Relazioni statistiche tra variabili quantitative: correlazione e regressione

A cura di

Francesco Fabi

ce3s

CENTRO STUDI
STATISTICI
E SOCIALI

Un esempio

Età in mesi	Altezza in cm
18	76.01
19	77.00
20	78.10
21	78.20
22	78.80
23	79.70
24	79.90
25	8110
26	8120
27	8180
28	82.80
29	83.50

L'obiettivo

Unità	Carattere X	Carattere Y
1	x_1	y_1
2	x_2	y_2
…	…	…
n	x_n	y_n

- Per le due variabili quantitative X e Y considerate, ci interessa valutare l'associazione, la relazione o il legame che più precisamente chiameremo correlazione.

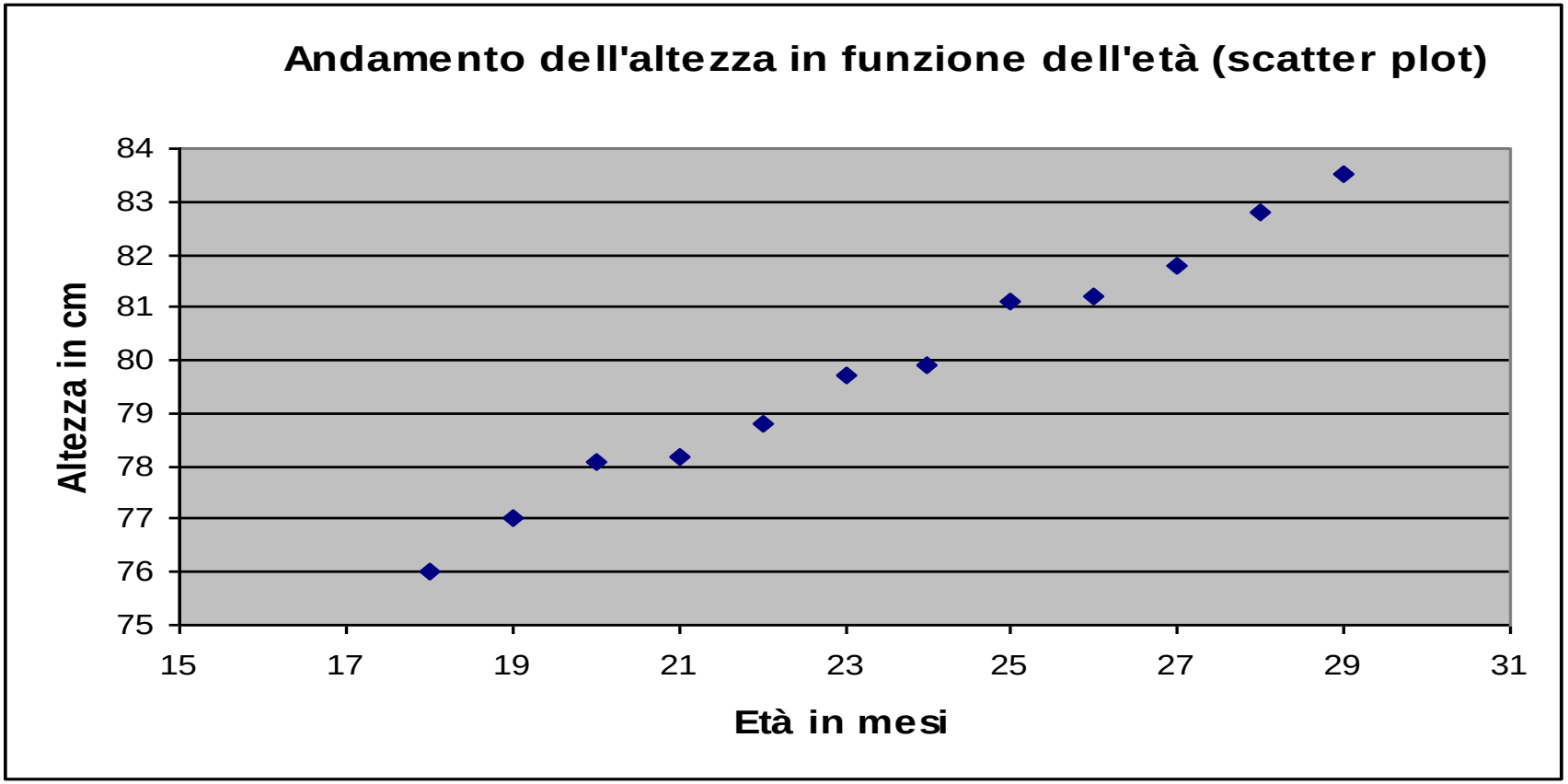
Visualizzazione: lo scatter plot

Per valutare ***qualitativamente*** l'esistenza di ***associazione (correlazione)*** tra due caratteri quantitativi, si procede ad una prima analisi grafica attraverso la costruzione di un **diagramma di dispersione o scatter plot**.

