Analisi di relazioni tra variabili
MODELLO DI REGRESSIONE LINEARE

SAPIENZA
UNIVERSITÀ DI ROMA

Prof. Annarita Vestri
Annarita.vestri@uniroma1.it

Correlazione: analizza se esiste una relazione tra due variabili (come e quanto due variabili variano insieme)

Regressione: analizza la forma della relazione tra variabili

Cosa intendiamo per regressione in statistica

In statistica, il termine regressione è stato utilizzato per la prima volta dal biologo inglese Francis Galton nel 1886, quando parlò di «regressione verso la media».
Nell'ambito dei suoi studi sull'ereditarietà dei caratteri, Galton raccolse le stature di 928 figli adulti e dei loro 205 genitori (maschi e femmine). Esaminando le altezze di genitori e figli, notò una relazione tra le due variabili: più alti erano i genitori, più alti erano i figli e viceversa. Partendo dalla statura media dei genitori ('mid parent's stature') scoprì che i figli più alti della media avevano genitori ancora più alti di loro e i figli più bassi della media avevano genitori ancora più bassi.
A questo fenomeno diede il nome di regressione verso la media.

REGRESSIONE

La regressione lineare è finalizzata all'analisi della dipendenza tra due variabili, delle quali
- una (**Y**) è a priori definita come **dipendente o effetto**,
- l'altra (**X**) è individuata come **indipendente o causa**.

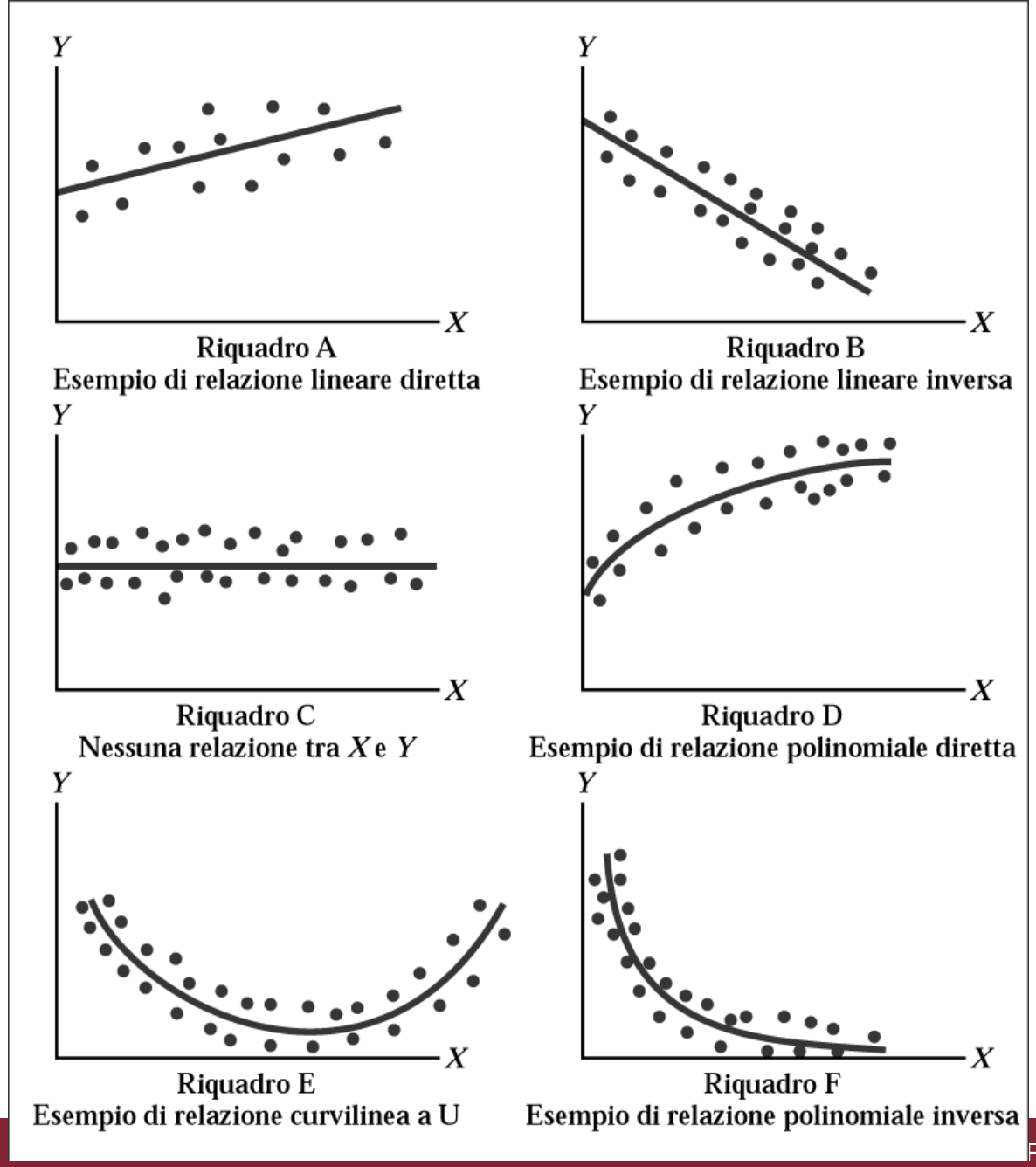
L'interesse della ricerca è rivolta essenzialmente all'**analisi delle cause** o allo **studio predittivo delle quantità medie di Y**, che si ottengono come risposta al variare di X.
Spesso anche nella ricerca ambientale, biologica e medica, la relazione di causa-effetto non ha una direzione logica o precisa: potrebbe essere ugualmente applicata nei due sensi, da una variabile all'altra.

Analisi di regressione

L’analisi di regressione è una tecnica che permette di analizzare la relazione lineare tra una variabile dipendente (o variabile di risposta) e una o più variabili indipendenti (o predittori).

