

Elementi di Matematica e Statistica

La Correlazione e il Modello di Regressione Semplice

Docente: Riccardo Ievoli
riccardo.ievoli@unife.it

Corso di Laurea in Biotecnologie

17/10/2025

Outline

- 1 La Correlazione
- 2 Il Modello di Regressione (semplice)

Introduzione agli indici di correlazione e di associazione

- Fino ad ora abbiamo trattato solo l'analisi univariata.
- Ma di solito si lavora con tante variabili ($p > 1$)
- Come è possibile combinare le informazioni disponibili?

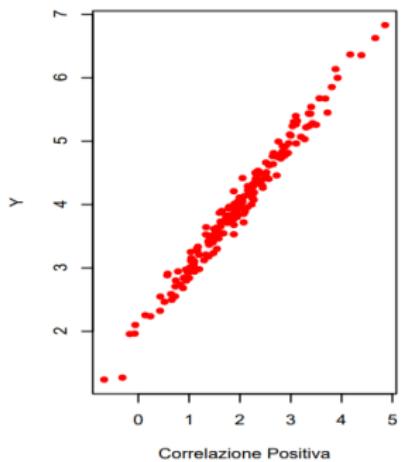
Possibili combinazioni (caso bivariato)

- ➊ Due variabili quantitative: **covarianza e correlazione (r)**
- ➋ Due variabili qualitative: associazione (indice χ^2)
- ➌ Qualitativa e quantitativa: scomposizione della devianza

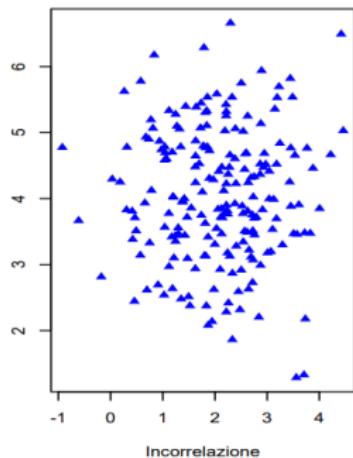
La Correlazione

Intuizione Grafica

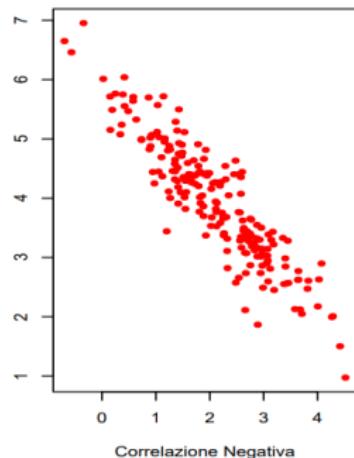
A



B



C



La Correlazione

Dipendenza e Indipendenza lineare

- Si ipotizza una relazione matematica tra i punti, in prima battuta **una retta**.
- Si parla di indipendenza lineare se la relazione tra due variabili (quantitative) può essere descritta mediante una retta parallela all'asse delle x (coefficiente angolare nullo).

Esistono due misure per esprimere tale relazione:

- Indice di correlazione di Pearson.
- Il coefficiente di regressione lineare.

Le suddette misure sono legate a quantità simili (Covarianza e varianze), ma hanno una interpretazione differente

La Correlazione

La Covarianza

La propensione di due variabili numeriche (ossia X e Y) nel variare insieme può essere misurata attraverso la **covarianza**:

$$\text{Cov}(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n}$$

Alcuni commenti

- La quantità al numeratore è detta anche CoDevianza
- $\text{Cov}(x, y) \in \mathbb{R}$, cioè può essere anche negativa (o nulla).
- $v^2(x) = \text{Cov}(x, x)$, ed inoltre $\text{Cov}(x, y) = \text{Cov}(y, x)$
- Risente dell'ordine di grandezza delle variabili
e delle unità di misura

La Correlazione

La Covarianza (2)

Così come per la varianza, la covarianza può essere calcolata utilizzando la **media dei prodotti** e il **prodotto delle medie**

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

Anche in questo caso è possibile evitare di passare per gli scarti dalle medie, e utilizzare soltanto tre “oggetti”: x , y e $x \cdot y$

La Correlazione

L'indice di Correlazione

Obiettivo: ottenere un indice indipendente (normalizzato) dalla combinazione di unità di misura

