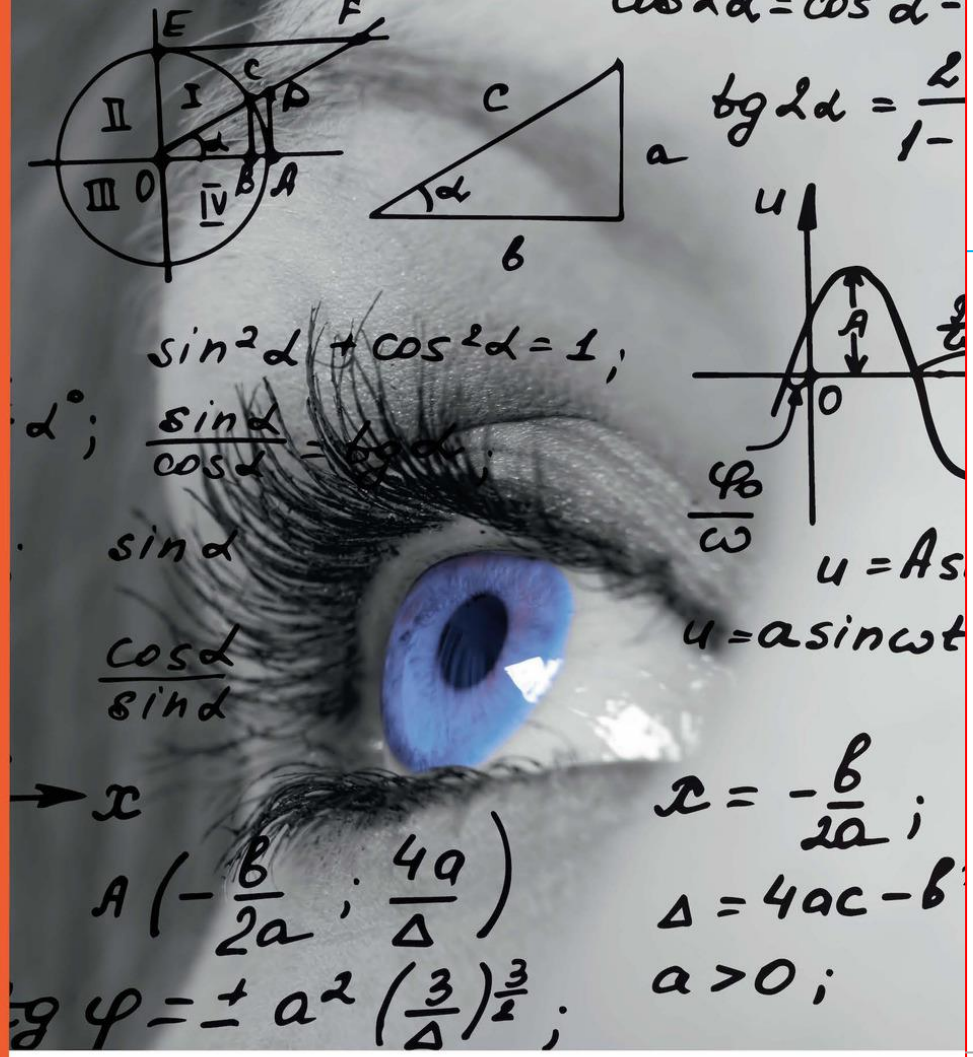
Statistica per le scienze sociali

Enrica Amaturo, Biagio Aragona, Maria Gabriella Grassia, Carlo Natale Lauro, Marina Marino

UTET UNIVERSITÀ



Statistica per le scienze sociali

Enrica Amaturo, Biagio Aragona,
Maria Gabriella Grassia, Carlo Natale Lauro,
Marina Marino

UTET UNIVERSITÀ

5 Analisi delle relazioni tra due caratteri

Capitolo a cura di
M.G. Grassia, C.N. Lauro, M.Marino

Analisi delle relazioni

Quale indice misura la relazione tra due variabili osservate?

Che tipo di variabili?

Che tipo di relazione supporre?

Indici statistici

		2 mutabili	1 variabile 1 mutabile	2 variabili
Indipendenza statistica	Legame bidirezionale	χ^2, Φ^2, V	χ^2, Φ^2, V	χ^2, Φ^2, V
	Legame unidirezionale	λ	λ	λ
Indipendenza in media	Legame bidirezionale			
	Legame unidirezionale		η^2	η^2
Indipendenza lineare	Legame bidirezionale			ρ
	Legame unidirezionale			Modello di regressione

Analisi delle relazioni

Quale indice misura la relazione tra due variabili osservate?

Che tipo di variabili?

Che tipo di relazione supporre?

Indici statistici

		2 mutabili	1 variabile 1 mutabile	2 variabili
Indipendenza statistica	Legame bidirezionale	χ^2, Φ^2, V	χ^2, Φ^2, V	χ^2, Φ^2, V
	Legame unidirezionale	λ	λ	λ
Indipendenza in media	Legame bidirezionale			
	Legame unidirezionale		η^2	η^2
Indipendenza lineare	Legame bidirezionale			ρ
	Legame unidirezionale			Modello di regressione

Associazione tra variabili
Variabili quantitative

Un caso importante nello studio delle relazioni tra due variabili, si ha quando entrambe le variabili sono quantitative

Se ci sono informazioni apriori che permettono di stabilire quale è l'antecedente logico e quale il conseguente, cioè si conosce il verso della dipendenza logica, allora si utilizza lo strumento della **regressione**

Quando invece non è possibile sapere quale variabile dipende dall'altra, si utilizza la **correlazione** che è un indice simmetrico

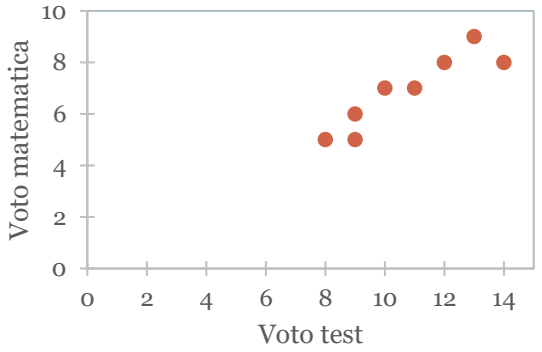
Associazione tra variabili
Variabili quantitative

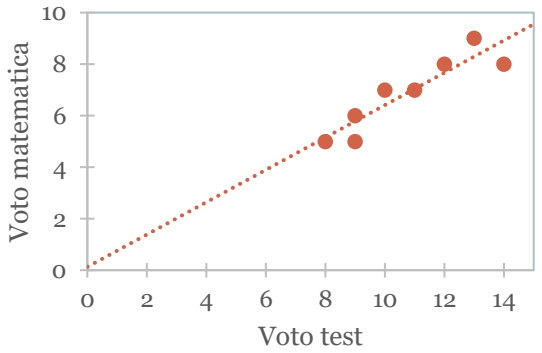
Per introdurre la regressione e la correlazione partiamo da un esempio, rappresentando congiuntamente due variabili *X* e *Y* con un grafico che si chiama **diagramma a dispersione** (detto anche *scatter)*

Voto ad un test di inizio anno scolastico e voto finale di matematica di 8 studenti

	Test	Voto
Studente 1	12	8
Studente 2	10	7
Studente 3	14	8
Studente 4	9	5
Studente 5	9	6
Studente 6	13	9
Studente 7	11	7
Studente 8	8	5

Scatter tra voto test di inizio anno scolastico e voto finale di matematica

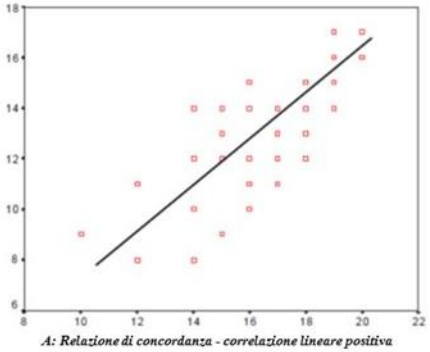




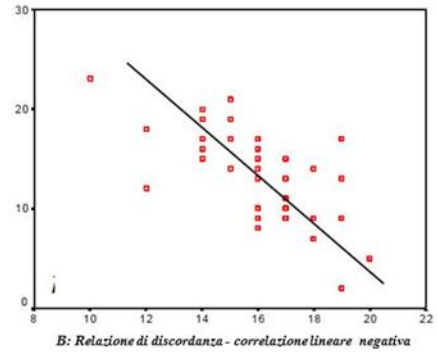
Associazione tra variabili
Variabili quantitative