- **Regressione semplice:** determinare la forma della relazione tra 2 variabili (una indipendente ed una dipendente)
- **Regressione multipla:** determinare la forma della relazione tra più variabili (più indipendenti ed una dipendente)

La scelta del modello matematico appropriato
è suggerita dal modo in cui si distribuiscono i valori delle due variabili
nel diagramma di dispersione



Analisi di regressione

L’analisi della regressione lineare è una metodologia asimmetrica che si basa sull’ipotesi dell’esistenza di una relazione di tipo causa-effetto tra una o più variabili indipendenti (o esplicative) e la variabile dipendente (o di criterio).

Lo studio di questa relazione può avere un duplice scopo:

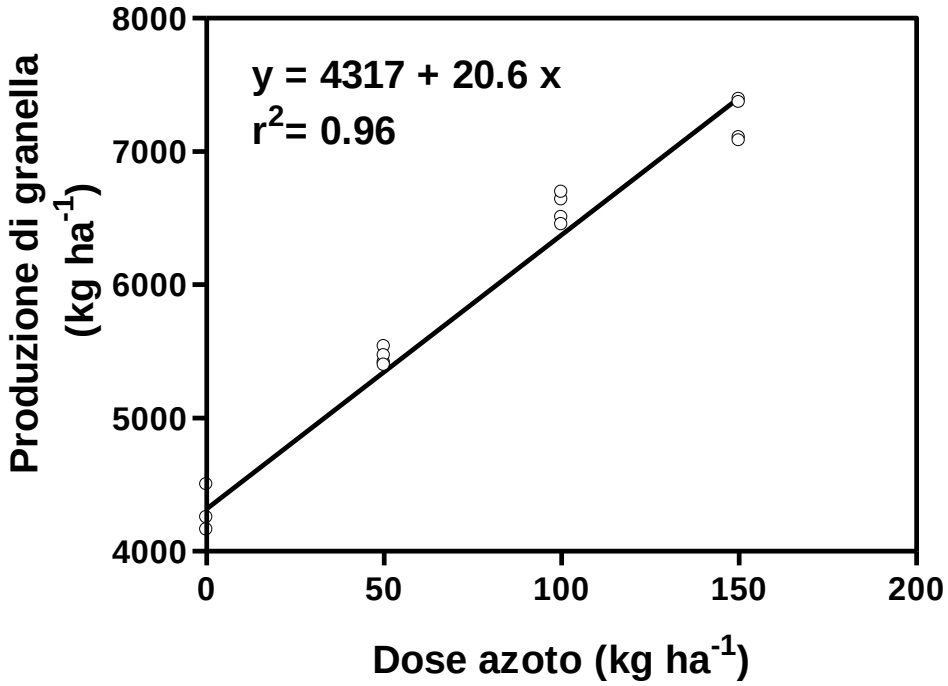
- esplicativo: comprendere e ponderare gli effetti delle variabili indipendenti (VI) sulla variabile dipendente (VD) in funzione di un determinato modello teorico;
- predittivo: individuare una combinazione lineare di variabili indipendenti per predire in modo ottimale il valore assunto dalla variabile dipendente.

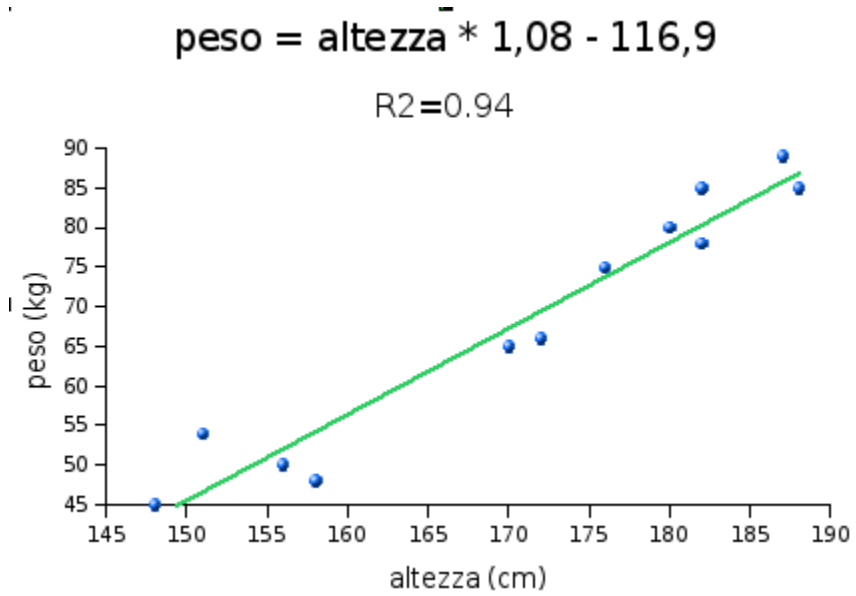
Perché è importante:

- Permette di costruire un modello funzionale della risposta di una variabile (effetto) rispetto ad un'altra (causa)
- Conoscendo la forma della relazione funzionale tra variabile indipendente e dipendente è possibile ***stimare*** il valore della variabile dipendente conoscendo quello della variabile indipendente (interpolazione) **nell'intervallo dei valori di X usato per la regressione**

Regressione lineare (semplice)

Nella regressione lineare la relazione tra variabili (*causa-effetto*) è rappresentata da una linea retta





