

Statistica e valutazione della Formazione

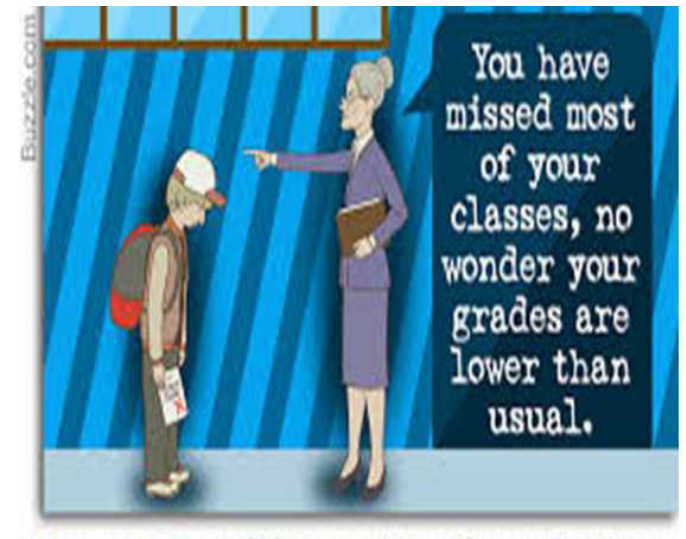
Slides 7

A.A. 2025-2026

Docente: ANNA LINA SARRA

Modulo 1: elementi di statistica descrittiva

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

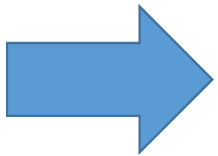


Partiamo da un esempio

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.

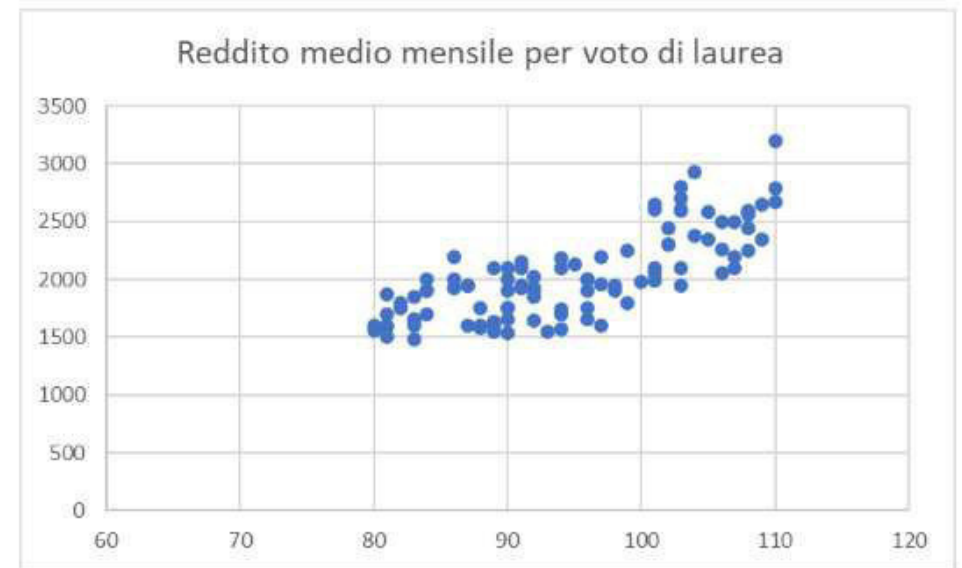
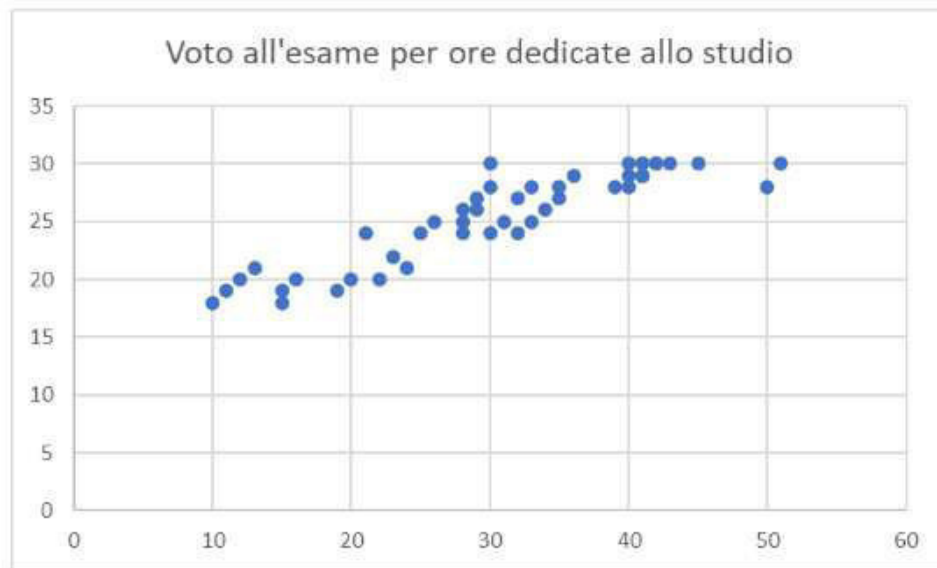
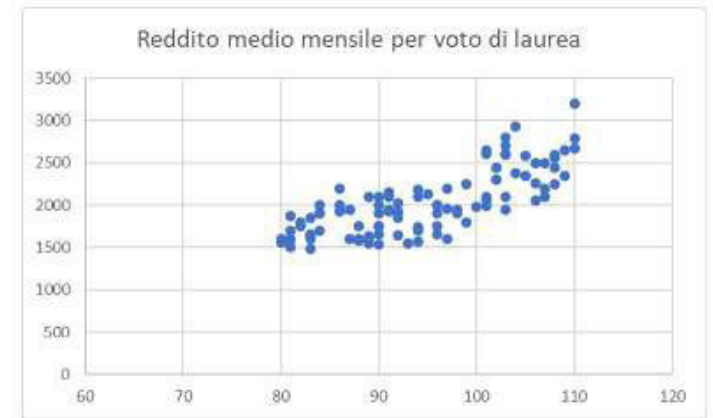


