Statistica sociale - 11

Prof. Antonio Mussino

a. a. 2016-2017

SAPIENZA
UNIVERSITÀ DI ROMA

L'elaborazione dei dati
(segue)

Il caso c): distribuzioni quanti-qualitative

- Ricordiamo che, nel caso c), abbiamo considerato le situazioni nelle quali:
  - la variabile ***indipendente*** è ***qualitativa*** e
  - quella ***dipendente*** è ***quantitativa***.

- Ad esempio "genere" è "frequenza della pratica nell'ultima settimana", oppure "numero di sport praticati".

- Si parla, in questo caso, di ***confronto fra medie***.

Confronto fra medie - 1

- Ovviamente più le medie differiscono al variare delle modalità della variabile indipendente più forte sarà l'intensità della relazione, ovvero l'associazione.
- Per misurare la forza dell'associazione definiamo le seguenti quantità:
- **Devianza Totale**, ovvero la media dei quadrati degli scarti dei singoli valori dalla media generale del collettivo;
- **Devianza Interna**, ovvero la media dei quadrati degli scarti dei singoli valori dalla loro media parziale di modalità;
- **Devianza dei Gruppi**, ovvero la media dei quadrati degli scarti delle singole medie dei gruppi dalla media generale del collettivo.

Confronto fra medie - 2

- Poiché vale la relazione:

Devianza Totale = Devianza Interna + Devianza Gruppi

- ovvero

$$1 = \frac{DT}{DT} = \frac{DI}{DT} + \frac{DG}{DT}$$

- il valore

$$\eta^2 = \frac{DG}{DT}$$

Dove η^2 ci dà la proporzione di devianza totale che è ***spiegata*** dalla variabile indipendente.

Confronto fra medie - 3

- Dove **η^2** ci dà la proporzione di devianza totale che è ***spiegata*** dalla variabile indipendente.
- **η^2** varia tra **0**, nessuna relazione fra le variabili, e **1**, relazione perfetta, ovvero tutta la devianza dipende dalla variabile indipendente (es. tutti gli uomini hanno dato una stessa risposta e così, diversa, tutte le donne!).
- In genere **η^2** non è molto elevato: è difficile andare oltre il 25% - 30%, spesso si è sul 10%.
- *Aumenta all'aumentare del numero di modalità.*
- È assimilabile al coefficiente **r^2**.

ANOVA - 1

- Anche in questo caso possiamo vedere gli aspetti inferenziali.
- Nel nostro data set consideriamo le variabili BMI (come indicatore dello stile di vita attiva e dell'alimentazione) e colore della pelle: vogliamo testare l'ipotesi che gli stili di vita siano diversi nelle diverse etnie che vivono ad Aracaju (Sergipe).
- Calcoliamo le statistiche di sintesi per il BMI:

Report BMI

colore della pelle	Media	N	Deviazione std.
bianca	24,4733	237	3,63455
gialla	23,8859	113	3,39628
marrone	24,9505	564	4,25270
nera	25,0485	223	3,78158
Totale	24,7644	1137	3,97074

ANOVA - 2

- L' η^2 in questo caso è pari a 0,08, quindi c'è, anche se ridotta, un'associazione fra le due variabili.
- Ma questa associazione è valida per tutta la popolazione da cui è stato estratto il campione degli intervistati, ovvero per i cittadini di Aracaju?
- La media generale è 24,8 e la misura della dispersione dei risultati di tutti i cittadini rispetto ad essa è lo s.q.m (la radice quadrata della ***varianza***).
- Ma quanta parte di questa varianza è dovuta alle differenze fra i risultati dei singoli cittadini appartenenti a un'etnia dalla media dell'etnia?
- E quanta alle differenze fra le (medie delle) etnie?

ANOVA - 3

- Utilizziamo la relazione fra le ***devianze*** (numeratore della varianza) precedentemente introdotta.
- La **Devianza tra i gruppi** è quella ***spiegata*** dai diversi stili di vita delle etnie.
- La **Devianza all'interno dei gruppi** è quella ***non spiegata***.
- Come già visto la **DG** è nulla in caso di assenza di relazione fra le variabili studiate.
- Nella **Tabella ANOVA** (***Analisi della Varianza***), che segue, sono riportate le informazioni utili per testare questa relazione.

ANOVA - 4

Tabella ANOVA BMI * colore della pelle

	Devianza	df	Media dei quadrati	F	Sig.
Fra gruppi	144,813	3	48,271	3,078	,027
Entro gruppi	17766,234	1133	15,681		
Totale	17911,047	1136			

- df ovvero "degrees of freedom" (gradi di libertà) sono i denominatori delle rispettive varianze, utilizzate come stime campionarie: (n-1) per la totale; (k-1) fra i gruppi; (n-k) interna ai gruppi*;
- le Medie dei quadrati sono le ***varianze*** stimate; se i due valori fossero simili (rapporto circa 1), allora non vi sarebbe l'effetto etnia e non potremmo rifiutare l'ipotesi nulla.
- Quindi calcoliamo questo rapporto che chiamiamo **F**, la cui distribuzione campionaria è nota ed è nelle appendici di ogni testo di Statistica (con df= (k-1) e (n-k)).

---

* qui n=1137; k= 4.

ANOVA - 5

❑ Oppure si può utilizzare il valore di significatività (**Sig.**) o ***p-value***, che rappresenta la probabilità di avere un rapporto **F** di questa dimensione se fosse vera l'ipotesi nulla.

❑ Ovvero, poiché in questo caso **p = .027**, questo risultato si verificherebbe 2,7 volte ogni 100 campioni estratti dalla popolazione di riferimento se fosse vera l'ipotesi nulla.

❑ Se quindi testiamo l'ipotesi nulla con un livello di errore del 5% potremmo rifiutarla (2,7<5).

❑ Al contrario, se ci fossimo posti un livello di errore dell'1%, i risultati non ci avrebbero permesso di rifiutarla.

