CLASE No. 25

Un ejemplo para estudiantes de estadística

Método de mínimos cuadrados
Supongamos que tenemos una colección de k puntos de coordenadas (x_i, y_i) con $1 \leq i \leq n$. Esto es, una conjunto de datos.

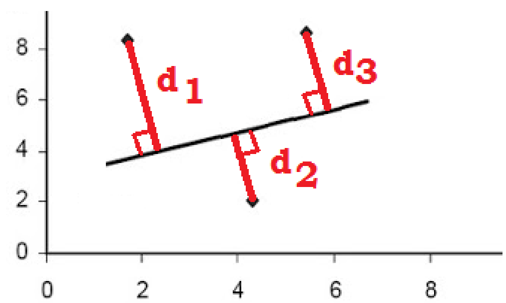
$$\{(x_1, y_1), (x_2, y_2), (x_3, y_3), \dots, (x_k, y_k)\}$$

y queremos encontrar la recta de ecuación

$$y = mx + n$$

que mejor ajuste a ellos.

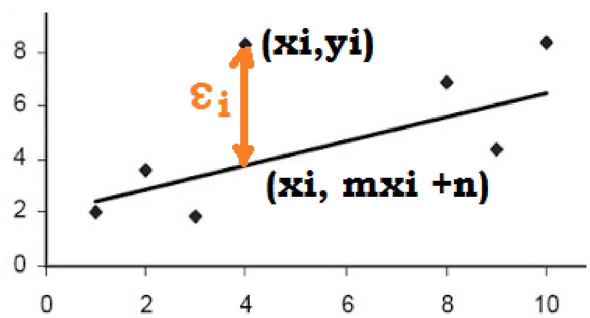
El criterio para elegirla será que la suma de las distancias de los puntos experimentales a la recta sea mínima.



Como ocupar la fórmula de la distancia de un punto a una recta puede resultan complicada en este caso (tenemos n puntos), utilizaremos alternativamente la resta de las ordenadas del punto experimental (y_i) menos la ordenada del punto teórico sobre la recta $(mx_i + n)$.

Llamaremos a esa diferencia ε_i:

$$\varepsilon_i = y_i - (mx_i + n)$$



Como además puede haber puntos experimentales por encima $(\varepsilon_i > 0)$ o por debajo $(\varepsilon_i < 0)$ de la recta, y esas diferencias, aunque grandes, podrían compensarse, utilizaremos sus cuadrados. De ahí el nombre del método, **método de los mínimos cuadrados**.

Por tanto, buscamos encontrar los valores de m y n que minimizar la suma,

$$\varepsilon = \sum_{i=1}^{k} (y_i - mx_i - n)^2$$

Con un poco de cálculo, y buscando mínimos, obtenemos …

$$\frac{d\varepsilon}{dm} = 0 \qquad \Rightarrow \qquad -2 \cdot \sum_{i=1}^{k} (y_i - mx_i - n)x_i = 0$$

$$\frac{d\varepsilon}{dn} = 0 \qquad \Rightarrow \qquad -2 \cdot \sum_{i=1}^{k} (y_i - mx_i - n) = 0$$

De donde obtenemos un sistema de ecuaciones en m y n:

$$\begin{cases} m \cdot \sum_{i=1}^{k} (x_i)^2 + n \cdot \sum_{i=1}^{k} (x_i) = \sum_{i=1}^{k} (x_i y_i) \\ m \cdot \sum_{i=1}^{k} (x_i) + n.k = \sum_{i=1}^{k} (y_i) \end{cases}$$

Simplifiquemos la notación, eliminando el símbolo de sumatoria.

$$\begin{cases} m \cdot S_{xx} + n \cdot S_x = S_{xy} \\ m \cdot S_x + n.k = S_y \end{cases}$$

Expresándolo en forma matricial, queda

$$\begin{pmatrix} S_{xx} & S_x \\ S_x & k \end{pmatrix} \cdot \begin{pmatrix} m \\ n \end{pmatrix} = \begin{pmatrix} S_{xy} \\ S_y \end{pmatrix}$$