Diagramma di dispersione

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 esprimere 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

$$\alpha = 3$$

$$\beta = 2$$

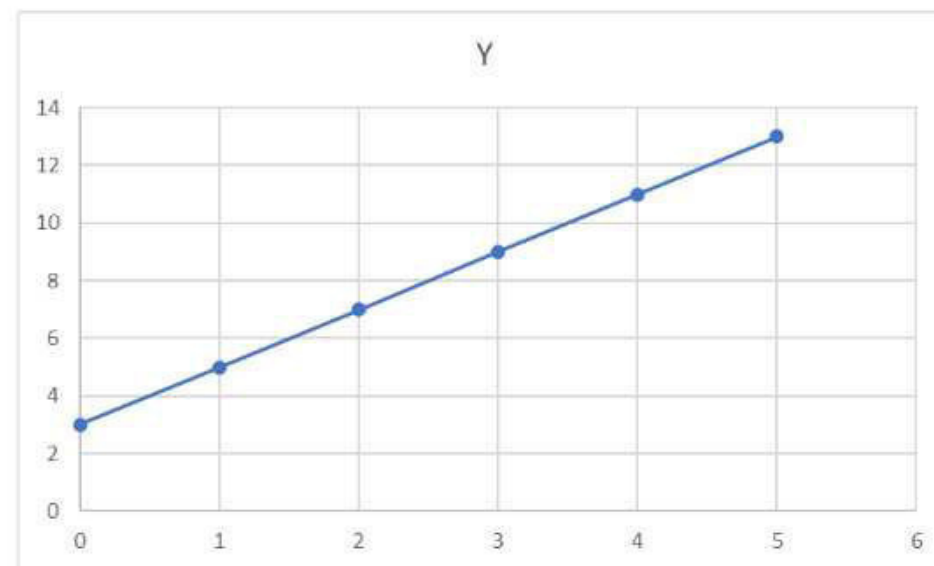
La nostra equazione lineare $Y = \alpha + \beta X$

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

$$\alpha = 5$$

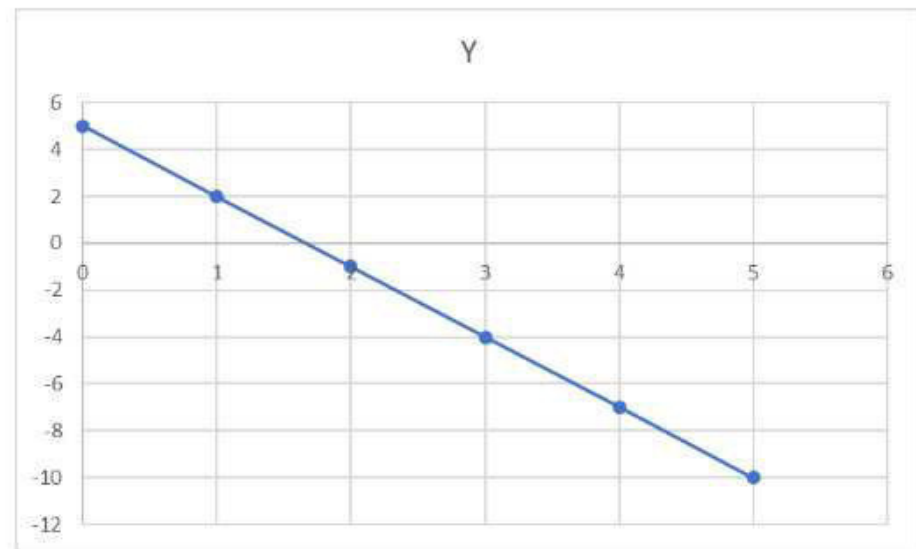
$$\beta = -3$$

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 covarianza tra le due variabili.

