LEZIONI DI STATISTICA MEDICA

Prof. Roberto de Marco

Lezione n.14

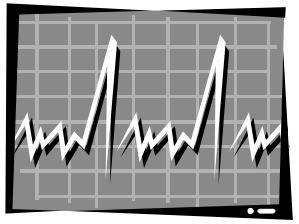
- Correlazione e regressione

Sezione di Epidemiologia & Statistica Medica
Università degli Studi di Verona

Siamo interessati a studiare la relazione che lega due variabili quantitative X e Y

Consumo di alcool  Incidenza di cirrosi epatica

Aumento della temperatura corporea  Frequenza cardiaca

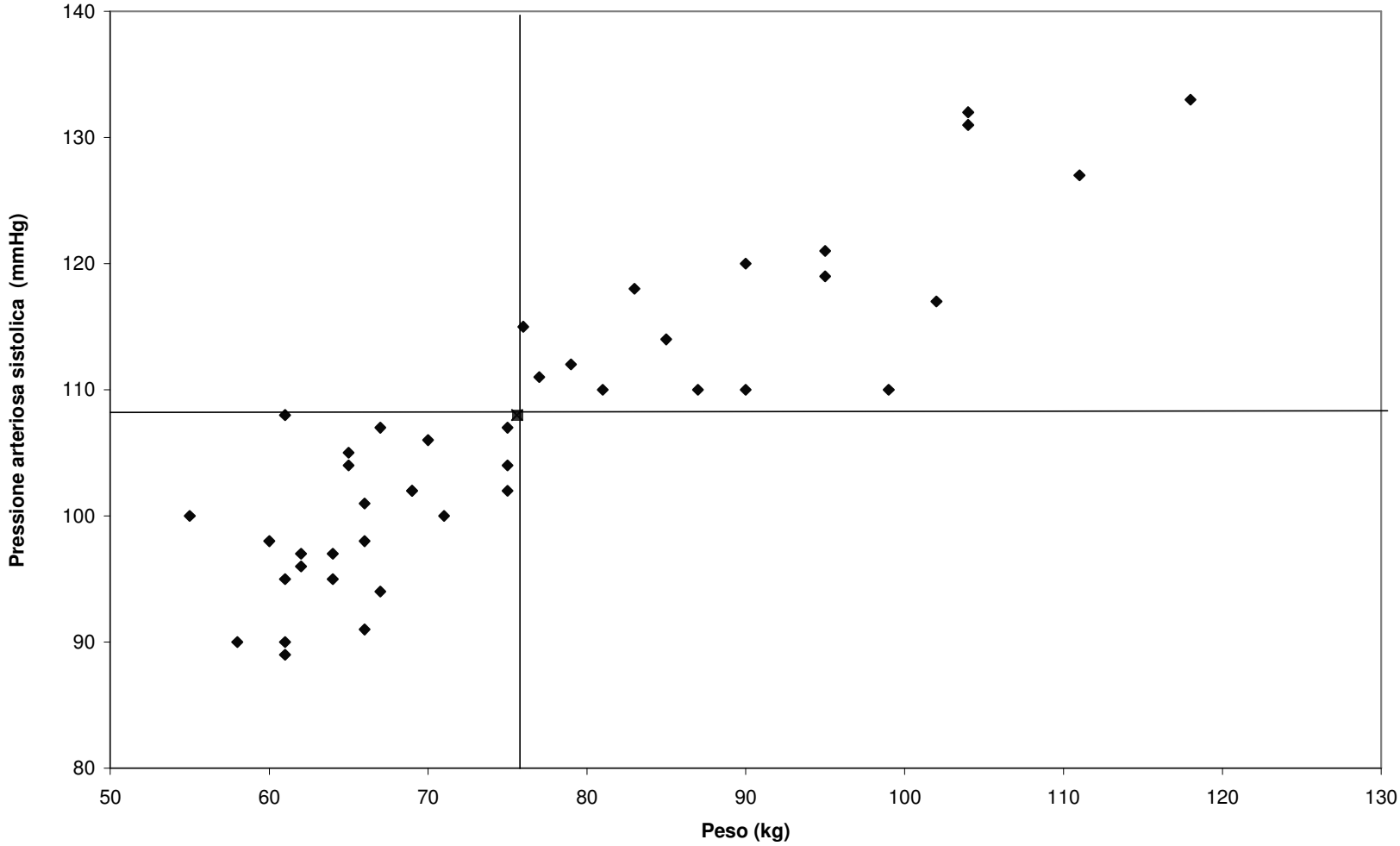


Lo scopo è sapere se variazioni nei livelli di una variabile inducono o si accompagnano a variazioni nei livelli dell'altra variabile

