RETTA DI REGRESSIONE LINEARE (RETTA AI MINIMI QUADRATI, LEAST SQUARES)

DATI

(x_1, y_1) (x_2, y_2) —— (x_N, y_N)

Sembra esserci una relazione lineare approssimata tra le x_i e le y_i, per $i = 1, \ldots, N$

$$\boxed{Y = mx + q}$$

<u>INCOGNITE</u> $m = ?$ $q = ?$

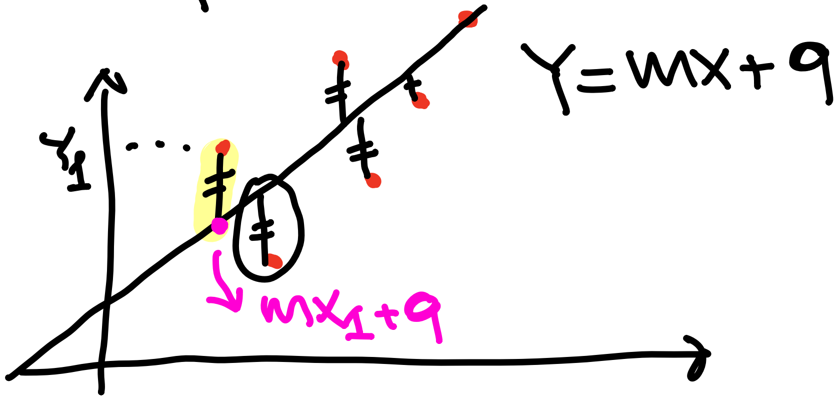
$$\begin{cases} Y_1 = m x_1 + q \\ Y_2 = m x_2 + q \\ \vdots \\ Y_N = m x_N + q \end{cases}$$

Ho un sistema lineare con solo 2 incognite e $N \gg 2$ equazioni.

$\gg$ molto più grande

In generale tale sistema <u>non</u> ammette soluzione. Fa eccezione il caso in cui le x_i e le Y_i

sono perfettamente allineate.



IDEA Cerchiamo la retta (il valore di m e quello di q) che rende minima la seguente quantità

$$L(m,q)=\sum_{i=1}^{N}(y_i-mx_i-q)^2$$

L si dice funzione di costo (cost function, loss function)

Per cercare il minimo si pone uguale a zero la derivata di L rispetto ad m e la derivata di L rispetto a q:

$$\begin{cases} \dfrac{\partial L}{\partial m} = 0 \\ \dfrac{\partial L}{\partial q} = 0 \end{cases}$$

Noi daremo il valore di m e q utilizzando dei concetti di statistica.

<u>DEF</u> Si dice <u>valor medio</u> di $x_1 - x_N$ la

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Analogamente

$$\bar{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i$$

Nel nostro esempio (Altezza padre/figlio)

$\overline{X} \simeq 179.85 \qquad \overline{Y} \simeq 145.2$

DEF Si definisce varianza di $x_1 — x_N$

$$\underbrace{S_x^2}_{\text{NOTAZIONE}} = \underbrace{\frac{1}{N}}_{\text{medio}} \sum_{i=1}^{N} \Big(\underbrace{x_i - \overline{X}}_{\text{scarto}}\Big)^{\overset{\text{quadratico}}{2}}$$

S_x^2 è uno scarto quadratico medio

La varianza fornisce informazioni su come le x_i sono distribuite intorno ad $\overline{X}$. Più S_x^2 è piccola e più i dati sono vicini ad $\overline{X}$.

Analogamente

$$S_Y^2 = \frac{1}{N} \sum_{i=1}^{N} (Y_i - \overline{Y})^2$$

Nel nostro esempio $S_x^2 \simeq 42.63$

DEF Si dice covarianza di

$x_1 \ldots x_N$ ed $y_1 \ldots y_N$

$$\underbrace{S_{xy}}_{\text{Notazione}} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})$$

Per il nostro esempio

$$S_{xy} \simeq 37.63$$

TEOREMA La retta ai minimi quadrati per i dati (x_i, y_i), $i = 1, \ldots, N$, è la retta di equazione

$$y = mx + q$$

con $\boxed{m = \frac{S_{XY}}{S_X^2}}$ e

$$\boxed{q = \overline{Y} - m\overline{X}}$$

<u>oss</u> $(\overline{X}, \overline{Y})$ appartiene alla retta di regressione lineare

---

$$A = \begin{pmatrix} 1 & x_1 \\ | & | \\ 1 & x_N \end{pmatrix}$$

$$A^T = \begin{pmatrix} 1 & \text{—} & 1 \\ x_1 & \text{—} & x_N \end{pmatrix}$$

$$(A^T A)\begin{pmatrix} m \\ q \end{pmatrix} = A^T \begin{pmatrix} Y_1 \\ | \\ Y_N \end{pmatrix}$$