Si riportano in ascissa le misure osservate $x_1, x_2,\ldots,x_n$ del **carattere X,** in ordinata le corrispondenti misure osservate $y_1, y_2,\ldots,y_n$ di **Y.**

Le **singole osservazioni** (x_i,y_j) vengono così rappresentate con dei punti su un piano cartesiano.

Lo scatter plot



Tipi di correlazione

- Se i **punti sono sparsi senza apparenti regolarità**

  $\Rightarrow$ **non c'è correlazione tra i caratteri**;
- Se appare una certa regolarità, ad esempio se:
  - punti con **ascissa piccola** hanno **ordinata piccola** e punti con **ascissa grande** hanno **ordinata grande** $\Rightarrow$ **c'è correlazione diretta** (positiva) tra i due caratteri
  - punti con **ascissa piccola** hanno **ordinata grande** e punti con **ascissa grande** hanno **ordinata piccola** $\Rightarrow$ **c'è correlazione inversa** (negativa) tra i due caratteri

Correlazione

Si abbiano **n** osservazioni congiunte di due caratteri, $\{(x_1,y_1),\ldots,(x_i,y_i),\ldots,(x_n,y_n)\}$, e siano μ_X e μ_Y le medie aritmetiche di X e Y. Se a modalità di X **maggiori** di μ_X sono associate modalità di Y **maggiori** di μ_Y e viceversa, allora per ogni unità statistica gli scarti $(x_i-\mu_X)$ e $(y_i-\mu_Y)$ tendono a essere concordi, e quindi i loro **prodotti**

$$(x_i - \mu_x)(y_i - \mu_y) > 0$$

sono con maggiore frequenza **positivi**.

La media di tali prodotti si chiama **covarianza** e sarà anche essa positiva se i prodotti tendono ad essere concordi.

Covarianza

Si dice covarianza dei due **caratteri X** e **Y** la funzione:

$$\sigma_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

n-1 e non n sono i gradi di libertà della covarianza perché un'osservazione è vincolata dal calcolo della media.

Covarianza

In altre parole si può dire che:

- **SE $\sigma_{xy} > 0$** $\Rightarrow$ mediamente a valori grandi (piccoli) di X corrispondono valori grandi (piccoli) di Y, e **Y è direttamente correlata ad X, (X e Y sono direttamente correlati)**;
- **SE $\sigma_{xy} < 0$** $\Rightarrow$ mediamente a valori grandi (piccoli) di X corrispondono valori piccoli (grandi) di Y, e **Y è inversamente correlata ad X, (X e Y sono inversamente correlati)**;
- **SE $\sigma_{xy} = 0$** $\Rightarrow$ **X e Y sono incorrelate o la loro relazione non è monotona.**

Correlazione:

Grado di associazione tra la variabile x e la variabile y è espresso dal coefficiente di correlazione lineare

→ Grado di relazione lineare tra x e y

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

Scarti standard delle due variabili dalle rispettive medie

Proprietà 1

- ρ_{xy} varia tra -1 e 1;
- ρ_{xy} è esattamente uguale a 1 se e solo se le due variabili sono perfettamente correlate positivamente e i punti $\mathbf{(x_i,y_i)}$ risultano allineati**;**
- ρ_{xy} è esattamente uguale a -1 se e solo se le due variabili sono perfettamente correlate negativamente e i punti (x_i,y_i) risultano allineati;
- Quanto più ρ_{xy} è prossimo ai valori estremi, tanto più è stretta l'associazione tra le due variabili.

Proprietà 2

- ρ_{xy} è un indice "normalizzato", cioè la cui grandezza ha un significato assoluto, è adimensionale.
- Opera direttamente sui valori osservati dei caratteri quantitativi senza passare a misure raggruppate in classi, che producono perdita di informazione;
- se il legame tra le variabili è non lineare ρ_{xy} **basso in valore assoluto (prossimo a 0) indica scarsa associazione fra le variabili**.

Regressione

- Se lo scatterplot evidenzia che i punti sono disposti attorno a una **retta crescente** o **decrescente**, si parla di **correlazione lineare**.
- In tal caso si può tracciare una **retta di regressione** (interpolante*) a partire dai dati.
- In ogni caso la costruzione di indici numerici è necessaria per valutare quantitativamente l'entità dell'associazione.