Se i valori di X e Y sono tali per cui i valori assunti da una variabile sono "indicativi" dei valori dell'altra

X e Y si dicono correlate

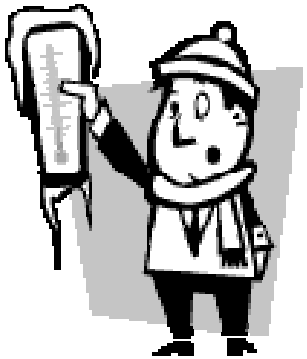
PER VERIFICARE GRAFICAMENTE LA CORRELAZIONE ⇨ DIAGRAMMA A DISPERSIONE



Esempio

Una tecnica utilizzata per ridurre i battiti cardiaci anormalmente rapidi è il "diving reflex" che consiste nell'immersione della faccia dei pazienti in acqua fredda. Il riflesso, mediato dalle basse temperature, è una risposta neurale involontaria che comporta una deviazione della circolazione del sangue ossigenato della pelle, muscoli e organi interni al cuore, polmoni e cervello.

Un ricercatore medico ha condotto un esperimento per verificare gli effetti di varie temperature fredde sugli effetti sulla diminuzione del battito cardiaco in 10 bambini:



bambino	temperatura dell'acqua (X, °F)	riduzione della frequenza del polso (Y, battiti/minuto)
1	68	2
2	65	5
3	70	1
4	62	10
5	60	9
6	55	13
7	58	10
8	65	3
9	69	4
10	63	6

- Frequenza cardiaca e temperatura sono correlate?
- Quanto?

COEFFICIENTE DI CORRELAZIONE r (DI PEARSON)

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

s_{xy} = codevianza tra X e Y

s_x = devianza di X

s_y = devianza di Y

$-1 \leq r_{xy} \leq 1$

Massima correlazione lineare negativa

Massima correlazione lineare positiva

$r_{xy} = 0$

Nessuna correlazione

Soluzione:

bambino	X	Y	X^2	Y^2	XY
1	68	2	4624	4	136
2	65	5	4225	25	325
3	70	1	4900	1	70
4	62	10	3844	100	620
5	60	9	3600	81	540
6	55	13	3025	169	715
7	58	10	3364	100	580
8	65	3	4225	9	195
9	69	4	4761	16	276
10	63	6	3969	36	378
TOTALE	635	63	40537	541	3835

$$SS_{xx} = (\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2) = (40537 - (1/10)\cdot 635^2) = 214.5$$

$$SS_{yy} = (\sum_{i=1}^{n} y_i^2 - \frac{1}{n}(\sum_{i=1}^{n} y_i)^2) = 541 - (1/10)\cdot 63^2 = 144.1$$

$$SS_{xy} = (\sum_{i=1}^{n} (x_i y_i) - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i) = 3835 - (1/10)\cdot 635 \cdot 63 = -165.5$$

$$r_{xy} = \frac{SS_{xy}}{\sqrt{SS_{xx} SS_{yy}}} = -\frac{165.5}{\sqrt{214.5 \cdot 144.1}} = -0.94$$

Test per il coefficiente di correlazione ρ

- la presenza di una forte correlazione non implica necessariamente una relazione causale tra le due variabili, ma solo l'esistenza di una forte associazione tra i dati.

- r è una stima campionaria di ρ e come tale è soggetta a fluttuazioni casuali: potremmo trovare alti valori di r, anche se la vera correlazione tra le due variabili fosse inesistente.

