DIMA

CORSO LAUREA MAGISTRALE IN **DIGITAL MARKETING**

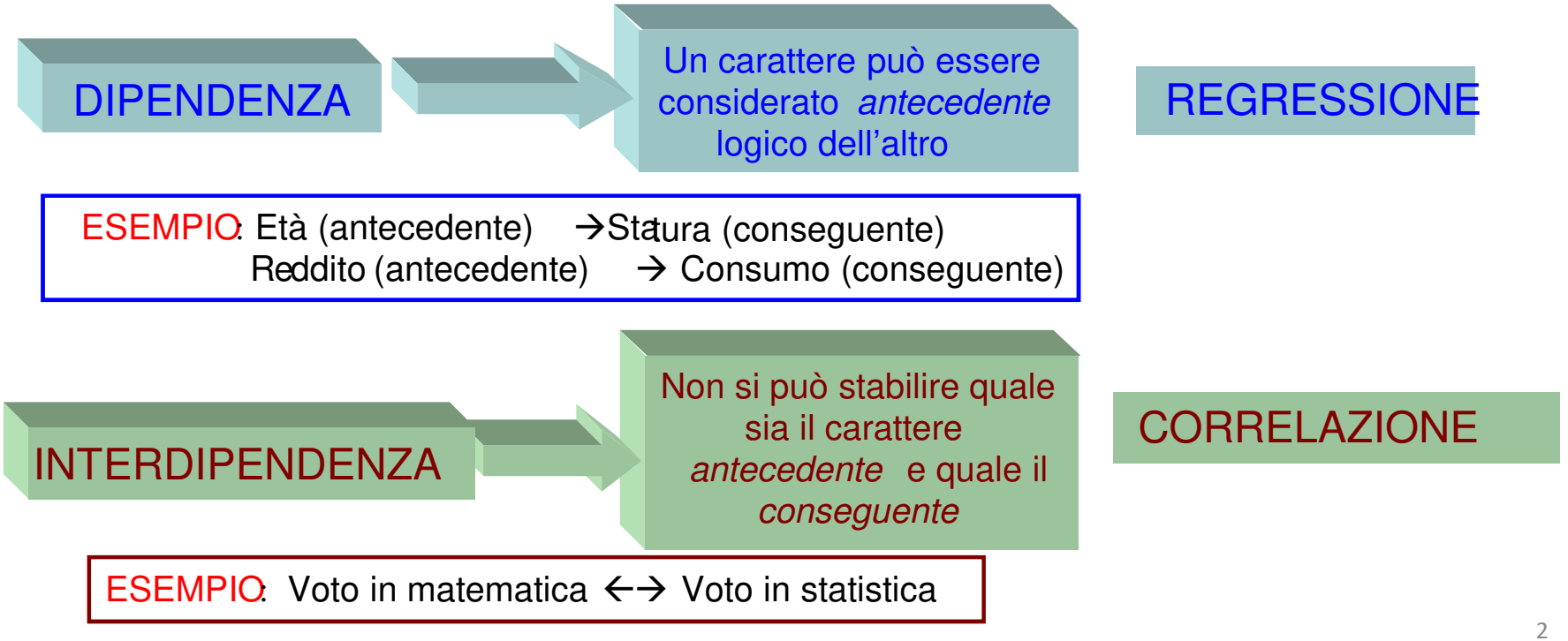
STATISTICA DESCRITTIVA

Prof. ssa ANNA LINA SARRA

Analisi delle relazioni fra i caratteri di una distribuzione doppia quantitativa

Data la distribuzione congiunta di due caratteri quantitativi, i quesiti a cui la statistica deve rispondere sono

- ✓esiste una relazione fra i due caratteri?
- ✓che tipo di relazione esiste?



- **Analisi delle distribuzioni doppie: analisi della correlazione**

Dai dati univariati ai dati bivariati

- In molte situazioni interessa studiare se esiste una relazione tra due variabili misurate sulle stesse unità.

Esempio:

– "*il voto di maturità è in relazione con la performance universitaria?*"

- Oppure si desidera prevedere il valore di una variabile conoscendo il valore di un'altra.

Esempio:

– "*conoscendo l'età del paziente, è possibile prevedere la sua pressione arteriosa?*"

- La statistica permette di rispondere a questo tipo di domande, con strumenti adatti alla natura delle variabili in esame. A tale scopo, per variabili quantitative, si tratteranno:
  - La CORRELAZIONE, che misura la dipendenza lineare tra due variabili;
  - La REGRESSIONE, che valuta la relazione lineare tra due variabili.

Correlazione

La correlazione misura l'associazione tra due variabili quantitative.

È lo strumento che si utilizza quando si hanno a disposizione coppie di valori di variabili. Permette di valutare come variano i valori di una variabile al variare dell'altra e viceversa.

- Esempi:
  - Numero di sigarette fumate in gravidanza e tasso di crescita del feto ⇒ all'aumentare del numero di sigarette fumate diminuisce il tasso di crescita (correlazione negativa).
  - Livello di colesterolo e BMI (Body Mass Index = peso (kg)/altezza^2 (m^2)) ⇒ tanto è maggiore il BMI quanto è maggiore il livello di colesterolo (correlazione positiva).
  - Il valor medio della temperatura (ambiente) e il BMI ⇒ non c'è motivo di pensare che la temperatura influenzi il BMI delle persone (assenza di correlazione).
- La relazione può essere valutata tramite:
  - Un grafico (grafico di dispersione)
  - Un indice che quantifica il grado di correlazione (coefficiente di correlazione)

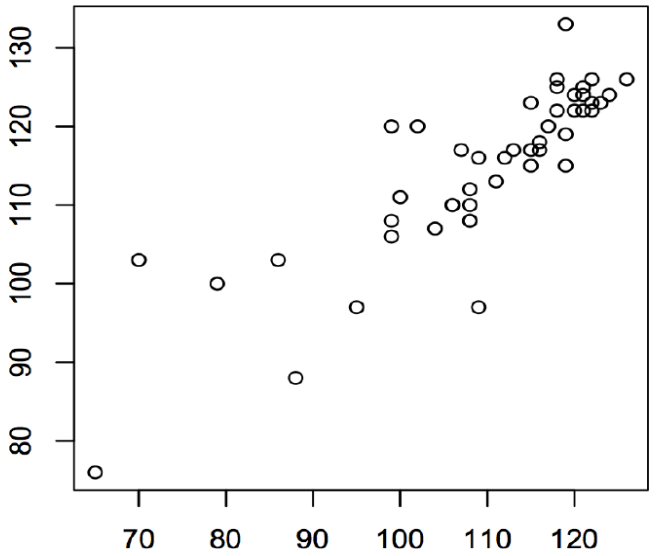
Diagramma a dispersione

Nello studio dell'associazione tra due variabili quantitative misurate sulle stesse unità statistiche, indicate con X e Y , è molto utile disegnare un grafico, il diagramma di dispersione, prima di procedere con altre analisi formali.

Nel grafico di dispersione le coppie

$$(x_1, y_1)\ (x_2, y_2)\ \ldots\ (x_n, y_n)$$

di valori di due variabili quantitative misurate sulle n unità sono rappresentati come punti di un piano cartesiano, i cui assi corrispondono alle due variabili.

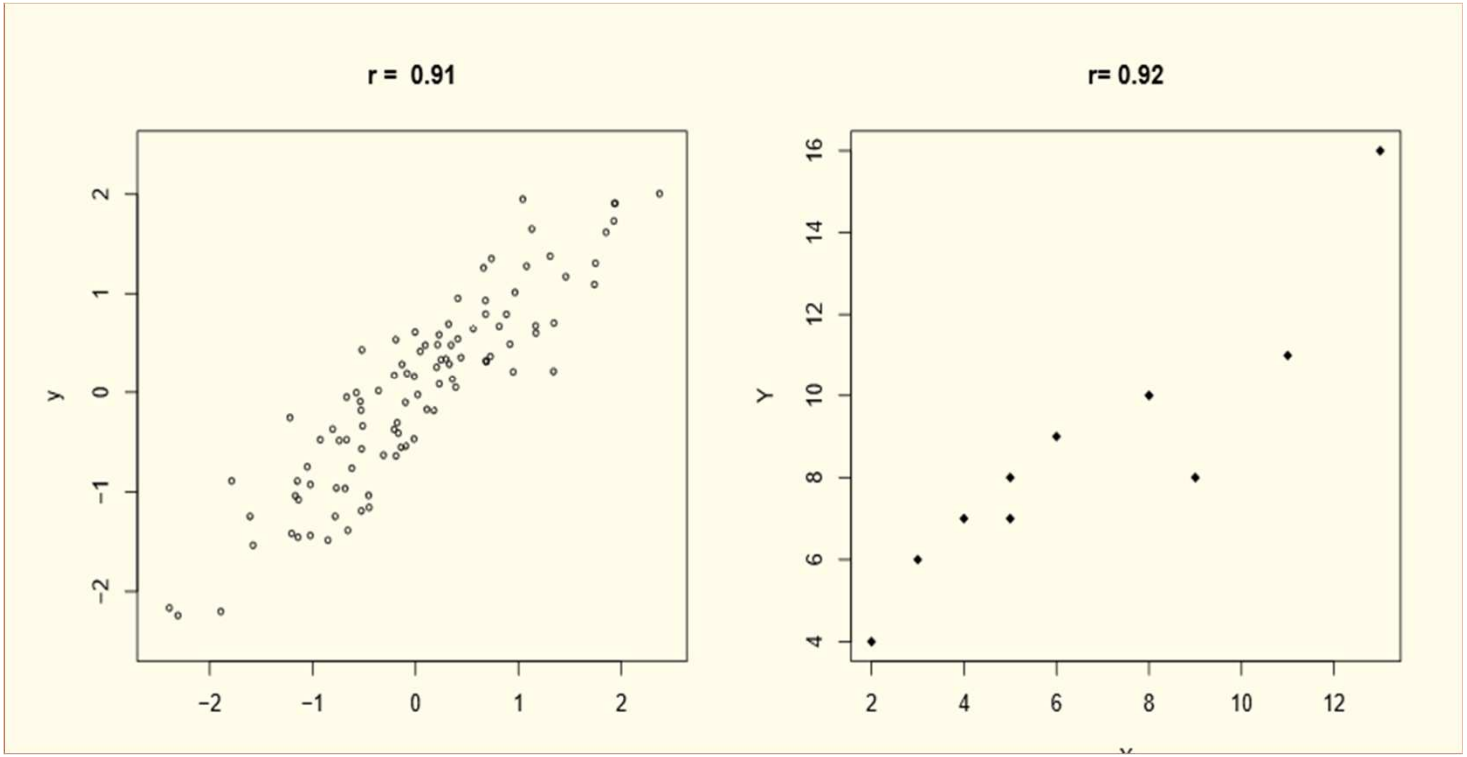


Correlazione

Data una **distribuzione doppia in forma disaggregata**, si dice che tra le due variabili *X* e *Y*

- vi è **correlazione positiva** o concordanza quando esse tendono a crescere (decrescere) insieme
- vi è **correlazione negativa** o discordanza quando al crescere di una variabile l'altra tende a decrescere.