Il caso b): variabili quantitative

- Chiudiamo con l'analisi del caso b), quello in cui si studia l'associazione fra due variabili quantitative, che potremo chiamare correttamente "relazione".
- Si parla, infatti, di ***correlazione*** e di ***regressione*** fra le variabili studiate.
- Le possibilità di analisi sono, in questo caso, molto ampie, cominciando dalla possibilità di partire dallo studio della rappresentazione grafica.

La correlazione - 1

- Prendiamo in considerazione una nuova matrice dei dati e consideriamo tre delle variabili quantitative: CORSA, FLESS e SALTO.

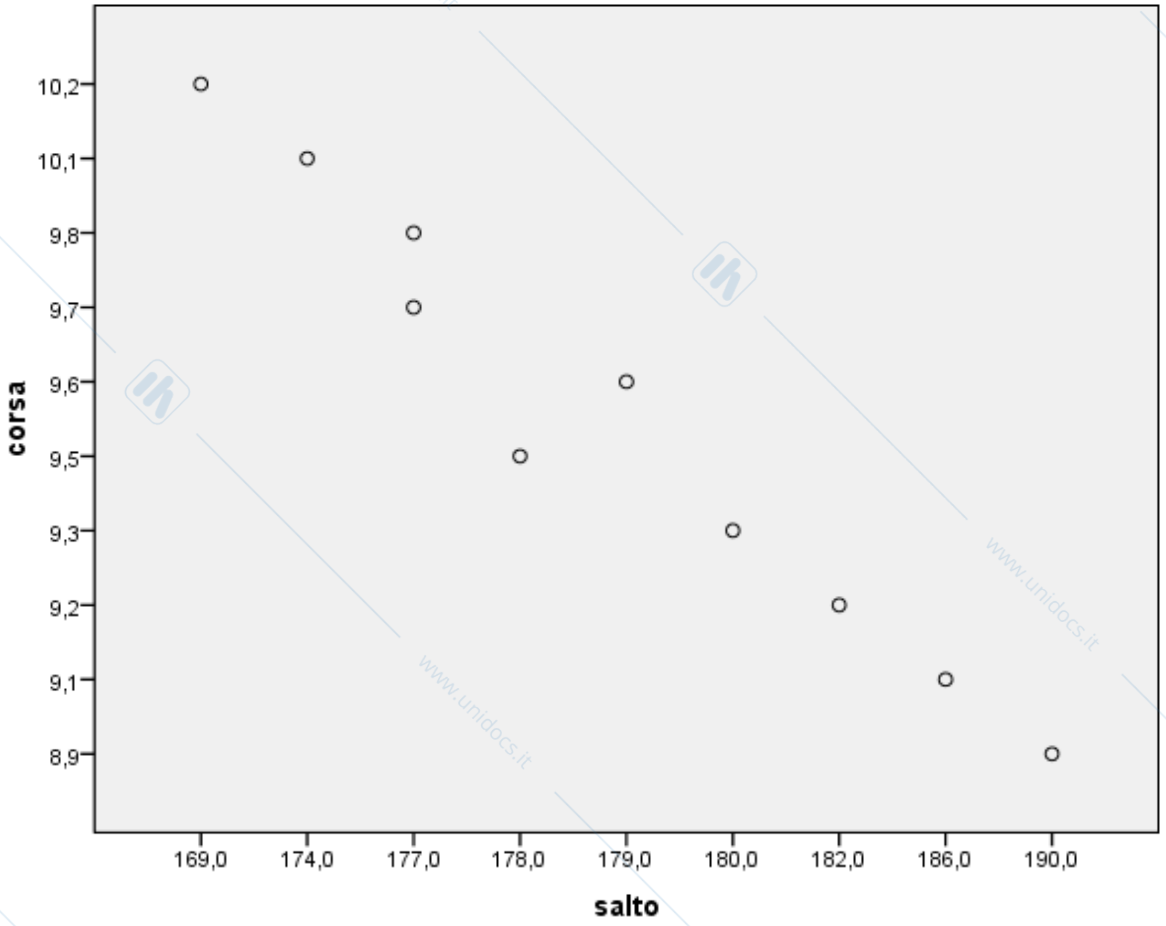
La matrice dei dati

Nome	Sport	corsa	salto	fless	spola	tapp	later	rank
Andrea	Calcio	9,8	177	6,2	17,9	38,6	Dx	6°
Carlo	Volley	10,2	169	10,2	18,2	38,4	Dx	22°
Enrico	Volley	9,5	178	11,9	17,6	38,1	Sn	3°
Gianni	Calcio	9,6	179	9,6	17,2	37,4	Dx	5°
Mario	Volley	9,2	182	6,4	16,8	36,2	Dx	10°
Mauro	Volley	9,1	186	10,1	16,4	37,4	Dx	7°
Nicola	Calcio	8,9	190	8,4	16,5	39,2	Sn	1°
Sandro	Atletica	9,3	180	10,4	17,0	39,6	Dx	12°
Silvano	Atletica	10,1	174	8,2	18,6	39,2	Dx	20°
Ugo	Atletica	9,7	177	8,4	17,9	38,1	Dx	14°