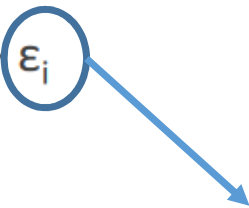
Modello di regressione lineare semplice

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).

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


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:

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

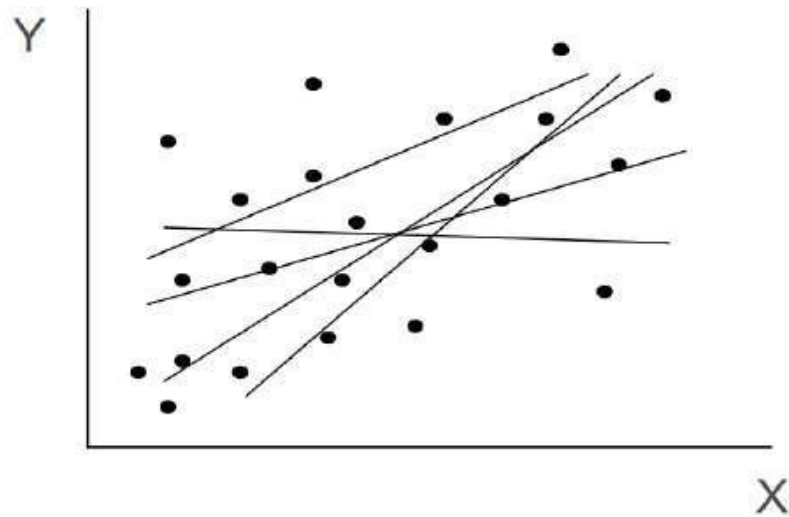
I residui

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 covarianza 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