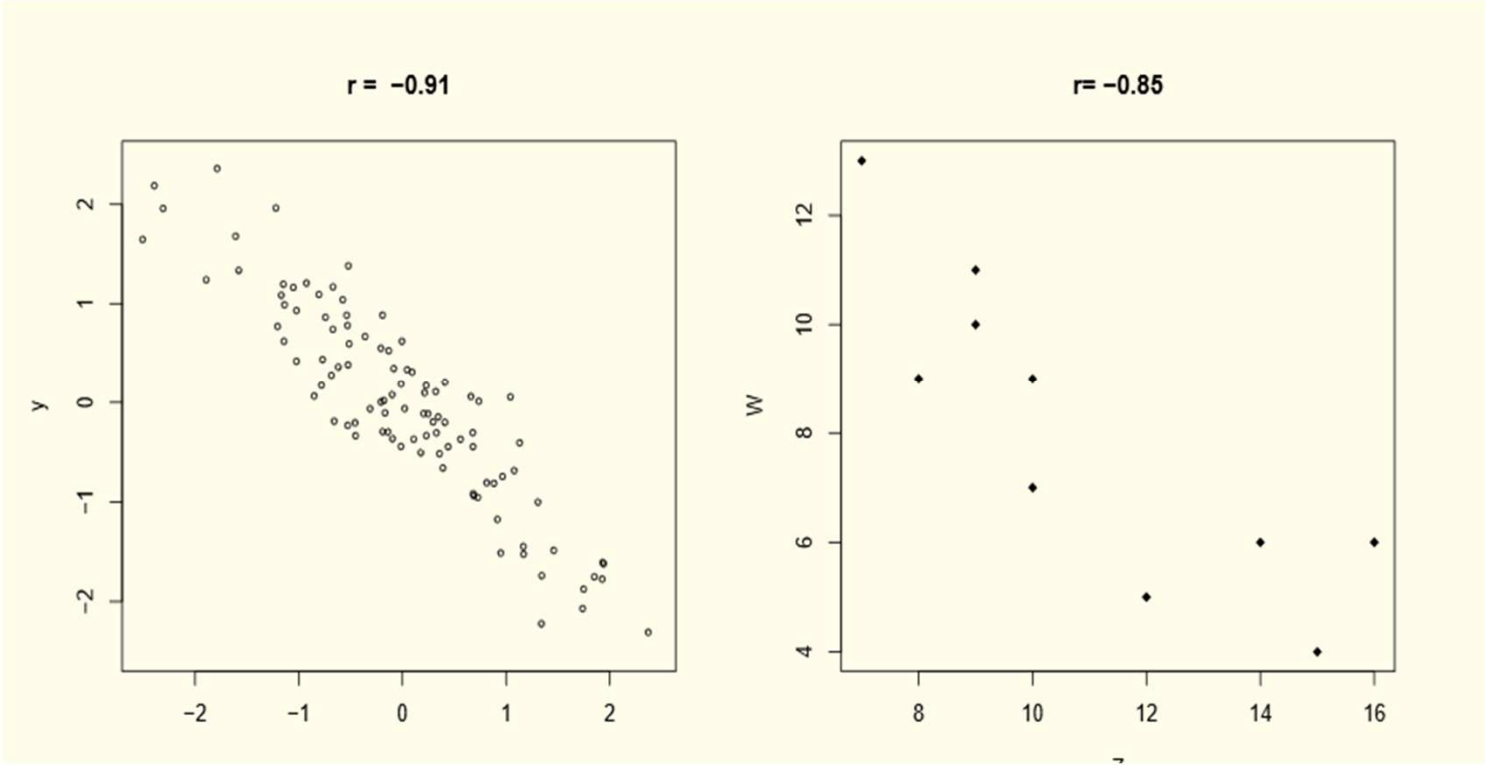
Esempio di correlazione positiva



All'aumentare di X aumenta anche Y, ciascuna variabile a modo suo e viceversa.

È una relazione lineare proporzionale.

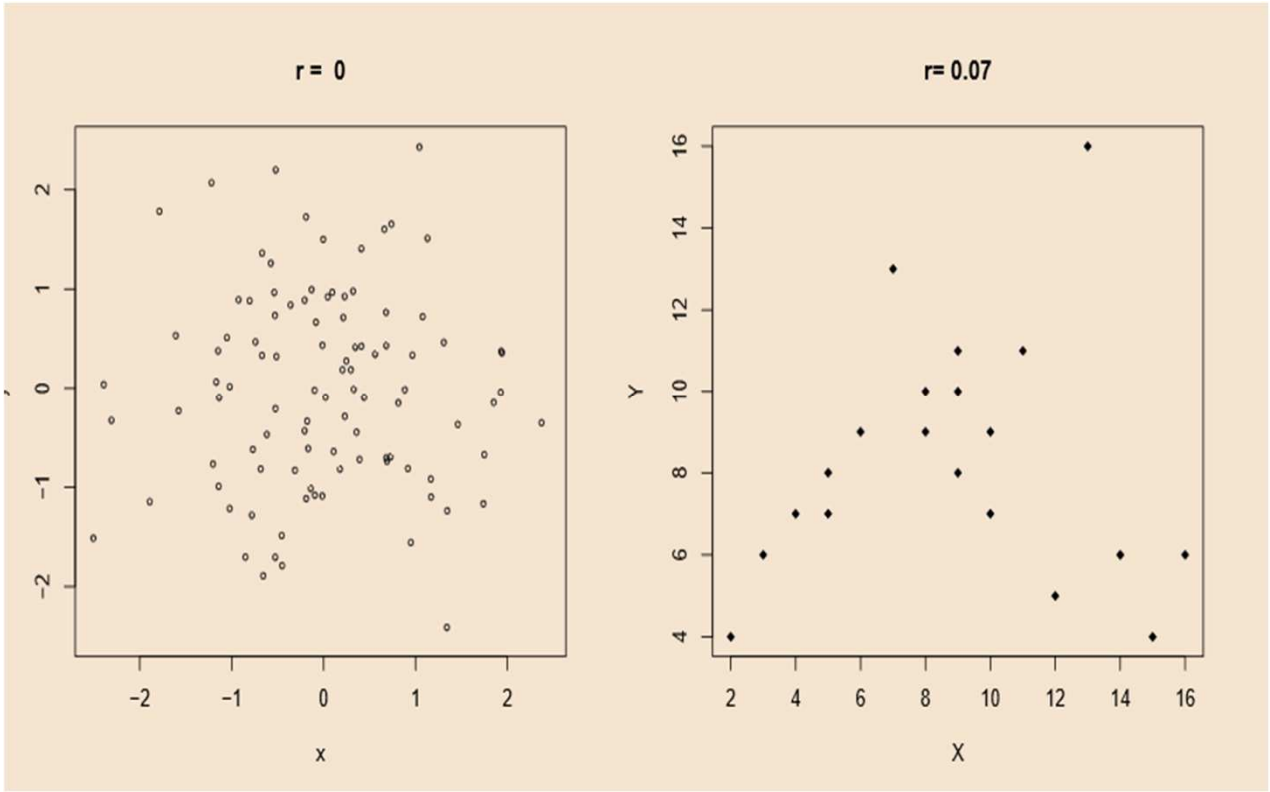
Esempio di correlazione negativa



All'aumentare di X diminuisce Y, ciascuna variabile a modo suo. E viceversa.

È una relazione lineare proporzionale.

Esempio di correlazione nulla



Non c'è alcun legame lineare fra X e Y.

Ciascuna
varia indipendentemente dall'altra

COVARIANZA

Per avere una valutazione analitica del grado di associazione tra due variabili quantitative, esiste un indice che misura la dispersione nel piano dei punti dal proprio centro : la **COVARIANZA**

La covarianza è un indice che esprime la quantità di varianza che due variabili hanno in comune.

La formula deriva da quella della varianza.

In formula

La covarianza è

$$cov(X,Y) = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{N}$$

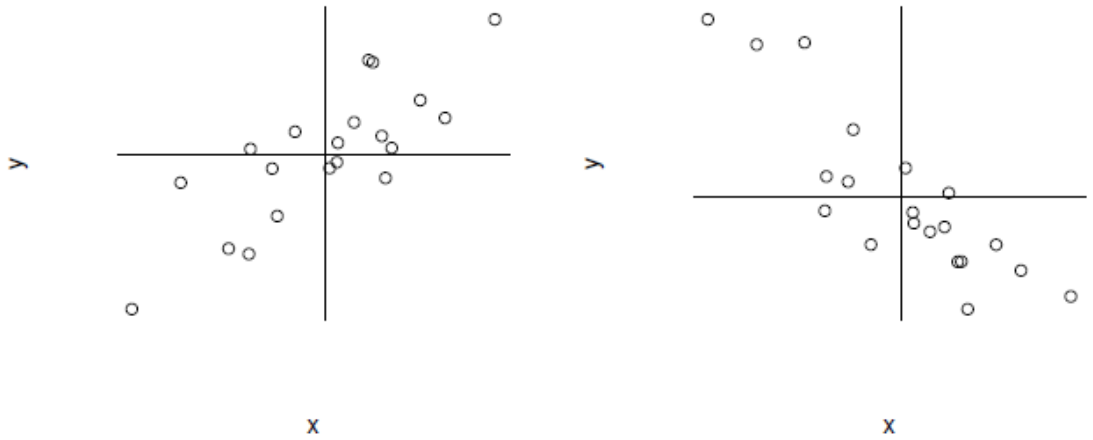
La varianza è

$$var(X) = \frac{\sum(X-\overline{X})^2}{N} = \frac{\sum(X-\overline{X})(X-\overline{X})}{N}$$

Notare la somiglianza tra le due formule

La covarianza

La covarianza, a differenza della varianza che è sempre positiva, **misura l'eventuale direzione del legame**, ovvero se le due variabili si muovono nella stessa direzione o in direzioni opposte. Il segno della covarianza riflette il senso crescente o decrescente dell'allineamento tendenziale.