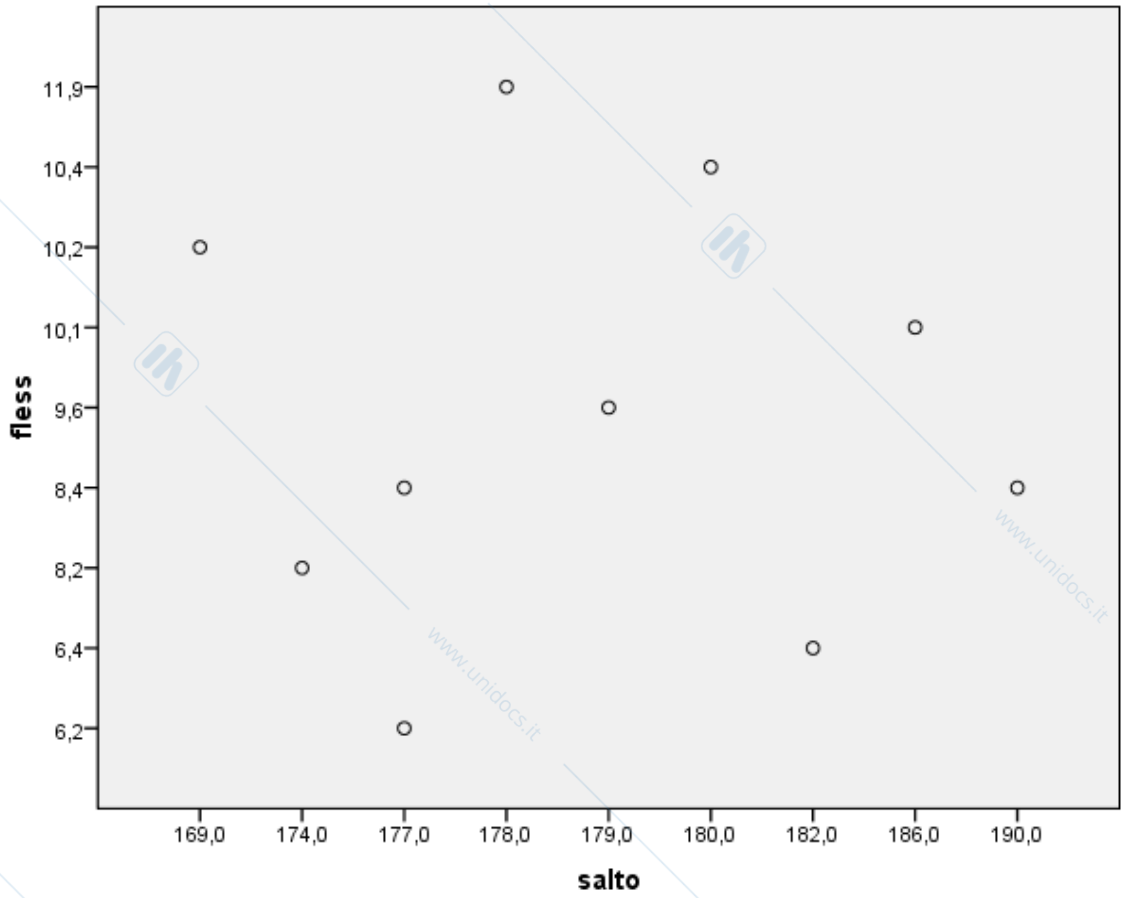
La correlazione - 2

- Poiché le variabili sono continue, si possono rappresentare su una retta di riferimento:
  - ✓CORSA sulla retta X_1
  - ✓SALTO sulla retta X_2
  - ✓FLESS sulla retta X_3
- Quindi le rette possono essere messe in relazione fra di loro su piani in coordinate cartesiane ortogonali (***diagrammi di dispersione***), ad esempio X_1 vs. X_2 e X_1 vs. X_3.