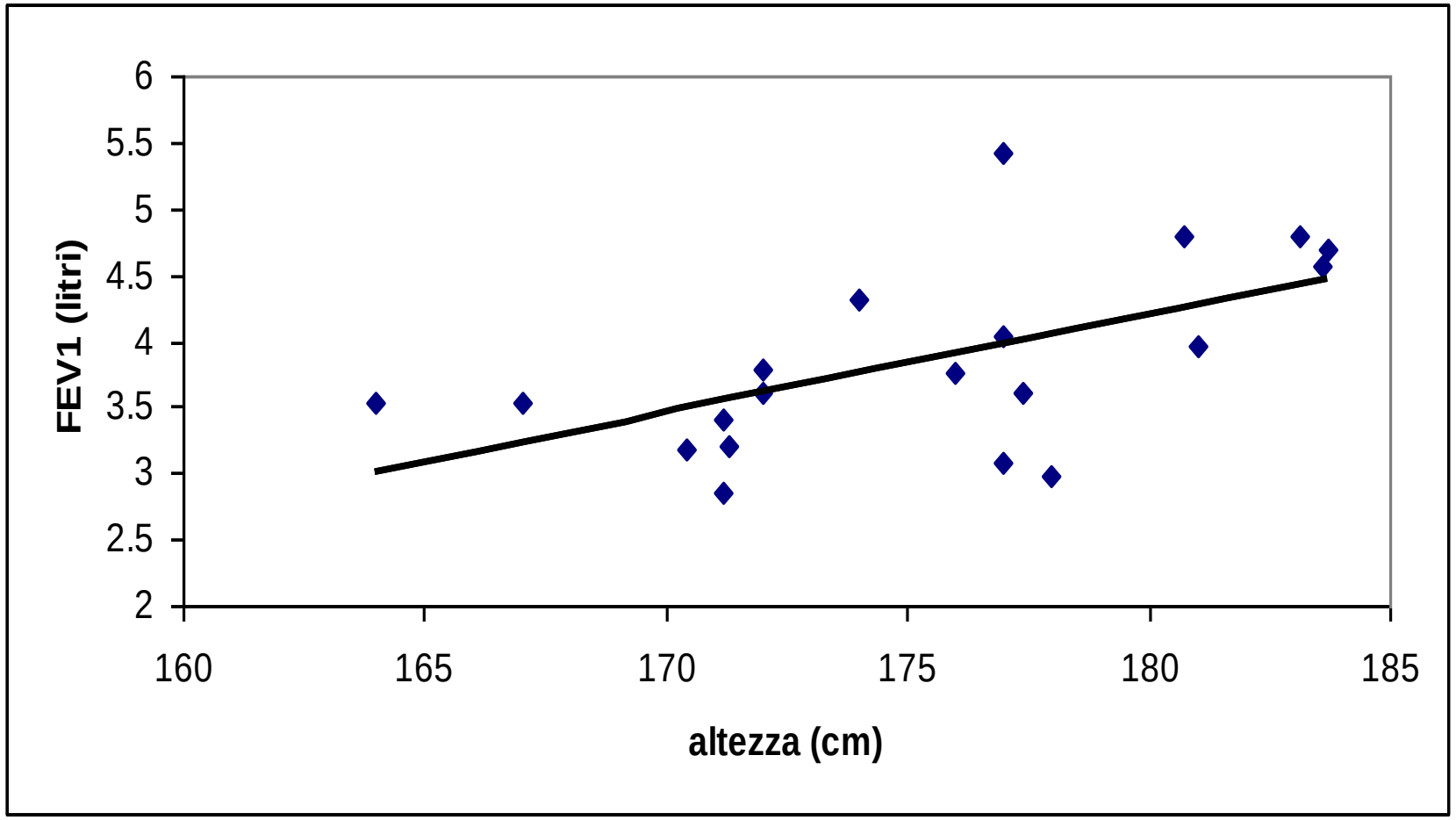
Se siamo indecisi su quale delle nostre variabili è dipendente e quale indipendente, allora l'analisi di regressione non è adatta!

WARNING!!!

Esempio: relazione tra FEV_1 (Volume espiratorio forzato) e altezza

altezza (cm)	FEV_1 (litri)	altezza (cm)	FEV_1 (litri)	altezza (cm)	FEV_1 (litri)
164.0	*3.54*	172.0	*3.78*	178.0	*2.98*
167.0	*3.54*	174.0	*4.32*	180.7	*4.80*
170.4	*3.19*	176.0	*3.75*	181.0	*3.96*
171.2	*2.85*	177.0	*3.09*	183.1	*4.78*
171.2	*3.42*	177.0	*4.05*	183.6	*4.56*
171.3	*3.20*	177.0	*5.43*	183.7	*4.68*
172.0	*3.60*	177.4	*3.60*		

La retta è la migliore rappresentazione della relazione tra le due variabili



Ma cosa rappresenta una regressione?

- Regredire una variabile sull'altra, significa spiegare il comportamento di una variabile mediante il comportamento di un'altra
- La retta di regressione esprime una tendenza; questo vuol dire che mediamente al variare della x_i *la* y_i assumerà certi valori (ricorda che c'è sempre un termine di errore!)
- Possiamo fare una considerazione di ordine generale:

  -la regressione rappresenta lo stesso concetto studiato con la media aritmetica;

  -l'errore standard (media dei quadrati degli errori) della retta di regressione equivale allo scarto quadratico medio.
- Il modello di regressione quindi esprime una misura di tendenza, alla quale viene associata una misura della variabilità (errore standard della regressione)

Regressione lineare semplice

Si cerca di trovare la retta che meglio interpola, che meglio si adatta alla nuvola di punti

La retta che meglio predice Y|X passa per la media di Y e X si deve cercare un criterio