- ① Siano X e Y due variabili numeriche con medie (osservate) \bar{x}, \bar{y}
- ② Sia $v^2(x) = n^{-1} \sum_{i=1}^n (x_i - \bar{x})^2$ (si può indicare anche con σ_x^2)
- ③ Sia $v^2(y) = n^{-1} \sum_{i=1}^n (y_i - \bar{y})^2$ (si può indicare anche con σ_y^2)
- ④ Infine: $v(x, y) = n^{-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$ (oppure σ_{xy})

Indice di correlazione

$$r = \frac{\text{Cov}(x, y)}{v(x)v(y)} = \frac{\sigma_{xy}}{\sigma_x\sigma_y}$$

La Correlazione

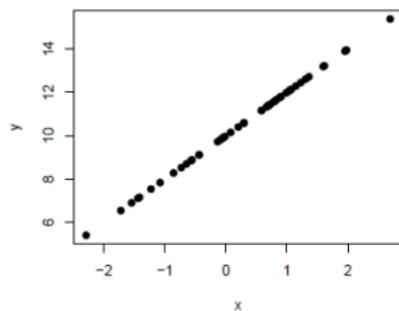
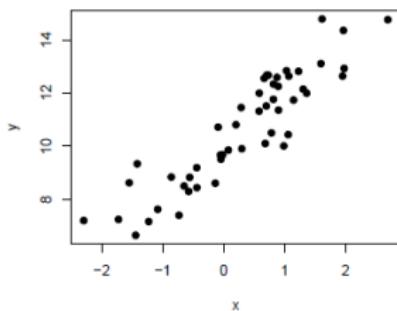
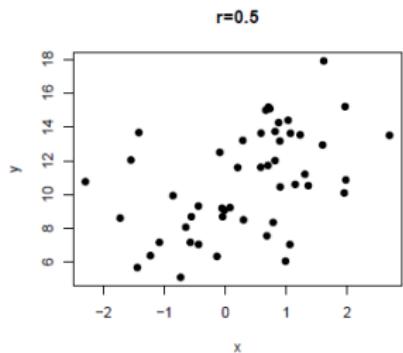
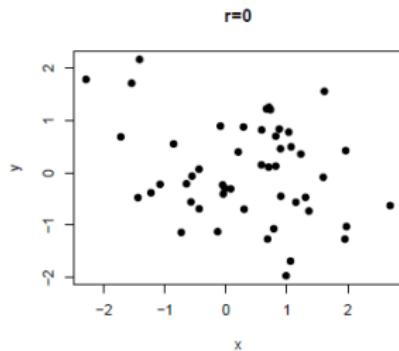
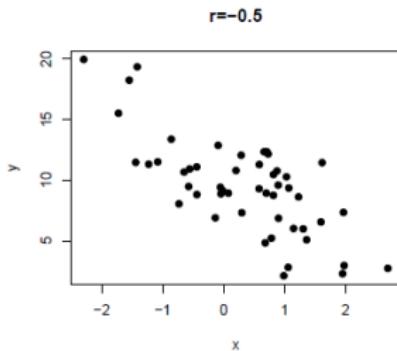
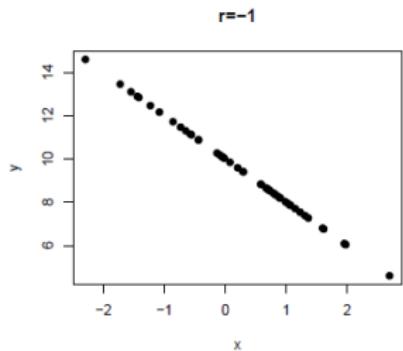
Caratteristiche dell'indice di Correlazione

L'indice r viene detto Indice (di correlazione) di Pearson

- $r = -1$: massima correlazione negativa
- $-1 < r < 0$: correlazione negativa
- $r = 0$: assenza di correlazione, ovvero **indipendenza lineare**
- $0 < r < 1$: correlazione positiva
- $r = 1$: massima correlazione positiva

