es de orden mayor de d[j], si de por sí la distancia d[j] es pequeña, al elevarla al cuadrado será aún menor. Además, todas las distancias quedarán positivas aunque los puntos queden por debajo de la recta.

Nota: la representación de $d[j]^2 = (y[j]-(a+b\ (s[j]))^2$ sería de la forma



por lo que nos aseguramos de que los valores que obtenemos nos indican el mínimo de la función.

Para obtener la derivada de *f,* es necesario derivar parcialmente respecto a cada uno de los coeficientes.

(1) Derivamos parcialmente respecto a *a:*

$$\frac{\partial f}{\partial a} = \frac{\partial}{\partial a}\left[\sum_{j=1}^{n} \left(y_j - (a + b \cdot s_j)\right)^2\right] = 2 \sum_{j=1}^{n} ((y_j - (a + b \cdot s_j))(-1) = 0$$

(2) Derivamos parcialmente respecto a *b:*

$$\frac{\partial f}{\partial b} = \frac{\partial}{\partial b}\left[\sum_{j=1}^{n}\left(y_j - (a + b \cdot s_j)\right)^2\right] = 2\sum_{j=1}^{n}((y_j - (a + b \cdot s_j))(-s_j) = 0$$

(siendo s_j; y_j ;n valores conocidos con los que vamos a aproximar a la función.

Con los resultados obtenidos podemos separar el sumatorio, simplificar el 2 (porque hemos igualado a 0 la derivada para obtener el mínimo de la función) y sacar factor común las constantes, de tal forma que nos quedaría lo siguiente:

(1)

$$\sum_{j=1}^{n} a + b\sum_{j=1}^{n} s_j = \sum_{j=1}^{n} y_j$$

(2)

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

Antes de continuar vamos a renombrar los sumatorios para que sea más fácil operar con ellos al calcular las constantes. Cuando lleguemos a la expresión final, los escribiremos de nuevo.

$$sx = \sum_{j=1}^{n} s_j$$

$$sy = \sum_{j=1}^{n} y_j$$

$$sx2 = \sum_{j=1}^{n} {s_j}^2$$

$$sxy = \sum_{j=1}^{n}(y_j \cdot s_j)$$

Con ésta simplificación podemos comenzar a calcular las constantes a y b, planteando un sistema de ecuaciones y resolviéndolo por el método de determinantes (aplicando el método de Cramer)

{n · a + b · sx = sy ¡¡**Ten en cuenta!!** que el sumatorio de 1 para j=1,n es igual a n
{a · sx + b · sxy=sxy
Planteamos el sistema de forma matricial:

$$\underbrace{\begin{pmatrix} n & sx \\ sx & sx2 \end{pmatrix}}_{A} \cdot \underbrace{\begin{pmatrix} a \\ b \end{pmatrix}}_{x} = \underbrace{\begin{pmatrix} sy \\ sxy \end{pmatrix}}_{F}$$

Calculamos el determinante de A: $|A| = n \cdot sx2 - (sx)^2$

Despejamos los valores de *a* y *b:*

$$a = \frac{\begin{vmatrix} sy & sx \\ sxy & sx^2 \end{vmatrix}}{n \cdot sx2 - (sx)^2} \rightarrow a = \frac{sy \cdot sx2 - sx \cdot sxy}{n \cdot sx2 - (sx)^2}$$

$$b = \frac{\begin{vmatrix} n & sy \\ sx & sxy \end{vmatrix}}{n \cdot sx2 - (sx)^2} \rightarrow b = \frac{n \cdot sxy - sx \cdot sy}{n \cdot sx2 - (sx)^2}$$

De esta forma hemos obtenido las fórmulas de los coeficientes de la recta de regresión r(x), que podemos expresar en función de los sumatorios:

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

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

Una vez hemos obtenido estos valores, simplemente hay que sustituirlos en r(x)=a+bx y ya podemos obtener el valor aproximado de la función para cualquier punto pedido.

Ejemplo práctico

El ejercicio se nos plantea de la siguiente forma:

x	0,10	0,80	1,20	1,20	1,70	2,50
y	-1,00	0,95	1,80	1,90	2,10	3,60

En donde tenemos seis puntos como podéis observar en la tabla superior (n= 6). Y lo que buscamos con este ejercicio es obtener la expresión de la recta de regresión y particularizar en x= 1.

Empezaremos creando una tabla con cuatro operaciones, para facilitar y agilizar mucho más rápido la operación de las fórmulas finales, pues como hemos visto en la práctica para calcular a y b solo necesitaremos tener en mente cuatro tipos de incógnitas, $\mathbf{s_j}$, $\mathbf{y_j}$, $\mathbf{s_j^2}$, $\mathbf{s_j \cdot y_j}$.

	s_j	y_j	s_j^2	$s_j \cdot y_j$
	0,10	-1,00	0,01	-0,10
	0,80	0,95	0,04	0,76
	1,20	1,80	1,44	2,16
	1,20	1,90	1,44	2,28
	1,70	2,10	2,89	3,57
	2,50	3,60	6,25	9,00
Suma (Σ)	7,50	9,35	12,67	17,67

Que si recurrimos a las fórmulas que hemos obtenido anteriormente:

$$\sum_{j=1}^{n} a + b\sum_{j=1}^{n} s_j = \sum_{j=1}^{n} y_j$$

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

Y lo sustituimos nos quedaría lo siguiente:

$$6a + 7{,}5b = 9{,}35$$

$$7{,}5a + 12{,}67b = 17{,}67$$

Que si resolvieramos esta ecuación nos va a dar que:

$$a = -0{,}7112$$

$$b = 1{,}8156$$

Dándonos que la recta de regresión sea igual a:

$$r(x) = -0{,}7112 + 1{,}8156x$$