- Sotto alcune assunzioni di normalità sulla distribuzione congiunta di X e Y nella popolazione, si può dimostrare che un test idoneo per testare:

$$\begin{cases} H_0: \rho = \rho_0 = 0 \\ H_1: \rho \neq \rho_0 \neq 0 \end{cases}$$

E' dato da: $$t = r\sqrt{\frac{n-2}{1-r^2}} \quad \sim t_{n-2}$$

CORRELAZIONE ⟸ Valutare il grado di associazione tra due variabili X e Y

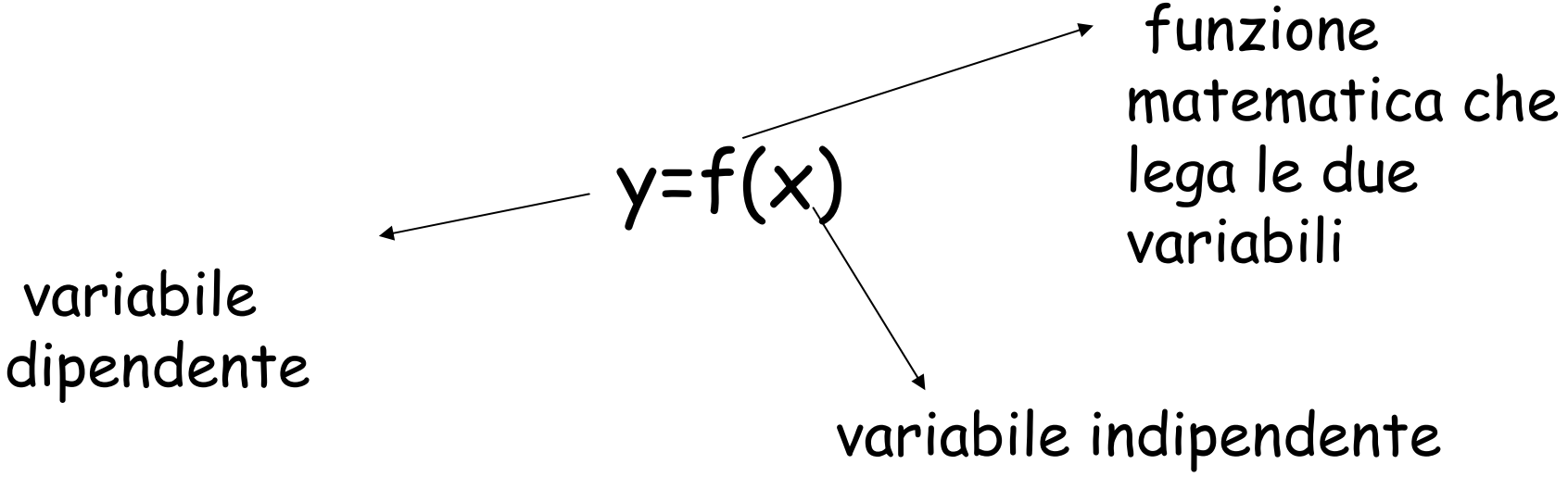
REGRESSIONE ⟸ Esprimere la relazione tra due variabili