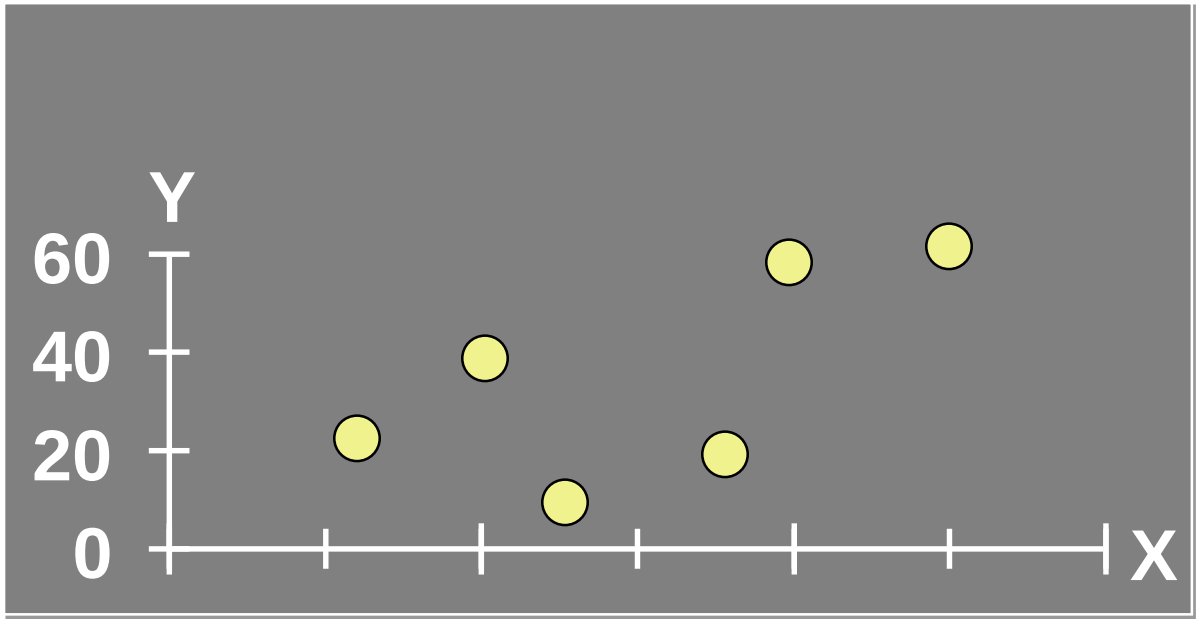
METODO DEI MINIMI QUADRATI

Si sceglie la retta che riduce al minimo la devianza residua

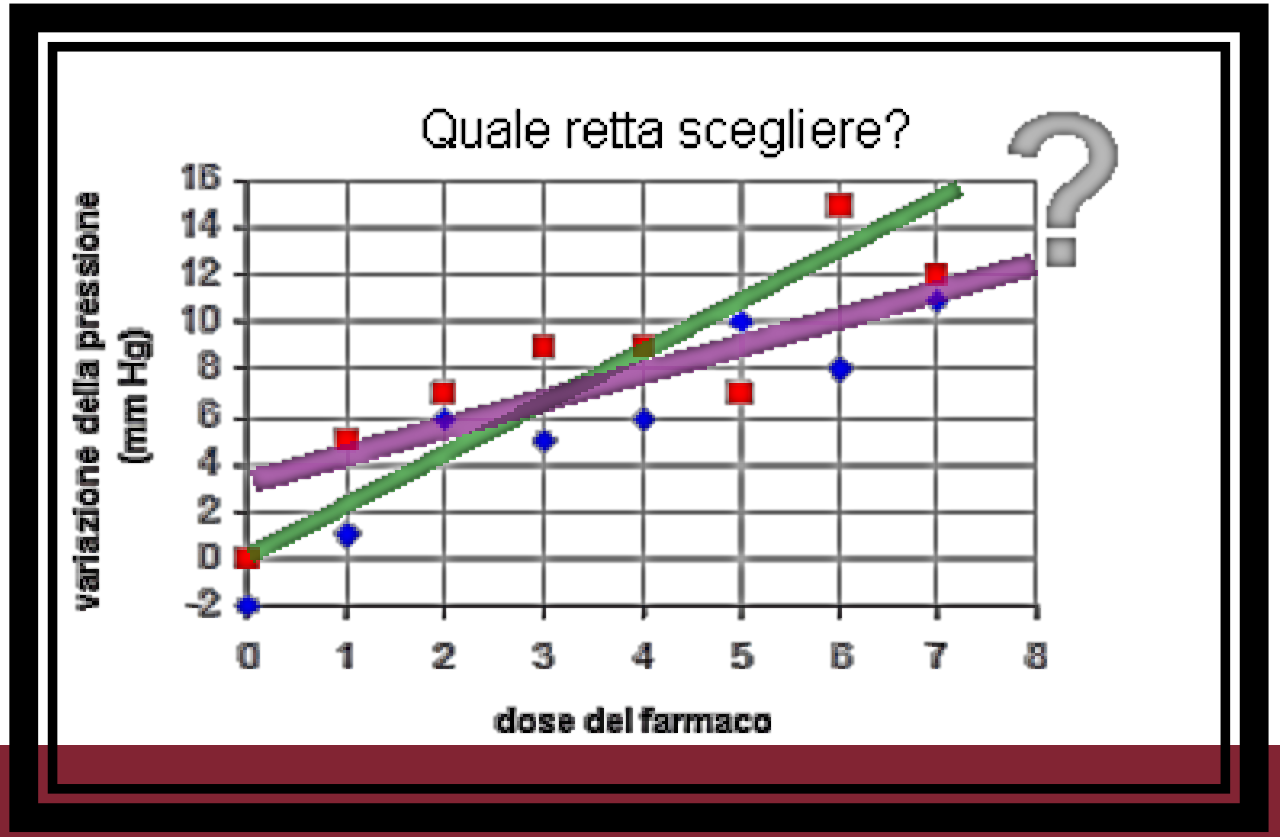
SSE = Σ(y- ŷ)²

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

- Disegnare tutte le coppie (X_i, Y_i)
- Come “fitterà” il modello?
- Come disegnare una retta tra i punti?
- Come determinare quale retta fitterà meglio?