Questo ragionamento intuitivo si formalizza nella definizione del concetto di **concordanza** e **discordanza** tra variabili quantitative

Tra due variabili *X* e *Y* esiste **concordanza** se al crescere di *X* anche *Y*, nel complesso, tende a crescere e se al diminuire di *X* anche *Y*, nel complesso, tende a diminuire. Si parla di **correlazione lineare positiva**



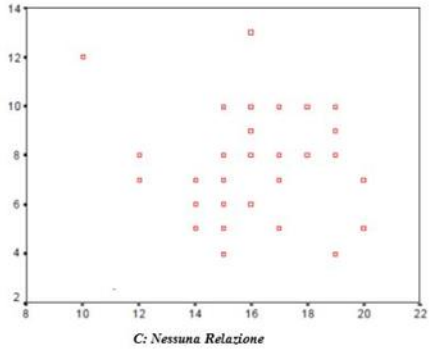
A: Relazione di concordanza - correlazione lineare positiva



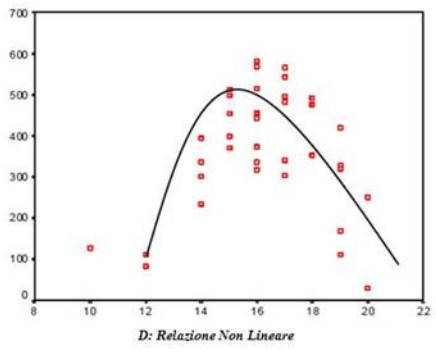
B: Relazione di discordanza - correlazione lineare negativa

C'è **discordanza** se al diminuire di *X* la variabile *Y*, nel complesso, tende a crescere e se al crescere di *X*, nel complesso, *Y* tende a diminuire. Si parla di **correlazione lineare negativa**

non esiste legame tra le due variabili



C: Nessuna Relazione



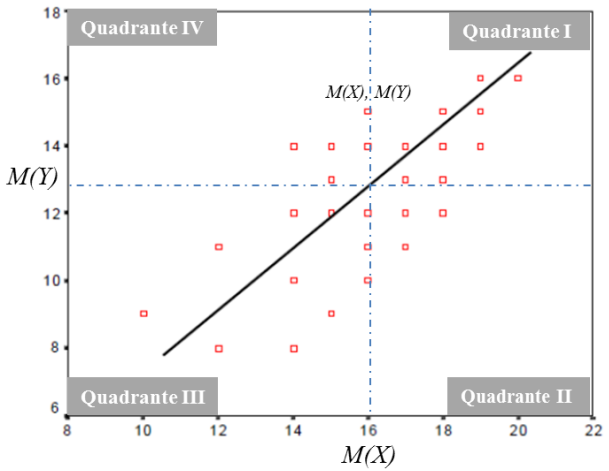
D: Relazione Non Lineare

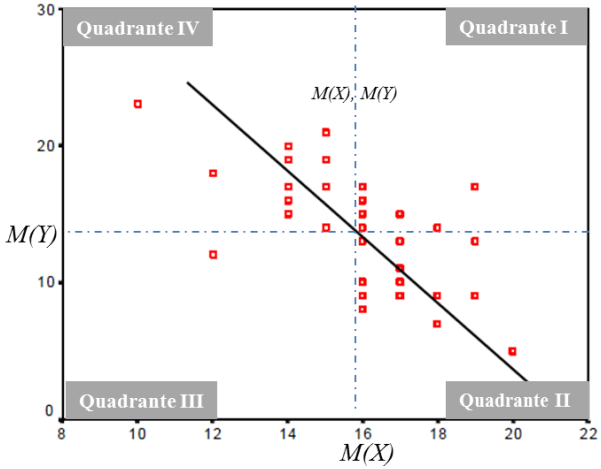
Esiste un legame non lineare tra le due variabili

Associazione tra variabili
Variabili quantitative

Questo ragionamento intuitivo si formalizza nella definizione del concetto di **concordanza** e **discordanza** tra variabili quantitative

Scatter tra le variabili X e Y. Divisione del piano in quadranti





Segni degli scarti nei singoli quadranti

Quadrante	Segno di ogni singolo Scarto $(x_i - M(X))$	Segno di ogni singolo Scarto $(y_i - M(Y))$
Quadrante I	+	+
Quadrante II	+	-
Quadrante III	-	-
Quadrante IV	-	+

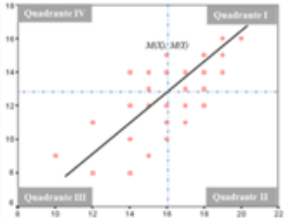
Associazione tra variabili
Variabili quantitative - Codevianza

La somma dei prodotti degli scarti è detta ***codevianza:***

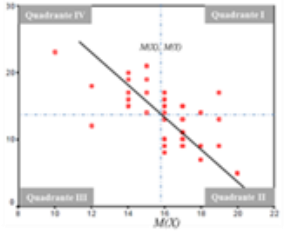
$$Codev\,(X;Y) = \sum_{i=1}^{N} \big(x_i - M(X)\big)\big(y_i - M(Y)\big)$$

Se prevalgono i prodotti tra scarti di segno uguale, la codevianza sarà positiva; se prevalgono i prodotti tra scarti di segno opposto, allora la codevianza sarà negativa.

$(+)\cdot(+)$ $(-)\cdot(-)$ concordanza ⇒ $Codev > 0$



$(+)\cdot(-)$ $(-)\cdot(+)$ discordanza ⇒ $Codev < 0$



Associazione tra variabili
Variabili quantitative - Covarianza

Dividendo per N (la numerosità del collettivo) si ottiene la **covarianza**:

$$Cov(X,Y) = \frac{\sum_{i=1}^{N}\big(x_i - M(X)\big)(y_i - M(Y))}{N}$$

La covarianza è anche definita come **media del prodotto degli scarti**

Se X ed Y sono indipendenti allora $Cov\ (X,Y) = \ 0$

Viceversa non è vero

Se $Cov\ (X,Y) = 0$ c'è ***assenza di dipendenza lineare*** fra le due variabili, ma ciò non esclude che ci sia una relazione di altro tipo

Esempio – covarianza

Voto ad un test di inizio anno scolastico e voto finale di matematica di 8 studenti

	Test (X)	Voto (Y)	X-M(X)	Y-M(Y)	(X-M(X))(Y-M(Y))
Studente 1	12	8	1.25	1.125	1.406
Studente 2	10	7	-0.75	0.125	-0.094
Studente 3	14	8	3.25	1.125	3.656
Studente 4	9	5	-1.75	-1.875	3.281
Studente 5	9	6	-1.75	-0.875	1.531
Studente 6	13	9	2.25	2.125	4.781
Studente 7	11	7	0.25	0.125	0.031
Studente 8	8	5	-2.75	-1.875	5.156
					19.750

M(X)	10.75
M(Y)	6.875
Codev(X;Y)	19.75
Cov(X;Y)	2.47

Cov(X;Y)>0 → Tra le due variabili *X* e *Y* esiste **concordanza**

Associazione tra variabili
Variabili quantitative - Correlazione

Si dimostra che il valore massimo per la covarianza è pari al prodotto degli scarti quadratici medi delle singole variabili (ovvero la radice quadrata del prodotto delle Varianze):

$$\max\ Cov\,(X,Y) = \sqrt{Var(X)Var(Y)} = \sigma_X \sigma_Y$$

Rapportando la codevianza al suo massimo si ottiene il **coefficiente $\boldsymbol{\rho}$** (rho dalla lettera greca) **di correlazione lineare di Bravais-Pearson**

$$\rho = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

$$oppure$$

$$\rho = \frac{Codev(X,Y)}{\sqrt{Dev(X)Dev(Y)}}$$

Associazione tra variabili
Variabili quantitative - Correlazione

Il coefficiente di correlazione lineare di Bravais-Pearson è un indice simmetrico della relazione lineare tra *Y* e *X* e misura ***l'interdipendenza lineare*** tra le stesse
Assume valori tra -1 e +1