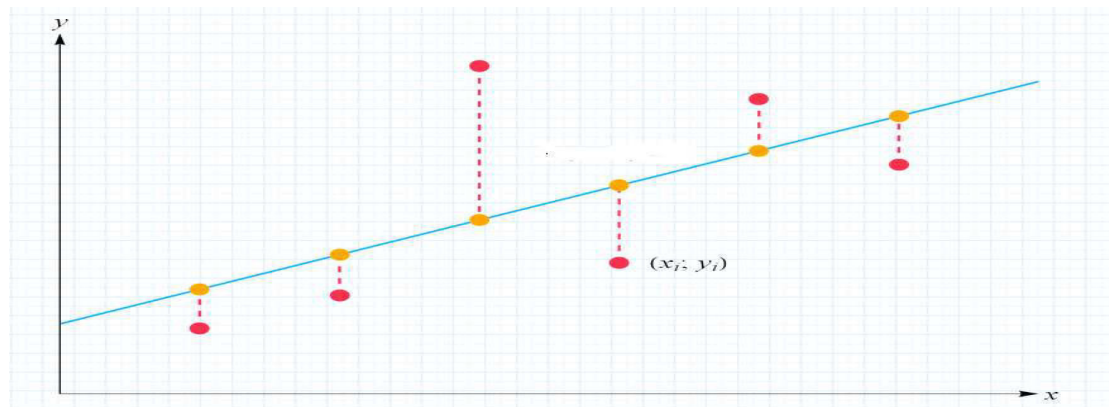
Metodo dei minimi quadrati

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 $(0, 0)$

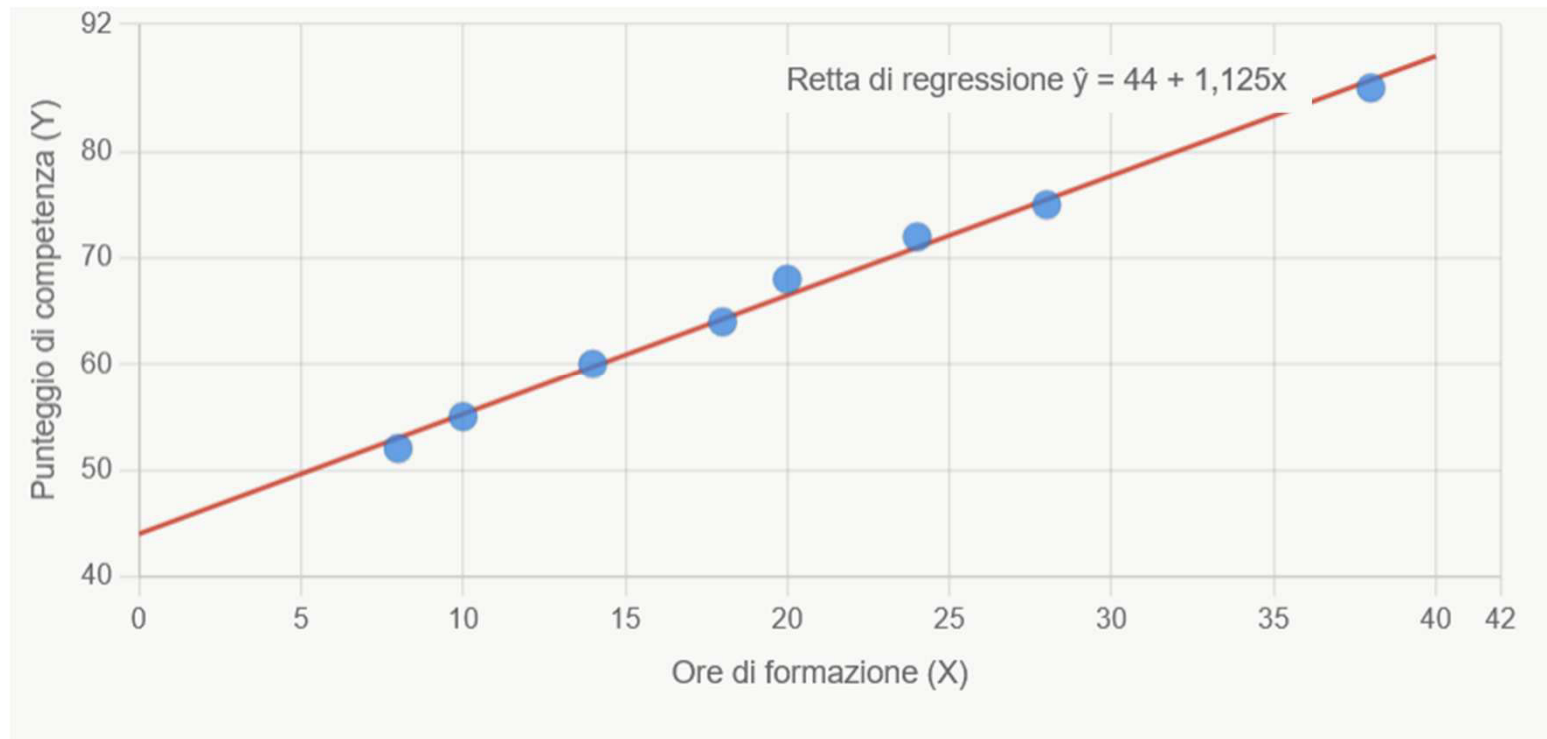
Esercizio

Un responsabile HR vuole verificare se esiste una relazione lineare tra le **ore di formazione erogate** (X) e il **punteggio di competenza** rilevato a fine corso (Y, scala 0–100) per 8 dipendenti.

i	x(i) — Ore di formazione	y(i) — Punteggio di competenza
1	8	52
2	10	55
3	14	60
4	18	64
5	20	68
6	24	72
7	28	75
8	38	86

Calcolo coefficienti di regressione — Metodo dei Minimi Quadrati							
Variabili: X = ore di formazione Y = punteggio di competenza							
i	$x(i)$ Ore formaz.	$y(i)$ Punteggio	$x(i) - Mx$	$y(i) - My$	$[x(i)-Mx]^2$	$[y(i)-My]^2$	$(x(i)-Mx) \cdot (y(i)-My)$
1	8	52	-12.00	-14.25	144.00	203.06	171.00
2	10	55	-10.00	-11.25	100.00	126.56	112.50
3	14	60	-6.00	-6.25	36.00	39.06	37.50
4	18	64	-2.00	-2.25	4.00	5.06	4.50
5	20	68	0.00	1.75	0.00	3.06	0.00
6	24	72	4.00	5.75	16.00	33.06	23.00
7	28	75	8.00	8.75	64.00	76.56	70.00
8	38	86	18.00	19.75	324.00	390.06	355.50
Totale	160	532	0	0	688.00	876.50	774.00
Media	20.00	66.50			86.00	109.56	96.75
Calcolo dei parametri							
σ^2_x (varianza di X)				86.0000			
σ^2_y (varianza di Y)				109.5625			
σ_{xy} (covarianza XY)				96.7500			
$\hat{b} = \sigma_{xy} / \sigma^2_x$				1.1250			
$\hat{a} = \bar{y} - \hat{b} \cdot \bar{x}$				44.0000			
Equazione retta: $\hat{y} = \hat{a} + \hat{b} \cdot x$				$\hat{y} = 44.00 + 1.125 \cdot x$			