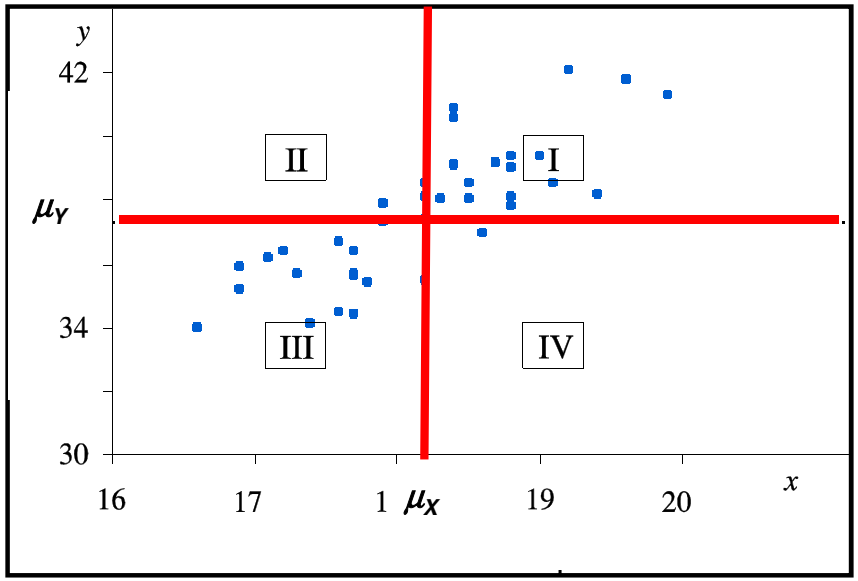


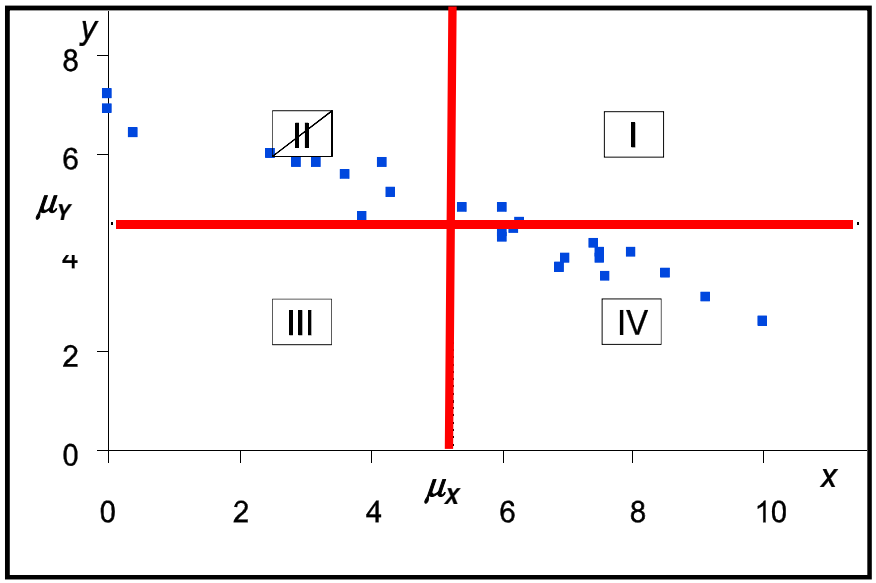
Interpretazione geometrica

$$cov(X,Y) = \frac{\sum(X - \overline{X})(Y - \overline{Y})}{N}$$

Concordanza: i punti osservati sono collocati in prevalenza nel I e nel III quadrante dei nuovi assi cartesiani aventi origine nel punto (μ_X, μ_Y).

Discordanza: i punti osservati sono collocati in prevalenza nel secondo e nel quarto quadrante dei nuovi assi cartesiani aventi origine nel punto (μ_X, μ_Y).





Osservazioni

- **Se X e Y sono concordi,** allora la covarianza assume **segno positivo**;
- **Se X e Y sono discordi,** allora la covarianza assume **segno negativo**;
- **Se la covarianza è nulla**, X e Y sono **indifferenti (incorrelati).**

COVARIANZA **misura assoluta di concordanza** tra due caratteri x e y ed è espressa in un'unità di misura pari al prodotto delle unità di misura dei due caratteri.

Campo di variazione della covarianza

La covarianza può assumere sia valori positivi che negativi.
In particolare vale che

$$-\sigma_x \sigma_y \leq COV(XY) \leq \sigma_x \sigma_y$$

Per ricercare un indice relativo che permetta di effettuare confronti significativi occorrerà **dividere la covarianza per il prodotto degli scarti quadratici medi di X e Y** L'indice così ottenuto prende valori in [-1,1] e viene detto **coefficiente di correlazione.**

Il coefficiente di correlazione lineare di Bravais-Pearson

In una **distribuzione doppia disaggregata**, il coefficiente di correlazione lineare di Bravais-Pearson è

$$r = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - \mu_X}{\sigma_X}.\frac{y_i - \mu_y}{\sigma_y}\right).$$

Scarti standardizzati della X e della Y

$$z_{x_i} = \frac{x_i - \mu_X}{\sigma_X}; \quad z_{y_i} = \frac{y_i - \mu_y}{\sigma_y}$$

Si può ottenere anche come

$$r = \frac{\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{N}(x_i - \mu_X)^2 \sum_{i=1}^{N}(y_i - \mu_y)^2}} = \frac{C_{xy}}{\sqrt{D_x D_y}}$$

dove il segno è determinato dalla codevianza

$$C_{xy} = \sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_y)$$

Interpretazione geometrica della formula

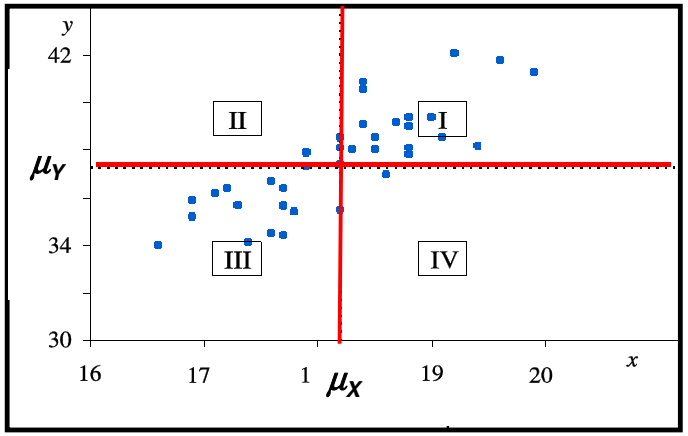
$$r = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - \mu_X}{\sigma_X}\cdot\frac{y_i - \mu_Y}{\sigma_Y}\right).$$

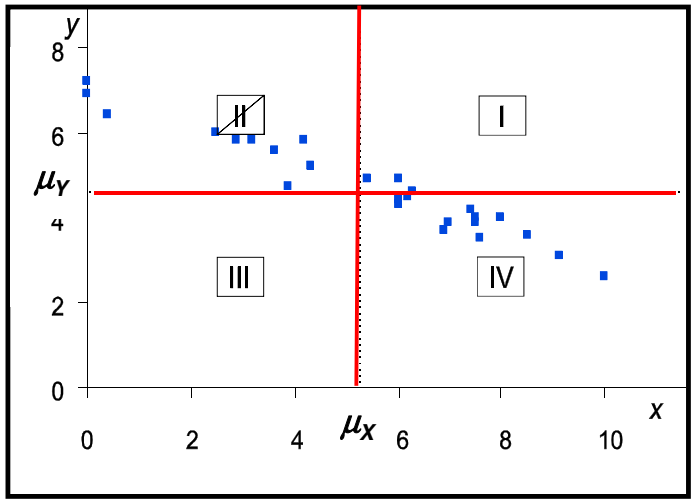
Ne segue che i prodotti $\dfrac{x_i - \mu_X}{\sigma_X}\cdot\dfrac{y_i - \mu_Y}{\sigma_Y}$

sono in prevalenza positivi nel caso di concordanza e prevalentemente negativi nel caso di discordanza. Cosicché la quantità

$$r = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - \mu_X}{\sigma_X}\cdot\frac{y_i - \mu_Y}{\sigma_Y}\right),$$

media di tali prodotti, è positiva in caso di concordanza e negativa nel caso di discordanza.





Coefficiente di correlazione

È un indice statistico che misura l'associazione (relazione) fra due variabili.

Misura come le due variabili si muovono assieme, ossia come co-relano.

Viene espresso come un valore che oscilla fra -1 e 1.

Coefficiente di correlazione

A. Riassunto numerico della forza della relazione fra due variabili

B. Permette di sostituire un diagramma a dispersione con un semplice indice.

È costituito da due parti:

Un segno che indica la direzione della relazione
Un numero fra 0 e 1 che indica la forza della relazione

1 indica una relazione perfetta, esprimibile tramite una formula matematica precisa

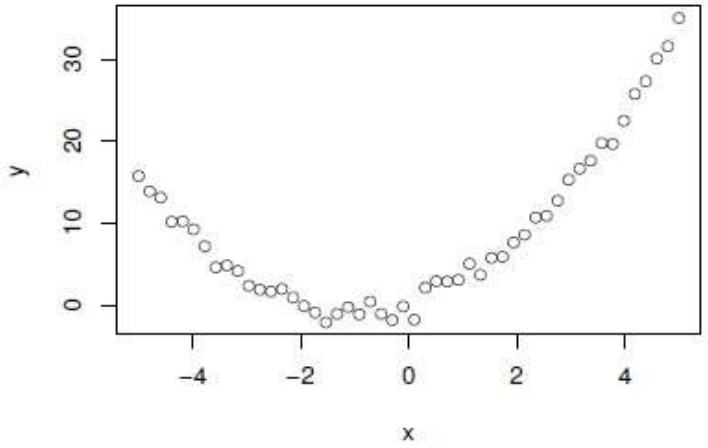
0 indica la mancanza di qualunque relazione fra le due variabile.

Guida all'interpretazione di r

- $-1 \leq r_{xy} \leq 1$
- $r_{xy} = +1$: correlazione positiva perfetta (tutti i punti su una retta: concordi)
- $r_{xy} = -1$: correlazione negativa perfetta (tutti i punti su una retta: discordi)
- $r_{xy} > 0$: correlazione positiva
- $r_{xy} < 0$: correlazione negativa
- $r_{xy} \cong 0$: assenza di relazione lineare

Quando tra X e Y non vi è una relazione lineare o essa è estremamente debole, il valore dell'indice r_{xy} è zero o circa zero, e le variabili sono dette incorrelate.

ATTENZIONE: Il coefficiente di correlazione misura una associazione lineare. Il valore $r_{xy} = 0$ non indica tuttavia un'assenza di relazione tra le due variabili. Può esserci una relazione curvilinea.



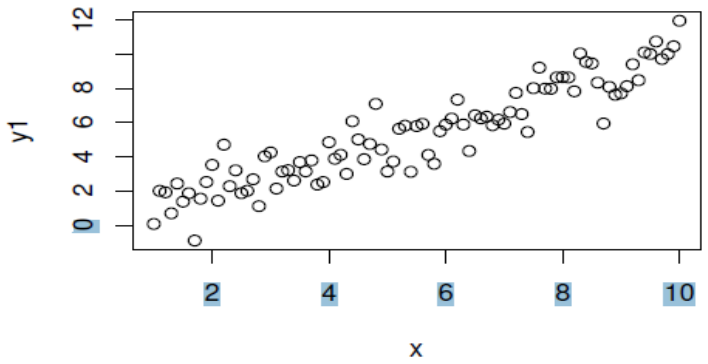
Guida all'interpretazione di r

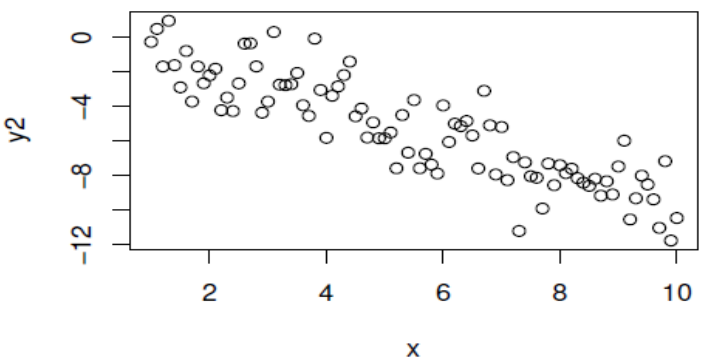
L'interpretazione si applica al valore della correlazione indipendentemente dal segno.