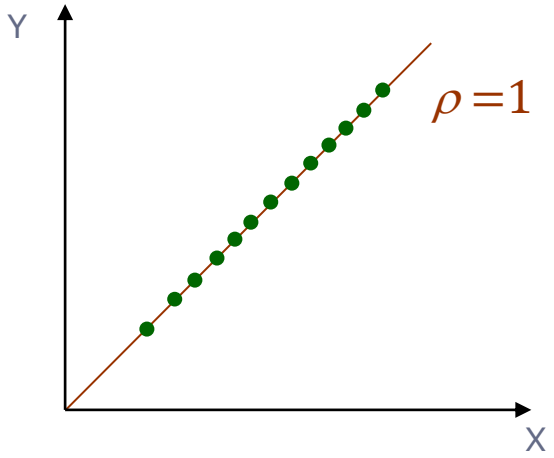
$$-1 \leq \rho \leq 1$$

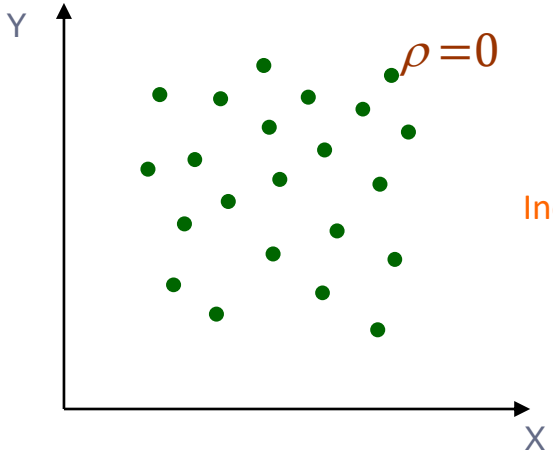
In particolare :

- $\rho = +1$; **concordanza perfetta**. Fra *X* e *Y* sussiste un perfetto legame lineare. I punti del diagramma di dispersione sono perfettamente allineati lungo una **retta crescente**
- $\rho > 0$; **concordanza**. Il grado di concordanza dipende dal valore assunto da ρ. I punti del diagramma di dispersione non sono perfettamente allineati, ma seguono un andamento crescente
- $\rho = 0$; **indipendenza lineare**. Fra *X* e *Y* sussiste un'indipendenza lineare
- $\rho < 0$; **discordanza**. Il grado di discordanza dipende dal valore assunto da ρ. I punti del diagramma di dispersione non sono perfettamente allineati, ma seguono un andamento decrescente
- $\rho = +1$; **discordanza perfetta**. Fra *X* e *Y* sussiste un perfetto legame lineare. I punti del diagramma di dispersione sono perfettamente allineati lungo una **retta decrescente**

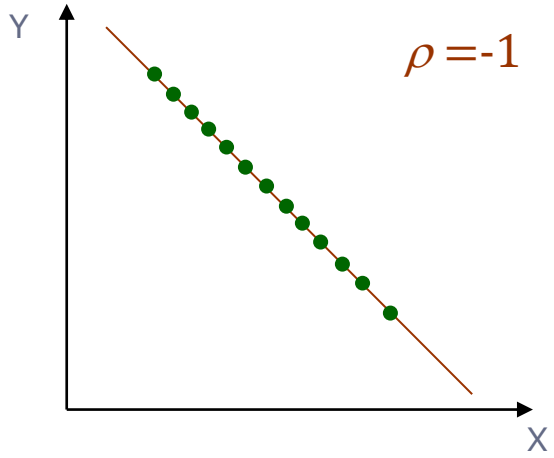
Associazione tra variabili
Variabili quantitative - Correlazione

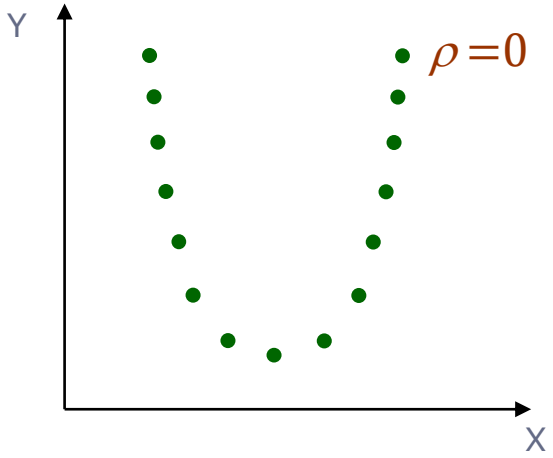
L'interpretazione di ρ





Indipendenza $\Rightarrow$ $\rho = 0$





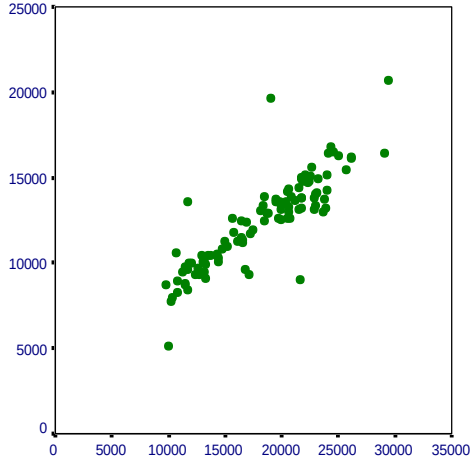
$\rho = 0 \not\Rightarrow$ Indipendenza

$\rho = 0 \Rightarrow$ Assenza di correlazione lineare

Associazione tra variabili
Variabili quantitative - Correlazione

L'interpretazione di ρ

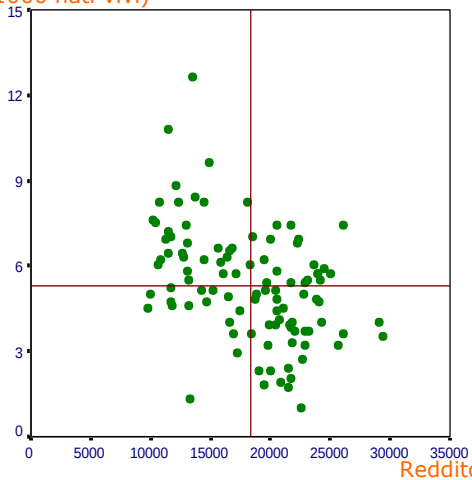
Consumi p.c.
(in €)



Reddito p.c.
(in €)

$\rho = 0{,}87$

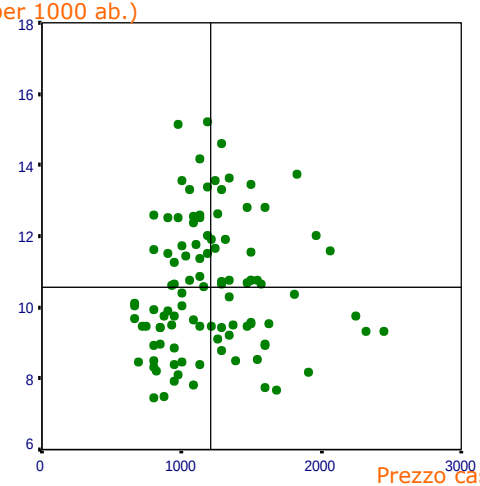
Mortalità infantile
(per 1000 nati vivi)



Reddito p.c.
(in €)

$\rho = -0{,}438$

Num. Morti
(per 1000 ab.)