RETTA DI REGRESSIONE



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

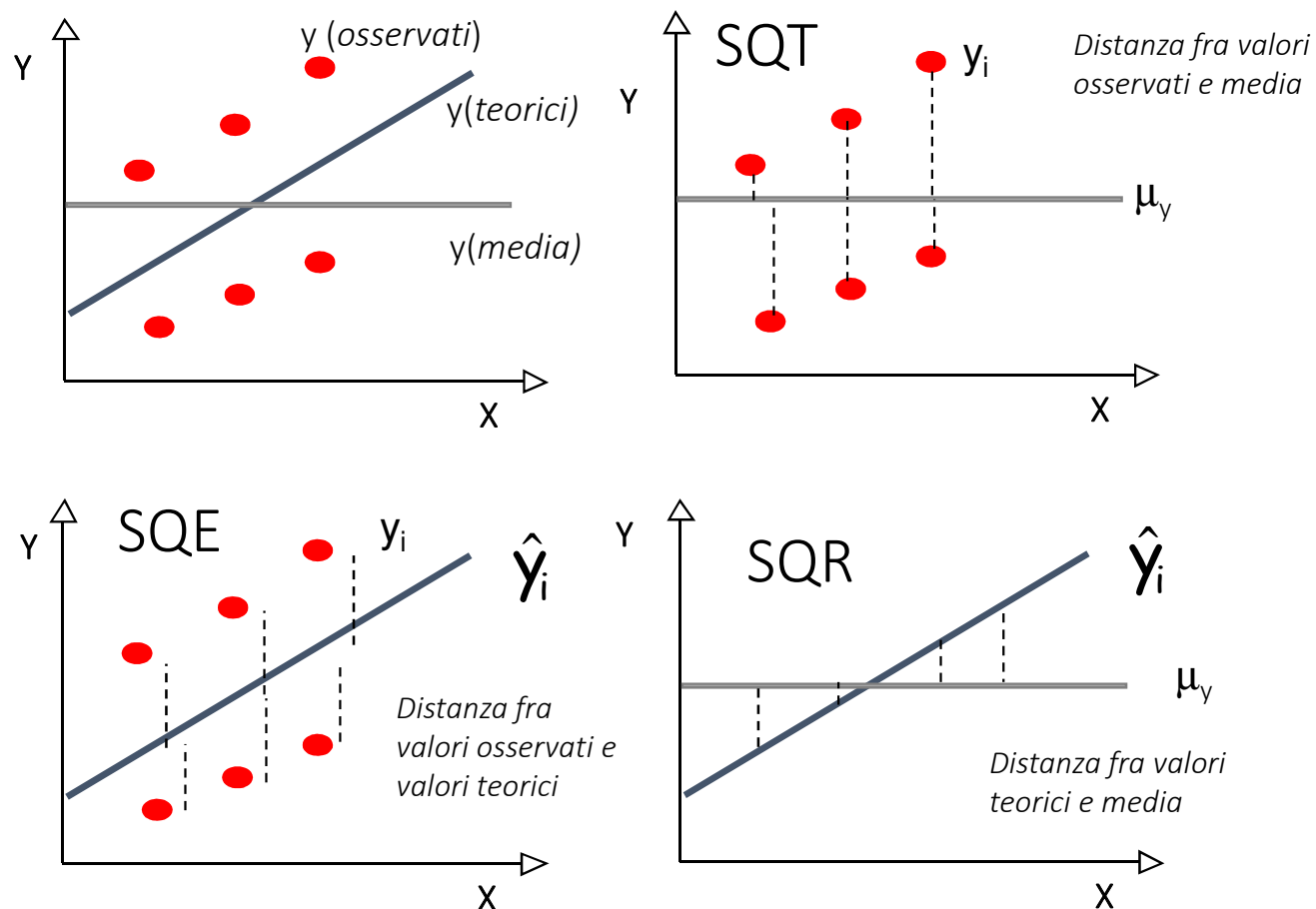
$$SQR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

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 retta di regressione.

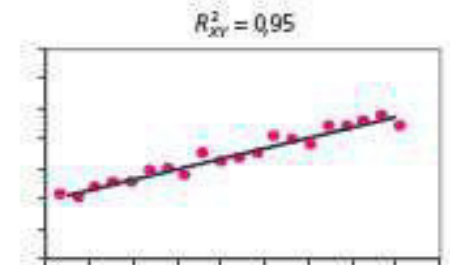
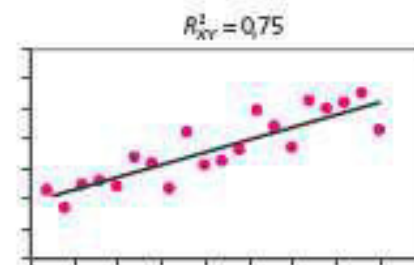
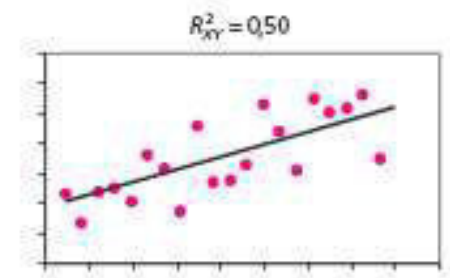
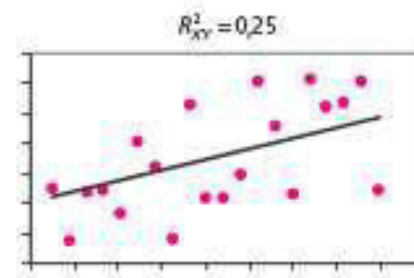
Il rapporto

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

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

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

$$R_{XY}^2 = \rho_{XY}^2$$



Proprietà del coefficiente di determinazione

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

- 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**.