- Disegnare tutte le coppie (X_i, Y_i)
- Come “fitterà” il modello?
- Come disegnare una retta tra i punti?
- Come determinare quale retta fitterà meglio?



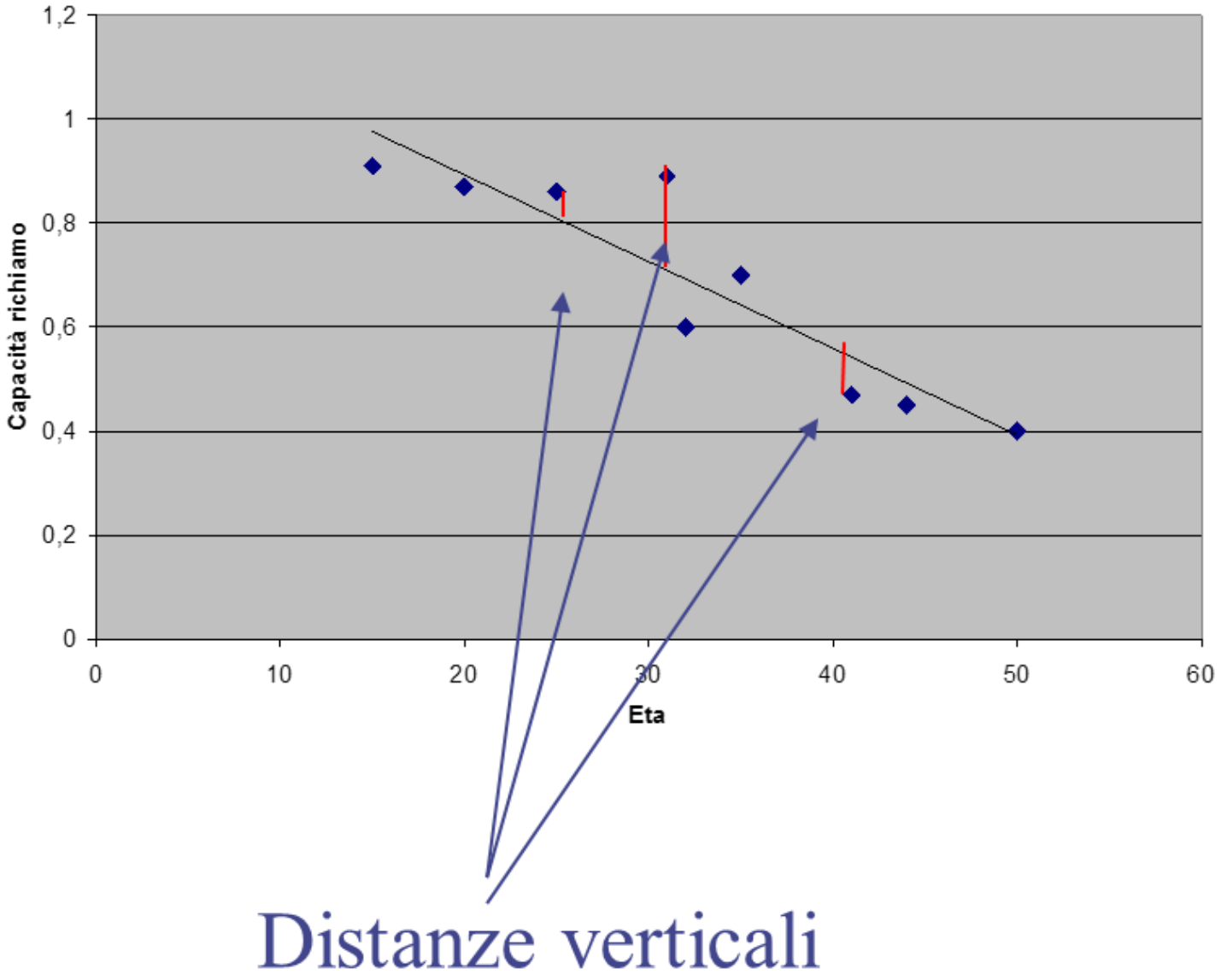
Il metodo dei minimi quadrati

- Metodo che ci consente di scegliere i coefficienti di regressione in modo che la retta di regressione è "*il più vicino possibile*" ai dati osservati.
- Come misuriamo la vicinanza ai dati?
- Analizziamo i **residui**: differenza tra i valori della variabile risposta osservati e quelli predetti dal modello

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

- Supponiamo di aver stimato β_0 con b_0 e β_1 con b_1: i valori predetti saranno pari a

$$\hat{Y}_i = b_0 + b_1 X_i$$



Distanze verticali

La linea di regressione ottimale si definisce come quella linea che minimizza le distanze verticali fra le osservazioni e la linea stessa

Alcune distanze sono “positive”
Alcune sono negative
L’ottimizzazione minimizza la somma degli scarti quadratici

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

Retta di regressione

$$\hat{Y}_i = b_0 + b_1 X_i$$

Pendenza

$$b_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{\sum \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum \left(X_i - \bar{X}\right)^2}$$

Intercetta

$$b_0 = \bar{Y} - b_1 \bar{X}$$

Nell'esempio precedente

$$FEV_1\ (litri) = -9.19 + 0.0744 \times altezza\ (cm)$$

coefficiente di regressione

intercetta

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$b_0 = \bar{y} - b\bar{x}$$

Interpretazione dei coefficienti

Pendenza (b_1)