Plot: CORSA vs. SALTO



Plot: FLESS vs. SALTO



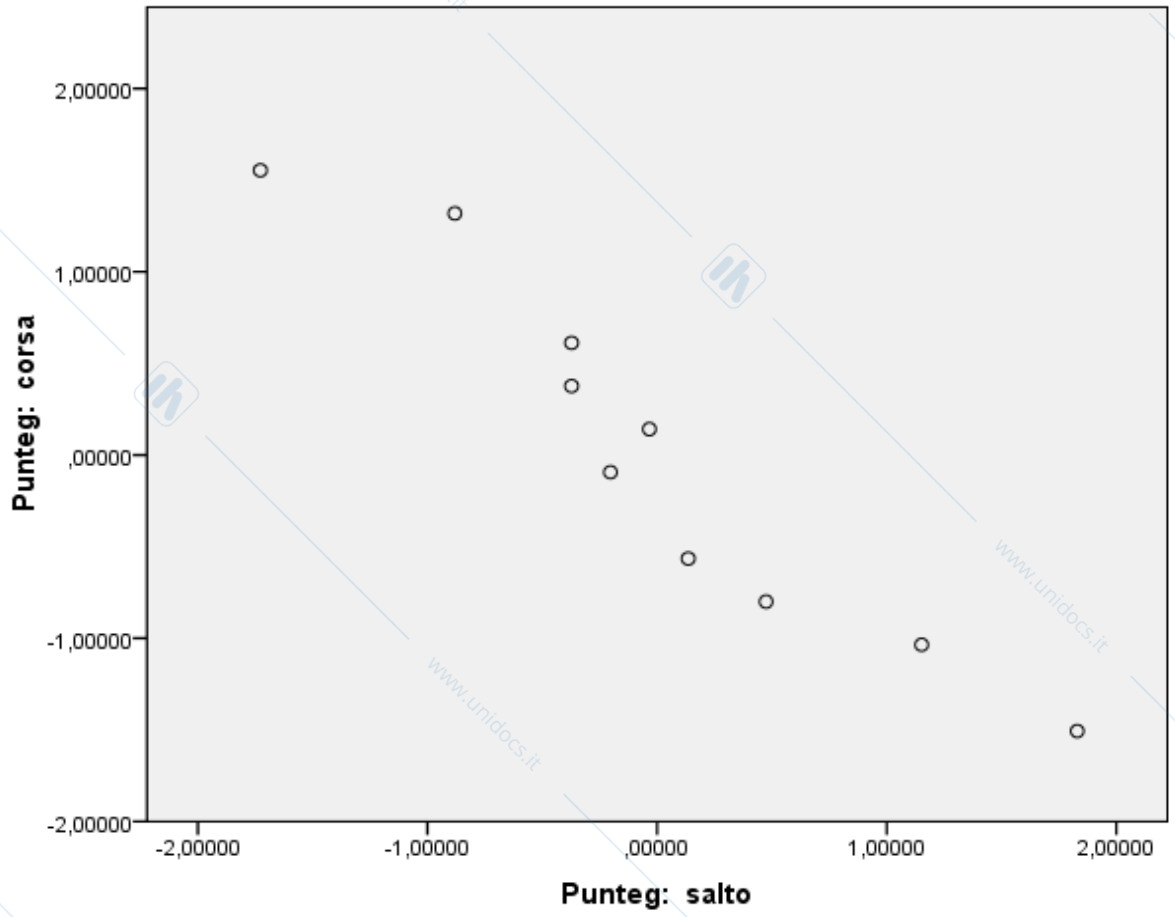
La correlazione - 3

- In questo modo su ogni dimensione unitaria (la retta è uno *spazio a una dimensione*) i risultati sono messi in ordine crescente e ogni allievo è rappresentato su un punto (coordinata).
- Mettendo in relazione due rette ogni allievo è rappresentato da un punto sul piano (*spazio a due dimensioni*), che si individua tramite le coordinate sulle rette.
- *Qualora le prove considerate fossero più di 2 (ad es. "p") lo spazio di riferimento sarebbe a "p" dimensioni.*

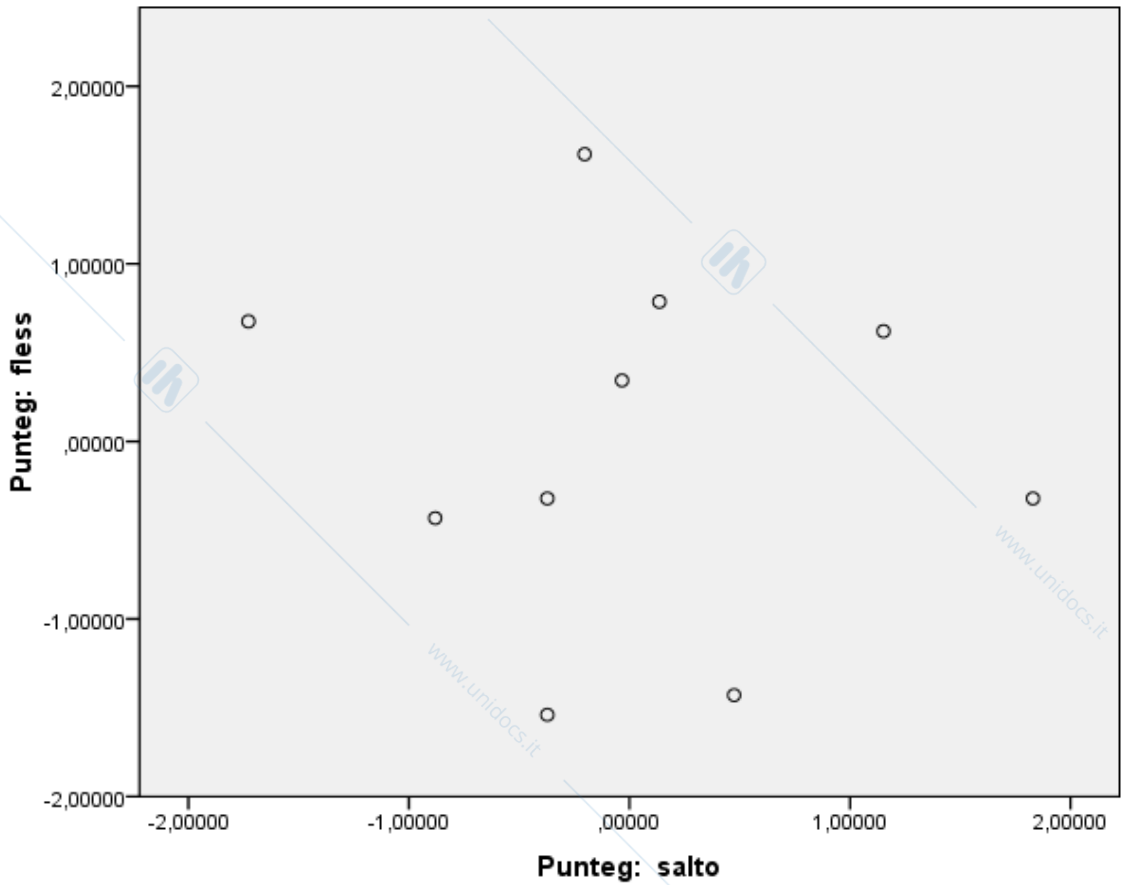
La correlazione - 4

- Possiamo vedere come le ***nuvole*** dei punti che rappresentano gli allievi si disperdono nel piano in maniera diversa, seguendo una certa regolarità nel primo caso e in maniera casuale nel secondo.
- Ma i due grafici possono essere fuorvianti, perché le unità di misura e/o la variabilità sono diverse.
- *Standardizziamo pertanto le tre variabili e vediamo il diverso risultato grafico.*