Indice di determinazione calcolo

L'indice di determinazione con le tre formule:

$$R^2 = \frac{SQR}{SQT} = \frac{870.74}{876.0} = 0.99$$

$$R^2 = 1 - \frac{SQE}{SQT} = 1 - \frac{5.24}{876.0} = 0.99$$

$$R^2 = r^2 = \frac{C^2_{XY}}{DX \cdot DY} = \frac{774^2}{688 \times 876} = 0.99$$

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

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

x_i	y_i	$\hat{y}_i$	$(y_i - \mu_y)^2$	$(\hat{y}_i - \mu_y)^2$	$(y_i - \hat{y}_i)^2$
8	52	53.00	210.25	182.25	1.00
10	55	55.25	132.25	126.56	0.06
14	60	59.75	42.25	45.56	0.06
18	64	64.25	6.25	5.06	0.06
20	68	66.50	2.25	0.00	2.25
24	72	71.00	30.25	20.25	1.00
28	75	75.50	72.25	81.00	0.25
38	86	86.75	380.25	410.06	0.56
		Totale	876.00	870.74	5.24
			SQT	SQR	SQE
$R^2 = SQR / SQT = 870.74 / 876.0 = 0.994$					

Domanda (1)

La retta di regressione lineare semplice è:

- a) Una retta che passa esattamente per tutti i punti del diagramma di dispersione
- b) La retta che minimizza la somma dei quadrati dei residui (metodo dei minimi quadrati)
- c) La retta che massimizza la covarianza tra X e Y
- d) La retta che passa per l'origine degli assi cartesiani

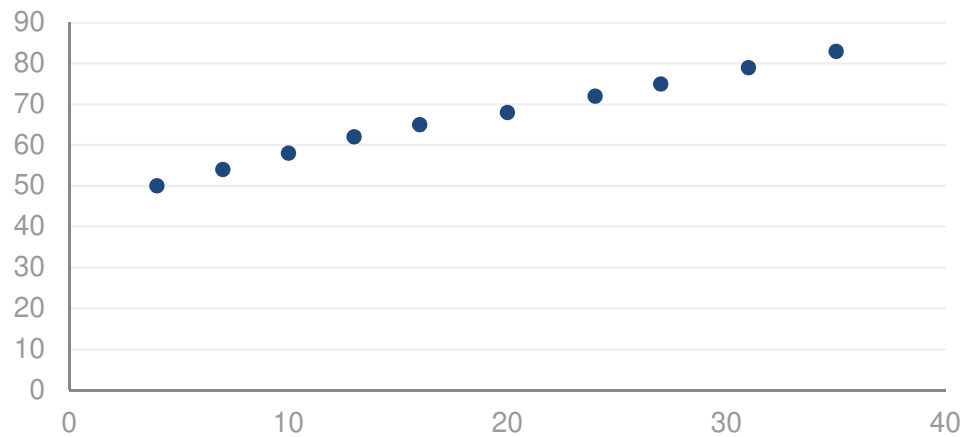
Domanda (2)

Esaminare i seguenti diagrammi di dispersione.

Per ciascun grafico indicare:

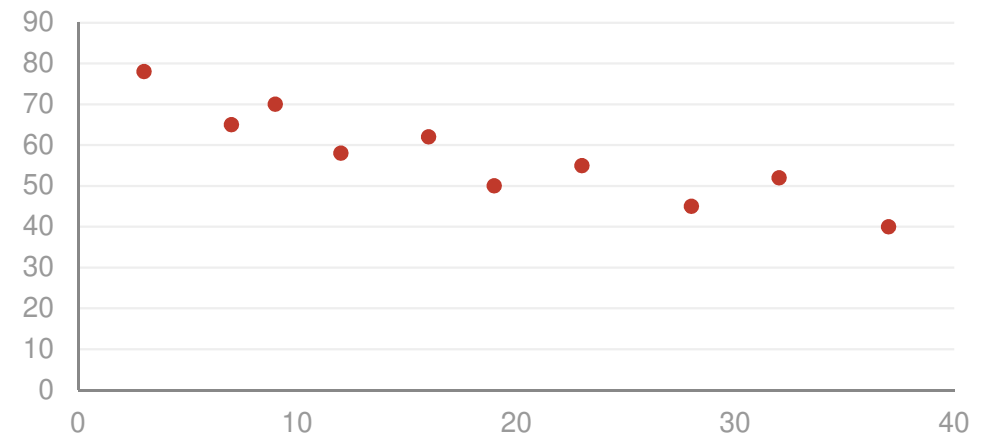
- il segno di $\hat{b}$ (positivo / negativo / zero)
- se la relazione appare forte o debole
- il valore approssimativo di R^2

Grafico A



Cosa notiamo?