Stima il cambiamento di Y per un'unità di incremento di X

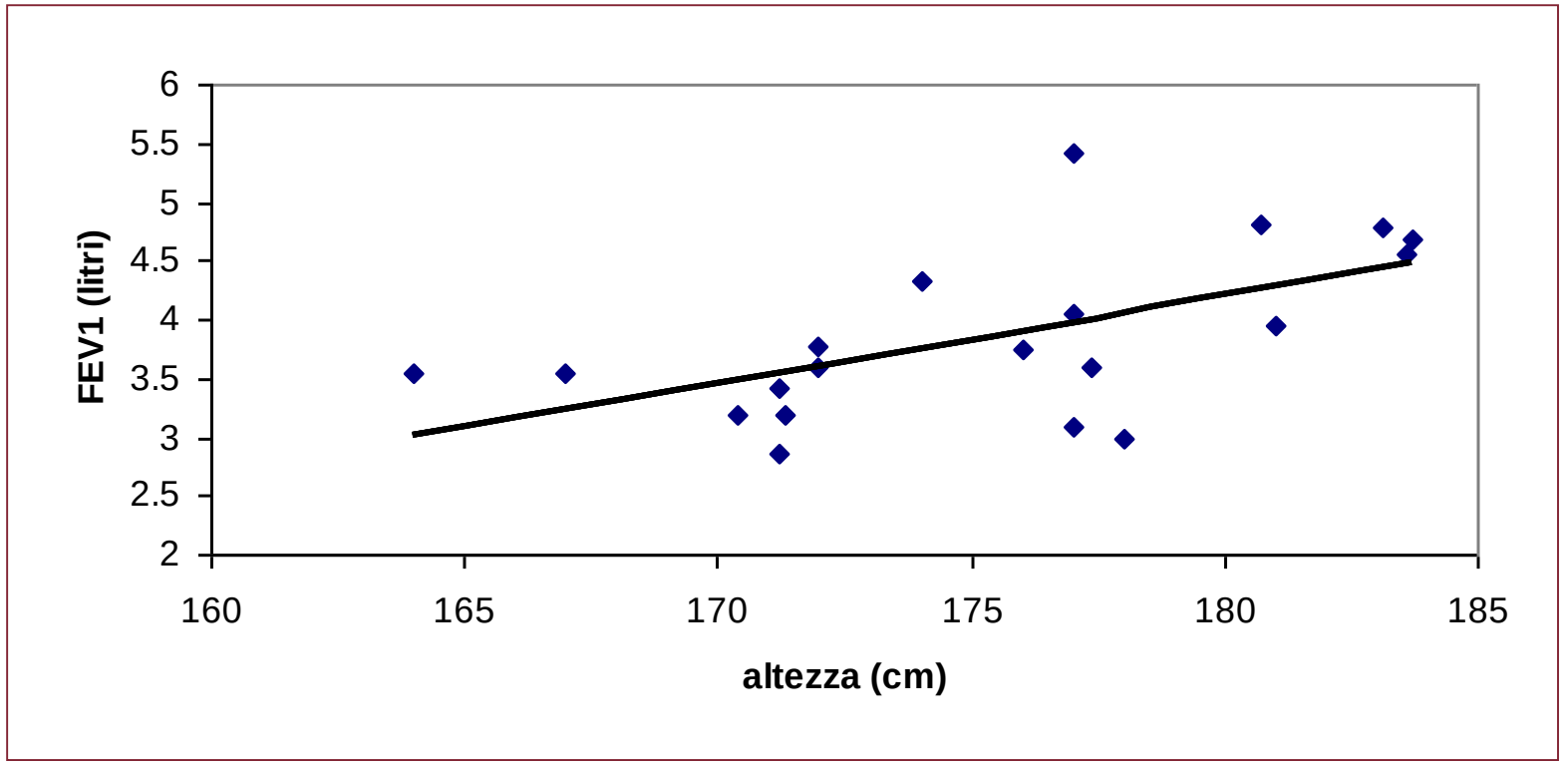
- se $b_1 = 0,07$, ci aspettiamo che la FEV1(*Y*) aumenti di 0,07 unità per ogni unità di incremento in altezza (*X*)

Interpretazione dei Coefficienti

Intercetta (b_0)

Valore medio di Y quando $X = 0$

- Se $b_0 = -9$, allora la FEV1(*Y*) sarà pari a -9 quando l'altezza (*X*) è uguale a 0

