Cos'è la regressione lineare?

La **regressione lineare** è un metodo matematico che permette di trovare una **retta** che "si avvicina il più possibile" ai punti di un grafico, per descrivere la relazione tra due variabili numeriche.

In pratica:

- Se abbiamo tanti dati che collegano una causa e un effetto (es: ore di studio e voto, chilometri percorsi e benzina consumata...)
- La regressione lineare cerca la retta che meglio "riassume" questa relazione
- Così possiamo **usare la retta per fare previsioni**: se so quante ore studio, posso stimare il voto che prenderò

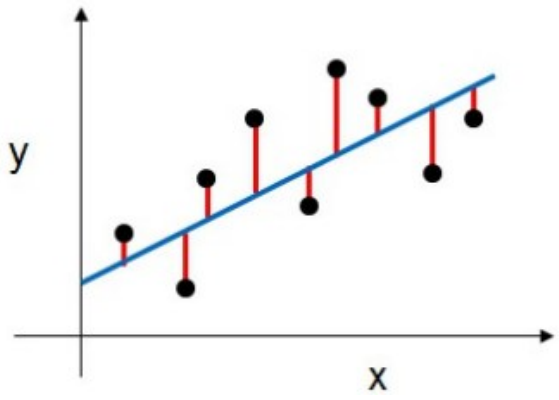
In sintesi:

La regressione lineare trova la linea che meglio collega i dati, e ci aiuta a prevedere un valore conoscendo un altro.

La linea viene chiamata linea di best fit

Possiamo tracciare infinite linee fra i punti dei dati, ma che cosa rende "migliore" una di esse?

Considerate il seguente grafico.



La formula della retta (modello):

$\hat{y} = ax + b$

$\hat{y}$: valore previsto

x: variabile indipendente(es: ore di studio)

a: coefficiente angolare (quanto cresce Y se X aumenta di 1)

b: intercetta (valore di Y quando $X = 0$)

Come si calcolano a e b

$$a = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2}$$

$$b = \overline{y} - a\overline{x}$$

Dove:

- x_i, y_i: i dati osservati
- $\overline{x}, \overline{y}$: le medie di X e Y

ECCO UN **ESEMPIO NUMERICO CONCRETO** DI REGRESSIONE LINEARE SEMPLICE, PASSO PASSO:

Dati raccolti:

Supponiamo di avere questi dati su **ore di studio** (X) e **voti al test** (Y):

Ore di studio (X)	Voto (Y)
2	65
4	70
6	75
8	80

---

1. Calcoliamo le medie

$$\overline{x} = \frac{2 + 4 + 6 + 8}{4} = 5$$

$$\overline{y} = \frac{65 + 70 + 75 + 80}{4} = 72.5$$

---

2. Calcoliamo a

$$a = \frac{(2-5)(65-72.5) + (4-5)(70-72.5) + (6-5)(75-72.5) + (8-5)(80-72.5)}{(2-5)^2 + (4-5)^2 + (6-5)^2 + (8-5)^2}$$

Calcolo numeratore:

- (2-5) x (65 -72.5) = (-3) * (-7.5) = 22.5
- (4-5) x (70-72.5) = (-1) * (-2.5) = 2.5
- (6-5) x (75-72.5) = 1 * 2.5 = 2.5
- (8-5) x (80-72.5) = 3 * 7.5 = 22.5

- **Totale numeratore:** 22.5 + 2.5 + 2.5 + 22.5 = **50**

Calcolo denominatore:

- $(2\text{-}5)^2 = 9$
- $(4\text{-}5)^2 = 1$
- $(6\text{-}5)^2 = 1$
- $(8\text{-}5)^2 = 9$
- **Totale denominatore:** 9 + 1 + 1 + 9 = **20**

Quindi:

$$a = \frac{50}{20} = 2.5$$

---

3. Calcoliamo b

$$b = \overline{y} - a \cdot \overline{x} = 72.5 - 2.5 \times 5 = 72.5 - 12.5 = 60$$

---

4. La retta trovata è:

$$\hat{y} = 2.5x + 60$$

---

5. Uso la retta per prevedere:

Ad esempio, se uno studente studia **7 ore**, il voto previsto è:

$$\hat{y} = 2.5 \times 7 + 60 = 17.5 + 60 = 77.5$$

Le formule della regressione lineare multipla con più variabili indipendenti sono più complesse ma **l'idea di base resta la stessa**: trovare i coefficienti che "fanno combaciare" il meglio possibile i dati con una superficie (un iperpiano) invece che una retta. Ve la risparmio 😉