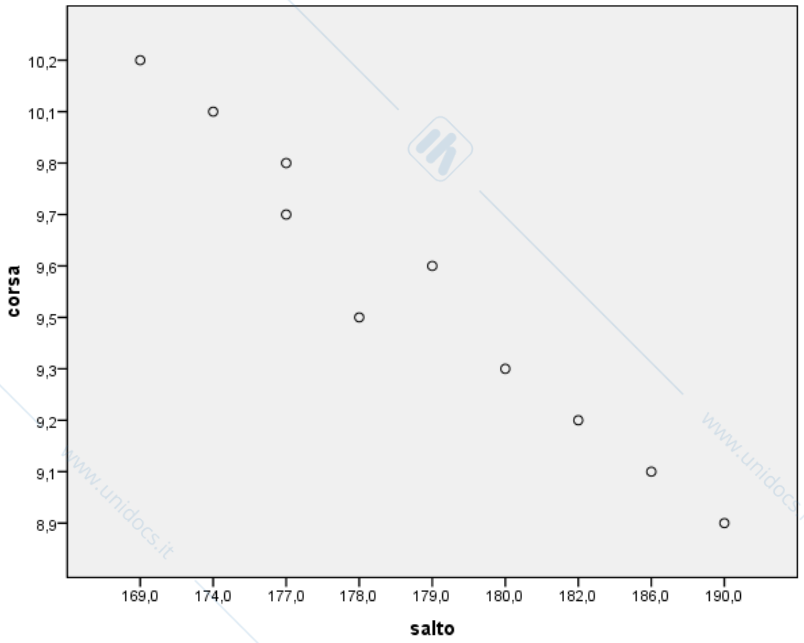
Plot ZCORSA vs. ZSALTO



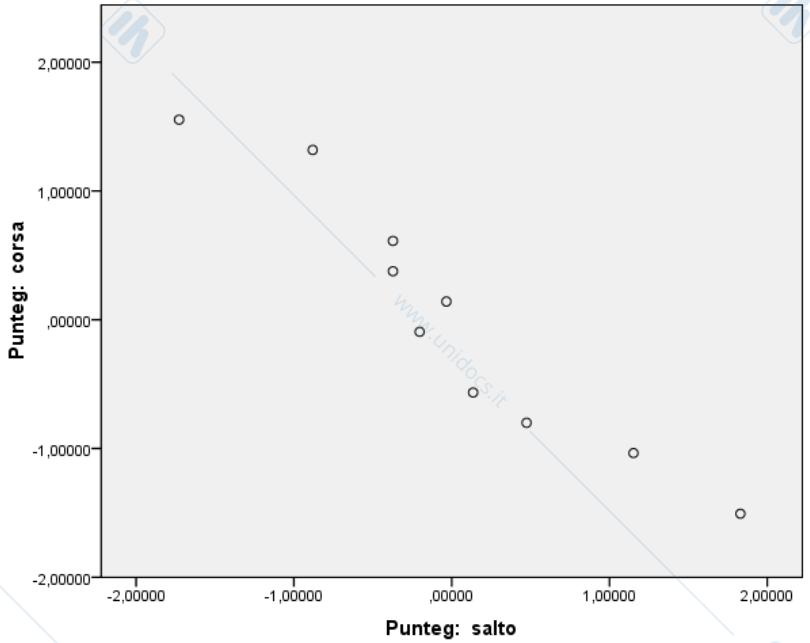
Plot: ZFLESS vs. ZSALTO



Plot CORSA vs. SALTO



Plot ZCORSA vs. ZSALTO



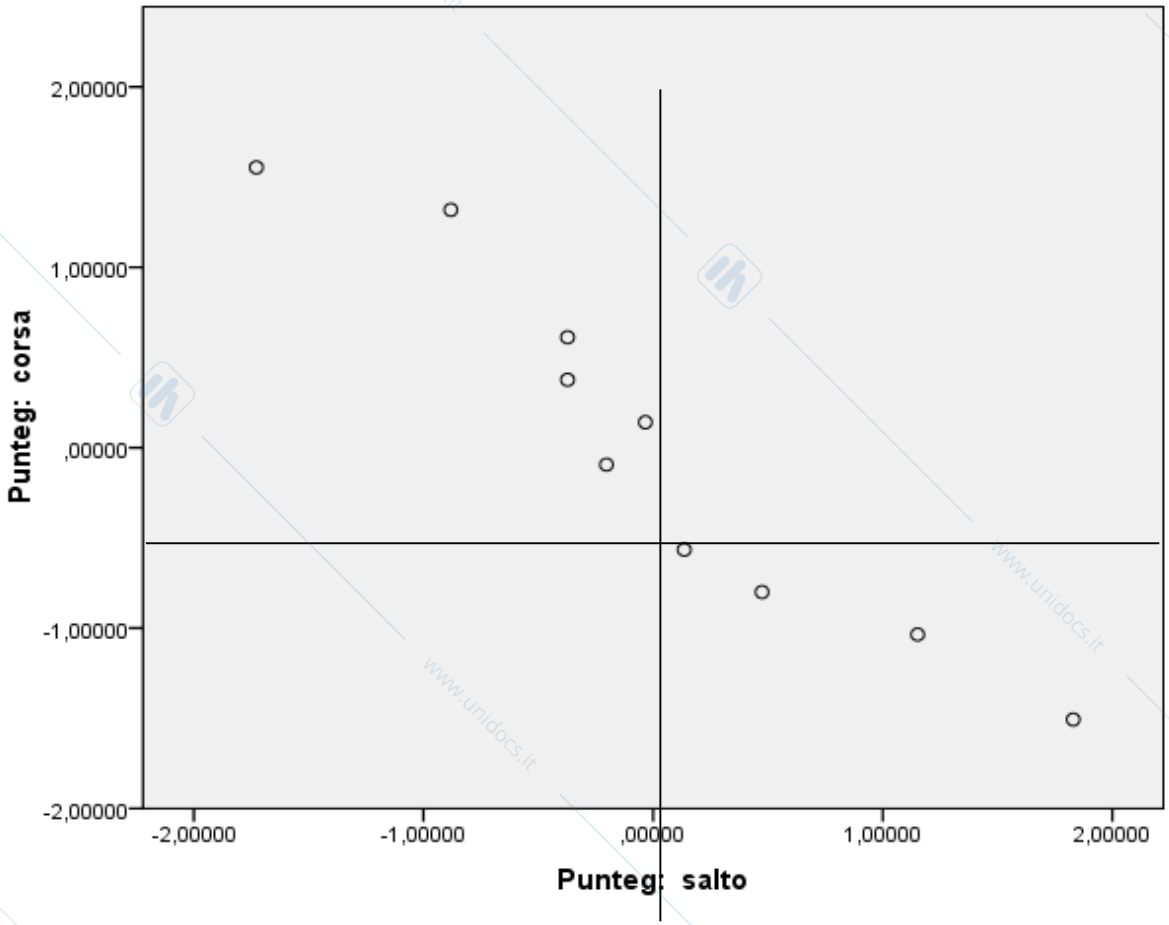
La correlazione - 5

- L’operazione di standardizzazione ha permesso di *centrare* le variabili, portando il centro del sistema di coordinate (0,0) a coincidere con i valori medi delle due variabili.
- Ha anche permesso di omogeneizzare la dispersione, per cui il peso delle due variabili nel determinare la forma della nuvola è uguale.

