La recta de regresión

Cuando una distribución estadística bidimensional presenta una correlación fuerte o muy fuerte, tiene sentido averiguar cuál es la recta que mejor representa a su conjunto de valores. Esto permite resumir toda la distribución y razonar cuál podría ser el valor de una variable que corresponda a un valor de la otra. Esa recta se denomina recta de regresión.

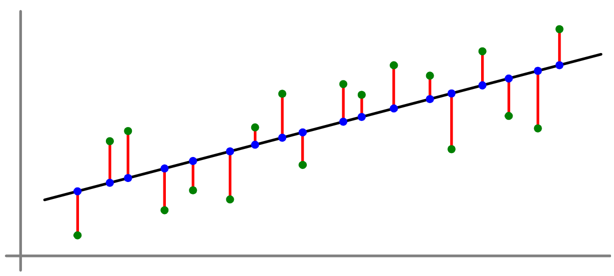
Método de los mínimos cuadrados

En 1795 el matemático alemán Karl Fiedrich Gauss (1777-1855) dio con las primeras ideas generales del método de los mínimos cuadrados, que en 1801 permitió continuar observando el planeta enano Ceres, tras perderse su trayectoria. Resultó que el método es óptimo para resolver muchos problemas. Consiste en averiguar los valores que hacen que cierta suma de cuadrados dé el menor resultado posible.

Definición de la recta de regresión

Consideramos una distribución estadística bidimensional con las variables X e Y, que consista en un número de parejas de valores (x_i,y_i). La recta de regresión tendrá ecuación explícita «$y=mx+q$». Llamamos $y(x_i)$ al valor de la ordenada de la recta para $x=x_i$, es decir: $y(x_i)=mx_i+q$. La recta de regresión es la que hace mínima la suma de los cuadrados de las distancias entre (x_i,y_i) y $(x_i,y(x_i))$.

Ilustración



- ✱ Mostramos en verde la nube de puntos de la distribución, los puntos (x_i,y_i).
- ✱ Mostramos en azul los puntos $(x_i,y(x_i))$.
- ✱ Señalamos en rojo los segmentos que unen (x_i,y_i) con $(x_i,y(x_i))$.
- ✱ La recta de regresión, dibujada en negro, es la que hace mínima la suma de los cuadrados de las longitudes de los segmentos en rojo.

Obtención de la recta de regresión

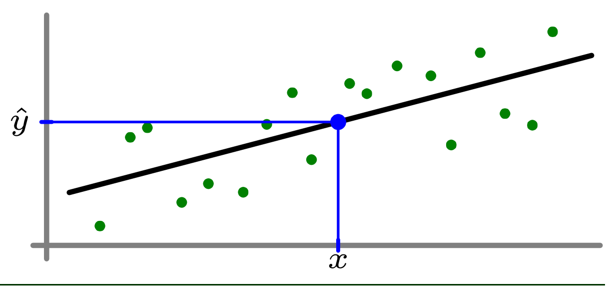
Mediante métodos que no se pueden explicar en enseñanza secundaria, se demuestra que la recta de regresión que hemos definido cumple dos propiedades:

- ✱ Pasa por el centro de gravedad de la distribución, el punto $(\overline{x},\overline{y})$.
- ✱ Su pendiente es $\frac{\sigma_{xy}}{\sigma_x^2}$, número llamado coeficiente de regresión.

Por tanto, su ecuación punto-pendiente es

$$\boxed{y-\overline{y} = \frac{\sigma_{xy}}{\sigma_x^2}(x-\overline{x})}$$

Uso de la recta de regresión



Esta recta se utiliza para calcular el valor aproximado que puede tener la variable Y para algún valor de la variable X que no sea uno de los datos. El valor de Y obtenido para un valor *x* se denomina **valor estimado**, y se representa $\hat{y}$ o bien $\hat{y}(x)$.