Prezzo casa al mq
(in €)

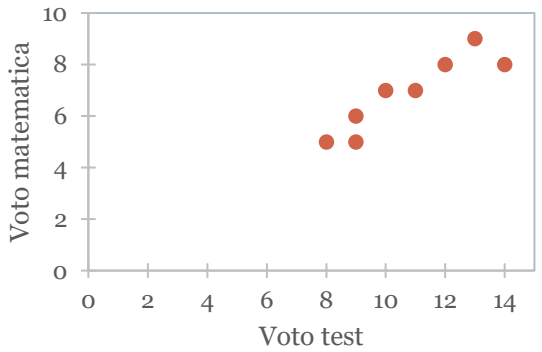
$\rho = 0{,}057$

Esempio – correlazione

Voto ad un test di inizio anno scolastico e voto finale di matematica di 8 studenti

	Test (X)	Voto (Y)	X-M(X)	Y-M(Y)	(X-M(X))(Y-M(X))	$(X-M(X))^2$	$(Y-M(Y))^2$
Studente 1	12	8	1.25	1.125	1.406	1.563	1.266
Studente 2	10	7	-0.75	0.125	-0.094	0.563	0.016
Studente 3	14	8	3.25	1.125	3.656	10.563	1.266
Studente 4	9	5	-1.75	-1.875	3.281	3.063	3.516
Studente 5	9	6	-1.75	-0.875	1.531	3.063	0.766
Studente 6	13	9	2.25	2.125	4.781	5.063	4.516
Studente 7	11	7	0.25	0.125	0.031	0.063	0.016
Studente 8	8	5	-2.75	-1.875	5.156	7.563	3.516
					19.750	31.500	14.875

M(X)	10.75
M(Y)	6.875
Codev(X;Y)	19.75
Cov(X;Y)	2.47
Var(X)	3.94
Var(Y)	1.86
sqm(X)	1.98
sqm(Y)	1.36



$$\rho = \frac{2.47}{1.98 \cdot 1.36} = 0.91$$

Analisi delle relazioni

Quale indice misura la relazione tra due variabili osservate?

Che tipo di variabili?

Che tipo di relazione supporre?

Indici statistici

		2 mutabili	1 variabile 1 mutabile	2 variabili
Indipendenza statistica	Legame bidirezionale	χ^2, Φ^2, V	χ^2, Φ^2, V	χ^2, Φ^2, V
	Legame unidirezionale	λ	λ	λ
Indipendenza in media	Legame bidirezionale			
	Legame unidirezionale		η^2	η^2
Indipendenza lineare	Legame bidirezionale			ρ
	Legame unidirezionale			Modello di regressione

Associazione tra variabili
Variabili quantitative

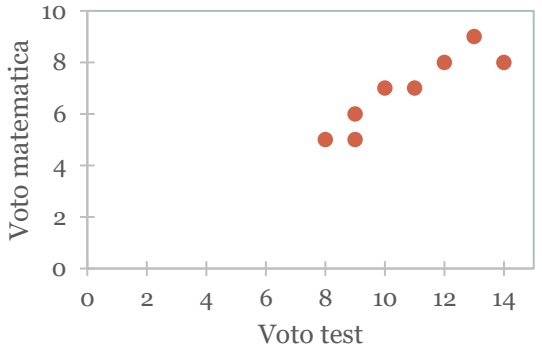
Per introdurre la regressione e la correlazione partiamo da un esempio, rappresentando congiuntamente due variabili *X* e *Y* con un grafico che si chiama **diagramma a dispersione** (detto anche *scatter)*

Voto ad un test di inizio anno scolastico e voto finale di matematica di 8 studenti

	Test	Voto
Studente 1	12	8
Studente 2	10	7
Studente 3	14	8
Studente 4	9	5
Studente 5	9	6
Studente 6	13	9
Studente 7	11	7
Studente 8	8	5

ρ= 0.91

Scatter tra voto test di inizio anno scolastico e voto finale di matematica



X ⟶ Voto test

Y ⟶ Voto matematica

Associazione tra variabili
Modello di Regressione lineare semplice

Con l'analisi di regressione lineare si cerca di individuare un **modello statistico** che può essere utilizzato per scopi descrittivi, interpretativi, previsivi. Si parla, infatti, di **modello di regressione**

Un modello statistico è una rappresentazione semplificata, ma (si auspica) soddisfacente, della realtà osservata

In genere è definito da una legge che lega le due variabili X e Y

Questa legge viene espressa da una funzione per cui si dice che tra X e Y sussiste una relazione funzionale.

Associazione tra variabili
Variabili quantitative

Per introdurre la regressione e la correlazione partiamo da un esempio, rappresentando congiuntamente due variabili *X* e *Y* con un grafico che si chiama **diagramma a dispersione** (detto anche *scatter)*

Voto ad un test di inizio anno scolastico e voto finale di matematica di 8 studenti

	Test	Voto
Studente 1	12	8
Studente 2	10	7
Studente 3	14	8
Studente 4	9	5
Studente 5	9	6
Studente 6	13	9
Studente 7	11	7
Studente 8	8	5

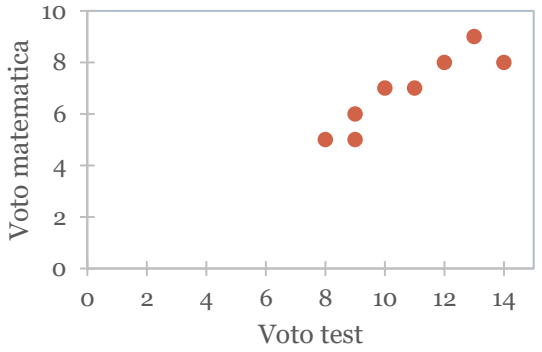
X → Voto test

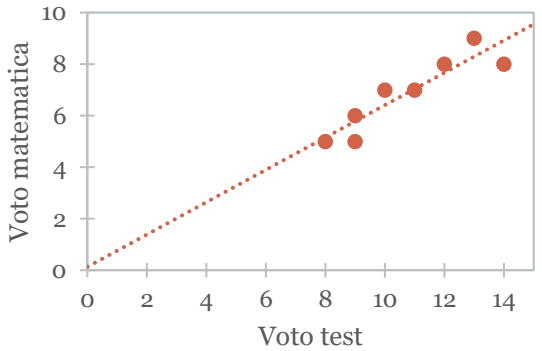
Y → Voto matematica

ρ= 0.91

?

Scatter tra voto test di inizio anno scolastico e voto finale di matematica





Associazione tra variabili
Modello di Regressione lineare semplice

Modello matematico

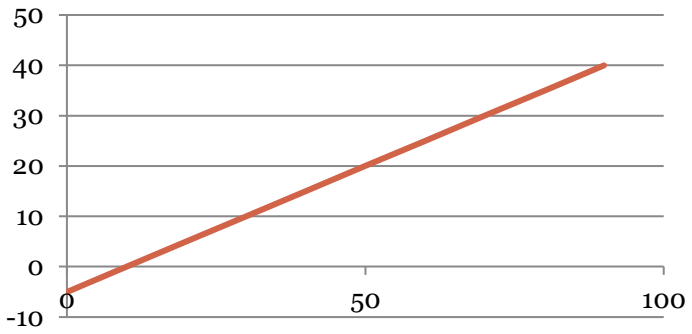
Una variabile ***Y*** è funzione di ***X*** se ad ogni valore di X corrisponde un unico valore di Y

In tal caso esiste ***una relazione funzionale***

Es: *X lunghezza del lato di un quadrato Y area del quadrato* $(Y = X^2)$

Una relazione funzionale ***lineare*** è

$$Y = a + bX$$



Y = variabile dipendente

X = variabile indipendente

a = Y-intercetta (il valore di Y quando X è pari a zero)

b = coefficiente angolare (l'incremento di Y per un incremento unitario di X)

Associazione tra variabili
Modello di Regressione lineare semplice

Relazione statistica

Negli studi empirici la relazione che lega **Y** e **X** non può mai essere una relazione matematica esatta perché ad ogni valore di X non corrisponderà mai un unico valore di Y

In tal caso si parla si ***relazione statistica***

Es: *X reddito procapite Y consumi familiari*

$$Y = f(X) + \varepsilon$$

$f(X)$ Definisce il contributo della variabile esplicativa X al valore della variabile di risposta

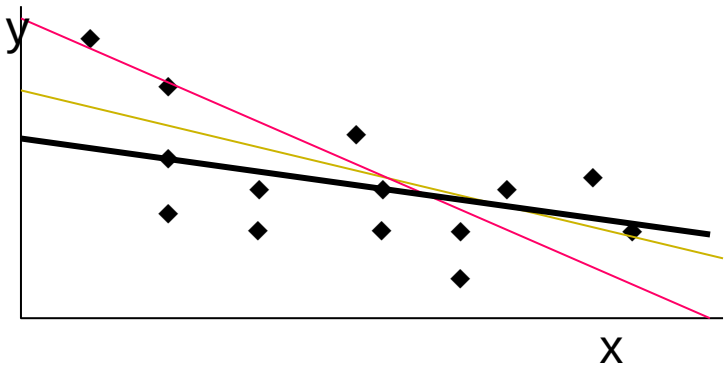
ε E' il ***residuo*** o l'***errore*** e giustifica la differenza tra il valore di Y per un valore fissato di X e il corrispondente valore $f(X)$

Nella regressione lineare semplice $f(X)$ è una funzione lineare

- ✓ La funzione di regressione che viene individuata esprime la relazione di dipendenza in media della variabile Y dalla variabile X
- ✓ Il valore medio dell'errore (ε) è nullo

Associazione tra variabili
Modello di Regressione lineare semplice

La domanda è:
Quale retta esprime meglio dati?