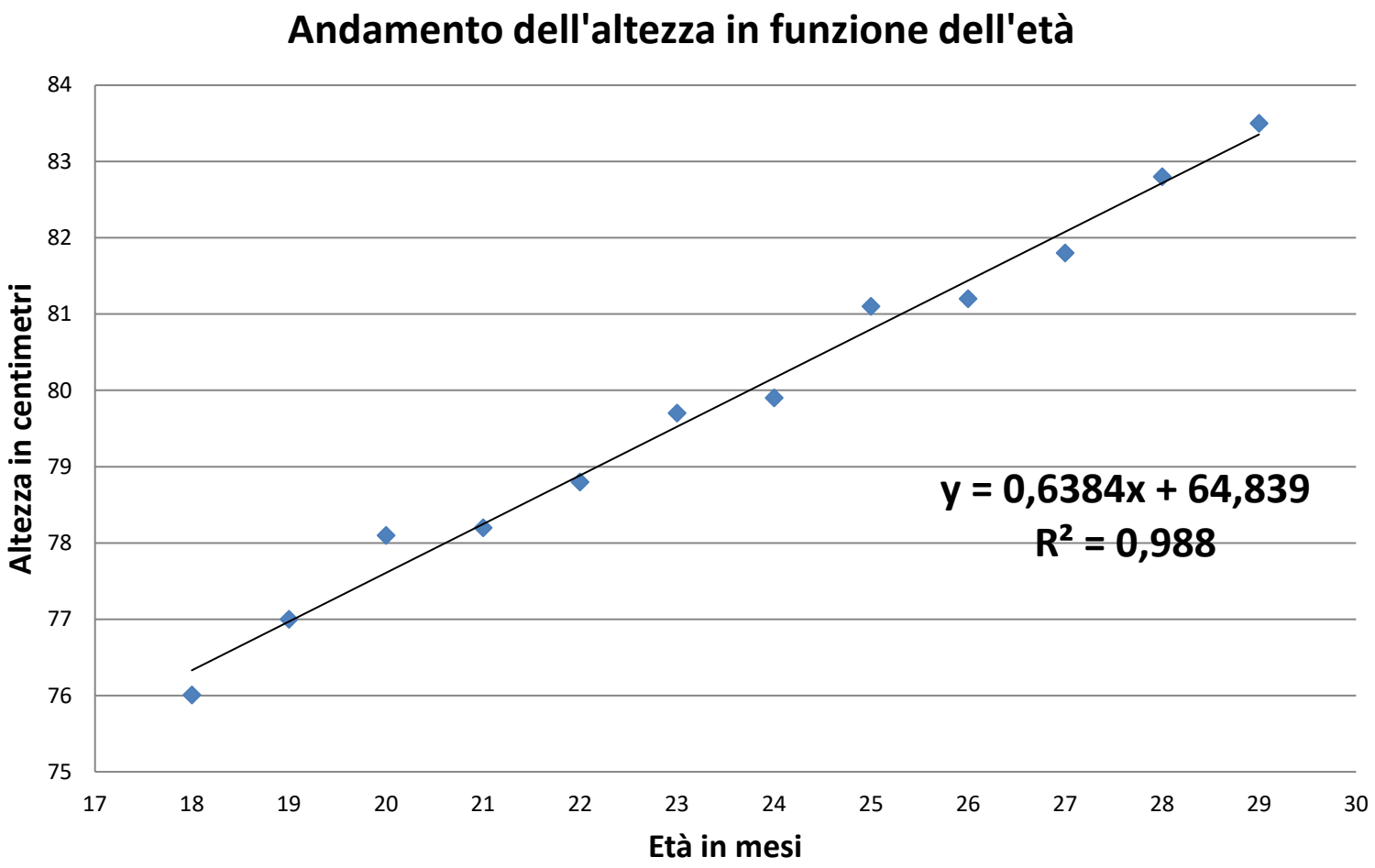
*Che approssima la tendenza dei dati

Regressione lineare semplice

Variabile indipendente X
Variabile dipendente Y

assume che la relazione tra x e y possa essere riassunta graficamente sotto forma di una retta

La retta interpolante

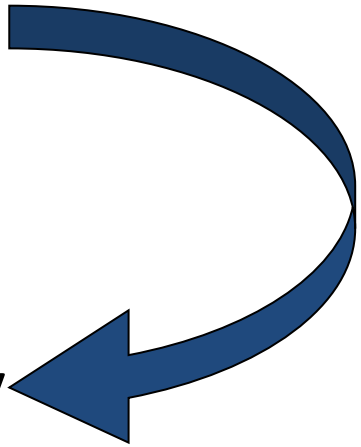


Significato della retta

- La retta che interpola i punti sperimentali rappresenta il modello teorico dei dati.
- Essa permette di prevedere, data l'età di un bambino, quale dovrebbe essere in media la sua altezza.
- I punti molto lontani dalla retta, sono delle osservazioni anomale.

Primo utilizzo

Predittività: Uso di x per predire i valori di y



Se il grado di correlazione tra x e y è alto

L’ Y osservato sulla retta sarà un buon predittore del y reale

Secondo utilizzo:correlazione

Il coefficiente di determinazione R^2

$$R^2 = (\rho_{XY})^2$$

- R^2 varia tra 0 e 1.
- R^2 prossimo a 1 $\Rightarrow$ *buon adattamento della retta di regressione ai dati osservati*.
- R^2 prossimo a 0 $\Rightarrow$ *cattivo adattamento della retta di regressione ai dati osservati*.