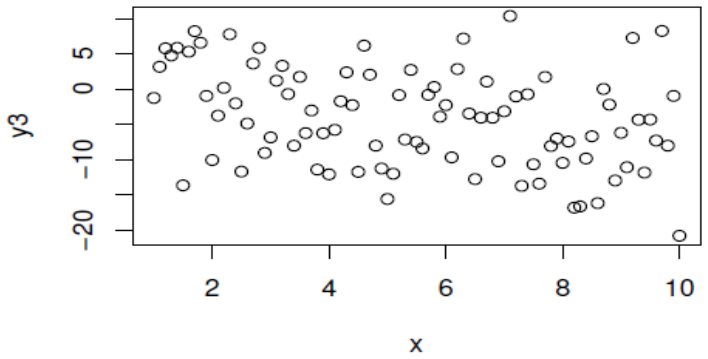
Valore di r	Correlazione	Relazione
0.00-0.20	Piccola	Molto poco intensa, quasi inesistente
0.20-0.40	Bassa	Piccola, appena appena apprezzabile
0.40-0.60	Regolare	Considerevole
0.60-0.80	Alta	Intensa
0.80-1.00	Molto alta	Molto intensa

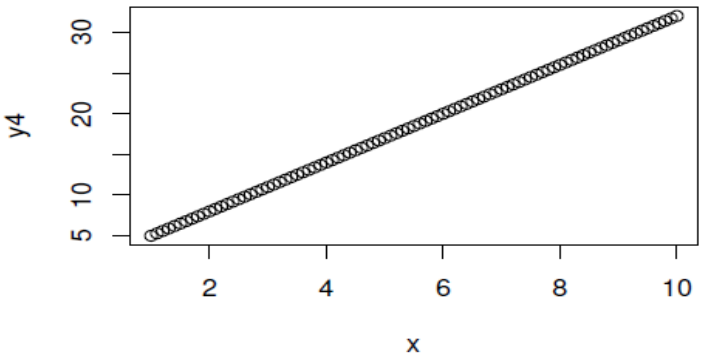
Il segno indica solo la relazione proporzionale (positiva) o inversamente proporzionale (negativa).

Quale correlazione……









Coefficiente di correlazione lineare di Bravais-Pearson: Esempio

La Tabella mostra i punteggi ottenuti da un gruppo di 10 studenti agli esami di un college (X) e ad un test di comprensione verbale (Y).

Studente	Esami di ammissione *X*	Test di comprensione verbale *Y*
A	52	49
B	28	34
C	70	45
D	51	49
E	49	40
F	65	50
G	49	37
H	49	49
I	63	52
J	32	32

Coefficiente di correlazione lineare di Bravais-Pearson: calcolo

$$r = \frac{\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{N}(x_i - \mu_X)^2 \sum_{i=1}^{N}(y_i - \mu_y)^2}}$$

- **indice di correlazione**

Esame di ammissione X	Test di comprensione verbale Y	$x_i - \mu_x$	$y_i - \mu_y$	$(x_i - \mu_x)^2$	$(y_i - y_x)^2$	$(x_i - \mu_x) * (y_i - \mu_Y)$
52	49	1.2	5.3	1.44	28.09	6.36
28	34	-22.8	-9.7	519.84	94.09	221.16
70	45	19.2	1.3	368.64	1.69	24.96
51	49	0.2	5.3	0.04	28.09	1.06
49	40	-1.8	-3.7	3.24	13.69	6.66
65	50	14.2	6.3	201.64	39.69	89.46
49	37	-1.8	-6.7	3.24	44.89	12.06
49	49	-1.8	5.3	3.24	28.09	-9.54
63	52	12.2	8.3	148.84	68.89	101.26
32	32	-18.8	-11.7	353.44	136.89	219.96
				1603.6	**484.1**	**673.4**

Le medie sono:

$\mu_X = 50.8$ $\mu_Y = 43.7$

$$r = \frac{673.4}{\sqrt{1603.6 * 484.1}} = 0.76$$

Il caso delle distribuzioni doppie di frequenze

Nel **caso delle distribuzioni di frequenze** abbiamo

$$r = \frac{\sum_{i=1}^{s}\sum_{j=1}^{t}(x_i - \mu_X)(y_j - \mu_Y)n_{ij}}{\sqrt{\sum_{i=1}^{s}(x_i - \mu_X)^2 n_{i0}\sum_{j=1}^{t}(y_j - \mu_Y)^2 n_{0j}}}.$$

N.B:

Quando uno o entrambi i caratteri sono divisi in intervalli, l'indice *r* si calcola prendendo i valori centrali di classe.

Il caso delle distribuzioni doppie di frequenze: calcolo di r

Distribuzione doppia di frequenze di un campione di coniugi classificati secondo l'età:

I numeri in rosso sono i valori centrali

Età marito	Età della moglie				Totale
	18-30	31-40	41-50	51-65	
	24	35.5	45.5	58	
20-30 25.0	14	0	0	0	14
31-40 35.5	5	23	0	0	28
41-50 45.5	0	5	17	1	23
51-65 58.0	0	0	9	26	35
Totale	19	28	26	27	100

μ_X	44.21
μ_Y	41.99
D_X	13984.55
D_Y	14569.49
C_{XY}	13153.46
r	0.92

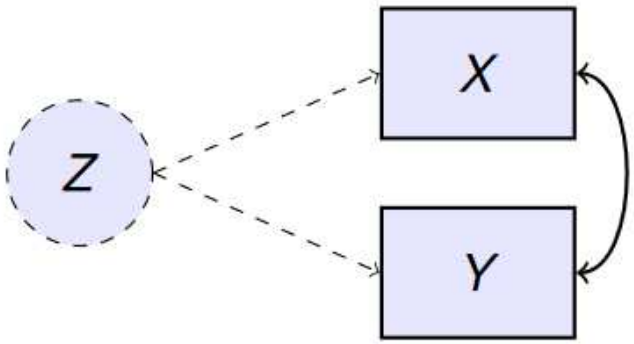
- **L'indice di correlazione è** $r = \frac{13153.46}{\sqrt{13984.55 \cdot 14569.49}} = 0.92$

Legame tra le variabili

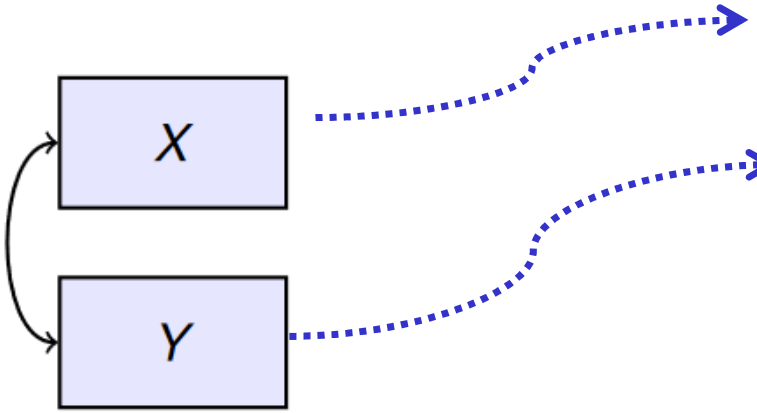
- È importante ricordare che se esiste una correlazione fra due variabili (che calcoliamo con r), questo indice non ci dà nessuna informazione sui legami di causa-effetto.

- Le due variabili "si muovono assieme". STOP!

- È possibile che esista una terza variabile che ha influenza su entrambe e che la correlazione che abbiamo calcolato sia dovuta a questa influenza.

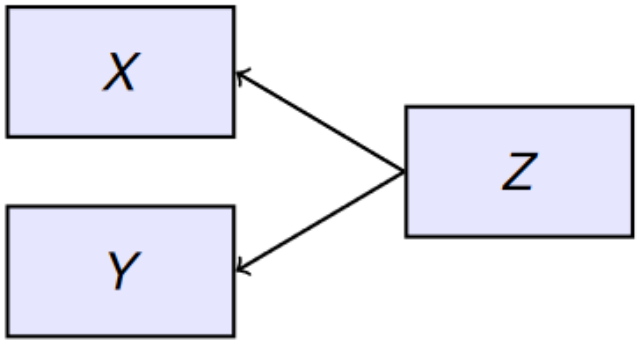


Correlazioni spurie



- X è il numero di vigili del fuoco mandato a spegnere un incendio
- Y è l'entità del danno prodotto dall'incendio

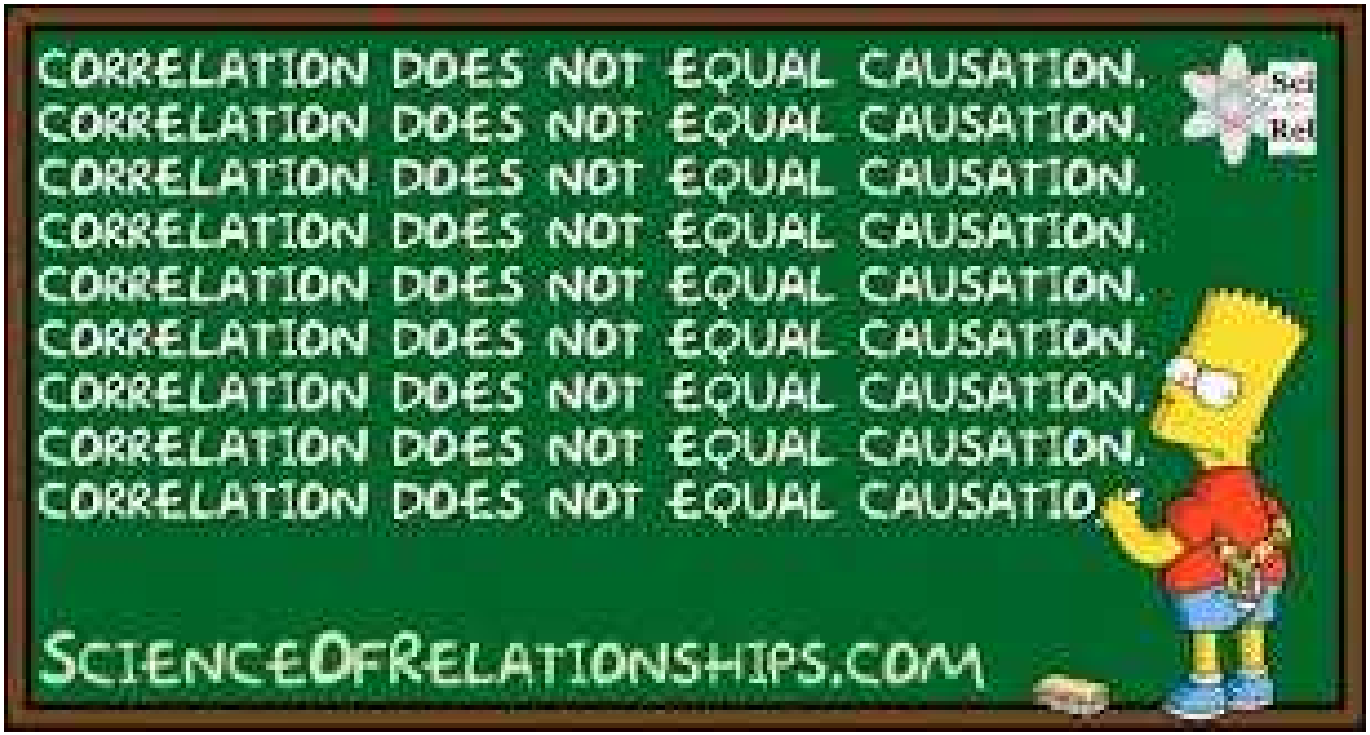
La loro correlazione vuol dire che più vigili del fuoco producono più danni?