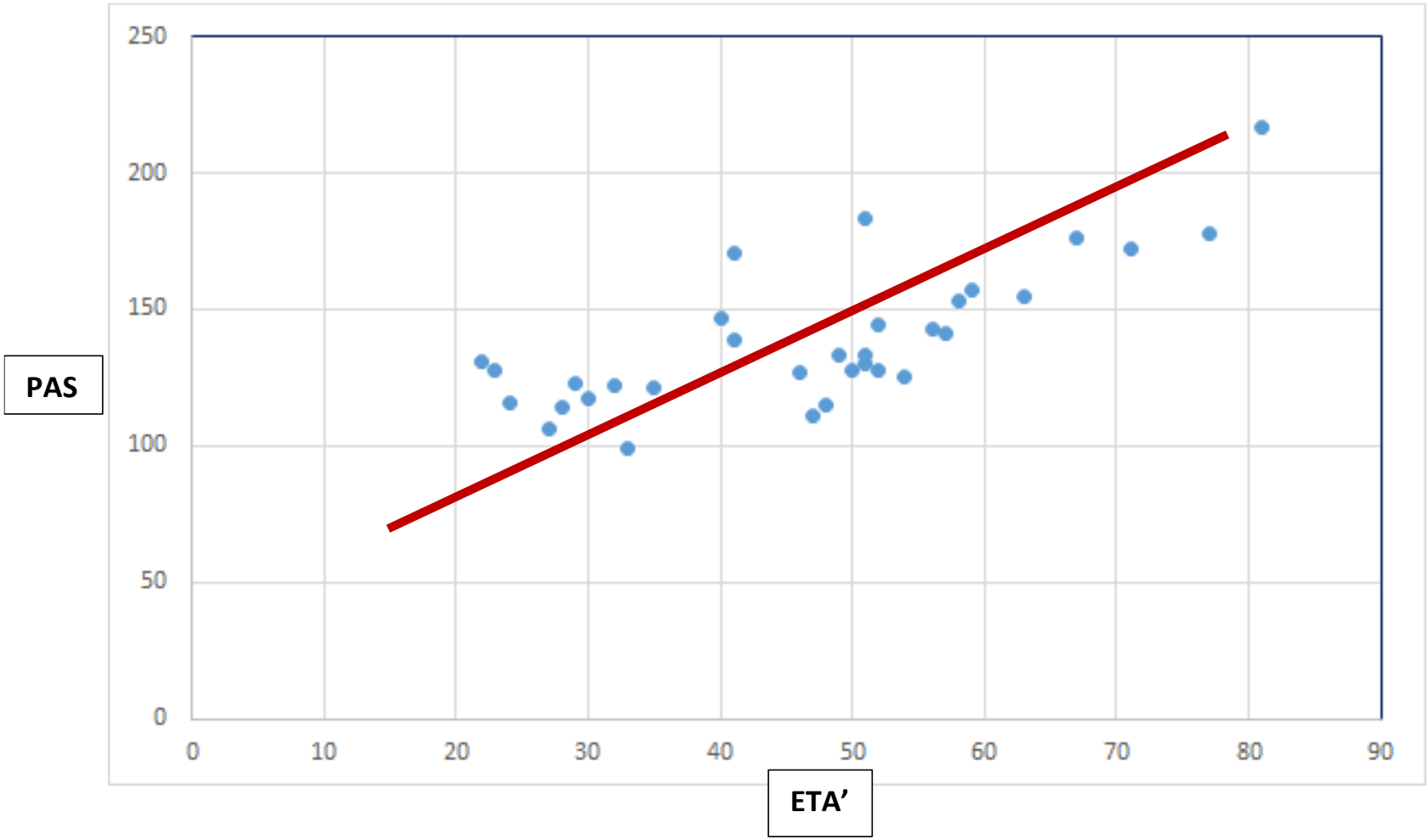


$$FEV_1 \text{ (litri)} = -9.19 + 0.0744 \times \text{altezza (cm)}$$

Se uno studente è alto 170 cm, il suo FEV_1 è 3.458 litri

Soggetto	Età	Pas	Soggetto	Età	Pas	Soggetto	Età	Pas
1	22	131	12	41	139	23	52	128
2	23	128	13	41	171	24	54	125
3	24	116	14	46	127	25	56	143
4	27	106	15	47	111	26	57	141
5	28	114	16	48	115	27	58	153
6	29	123	17	49	133	28	59	157
7	30	117	18	50	128	29	63	155
8	32	122	19	51	183	30	67	176
9	33	99	20	51	130	31	71	172
10	35	121	21	51	133	32	77	178
11	40	147	22	52	144	33	81	217

DI QUANTO VARIA LA PRESSIONE SISTOLICA AL VARIARE DELL'ETA?
LA RELAZIONE TRA QUESTE DUE VARIABILI E' TENDENZIALMENTE LINEARE?


PAS
250
200
150
100
50
0
0
10
20
30
40
50
60
70
80
90
ETA’

Calcoli:

- n = 33 $\sum x$ =1545 $\sum y$ =4583 $\sum xy$= 223144
- $\sum x^2$= 80019 $\sum y^2$= 657945 $\sum xi/n$ = 46,82 $\sum yi/n$ =138,88
- dev x =80019 – (46,82) ² * 33 = 7684,909
- dev y = 657945- (138,88)2 / 33 = 21463,52
- codev xy = 224111- (46,82) (138,88) * 33= 9543,273
- by|x = (9543,273) / (7684,909) = 1,24
- a = $\sum yi/n$ - by|x $\sum xi/n$ = 80.82
- **y = 80.82 + 1.24 x (retta)**