Y resolviendo el sistema por Cramer, obtenemos:

$$\Delta = \begin{vmatrix} S_{xx} & S_x \\ S_x & k \end{vmatrix}$$

$$\Delta_m = \begin{vmatrix} S_{xy} & S_x \\ S_y & k \end{vmatrix}$$

$$\Delta_n = \begin{vmatrix} S_{xx} & S_{xy} \\ S_x & S_y \end{vmatrix}$$

Finalmente,

$$\boldsymbol{m = \frac{k \cdot S_{xy} - S_x \cdot S_y}{k \cdot S_{xx} - S_x \cdot S_x}} \qquad\qquad \boldsymbol{n = \frac{S_{xx} - S_{xy} \cdot S_x}{k \cdot S_{xx} - S_x \cdot S_x}}$$

Un ejemplo con matrices

Se sabe que el 1er. semestre de la carrera lo terminan el 80% de los alumnos inscritos, el 2º. Semestre 72% y el 3er. semestre 75%. Con esos datos extrapole cuantos estudiantes terminarán la carrera si ésta tiene 8 semestres.

Solución:

Tenemos una colección de tres puntos:

semestre	terminan
1	0,80
2	0,72
3	0,75

Tenemos tres puntos. Y escribamos las tres igualdades:

$$y_1 = mx_1 + n$$
$$y_2 = mx_2 + n$$
$$y_3 = mx_3 + n$$

En general, sería:

$$\begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ x_3 & 1 \end{pmatrix} \cdot \begin{pmatrix} m \\ n \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}$$

Que en forma matricial, puede expresarse como:

$$\boldsymbol{A} \cdot \boldsymbol{X} = \boldsymbol{B}$$

Multipliquemos a ambos lados de la igualdad por la matriz transpuesta $\boldsymbol{A^t}$

$$\boldsymbol{A^t} \cdot \boldsymbol{A} \cdot \boldsymbol{X} = \boldsymbol{A^t} \cdot \boldsymbol{B}$$

$$\begin{pmatrix} x_1 & x_2 & x_3 \\ 1 & 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ x_3 & 1 \end{pmatrix} \cdot \begin{pmatrix} m \\ n \end{pmatrix} = \begin{pmatrix} x_1 & x_2 & x_3 \\ 1 & 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}$$

Operando, nos quedará:

$$\begin{pmatrix} S_{xx} & S_x \\ S_x & k \end{pmatrix} \cdot \begin{pmatrix} m \\ n \end{pmatrix} = \begin{pmatrix} S_{xy} \\ S_y \end{pmatrix}$$

que es el sistema que antes resolvimos por Cramer.

Apliquemos el procedimiento al ejemplo propuesto:

Será:

$$1m + n = 0{,}80$$
$$2m + n = 0{,}72$$
$$3m + n = 0{,}75$$

En forma matricial:

$$\begin{pmatrix} 1 & 1 \\ 2 & 1 \\ 3 & 1 \end{pmatrix} \cdot \begin{pmatrix} m \\ n \end{pmatrix} = \begin{pmatrix} 0{,}80 \\ 0{,}72 \\ 0{,}75 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 2 & 3 \\ 1 & 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 \\ 2 & 1 \\ 3 & 1 \end{pmatrix} \cdot \begin{pmatrix} m \\ n \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} 0{,}80 \\ 0{,}72 \\ 0{,}75 \end{pmatrix}$$

$$\begin{pmatrix} 14 & 6 \\ 6 & 3 \end{pmatrix} \cdot \begin{pmatrix} m \\ n \end{pmatrix} = \begin{pmatrix} 4{,}49 \\ 2{,}27 \end{pmatrix}$$

Obtenemos,

$$f(x) = -0{,}0025x + 0{,}806$$

En el octavo semestre será:

$$f(8) = -0{,}0025 \cdot 8 + 0{,}806$$

$$\boldsymbol{f(8) = 0,606}$$

Terminarán la carrera aproximadamente el 61% de los estudiantes.