⟋ **Attraverso un**

MODELLO che permette di chiarire o stabilire una <u>RELAZIONE FUNZIONALE</u>

REGRESSIONE

- permette di esprimere la relazione tra due variabili con un modello funzionale



SCOPO ⇒ descrittivo

⇒ predittivo

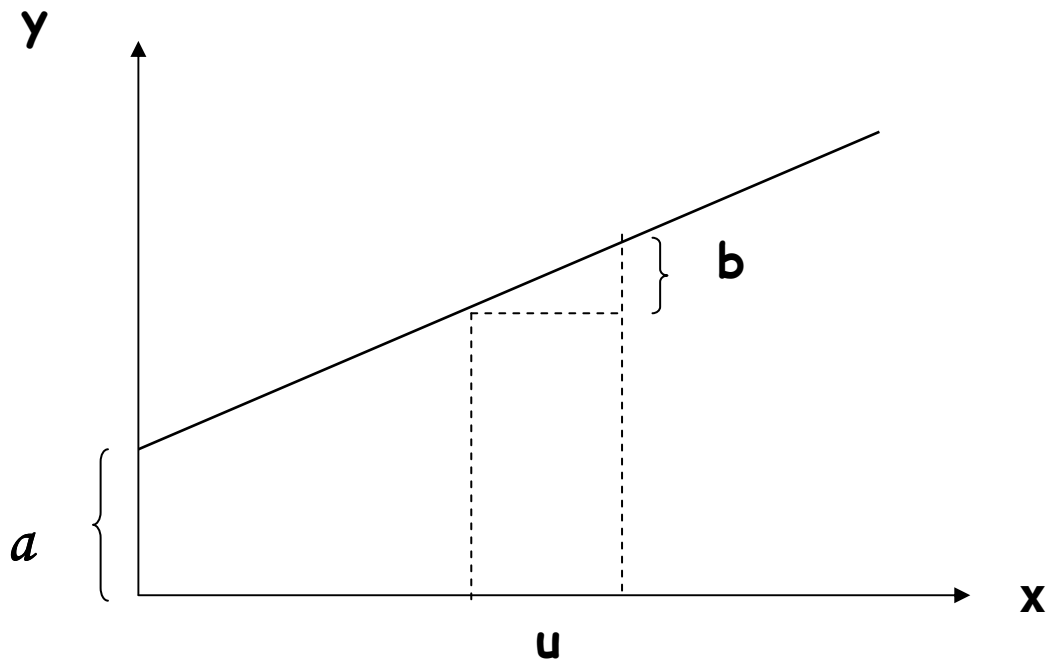
Regressione lineare

Ampiamente usata in ambito biomedico

intercetta

Coefficiente angolare

$$y=a+b(x)$$



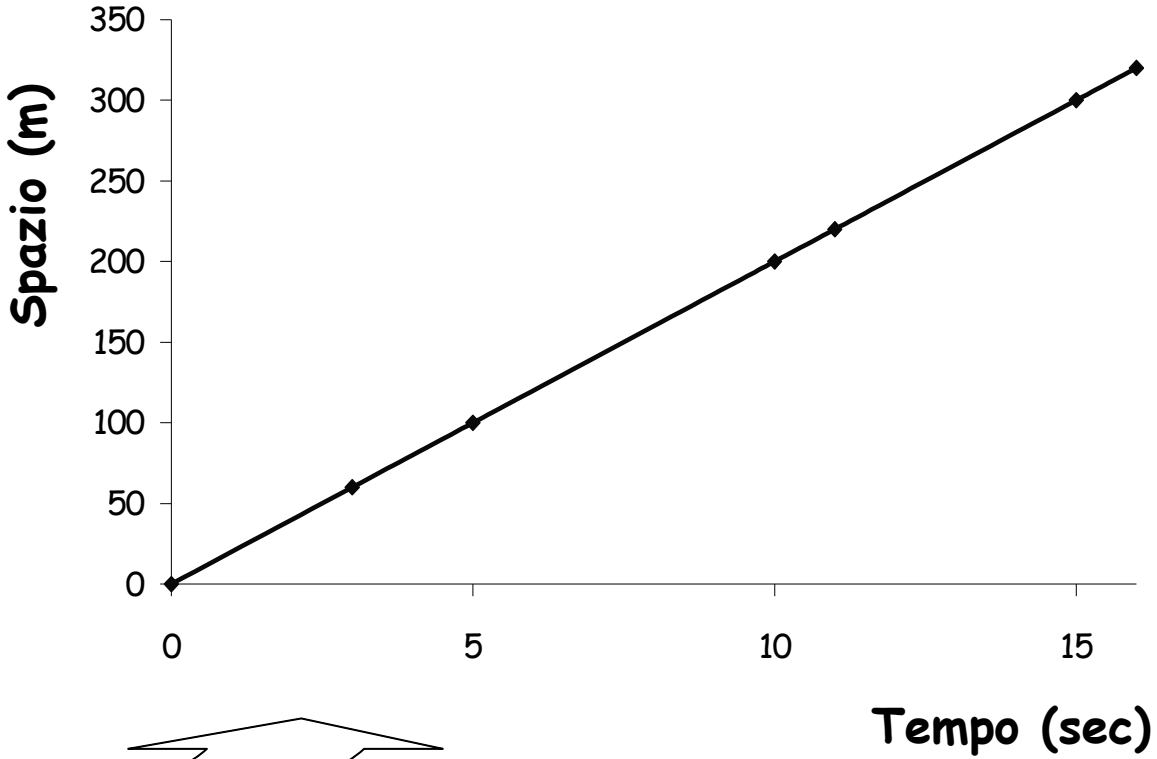
Esempio di relazione lineare

Si abbiano due variabili X e Y.