Indice di determinazione

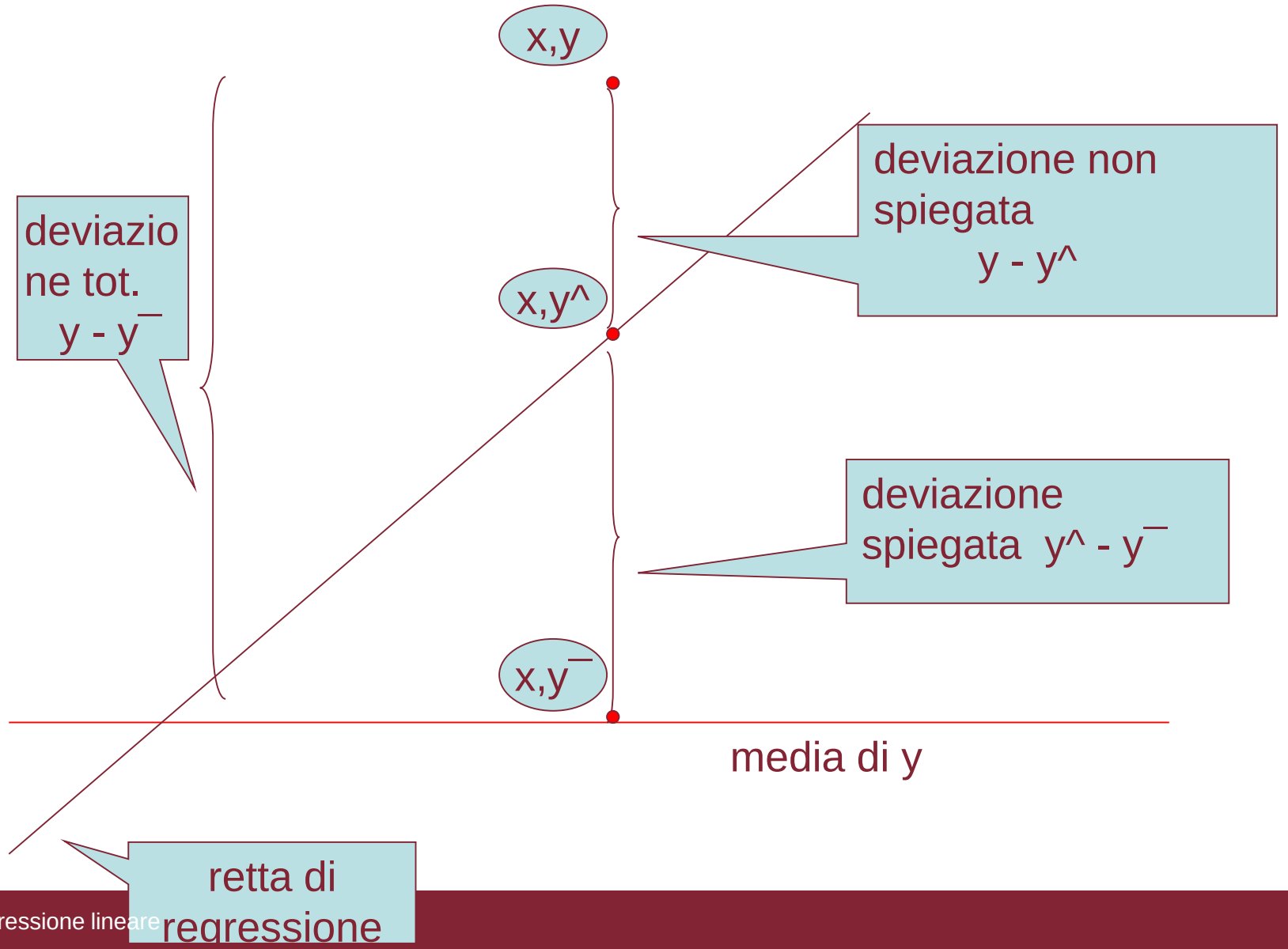
Dopo aver calcolato i parametri della retta di regressione è importante valutare il grado di accostamento tra i valori osservati e quelli teorici

Indice di determinazione R^2 che esprime quanta parte della devianza totale di y è spiegata dalla retta di regressione

Rapporto tra la devianza totale del modello e la devianza totale della y

$$R^2 = \frac{\sigma^2_{modello}}{\sigma^2_y} = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}$$

Variabilità della y spiegata e non spiegata dal modello



Coefficiente di determinazione

- SST=SSR+SSE
- Per valutare quanto la retta di regressione si adatta bene ai dati utilizziamo il coefficiente di determinazione R^2 che misura la quota di variazione totale spiegata dalla regressione, ovvero:

$$R^2 = \frac{\text{Variazione Spiegata (SSR)}}{\text{Variazione Totale (SST)}} = 1 - \frac{\text{SSE}}{\text{SST}} =$$

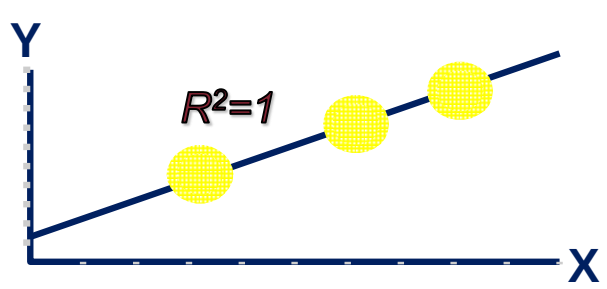
$$= \frac{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2 - \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}$$

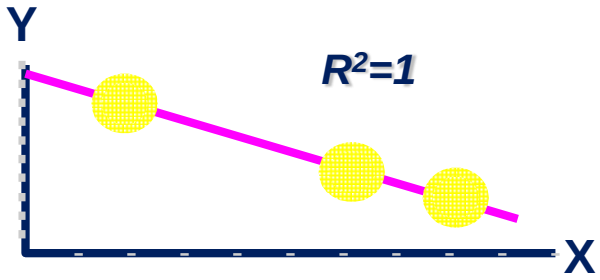
$0 \leq R^2 \leq 1$

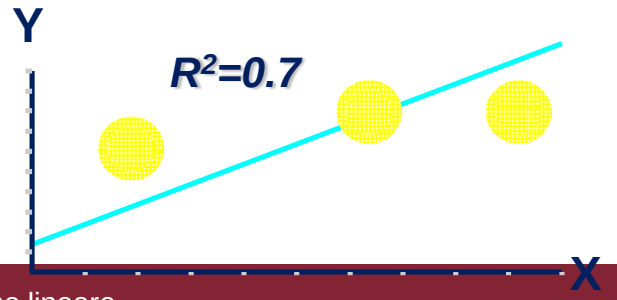
Coefficiente di determinazione

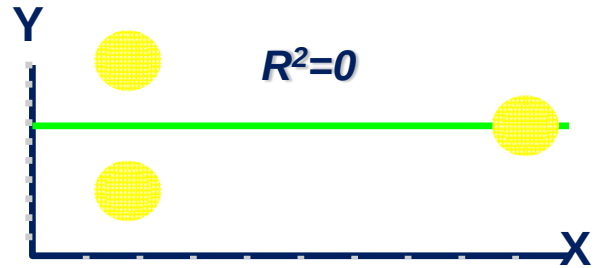
$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$

- Se il fit è perfetto SSE=0 e R^2=1
- Se il fit non è buono SSR=0 e R^2=0 (SSE=SST)









Indice di determinazione

$$\mathbf{0 \leq R^2 \leq 1}$$

Es. $R^2=0.39$ significa che il 39% della variazione FEV1 è spiegata dall'altezza, cioè esiste una variabilità comune del 39% mentre il rimanente 61% di variabilità è legato ad altri fattori che non possiamo spiegare con la sola regressione