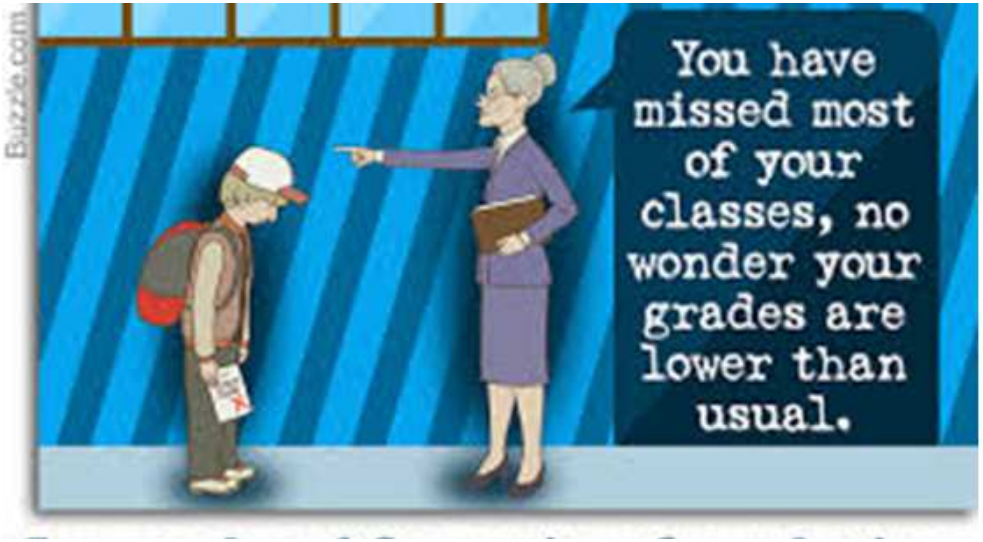


Nel momento in cui si identifica una variabile antecedente ad entrambe, la correlazione spuria acquista senso.

Z è l'ampiezza dell'incendio

Più ampio l'incendio, più vigili del fuoco vengono inviati a spegnerlo più ampio l'incendio, più danni prodotti


CORRELATION DOES NOT EQUAL CAUSATION.
CORRELATION DOES NOT EQUAL CAUSATION.
CORRELATION DOES NOT EQUAL CAUSATION.
CORRELATION DOES NOT EQUAL CAUSATION.
CORRELATION DOES NOT EQUAL CAUSATION.
CORRELATION DOES NOT EQUAL CAUSATION.
CORRELATION DOES NOT EQUAL CAUSATION.
CORRELATION DOES NOT EQUAL CAUSATION.
CORRELATION DOES NOT EQUAL CAUSATIO
SCIENCEOFRELATIONSHIPS.COM



Analisi delle distribuzioni doppie quantitative: dipendenza -regressione

Partiamo da un esempio

Partiamo da un caso concreto con due variabili (dipendente e indipendente) di tipo quantitativo.

Possiamo chiederci che relazione esiste tra il numero di ore dedicate alla preparazione di un esame e il voto conseguito all'esame stesso.

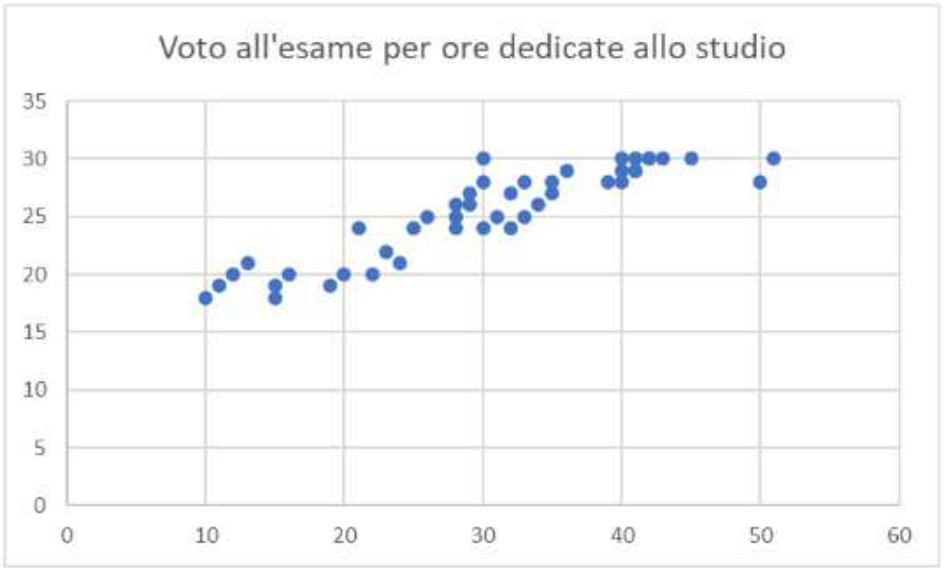
Oppure potremmo chiederci se esiste una relazione tra il voto conseguito all'esame di laurea e il reddito mensile medio.

In entrambi i casi abbiamo definito una variabile dipendente (il voto all'esame e il reddito mensile medio) e una variabile indipendente (il numero di ore dedicate alla preparazione di un esame e il voto conseguito all'esame di laurea).

Diagramma di dispersione

Il diagramma di dispersione è la rappresentazione grafica di una possibile relazione tra due variabili.

Sull'asse X troviamo la variabile indipendente e sull'asse Y la variabile dipendente. L'insieme dei punti che si crea indica come covariano (variano insieme) le due variabili.



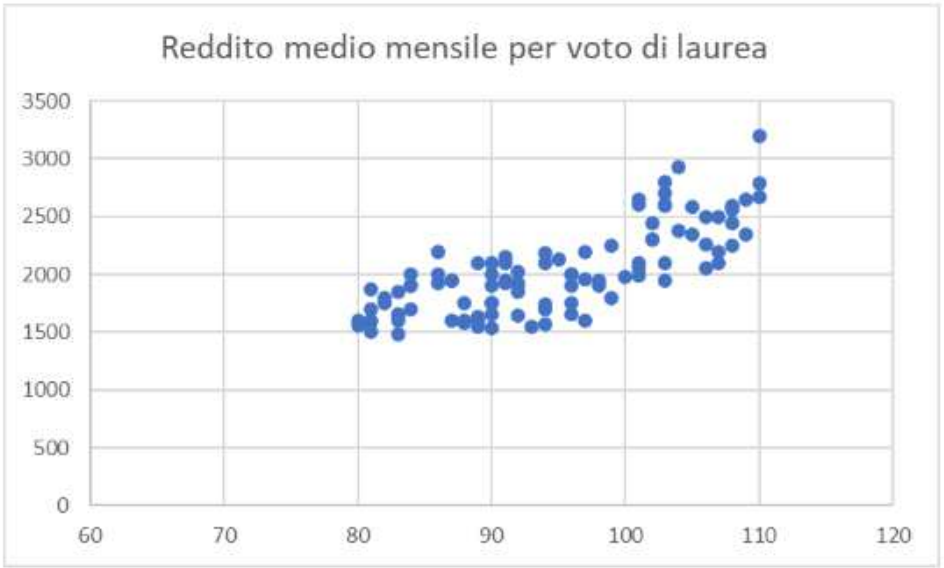


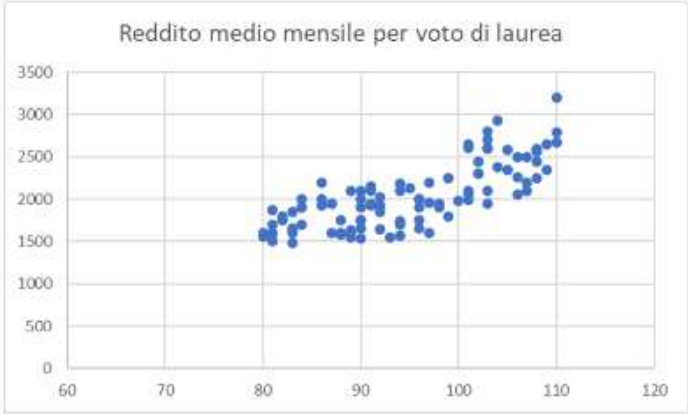
Diagramma di dispersione

L'osservazione del diagramma di dispersione ci consente di trarre **alcune conclusioni:**

o Le due variabili covariano in modo sistematico, ovvero sono legate da una relazione,

o La distribuzione della «nuvola di punti» dalla sinistra in basso alla destra in alto ci indica che la relazione è positiva, ovvero al crescere della variabile indipendente cresce anche la variabile dipendente.

o La disposizione dei punti ci suggerisce che siamo in presenza di una relazione lineare, ovvero la variabile Y tende a variare sempre nella stessa direzione e nella stessa misura al variare di X.



Non ci aiuta a misurare l'effetto causale, ovvero a quantificare la variazione della variabile dipendente al variare di quella indipendente.

Equazione lineare

Se vogliamo quantificare l'intensità della relazione tra le due variabili, la dobbiamo esprimere attraverso un'equazione matematica. Ogni equazione è definita dalla sua forma funzionale e dai valori che assumono i suoi parametri.

Parlando di forma funzionale prenderemo in considerazione solo quella lineare, che possiamo esprime semplicemente così:

$$Y = \alpha + \beta X$$

Il valore di Y è dato dal parametro α, che è costante, più X moltiplicato per il parametro β.