La Correlazione

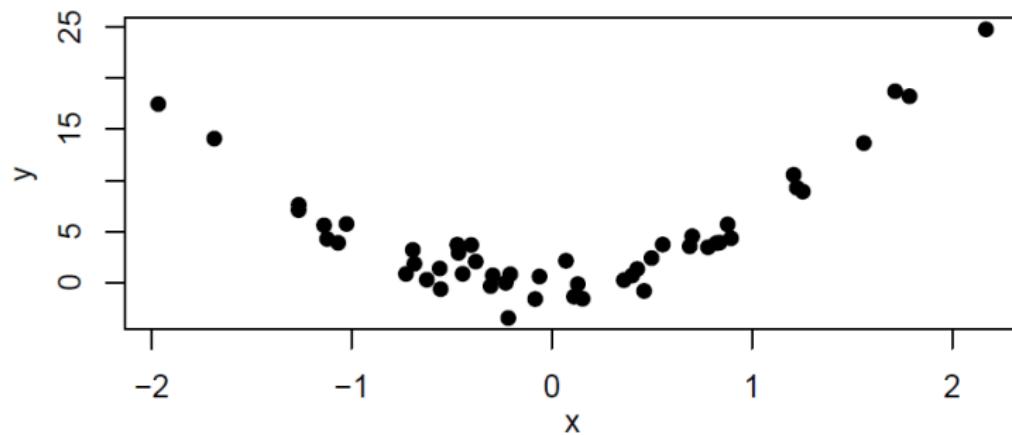
Esempi Grafici



La Correlazione

Possibili problematiche e riflessioni

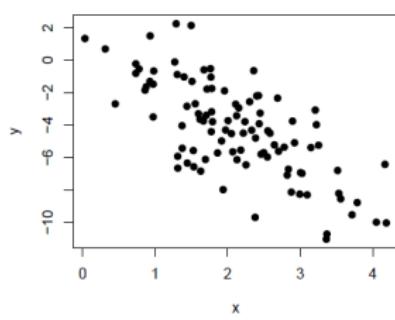
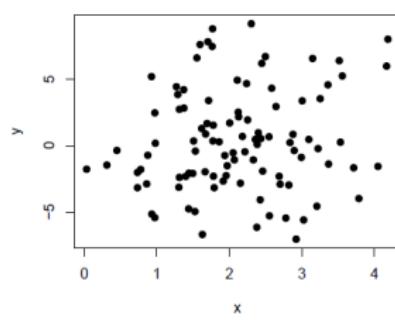
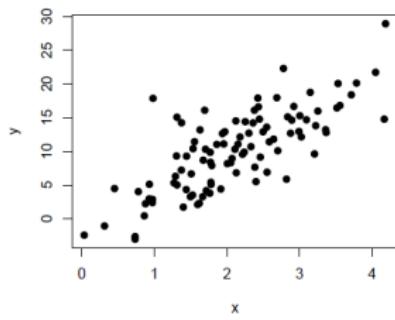
- Non Linearità della relazione tra le variabili (si veda l'esempio)
- Correlazione spuria:
<https://www.tylervigen.com/spurious-correlations>
- La correlazione è...un gioco?
<https://www.guessthecorrelation.com/>



La Correlazione

Esercizio 1

Si considerino i seguenti diagrammi a dispersione



- Sono noti i 3 valori assunti dall'indice di correlazione lineare di Pearson: $r = 0,13$; $r = -0,71$ e $r = 0,79$.
- È possibile stabilire, senza fare calcoli, a quale figura corrisponde ciascun indice?

La Correlazione

Esercizio 2

In uno studio sull'Epatite A nei soggetti con età maggiore di 40 anni si dispone delle seguenti variabili: X =età (anni), Y =Bilirubina (mg/dL) per $n = 10$ soggetti. Calcolare il coefficiente di correlazione r sulla base dei dati seguenti.

X	Y
78	7,5
72	12,9
81	14,3
59	8,0
64	14,1
48	10,9
46	12,3
42	1,0
58	5,2
52	5,1

Il modello di regressione

Dalla correlazione alla regressione

L'indice di correlazione (r) misura l'intensità di una relazione **simmetrica**.

Può essere utile misurare l'entità di una relazione logica tra le due variabili, ipotizzando Y come variabile di risposta in funzione di X , ossia la esplicativa.