- Il problema consiste nel trovare la retta di regressione che passi attraverso la nuvola dei punti avvicinandosi il più possibile ad essi (interpolazione dei dati)
- In altri termini occorre individuare una retta che per ogni x_i restituisca un valore di y_i prossimo ai valori osservati

Associazione tra variabili
Modello di Regressione lineare semplice

Indichiamo con $\hat{y}_i = \hat{a} + \hat{b}x_i$ il valore di Y fornito dalla retta stimata in corrispondenza di x_i, dove $\hat{a}$ e $\hat{b}$ sono i **coefficienti di regressione**

$\hat{a}$

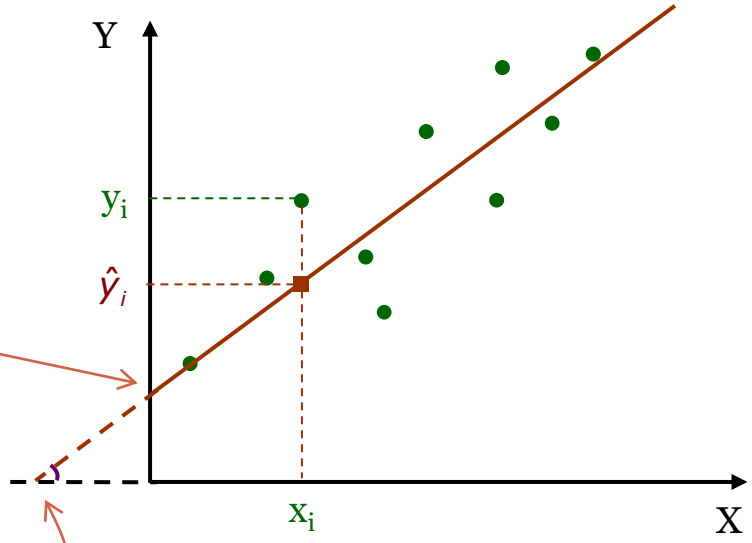
E' l'*intercetta* sull'asse delle ordinate. Può essere interpretato come il valore di Y per X=0 (quando ciò ha senso).

$\hat{b}$

E' il *coefficiente angolare* della retta di regressione in quanto funzione dell'angolo che la retta forma con l'asse delle ascisse.

Esprime dunque la *pendenza* (positiva, negativa o nulla) della retta.

Esprime anche quanto varia la variabile Y al variare unitario della variabile X.



Associazione tra variabili
Modello di Regressione lineare semplice

Indichiamo con $\hat{y}_i = \hat{a} + \hat{b}x_i$ il valore di Y fornito dalla retta stimata in corrispondenza di x_i, dove $\hat{a}$ e $\hat{b}$ sono i **coefficienti di regressione**

Diremo che una retta ha un migliore **adattamento ai dati osservati** di un'altra, se fissati i valori dei coefficienti di regressione, complessivamente i residui $\varepsilon_i = y_i - \hat{y}_i$ sono più piccoli

?

$$\sum_{I=1}^{N} \varepsilon_i = \sum_{I=1}^{N} (y_i - \hat{y}_i) = 0$$

$$\sum_{I=1}^{N} \varepsilon_i{}^2 = \sum_{I=1}^{N} (y_i - \hat{y}_i)^2 = min$$

Associazione tra variabili
Modello di Regressione lineare semplice

Metodo dei minimi quadrati- ricercare i coefficienti di regressione che rendono minima la funzione di perdita

$$G(\hat{a}, \hat{b}) = \sum_{i=1}^{n} (y_i \underbrace{- \hat{a} - \hat{b}x_i}_{\hat{y}_i})^2 = min$$

Associazione tra variabili
Modello di Regressione lineare semplice

Metodo dei minimi quadrati- ricercare i coefficienti di regressione che rendono minima la funzione di perdita

$$G(\hat{a},\hat{b}) = \sum_{i=1}^{n} (y_i - \underbrace{\hat{a} - \hat{b}x_i}_{\hat{y}_i})^2 = min$$

Derivando rispetto a $\hat{a}$ e $\hat{b}$ e ponendo le derivate parziali uguali a zero, otteniamo la formula

$$\frac{\partial G}{\partial \hat{a}} = -2\sum_{1=1}^{n}(y_i - \hat{a} - \hat{b}x_i) = 0$$

$$\frac{\partial G}{\partial \hat{b}} = -2\sum_{1=1}^{n}(y_i - \hat{a} - \hat{b}x_i)x_i = 0$$

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

$$b = \frac{Cov(XY)}{Var(X)}$$

Associazione tra variabili
Modello di Regressione lineare semplice

Metodo dei minimi quadrati- ricercare i coefficienti di regressione che rendono minima la funzione di perdita

$$G(\hat{a},\hat{b}) = \sum_{i=1}^{n} (y_i - \underbrace{\hat{a} - \hat{b}x_i}_{\hat{y}_i})^2 = min$$

La retta costruita con i valori di a e b ottenuti dalla risoluzione del sistema, sarà dunque quella più "vicina" ai punti, ossia quella che rende minima la somma dei quadrati delle distanze tra valori osservati e valori teorici della variabile dipendente y.

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

$$b = \frac{Cov(XY)}{Var(X)}$$

Associazione tra variabili
Modello di Regressione lineare semplice

X ⟶ Variabile indipendente

Y ⟶ Variabile dipendente

$$Y = f(X;\theta) + e$$

• Decidiamo di rappresentare la nube di punti con una *funzione* che passi *tra* i punti stessi;

• Tra tutte le possibili funzioni, scegliamo la funzione lineare, y=a+bx ;

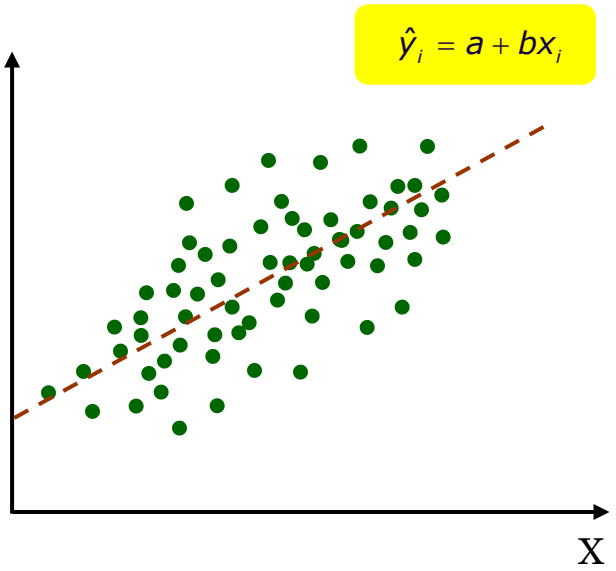
• Tra tutte le infinite possibili rette, scegliamo quella che ottimizza un criterio che definiamo arbitrariamente, per esempio *quella che minimizza la somma dei quadrati degli scarti tra valori osservati e valori teorici*:

$$S = \sum_i (y_i - \hat{y}_i)^2 = \sum_i (y_i - a - bx)^2 = min$$

• Il *metodo dei minimi quadrati* ci consente di ottenere le soluzioni di questo problema, soluzioni che rappresentano i parametri della retta:

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

$$b = \frac{Cov(XY)}{Var(X)}$$



• Sostituendo questi valori nell'equazione $\hat{y}_i = a + bx_i$, per ogni valore dato di X otterremo il corrispondente valore teorico di Y.