Equazione lineare

Facciamo un esempio concreto:

Ho i due parametri α e β
α = 3
β = 2

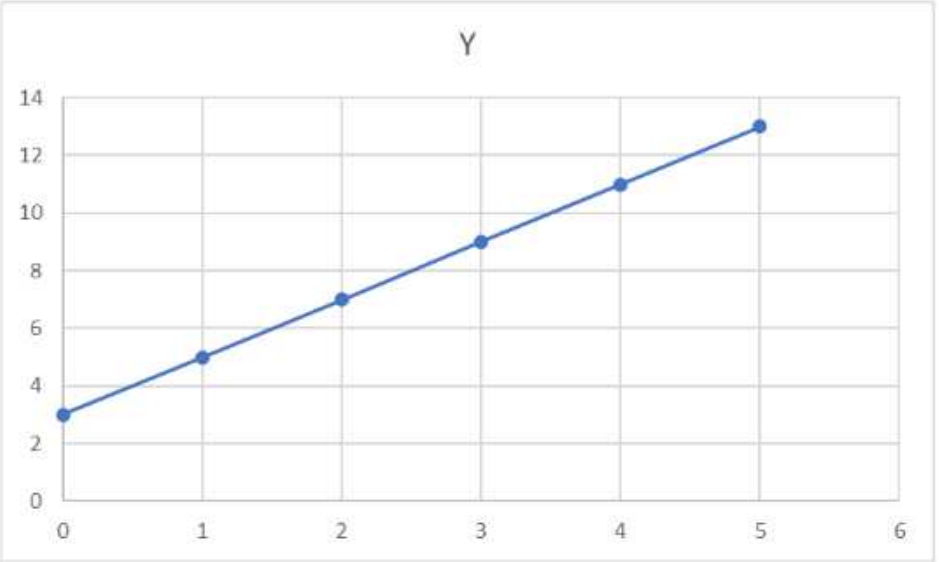
La nostra equazione lineare $Y = \alpha + \beta X$

diventa $Y = 3 + 2X$

X	Y
0	3
1	5
2	7
3	9
4	11
5	13

Equazione lineare

X	Y
0	3
1	5
2	7
3	9
4	11
5	13



Equazione lineare

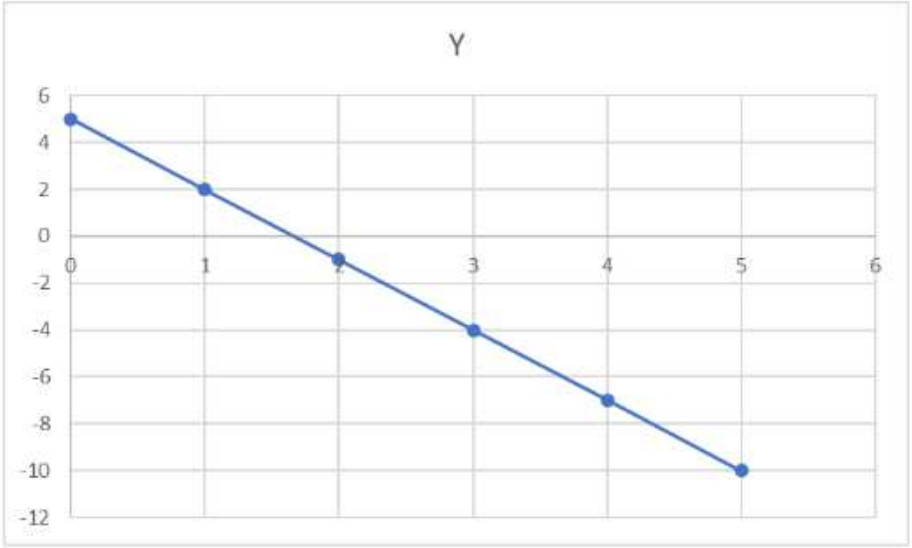
Facciamo un esempio concreto
$\alpha = 5$
$\beta = -3$

La nostra equazione lineare $Y = \alpha + \beta X$ diventa $Y = 5 - 3X$

X	Y
0	5
1	2
2	-1
3	-4
4	-7
5	-10

Equazione lineare

X	Y
0	5
1	2
2	-1
3	-4
4	-7
5	-10



Equazione lineare

Abbiamo espresso la relazione tra X e Y con una linea retta.

Come influiscono i due parametri α e β?

α stabilisce la distanza dall'asse orizzontale, ovvero il valore di Y in corrispondenza dello 0 della X. Questo parametro viene definito anche come intercetta o costante.

β determina l'inclinazione o coefficiente angolare e ci dice di quanto varia la Y al variare di X. Questo valore ci spiega l'intensità dell'effetto della variabile indipendente sulla variabile dipendente.

Se β è positivo ci troviamo di fronte a una relazione diretta, mentre se è negativo la relazione è inversa.

Equazione lineare

Abbiamo espresso la relazione tra X e Y con una linea retta $Y = \alpha + \beta X$.

Tuttavia non è possibile rappresentare esattamente con un'equazione lineare una relazione complessa, come, nel nostro esempio, quella tra reddito e voto di laurea.

La relazione tra due variabili non può essere rappresentata esattamente da un'equazione lineare, tuttavia non possiamo negare che la nuvola di punti ci suggerisca una tendenza precisa, ovvero al crescere delle ore di studio aumenta il voto all'esame, oppure al crescere del voto di laurea corrisponda un reddito medio più elevato.

Un'equazione lineare ci aiuta a stimare i due parametri α e β, e ad approssimare la covariazione tra le due variabili.

Modello di regressione lineare semplice

La formula che approssima la relazione tra X e Y con una linea retta è la seguente:

$$\hat{Y}_i = \alpha + \beta X_i$$

$\hat{Y}_i$ rappresenta il valore atteso sulla base dei parametri stimati α e β e non quello osservato.

Se vogliamo esprimere i valori osservati di Y dobbiamo aggiungere all'equazione lineare un ulteriore elemento ε_i, che rappresenta la componente erratica (gli errori di predizione).

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

ε_i rappresenta la differenza tra il valore osservato e il valore atteso, derivante dal modello di regressione lineare, ovvero:

$$\varepsilon_i = Y_i - \hat{Y}_i$$

se sostituiamo $\hat{Y}_i$ con $\alpha + \beta X_i$ (visto che $\hat{Y}_i = \alpha + \beta X_i$) otteniamo che:

$$\varepsilon_i = Y_i - \alpha - \beta X_i$$

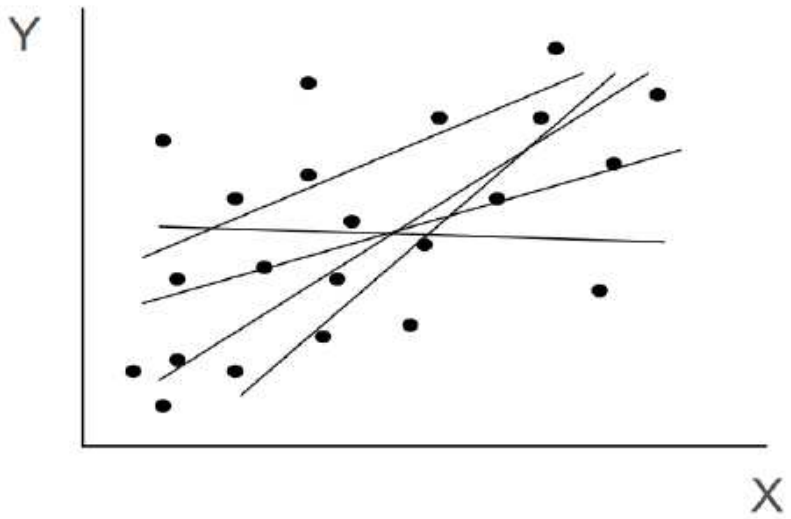
I residui

Gli errori di predizione vengono chiamati **residui**, dal momento che corrispondono a quella parte di Y che non viene spiegata dall'effetto lineare di X.

ε_i rappresenta:

- l'influenza su Y di tutti i fattori casuali che non sono stati introdotti nel modello di regressione lineare utilizzato.
- Il fatto che la relazione tra X e Y non è detto sia perfettamente lineare.
- Nelle scienze sociali i comportamenti umani sono di consueto caratterizzati da una componente di casualità che nessun modello di regressione, anche il più sofisticato, sarebbe in grado di stimare con precisione il valore di Y.

Come scegliere la retta migliore?



Per un insieme di punti possono passare infinite rette!

Come scegliere la retta "migliore"?

Metodo dei Minimi Quadrati

Scelta della retta migliore

Abbiamo chiarito che l'obiettivo della regressione lineare è stimare i valori dei parametri α e β che consentono di approssimare nel modo migliore la covariazione tra X e Y.

Questo significa che la retta di regressione migliore è quella che **minimizza i valori osservati di Y e quelli predetti attraverso il modello.**

Dal momento che la differenza tra valori attesi e valori osservati:

$\varepsilon_i = Y_i - \hat{Y}_i$

La migliore retta di regressione è quella che minimizza gli errori di predizione.

Scelta della retta migliore

Se proviamo a sommare gli errori da una determinata retta scopriremo che si annullano, ovvero

$$\sum_{i=1}^{N} \varepsilon_i = 0$$

Possiamo elevare al quadrato gli scarti e la migliore retta di regressione è quella che minimizza il quadrato della somma dei residui.

Questo significa che rende minima la seguente quantità:

$$\sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{N} \varepsilon_i^2$$

I valori dei parametri α e β che soddisfano questo criterio sono detti **stime dei minimi quadrati**.

Ipotesi da soddisfare per stimare il modello di regressione

1) media attesa degli errori è nulla (scarti positivi e negativi si compensano)

2) la varianza degli errori è costante al variare delle osservazioni (omoschedasticità)

3) la covarianza degli errori è nulla, gli errori si assumono incorrelati e quindi indipendenti fra loro

4) la variabile indipendente (x) non è affetta da errori di misura

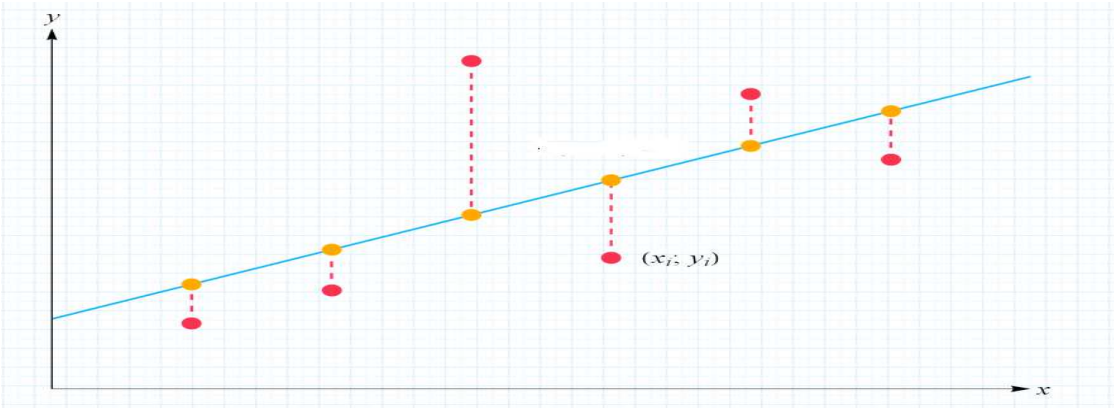
Metodo dei minimi quadrati

Il criterio detto dei minimi quadrati prevede di valutare la bontà del modello sulla base della somma dei quadrati di tutti errori di stima commessi:

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

La retta migliore, secondo questo criterio, è quella che minimizza la somma dei quadrati degli scarti dei valori stimati da quelli osservati, detti anche residui della regressione.

- Perché proprio il quadrato dei residui ?
  - per evitare che residui positivi e negativi si compensino
  - il valore assoluto è matematicamente più scomodo da gestire e non sempre porta ad una soluzione univoca
  - il quadrato dà peso maggiore agli scarti più grandi, che sono anche quelli che ci disturbano di più: è meglio fare tanti piccoli errori che non un errore molto grosso



I coefficienti di regressione

Tralasciamo la dimostrazione e vediamo qual è la formula che ci consente di stimare i due parametri.