Tale relazione lineare sarà descritta dalla seguente retta:

$$y = b_0 + b_1 x$$

dove b_0 è l'intercetta e b_1 il coefficiente angolare. Ovviamente la relazione non è esatta: quindi si può scrivere

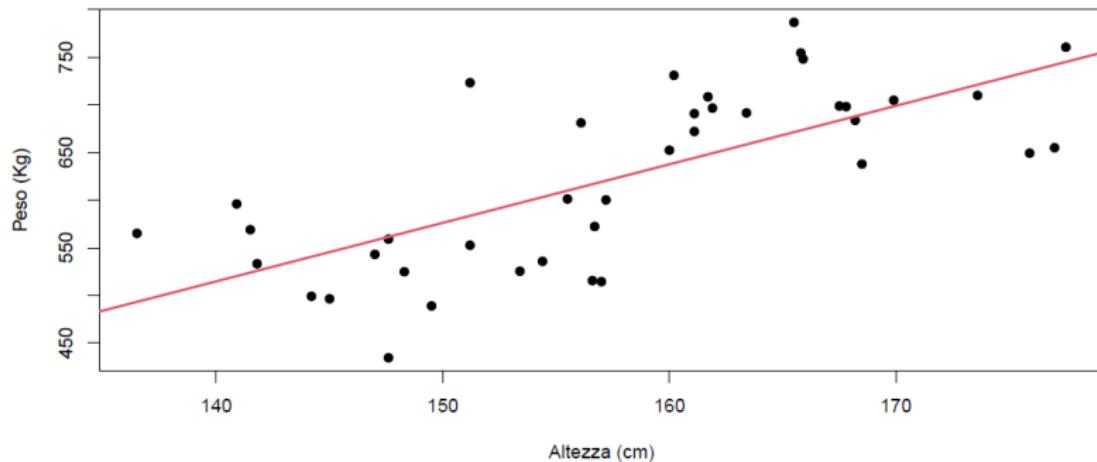
$$y_i = b_0 + b_1 x_i + \varepsilon_i \quad i = 1, \dots, n$$

dove ε è un termine di errore.

Il modello di regressione

La retta di regressione

- L'obiettivo è determinare una retta in grado di interpolare i punti campionari.
- Esempio. y : peso; x : altezza di 40 cavalli



Il modello di regressione

Calcolo statistico dei valori dei parametri

- Per ottenere i valori di b_0 e b_1 è necessario fissare un criterio di ottimo
- Solitamente si vuole minimizzare la somma quadratica degli scarti tra punti e retta (cd. metodo dei minimi quadrati).
- Si ricavano così i seguenti parametri:

$$b_1 = \frac{\text{Cov}(x, y)}{v^2(x)}; \quad b_0 = \bar{y} - b_1 \bar{x}$$

- b_1 è detto coefficiente di regressione e misura la “forza” della relazione lineare tra Y e X. Indica di quanto varia Y (in media) se X cresce di una unità.

Il modello di regressione

Il Coefficiente di Regressione

I valori che il coefficiente di regressione può assumere sono sintetizzabili come segue:

- Se $b_1 = 0$ la retta è orizzontale e indica assenza di relazione tra le variabili, ossia indipendenza lineare
- Se $b_1 > 0$ è presente una correlazione positiva tra le variabili.
- Se $b_1 < 0$ è presente una correlazione negativa tra le variabili.

Il modello di regressione

Il Coefficiente di Determinazione Lineare

Per valutare la bontà di adattamento della retta ai dati, si utilizza l'indice di **determinazione lineare** R^2 .

Nel caso di regressione **semplificata** (una variabile dipendente e una indipendente), esso coincide con il quadrato di r :

$$R^2 = r^2 = \left(\frac{\sigma_{xy}}{\sigma_x \sigma_y} \right)^2$$

R^2 assume valori compresi tra 0 (indipendenza lineare) e 1 (perfetta dipendenza lineare).

Il modello di regressione

Esercizio

Sulla base dei dati disponibili, qual è intensità della relazione tra Consumo di Gas ($\text{ft}^3 \cdot 1000$, variabile dipendente Y) e Temperatura Esterna ($^{\circ}\text{C}$, variabile indipendente X) in una abitazione?

X	Y
0,0	7,2
0,0	6,9
0,4	6,4
2,5	6,0
2,9	5,8
3,2	5,8
3,6	5,6
3,9	4,7
4,3	5,2
6,0	4,9

Riassunto: cosa abbiamo imparato oggi?

- ① La Covarianza
- ② L'indice di Correlazione
- ③ La retta di regressione
- ④ Il coefficiente di regressione
- ⑤ L'indice di determinazione lineare