Grafico B



Domanda (3)

Se il coefficiente angolare $\hat{b}$ della retta di regressione è uguale a zero:

- a) La variabile X non ha alcun effetto lineare su Y
- b) La retta di regressione è parallela all'asse delle ascisse (X)
- c) R^2 è uguale a 1
- d) La covarianza tra X e Y è positiva

Domanda (4)

Se il coefficiente di determinazione R^2 è uguale a 1:

- Tutti i punti del diagramma di dispersione giacciono esattamente sulla retta di regressione
- La somma dei quadrati di regression SQR è uguale a zero
- La retta di regressione spiega il 100% della variabilità di Y
- Il coefficiente di correlazione r è uguale a 100

Quali affermazioni sono vere?

Domanda (5)

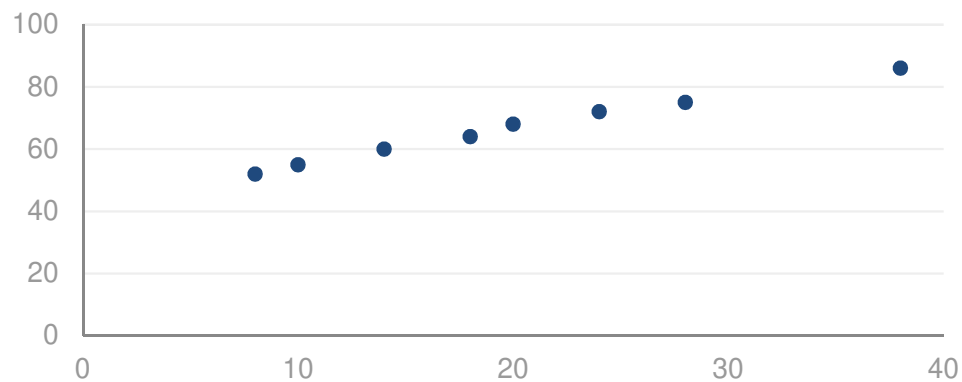
Dati raccolti da n = 8 dipendenti di un'azienda (anno 2025):

- Ore di formazione (X)
- Punteggio di competenza (Y, scala 0–100)

Valutare la relazione tra le seguenti coppie:

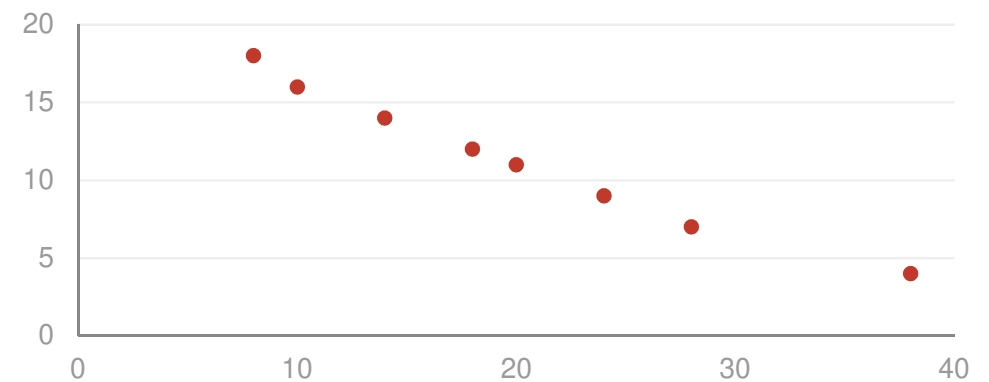
- Ore di formazione — Punteggio di competenza
- Ore di formazione — Ore di assenza dal lavoro (variabile aggiuntiva)

Ore formazione – Punteggio



MODULO 1. Elementi di statistica descrittiva

Ore formazione – Ore assenza



Domanda (6)

Completare le seguenti affermazioni sulla scomposizione della devianza:

$$SQT = SQR + \underline{\hspace{2cm}}$$

$$SQR = \sum (\hat{y}_i - \underline{\hspace{2cm}})^2$$

$$SQE = \sum (y_i - \underline{\hspace{2cm}})^2$$

$$R^2 = SQR / \underline{\hspace{2cm}}$$

$$\text{Se } SQE = 0 \rightarrow R^2 = \underline{\hspace{2cm}}$$

Domanda (7)