La correlazione - 6

- Tornando al grafico ZCORSA vs. ZSALTO, si può osservare come i ragazzi con risultati inferiori alla media nella corsa li abbiano ottenuti superiori alla media nel salto e viceversa: la nuvola dei punti si distribuisce unicamente nel II e IV quadrante del piano.

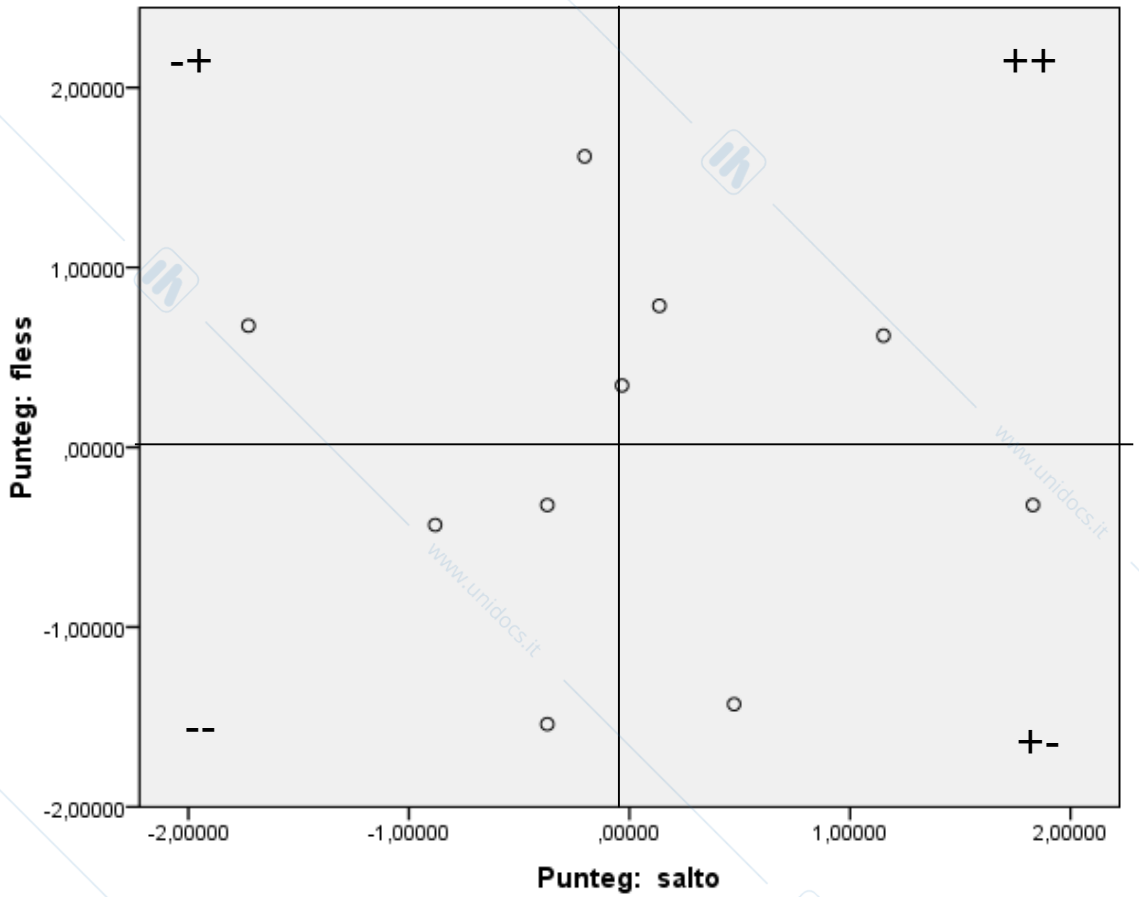
Plot ZCORSA vs. ZSALTO



La correlazione - 7

- Ovviamente è un fatto tecnico: le graduatorie dei risultati sono inverse (tempi e lunghezze). Quindi fra le due prove c'è discordanza tecnicamente e concordanza logicamente: gli allievi più bravi nella prima lo sono anche nella seconda.
- Invece nel grafico ZFLESS vs. ZSALTO non si individua nessuna relazione tendenziale fra i risultati nelle due prove, osservabile dalla dispersione della nuvola dei punti: questi si collocano in maniera uniforme nei quattro quadranti.

Plot: ZFLESS vs. ZSALTO



La correlazione - 8

- Siamo pronti a definire una misura sintetica delle relazioni che abbiamo presentato: il ***coefficiente di correlazione lineare di Pearson***.

$$-1 \leq \rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{n}(x_i - \mu_x)^2}\sqrt{\sum_{i=1}^{n}(y_i - \mu_y)^2}} \leq +1$$

La correlazione - 9

- Questo indice è la media dei prodotti degli scarti standardizzati fra due variabili e, proprio per la standardizzazione, può assumere un *range* di valori compreso fra:
  - **-1** tra le due variabili vi è perfetta correlazione negativa, quindi discordanza, e
  - **+1** tra le due variabili vi è perfetta correlazione positiva, quindi concordanza, i punti sono allineati su una retta, passando per
  - **0** non vi è correlazione di tipo lineare, il che non vuol dire che non ci sia alcuna relazione.