Associazione tra variabili
Modello di Regressione lineare semplice

Retta dei minimi quadrati

$$\hat{y}_i = \hat{a} + \hat{b}x_i$$

$$\hat{a} = \bar{Y} - \hat{b}\bar{X}$$

$$\bar{Y} = \hat{a} + \hat{b}\bar{X}$$

E' l'*intercetta* sull'asse delle ordinate. Può essere interpretato come il valore di Y per X=0 (quando ciò ha senso).

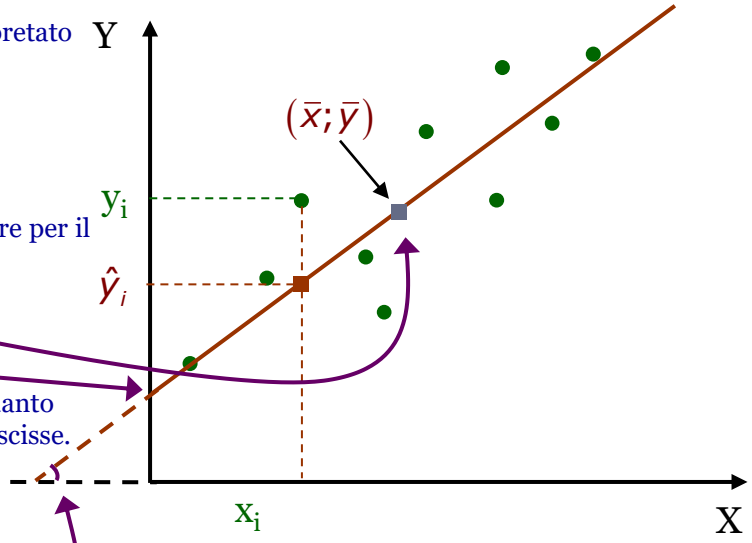
Il punto di coordinate $(\bar{x};\bar{y})$ è un punto della retta di regressione. La retta di regressione passa, dunque, sempre per il baricentro della nube.

$$\hat{b} = \frac{Cov(XY)}{Var(X)}$$

E' il *coefficiente angolare* della retta di regressione in quanto funzione dell'angolo che la retta forma con l'asse delle ascisse.

Esprime dunque la *pendenza* (positiva, negativa o nulla) della retta.

Esprime anche quanto varia la variabile Y al variare unitario della variabile X.



Associazione tra variabili
Modello di Regressione lineare semplice

Retta dei minimi quadrati

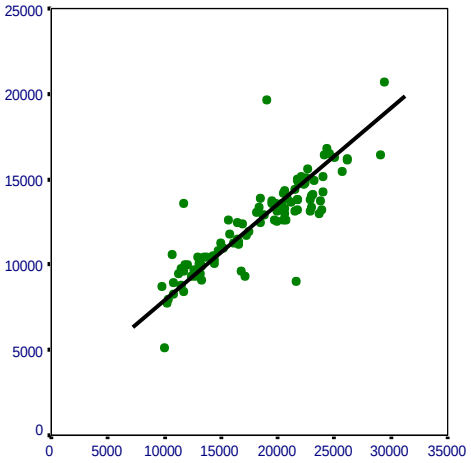
N.B. il segno di $\hat{b}$ è quello della codevianza fra le due variabili

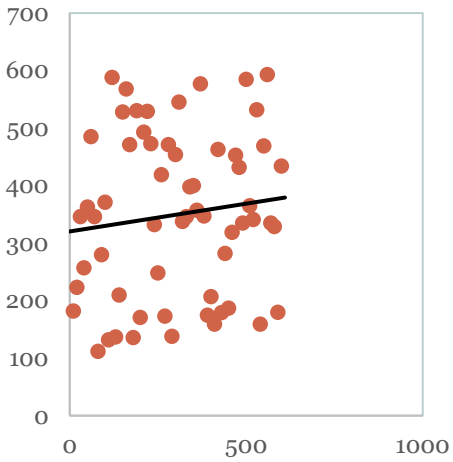
Ciò significa:

- se c'è concordanza tra le due variabili, $\hat{b}$ sarà maggiore di zero e la pendenza della retta sarà positiva
- se c'è discordanza tra le due variabili, $\hat{b}$ sarà minore di zero e la pendenza della retta sarà negativa
- se la codevianza è nulla, ovvero in caso di indipendenza lineare tra le variabili, $\hat{b}$ sarà uguale a zero e la retta sarà parallela all'asse delle ascisse

Associazione tra variabili
Modello di Regressione lineare semplice

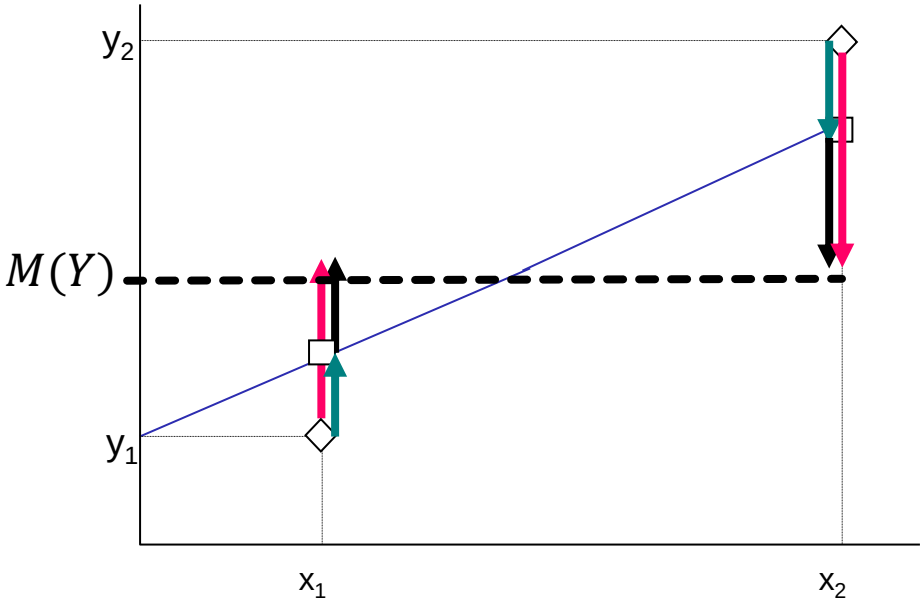
Il metodo dei minimi quadrati produce una modello di regressione lineare anche quando non c'è una relazione lineare tra X ed Y





E' importante, perciò, valutare la bontà di adattamento della retta alla nuvola dei punti

Associazione tra variabili
Modello di Regressione lineare semplice



Consideriamo due punti (x_1, y_1) e (x_2, y_2)

Devianza Totale in Y =

Devianza espressa dalla retta di regressione

\+ Devianza dell'errore

$$(y_1 - M(Y))^2 + (y_2 - M(Y))^2 = (\hat{y}_1 - M(Y))^2 + (\hat{y}_2 - M(Y)) + (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2$$

Associazione tra variabili
Modello di Regressione lineare semplice

Coefficiente di determinazione lineare

$$R^2 = \frac{Dev_{reg}}{Dev\ (Y)} = 1 - \frac{Dev_{res}}{Dev\ (Y)}$$

Associazione tra variabili
Modello di Regressione lineare semplice

Coefficiente di determinazione lineare

$$R^2 = \frac{Dev_{reg}}{Dev\ (Y)} = 1 - \frac{Dev_{res}}{Dev\ (Y)}$$

L'indice di determinazione varia tra 0 ed 1

- E' pari a 1 quando la variabilità totale di Y è totalmente spiegata dalla la retta di regressione e quindi la Devianza Totale coincide con la Devianza di Regressione.
- È pari a zero quando la variabilità totale di *Y* non è per nulla spiegata dalla la retta di regressione e quindi la Devianza Totale coincide con la Devianza Residua
- Tanto più R^2 si approssima ad 1, tanto più la bontà di adattamento del modello ai dati è buona

Associazione tra variabili
Modello di Regressione lineare semplice

Coefficiente di determinazione lineare

Nel caso della regressione lineare semplice

$$R^2 = \left[\frac{Codev(X,Y)}{Dev\,(X\,)Dev\,(Y)}\right]^2 = \rho^2$$

Esempio 1 – Modello di Regressione

Voto ad un test di inizio anno scolastico e voto finale di matematica di 8 studenti