Retta di regressione stimata sui dati formazione: $\hat{y} = 44,00 + 1,125 \cdot x$
dove x = ore di formazione e $\hat{y}$ = punteggio di competenza atteso

Un dipendente ha seguito 32 ore di formazione. Qual è il punteggio di competenza atteso?

a) $\hat{y} = 44,00 + 1,125 \cdot 32 = 80,00$

b) $\hat{y} = 44,00 + 1,125 \cdot 32 = 76,00$

c) $\hat{y} = 44,00 \cdot 32 = 1.408$

d) Non è possibile fare previsioni con la retta di regressione

Domanda (8)

Quali tra le seguenti affermazioni descrivono correttamente la differenza tra correlazione e regressione:

- a) Il coefficiente di correlazione r misura la forza e la direzione della relazione lineare tra X e Y , senza distinguere quale variabile dipende dall'altra
- b) La correlazione permette di fare previsioni su Y dato un valore di X
- c) La regressione e la correlazione forniscono sempre la stessa informazione e sono intercambiabili
- d) La regressione assegna un ruolo asimmetrico alle variabili: X è indipendente e Y dipendente

Domanda (1)

Il **diagramma di dispersione** permette di:

- a) visualizzare i dati di due variabili espresse entrambe in funzione della stessa variabile X, collocata sull'asse delle ascisse
- b) visualizzare le coppie dei valori rilevati i cui dati sono posti per una variabile sull'asse X e per l'altra variabile sull'asse Y
- c) Visualizza i dati di due variabili espresse entrambe in funzione della stessa variabile Y, collocate sull'asse delle ordinate

Domanda (2)

Qual è il **metodo per trovare i parametri della retta di regressione**

- 1) il metodo dei minimi quadrati
- 2) il metodo della covarianza
- 3) il metodo della varianza

Domanda (3)

Se **la covarianza di due variabili X e Y è positiva**, allora il coefficiente angolare della retta di regressione che esprime Y in funzione di X è:

- 1) certamente positivo
- 2) certamente nullo
- 3) certamente negativo
- 4) le informazioni date non sono sufficienti

Domanda (4)

Nell'espressione di una retta: $y=a+bx$, il parametro a viene chiamato

1. Intercetta
2. Coefficiente angolare
3. Parametro medio

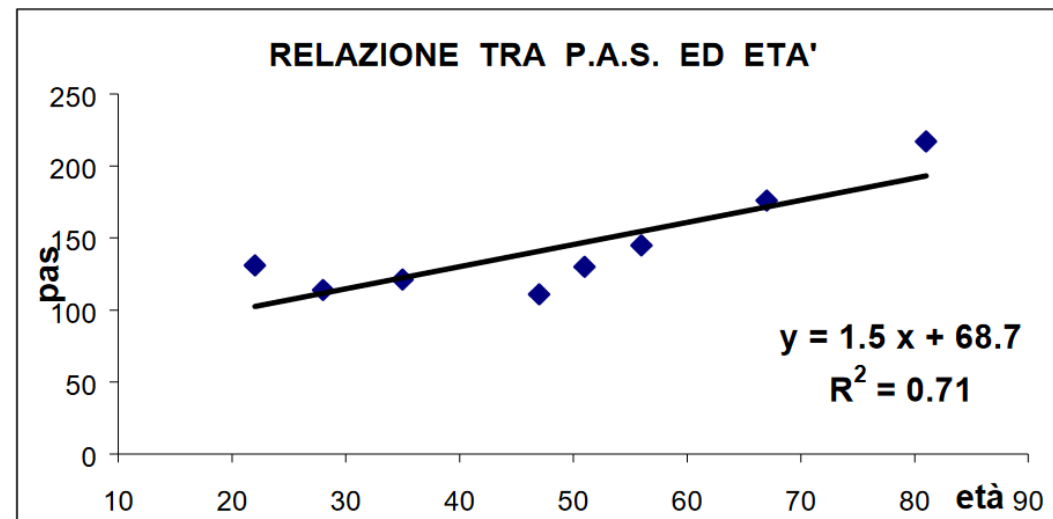
Cosa indica?

Domanda (5)

Allo scopo di studiare la relazione tra le variabili “età” e “pressione sistolica” (PAS) si rilevano, ad esempio, i valori di questi due caratteri su un campione di 8 soggetti sani.

ETA' (anni)	22	28	35	47	51	56	67	81
PAS (mm Hg)	131	114	121	111	130	145	176	217

Come si interpretano i parametri della retta di regressione riportati nel grafico ?



Domanda (6)

Nel modello di regressione, **l'indice di determinazione**

1. È pari al rapporto tra la devianza residua e la devianza totale
2. È pari al rapporto tra la devianza di regressione e la devianza totale
3. È tale che per un valore pari a 0.5 i punti sono allineati sulla retta di regressione