La correlazione - 10

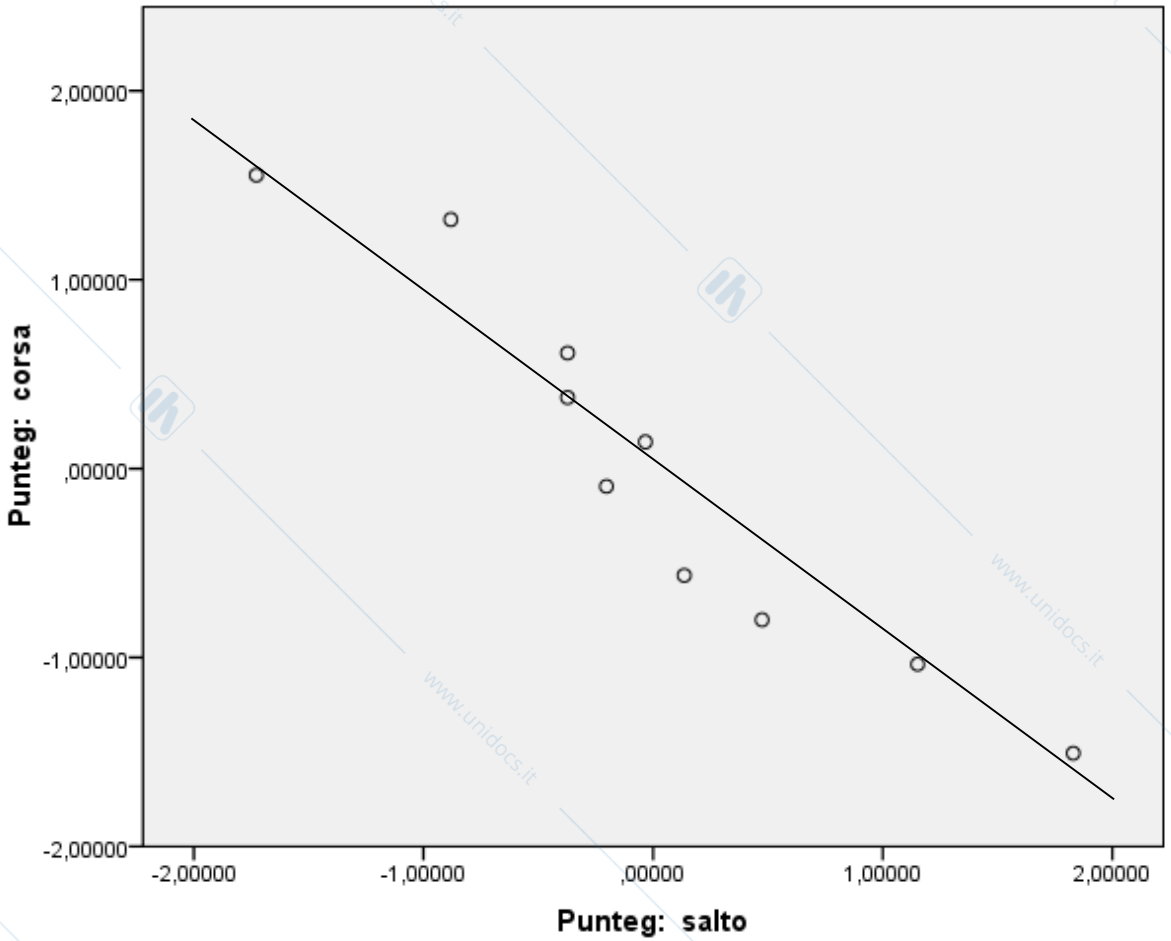
- Tra **-1** e **0** e, simmetricamente, tra **0** e **+1** vi sono infiniti valori, che indicano correlazione
  - bassa fra **0** e **-0,29** o fra **0** e **+0,29**;
  - media fra **-0,30** e **-0,59** o fra **+0,30** e **+0,59**;
  - alta fra **-0,60** e **-0,99** o fra **+0,60** e **+0,99**.

- Altra strategia va usata se si utilizza un coefficiente di correlazione campionaria per testare un ipotesi sulla correlazione in una popolazione (inferenza).

La correlazione -11

- Nel caso di correlazione alta, ossia quando le variabili hanno in comune qualcosa, possiamo cercare il significato di questa "comunalità": ad esempio cosa c'è in comune tra CORSA e SALTO?
- Questa "comunalità" può essere l'espressione di una variabile ***latente***, ossia non direttamente osservabile, che corrisponde a una retta (in quanto la correlazione misurata è quella lineare): questa retta è quella che passa (interpola) per la nuvola dei punti.
- La variabile latente è anche interpretabile, secondo la "Teoria dell'allenamento", come una capacità dell'atleta: la ***forza rapida***.

Plot ZCORSA vs. ZSALTO



La correlazione - 12

- Le ultime considerazioni aumentano di importanza quando si aumentano le dimensioni di riferimento, quando cioè le variabili poste a confronto simultaneamente sono più di tre e non è possibile rappresentare la nuvola dei punti in uno spazio per noi “leggibile”.

- Diviene così necessario cercare un ***punto di vista*** ottimale di dimensioni ridotte per leggere e sintetizzare i dati.

La correlazione - 13

- Pensiamo a una batteria di 30 prove: sarà necessario sintetizzare i risultati in un numero ridotto di prove per ***ordinare*** e ***classificare*** gli allievi, per vedere quali prove siano simili, quali attendibili, quali valide e quali dimensioni latenti soggiacciano alle loro prestazioni.

- Questo è il compito delle tecniche di **Analisi Multivariata**, sempre in un'ottica esplorativa (Analisi delle Componenti Principali, Analisi delle Corrispondenze Multiple, Cluster Analysis).

La regressione - 1

- Va tenuto presente che, finora, si sono studiate le relazioni fra due variabili seguendo un modello ***simmetrico***:
  - non si è considerata tecnicamente l'eventuale gerarchia tra le variabili stesse, ovvero non si è considerata una delle due indipendente e l'altra dipendente dai risultati della prima.
- Questo non toglie la possibilità di una gerarchia logica fra le due, ma nella correlazione le due variabili giocano un ruolo simmetrico.