I parametri del modello vengono chiamati anche coefficienti di regressione:

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

$$\hat{b} = \frac{Cov(x, y)}{\sigma^2(x)}$$

Intepretazione dei coefficienti di regressione

Il segno del coefficiente di regressione dipende da quello della covarianza

$$\hat{b} = \frac{\sigma_{xy}}{\sigma_x^2}$$

e indica quindi se la relazione è diretta o inversa

- Ricordiamo che, nell'equazione della retta $y = \hat{a} + \hat{b}\,x$
  b rappresenta il coefficiente angolare, cioè l'inclinazione della retta

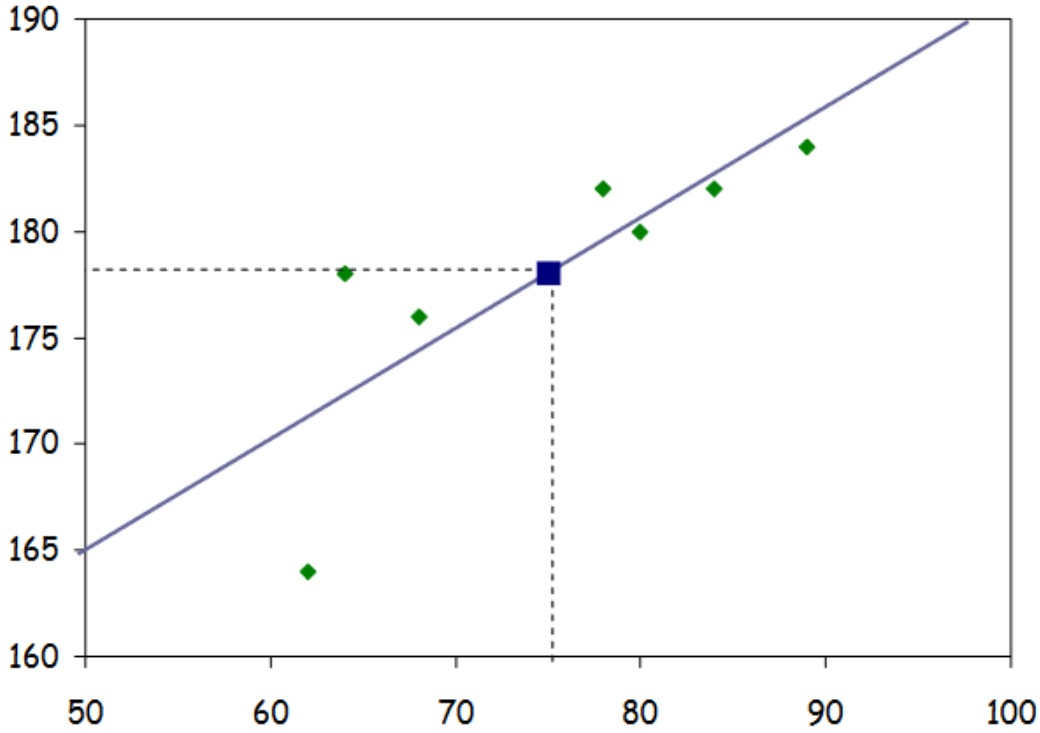
b > 0	relazione diretta, cioè y cresce al crescere di x	la retta è crescente, inclinazione positiva
b = 0	assenza di relazione (lineare), y non varia al variare di x	la retta è orizzontale, inclinazione nulla
b < 0	relazione inversa, cioè y diminuisce all'aumentare di x	la retta è decrescente, inclinazione negativa

- Il valore assoluto di b indica di quanto varia la Y al variare di una unità della X
- Il coefficiente a rappresenta l'intercetta della retta con l'asse Y:
  indica quanto vale Y quando X vale 0; quando a = 0, la retta passa per l'origine degli assi cartesiani, cioè per il punto di coordinate (0, 0)

Esercizio

Determiniamo la retta di regressione ai minimi quadrati per la relazione tra le variabili Y=altezza e X=peso:

i	x(i)	y(i)
1	62	164
2	64	178
3	68	176
4	75	178
5	78	182
6	80	180
7	84	182
8	89	184



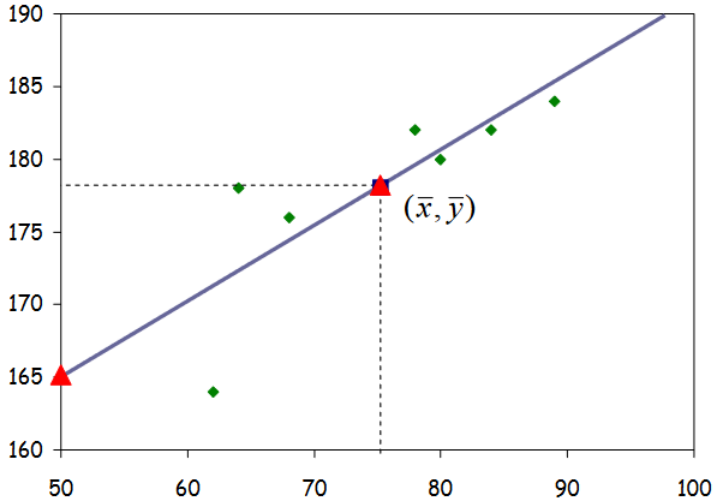
Calcolo dei coefficienti di regressione

i	x(i)	y(i)	x(i)-Mx	y(i)-My	[x(i)-Mx]²	[y(i)-My]²	(x(i)-Mx)(y(i)-My)
1	62	164	-13,00	-14,00	169,00	196,00	182,00
2	64	178	-11,00	0,00	121,00	0,00	0,00
3	68	176	-7,00	-2,00	49,00	4,00	14,00
4	75	178	0,00	0,00	0,00	0,00	0,00
5	78	182	3,00	4,00	9,00	16,00	12,00
6	80	180	5,00	2,00	25,00	4,00	10,00
7	84	182	9,00	4,00	81,00	16,00	36,00
8	89	184	14,00	6,00	196,00	36,00	84,00
Totale	600	1424	0	0	650,00	272,00	338,00
Media	75,00	178,00			81,25	34,00	42,25

$$\sigma_x^2 = 81{,}25 \quad \sigma_y^2 = 34 \quad \sigma_{xy} = 42{,}25$$

$$b = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{42{,}25}{81{,}25} = 0{,}52 \qquad a = \bar{y} - b\,\bar{x} = 178 - 0{,}52 \cdot 75 = 139$$

Equazione della retta di regressione



$$\hat{b} = \frac{\sigma_{x\,y}}{\sigma_x^2} = 0{,}52$$

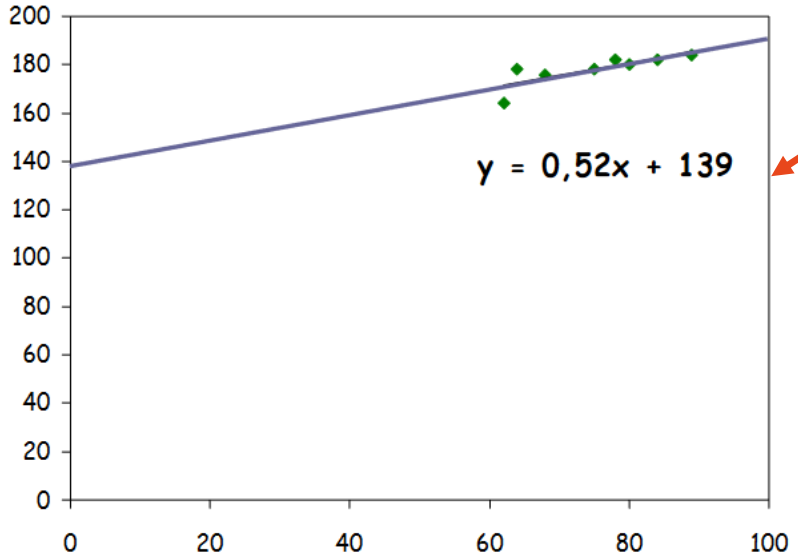
$$\hat{a} = \bar{y} - \hat{b}\,\bar{x} = 139$$

$$\hat{y} = 139 + 0{,}52\,x$$

Per disegnare la retta di regressione, è sufficiente determinarne due punti a scelta: uno in realtà lo conosciamo già: è il punto medio (baricentro) del sistema per l'altro scegliamo ad es. x = 50, da cui y = 139 + 0,52 * 50 = 165

Equazione della retta di regressione

Disegniamo il grafico e la retta di regressione a partire dall'origine degli assi



La retta interseca l'asse delle ordinate proprio nel punto 139

Adattamento della retta di regressione ai dati

L'analisi di regressione include la verifica dell'idoneità del modello a rappresentare la relazione statistica tra le variabili *Y* e *X*.

A questo fine, viene introdotto un apposito indice che misura la bontà dell'adattamento della retta di regressione ai punti osservati, per la cui costruzione ci si avvale della scomposizione della devianza.

$$D_Y = \sum_{i=1}^{N}(y_i - \bar{y})^2 = \sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

$$SQT = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

Somma totale dei quadrati

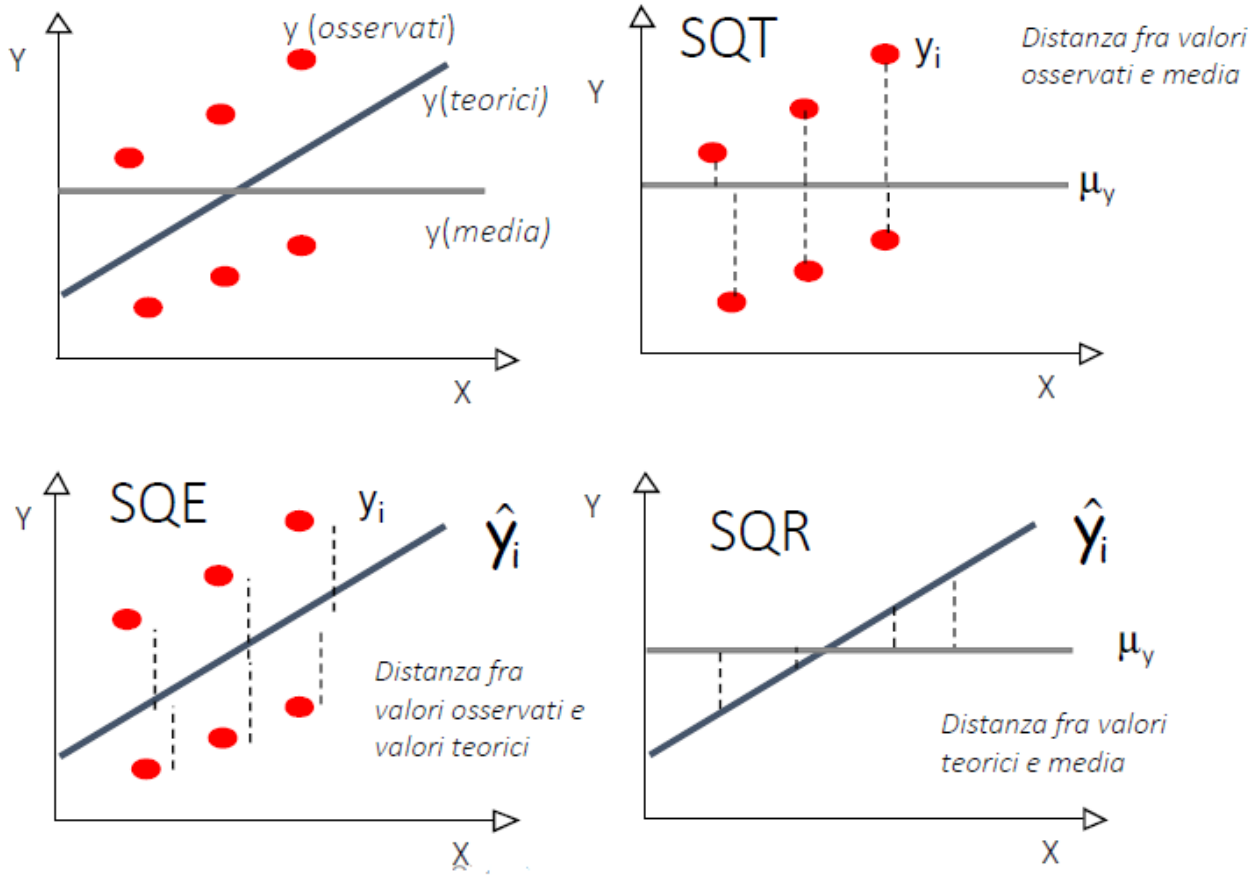
$$SQR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