	Test (X)	Voto (Y)	X-M(X)	Y-M(Y)	(X-M(X))(Y-M(X))	$(X-M(X))^2$	$(Y-M(Y))^2$
Studente 1	12	8	1.25	1.125	1.406	1.563	1.266
Studente 2	10	7	-0.75	0.125	-0.094	0.563	0.016
Studente 3	14	8	3.25	1.125	3.656	10.563	1.266
Studente 4	9	5	-1.75	-1.875	3.281	3.063	3.516
Studente 5	9	6	-1.75	-0.875	1.531	3.063	0.766
Studente 6	13	9	2.25	2.125	4.781	5.063	4.516
Studente 7	11	7	0.25	0.125	0.031	0.063	0.016
Studente 8	8	5	-2.75	-1.875	5.156	7.563	3.516
					19.750	31.500	14.875

M(X)	10.75
M(Y)	6.875
Codev(X;Y)	19.75
Cov(X;Y)	2.47
Var(X)	3.94
Var(Y)	1.86
sqm(X)	1.98
sqm(Y)	1.36

Esempio 1 – Modello di Regressione

Voto ad un test di inizio anno scolastico e voto finale di matematica di 8 studenti

	Test (X)	Voto (Y)	X-M(X)	Y-M(Y)	(X-M(X))(Y-M(X))	$(X-M(X))^2$	$(Y-M(Y))^2$
Studente 1	12	8	1.25	1.125	1.406	1.563	1.266
Studente 2	10	7	-0.75	0.125	-0.094	0.563	0.016
Studente 3	14	8	3.25	1.125	3.656	10.563	1.266
Studente 4	9	5	-1.75	-1.875	3.281	3.063	3.516
Studente 5	9	6	-1.75	-0.875	1.531	3.063	0.766
Studente 6	13	9	2.25	2.125	4.781	5.063	4.516
Studente 7	11	7	0.25	0.125	0.031	0.063	0.016
Studente 8	8	5	-2.75	-1.875	5.156	7.563	3.516
					19.750	31.500	14.875

M(X)	10.75
M(Y)	6.875
Codev(X;Y)	19.75
Cov(X;Y)	2.47
Var(X)	3.94
Var(Y)	1.86
sqm(X)	1.98
sqm(Y)	1.36

$$b = \frac{Cov(XY)}{Var(X)} = \frac{2.47}{3.94} = 0.627$$

$$a = \bar{y} - b\bar{x} = 6.875 - 0.627 * 10.75 = 0.135$$

Esempio 1 – Modello di Regressione

Voto ad un test di inizio anno scolastico e voto finale di matematica di 8 studenti

	Test (X)	Voto (Y)	X-M(X)	Y-M(Y)	(X-M(X))(Y-M(X))	$(X-M(X))^2$	$(Y-M(Y))^2$
Studente 1	12	8	1.25	1.125	1.406	1.563	1.266
Studente 2	10	7	-0.75	0.125	-0.094	0.563	0.016
Studente 3	14	8	3.25	1.125	3.656	10.563	1.266
Studente 4	9	5	-1.75	-1.875	3.281	3.063	3.516
Studente 5	9	6	-1.75	-0.875	1.531	3.063	0.766
Studente 6	13	9	2.25	2.125	4.781	5.063	4.516
Studente 7	11	7	0.25	0.125	0.031	0.063	0.016
Studente 8	8	5	-2.75	-1.875	5.156	7.563	3.516
					19.750	31.500	14.875

M(X)	10.75
M(Y)	6.875
Codev(X;Y)	19.75
Cov(X;Y)	2.47
Var(X)	3.94
Var(Y)	1.86
sqm(X)	1.98
sqm(Y)	1.36

$$b = \frac{Cov(XY)}{Var(X)} = \frac{2.47}{3.94} = 0.627$$

$$a = \bar{y} - b\bar{x} = 6.875 - 0.627 * 10.75 = 0.135$$

$$\hat{y}_i = 0.135 + 0.627x_i$$

Esempio 2 – Modello di Regressione

Il responsabile commerciale di un'azienda paga alcune stazioni radio locali per mandare in onda per una settimana un messaggio pubblicitario relativo all'immissione sul mercato di un nuovo prodotto. Poiché le stazioni richiedono compensi diversi, esiste una variabilità nel numero di messe in onda del messaggio pubblicitario.

Stazioni radio	Messaggi al giorno X	Vendite (in milioni) Y
Fox	4	15
FXZ	2	8
Power	5	21
Lizard	6	24
Rodeo	3	17

1. Determinare una misura dell'eventuale associazione tra la frequenza dei messaggi pubblicitari e le vendite del prodotto.

2. Determinare se esiste una misura dell'eventuale dipendenza tra la frequenza dei messaggi pubblicitari e le vendite del prodotto.

Esempio 2 – Modello di Regressione

Stazioni radio	Messaggi X	Vendite Y	$X_i - \mu_x$	$Y_i - \mu_y$	$(X_i - \mu_x)(Y_i - \mu_y)$	$(X_i - \mu_x)^2$	$(Y_i - \mu_y)^2$
Fox	4	15	0	-2	0	0	4
FXZ	2	8	-2	-9	18	4	81
Power	5	21	1	4	4	1	16
Lizard	6	24	2	7	14	4	49
Rodeo	3	17	-1	0	0	1	0
Totale	20	85			36	10	

$$\mu_x = 4 \qquad \sigma_x = 1.414$$

$$\mu_y = 17 \qquad \sigma_y = 5.48$$

$$\sigma_{xy} = 7.2$$

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \Rightarrow \rho_{xy} = \frac{7.2}{1.414 \cdot 5.48} = \frac{7.2}{7.748} = 0.93$$

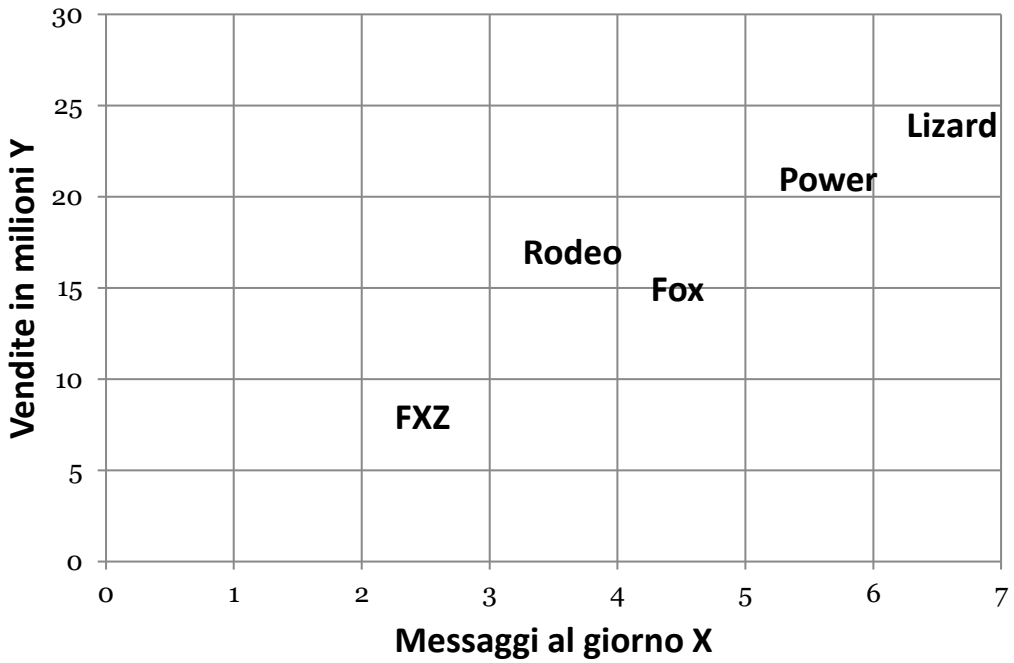
Esempio 2 – Modello di Regressione

$$\rho_{xy} = 0.93$$

Sussiste un forte legame lineare positivo tra la frequenza dei messaggi pubblicitari e le vendite del prodotto.