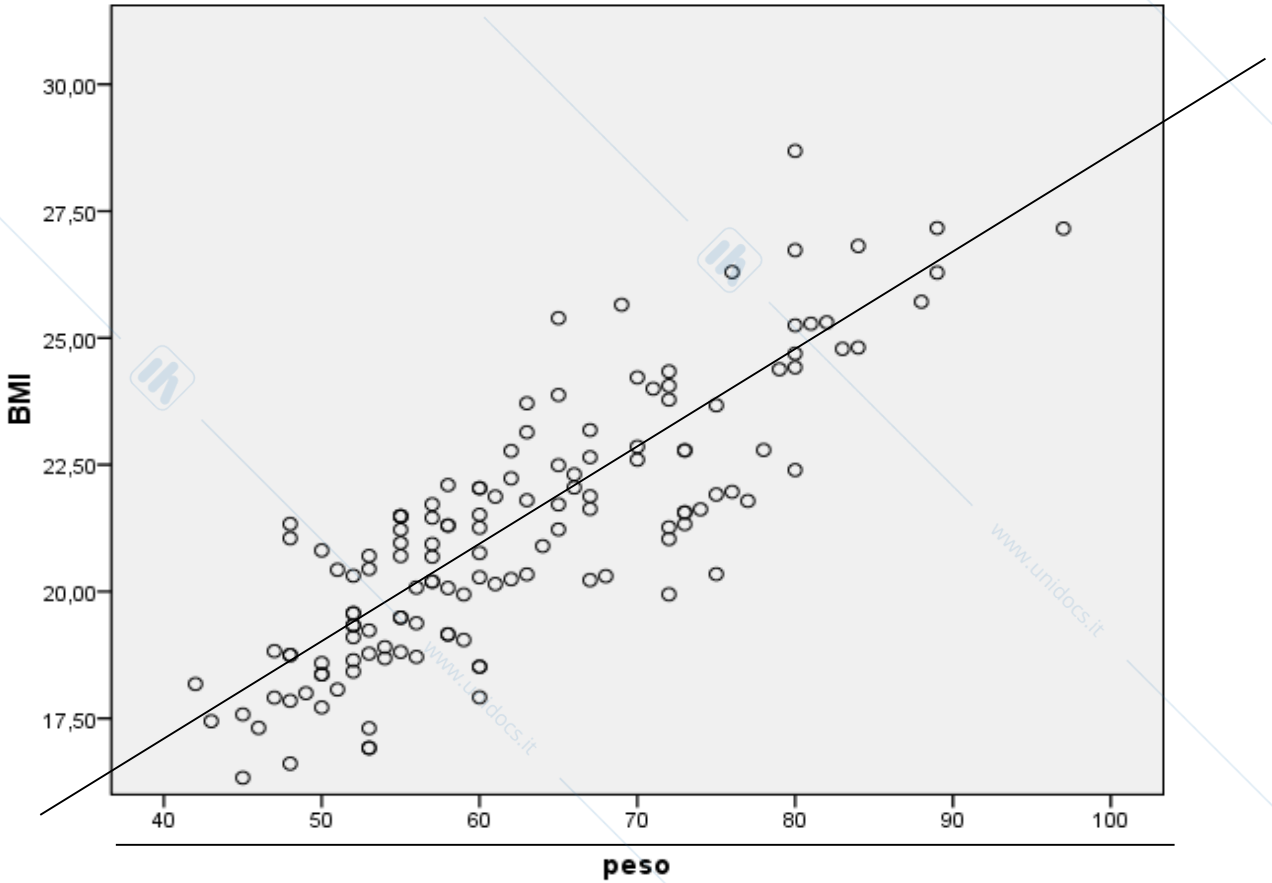
La regressione - 2

- Se c'è una correlazione, almeno "media" (oppure "significativa" in un'ottica inferenziale) e i risultati di una delle due variabili possono essere considerati dipendenti da quelli dell'altra, allora si può utilizzare il modello ***asimmetrico*** della **regressione**.

$$Y = a + b\,X$$

- Trovando i valori di **a** (intercetta sull'asse Y) e **b** (coefficiente angolare che indica l'inclinazione della retta) si possono prevedere i valori della **Y** al variare di quelli della **X**, anche in casi non osservati.

Plot: BMI vs. PESO



Ulteriori approfondimenti

Anomalie della Correlazione

- Abbiamo detto che **r** è il coefficiente di correlazione ***lineare*** (di Bravais Pearson): se però la relazione fra le variabili c'è, ma non è lineare (ad esempio è parabolica), **r=0**!

- Nel caso siano presenti valori estremi (***outlier***) la correlazione risulterà ***fittizia*** o ***soppressa***!!!!

- È necessario, quindi, studiare sempre il diagramma disispersione, per evitare questi errori!!!!

La fallacia ecologica

- Nel caso in cui le unità di analisi, su cui siano rilevate le due variabili messe in relazione, siano ***aggregati*** di individui (comune, municipio, regione e così via), si parla di ***correlazione ecologica***.
- In realtà il ricercatore vorrebbe conoscere la ***correlazione individuale*** (tra gli individui), ma questo non è possibile perché si è nel campo delle ***analisi secondarie*** (su dati rilevati da altri).
- **Pertanto non si dovrebbe mai interpretare una correlazione ecologica come correlazione individuale, da cui il termine di *fallacia ecologica.***
- Vedremo alcuni esempi nella seconda parte del corso.

La relazione spuria

- È il caso di "***presenza di covariazione, pur in assenza di causazione***"
- Esempio: correlazione fra numero di autopompe antincendio intervenute ed entità dei danni (in realtà dovuta alla dimensione dell'incendio.

- Anche questo caso sarà più facilmente trattabile con l'introduzione di più variabili!