Somma dei quadrati della regressione

$$SQE = \sum_{i=1}^{n}\hat{e}_i^2$$

Somma dei quadrati degli errori

Scomposizione della devianza nel modello di regressione: interpretazione grafica



Coefficiente di determinazione

Dalla relazione SQT=SQR+SQE si può definire un indice che misura la bontà di adattamento della ret di regressione.

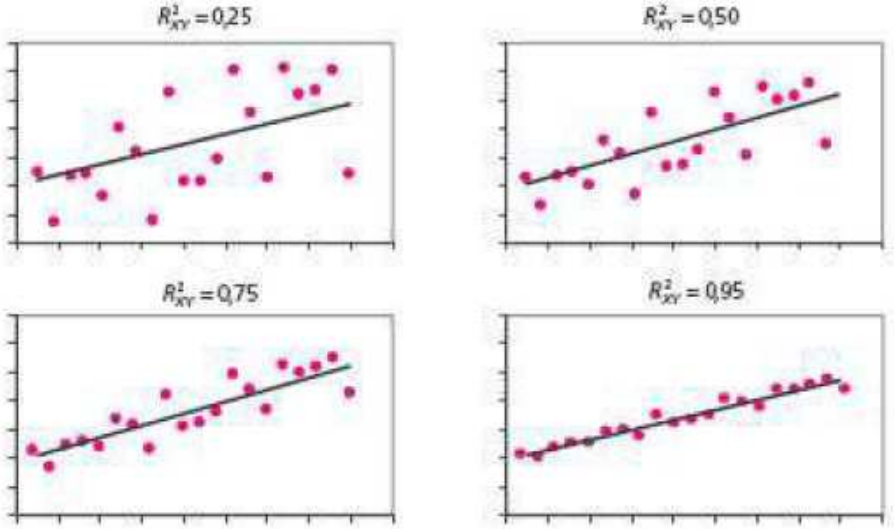
Il rapporto

$$R^2_{XY} = \frac{SQR}{SQT} = 1 - \frac{SQE}{SQT}$$

è detto coefficiente di determinazione e indica la proporzione di variabilità di Y spiegata dalla variabi esplicativa X, attraverso il modello di regressione.

Si può dimostrare che il coefficiente di determinazione corrisponde al quadrato del coefficiente di correlazione lineare:

$$R^2_{XY} = \rho^2_{XY}$$



Proprietà del coefficiente di determinazione

$$R^2_{XY} = \frac{SQR}{SQT} = 1 - \frac{SQE}{SQT}$$

- Assume valori nell'intervallo [0, 1].
- Raggiunge il minimo se e solo se *SQR* = 0, cioè se e solo se la retta di regressione è parallela all'asse delle ascisse.
- Raggiunge il massimo se e solo se *SQE* = 0, circostanza che si verifica se e solo se i punti osservati giacciono su una retta.
- Rappresenta la frazione della variabilità totale di Y spiegata dalla retta di regressione.

Esempio di relazione statistica

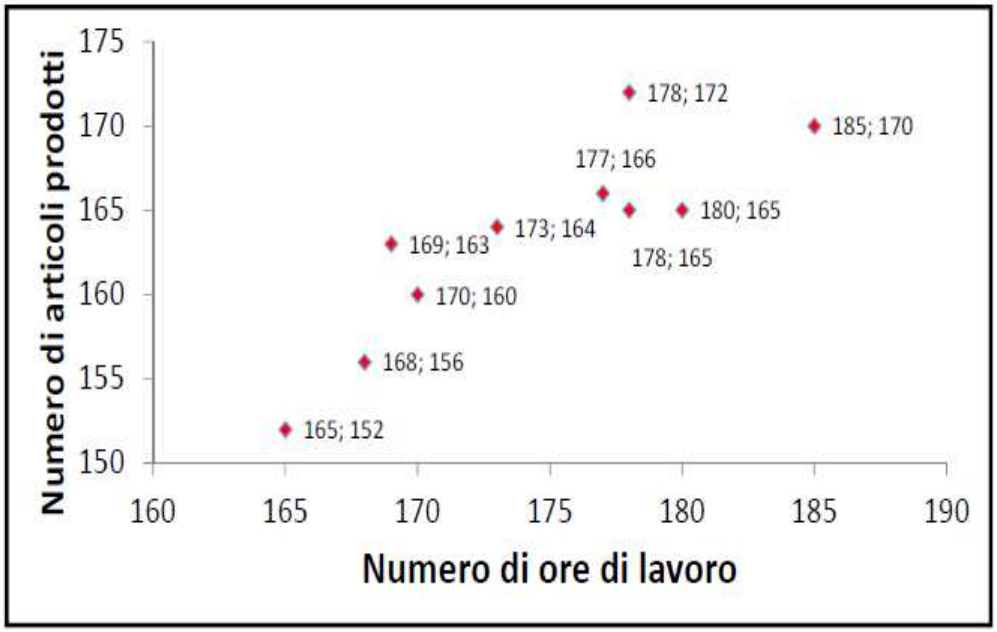
Nella tabella sono riportati il numero di ore lavorate ed il numero di articoli prodotti da 10 artigiani nel corso di un mese.

Vogliamo stabilire se la retta è una funzione adatta a esprimere il legame associativo tra il numero di articoli prodotti ed il numero di ore di lavoro.

Numero di ore di lavoro	Numero di articoli prodotti
173	164
178	172
169	163
170	160
177	166
178	165
180	165
185	170
165	152
168	156

Grafico di dispersione

Numero di ore di lavoro	Numero di articoli prodotti
173	164
178	172
169	163
170	160
177	166
178	165
180	165
185	170
165	152
168	156



L'andamento dei punti suggerisce che la relazione statistica che lega il numero di prodotti al numero di ore lavorate può essere espressa da una retta.

Parametri della retta di regressione

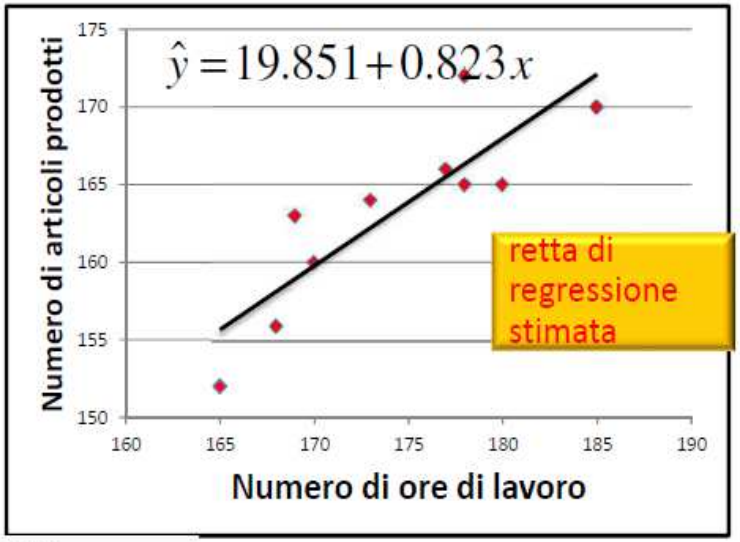
$$b_1 = \frac{\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_Y)}{\sum_{i=1}^{N}(x_i - \mu_X)^2}$$

$$b_0 = \mu_Y - b_1\mu_X$$

$(x_i - \mu_X).(y_i - \mu_Y)$

μ_X	= 174.30
μ_Y	= 163.30

x_i	y_i	$x_i - \mu_X$	$y_i - \mu_Y$	$(x_i - \mu_X)^2$	
173	164	-1.30	0.70	1.69	-0.91
178	172	3.70	8.70	13.69	32.19
169	163	-5.30	-0.30	28.09	1.59
170	160	-4.30	-3.30	18.49	14.19
177	166	2.70	2.70	7.29	7.29
178	165	3.70	1.70	13.69	6.29
180	165	5.70	1.70	32.49	9.69
185	170	10.70	6.70	114.49	71.69
165	152	-9.30	-11.30	86.49	105.09
168	156	-6.30	-7.30	39.69	45.99
			Totale	356.10	293.10



$$b_1 = \frac{C_{XY}}{D_X} = \frac{293.10}{356.10} = 0.823$$

$$b_0 = \mu_y - b_1\mu_x = 163.30 - 0.823 \cdot 174.30 = 19.851$$

Indice di determinazione calcolo

L'indice di determinazione con le tre formule:

$$R^2 = \frac{SQR}{SQT} = \frac{241.20}{326.10} = 0.74$$

$$R^2 = 1 - \frac{SQE}{SQT} = 1 - \frac{84.85}{326.10} = 0.74$$

$$R^2 = r^2 = \frac{C_{XY}^2}{D_X D_y} = \frac{293.10^2}{356.10 * 326.10} = 0.74$$

$$R^2 = \frac{SQR}{SQT} = \frac{\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

$$R^2 = 1 - \frac{SQE}{SQT} = \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

x_i	y_i	$\hat{y}_i$	$(y_i - \mu_Y)^2$	$(\hat{y}_i - \mu_Y)^2$	$(y_i - \hat{y}_i)^2$
173	164	162.23	0.49	1.145	3.13
178	172	166.35	75.69	9.27	31.98
169	163	158.94	0.09	19.03	16.50
170	160	159.76	10.89	12.52	0.06
177	166	165.52	7.29	4.94	0.23
178	165	166.35	2.89	9.27	1.81
180	165	167.99	2.89	22.01	8.95
185	170	172.11	44.89	77.55	4.44
165	152	155.65	127.69	58.58	13.29
168	156	158.12	53.29	26.88	4.47
		Totale	**326.10**	**241.20**	**84.85**
			SQT	*SQR*	*SQE*