Stazioni radio	Messaggi X	Vendite Y
Fox	4	15
FXZ	2	8
Power	5	21
Lizard	6	24
Rodeo	3	17

Diagramma di dispersione



Esempio 2 – Modello di Regressione

Stazioni radio	Messaggi X	Vendite Y	$X_i - \mu_x$	$Y_i - \mu_y$	$(X_i - \mu_x)(Y_i - \mu_y)$	$(X_i - \mu_x)^2$	$(Y_i - \mu_y)^2$
Fox	4	15	0	-2	0	0	4
FXZ	2	8	-2	-9	18	4	81
Power	5	21	1	4	4	1	16
Lizard	6	24	2	7	14	4	49
Rodeo	3	17	-1	0	0	1	0
Totale	20	85			36	10	

$\mu_x = 4 \qquad \sigma_x = 1.414$

$\mu_y = 17 \qquad \sigma_y = 5.48$

$\sigma_{xy} = 7.2$

X → Variabile indipendente → Messaggi pubblicitari

Y → Variabile dipendente → Vendite del prodotto

$$b_1 = \frac{Cov(XY)}{Var(X)} \rightarrow b_1 = \frac{7.2}{2} = 3.6$$

$$b_0 = \bar{y} - b_1\bar{x} \rightarrow b_0 = 17 - 3.6 * 4 = 2.6$$

$$Vendite = 2.6 + 3.6 \cdot Messaggi$$

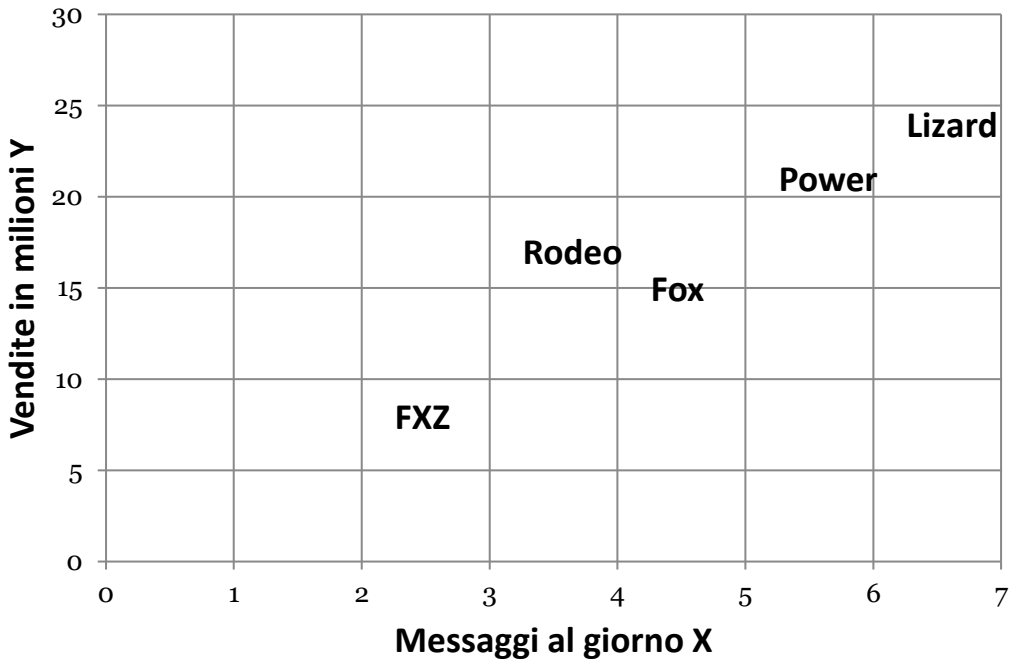
Esempio 2 – Modello di Regressione

$$\rho_{xy} = 0.93$$

Sussiste un forte legame lineare positivo tra la frequenza dei messaggi pubblicitari e le vendite del prodotto.

Stazioni radio	Messaggi X	Vendite Y
Fox	4	15
FXZ	2	8
Power	5	21
Lizard	6	24
Rodeo	3	17

Diagramma di dispersione

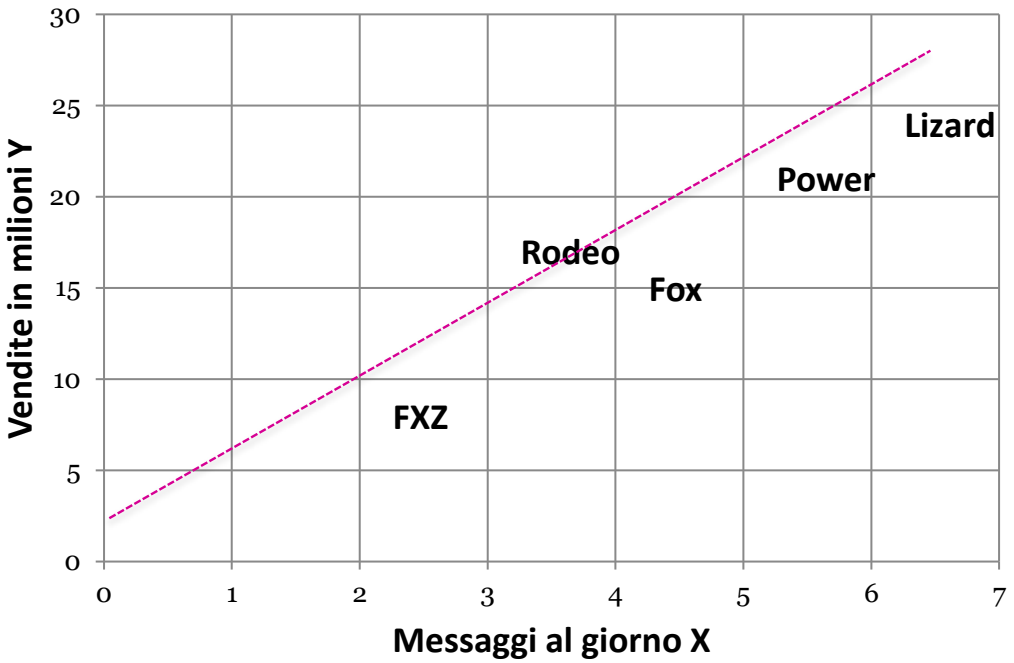


Esempio 2 – Modello di Regressione

$$\rho_{xy} = 0.93$$

Sussiste un forte legame lineare positivo tra la frequenza dei messaggi pubblicitari e le vendite del prodotto.

Stazioni radio	Messaggi X	Vendite Y
Fox	4	15
FXZ	2	8
Power	5	21
Lizard	6	24
Rodeo	3	17



$$Vendite = 2.6 + 3.6 \cdot Messaggi$$

Esempio 2 – Modello di Regressione

$\mu_y = 17$

$$Vendite = 2.6 + 3.6 \cdot Messaggi$$

Stazioni radio	Messaggi X	Vendite Y	$(\hat{y} - \mu_y)^2$	$(y - \mu_y)^2$
Fox	4	15	0	4
FXZ	2	8	51.84	81
Power	5	21	12.96	16
Lizard	6	24	51.84	49
Rodeo	3	17	12.96	0
Totale			129.6	150

$$R^2 = \frac{Dev(\hat{Y})}{Dev(Y)} = \frac{\sum_i (\hat{y} - \mu_y)^2}{\sum_i (y - \mu_y)^2}$$

$$R^2 = \frac{Dev(\hat{Y})}{Dev(Y)} = \frac{129.6}{150} = 0.86$$

$\hat{y}_{Fox} = 2.6 + 3.6 \cdot 4 = 17$

$\hat{y}_{Fxz} = 2.6 + 3.6 \cdot 2 = 9.8$

$\hat{y}_{Power} = 2.6 + 3.6 \cdot 5 = 20.6$

$\hat{y}_{Lizard} = 2.6 + 3.6 \cdot 6 = 24.2$

$\hat{y}_{Rodeo} = 2.6 + 3.6 \cdot 3 = 13.4$

Video

✓ *Correlazione*

- https://www.youtube.com/watch?v=S-j6T-GAjAI
- https://www.youtube.com/watch?v=NPwiJc4oUtw&t=37s (Excel)

✓ *Modello di regressione*

https://www.youtube.com/watch?v=39Y66O6wRGM
https://www.youtube.com/watch?v=DZpW5A_iNXw (Excel)

Video suggeriti

https://www.youtube.com/watch?v=iCYhc-iNKa8

https://www.youtube.com/watch?v=zNERNwVoJAw