X è il tempo (in secondi) a cui viene osservato un corpo.

Y è lo spazio (in metri) che il corpo ha percorso da un certo punto.

X	Y
0	0
3	60
5	100
10	200
11	220
15	300
16	320



$y=20x$

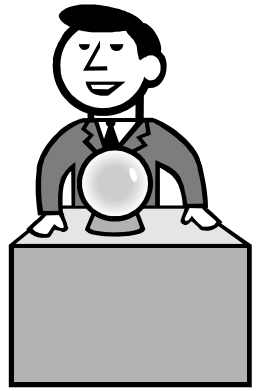
- La variabilità di Y è completamente spiegata dalla retta

- La retta descrive perfettamente i dati e individua la "legge" che li ha prodotti (*legge del moto uniforme*)

- Il coefficiente angolare (b=20) rappresenta l'incremento nello spazio per incremento unitario nel tempo (la velocità) ed è misurato come metri al secondo

- Tale modello è completamente deterministico:

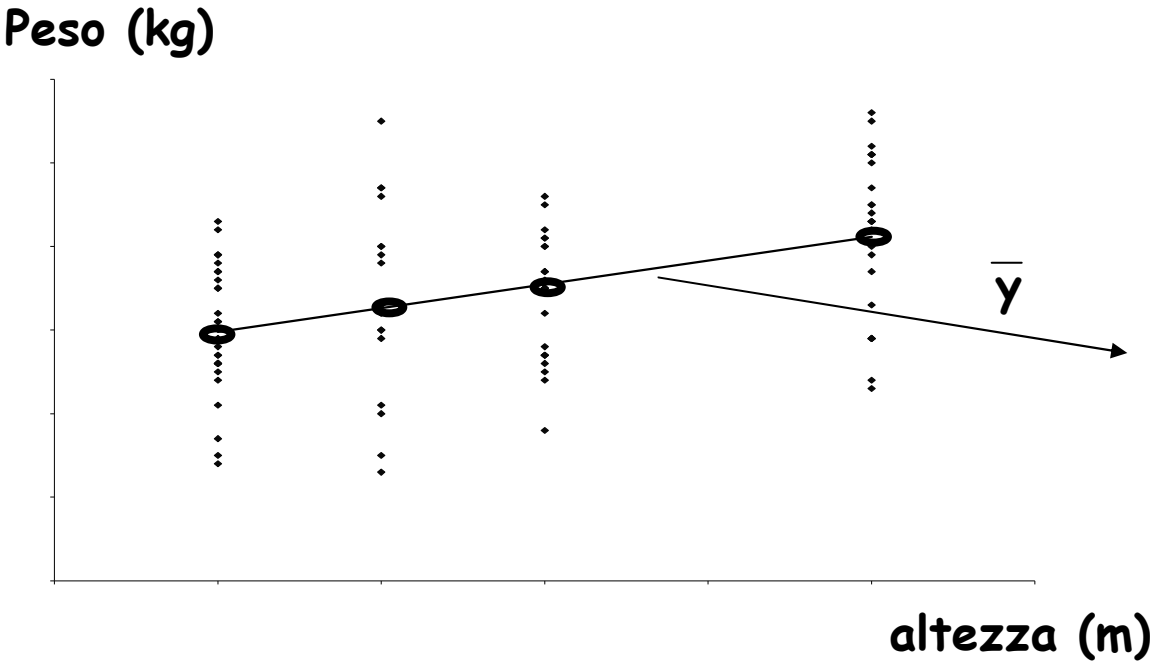
  noti i valori di x, si possono predire esattamente i valori di Y



- In biologia e medicina, la relazione tra variabili non è sempre perfettamente lineare

MA Il modello lineare permette di approssimare la descrizione del fenomeno

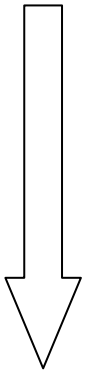
Esempio: relazione tra peso e altezza



- Per ogni altezza esiste un range di pesi ⇨ Variabilità biologica + Errore di misura

- In media il peso cresce linearmente con l'altezza

- Il luogo geometrico delle medie di Y per dati valori di X è detto CURVA DI REGRESSIONE DI Y SU X



se curva=retta

RETTA DI REGRESSIONE DI y SU x

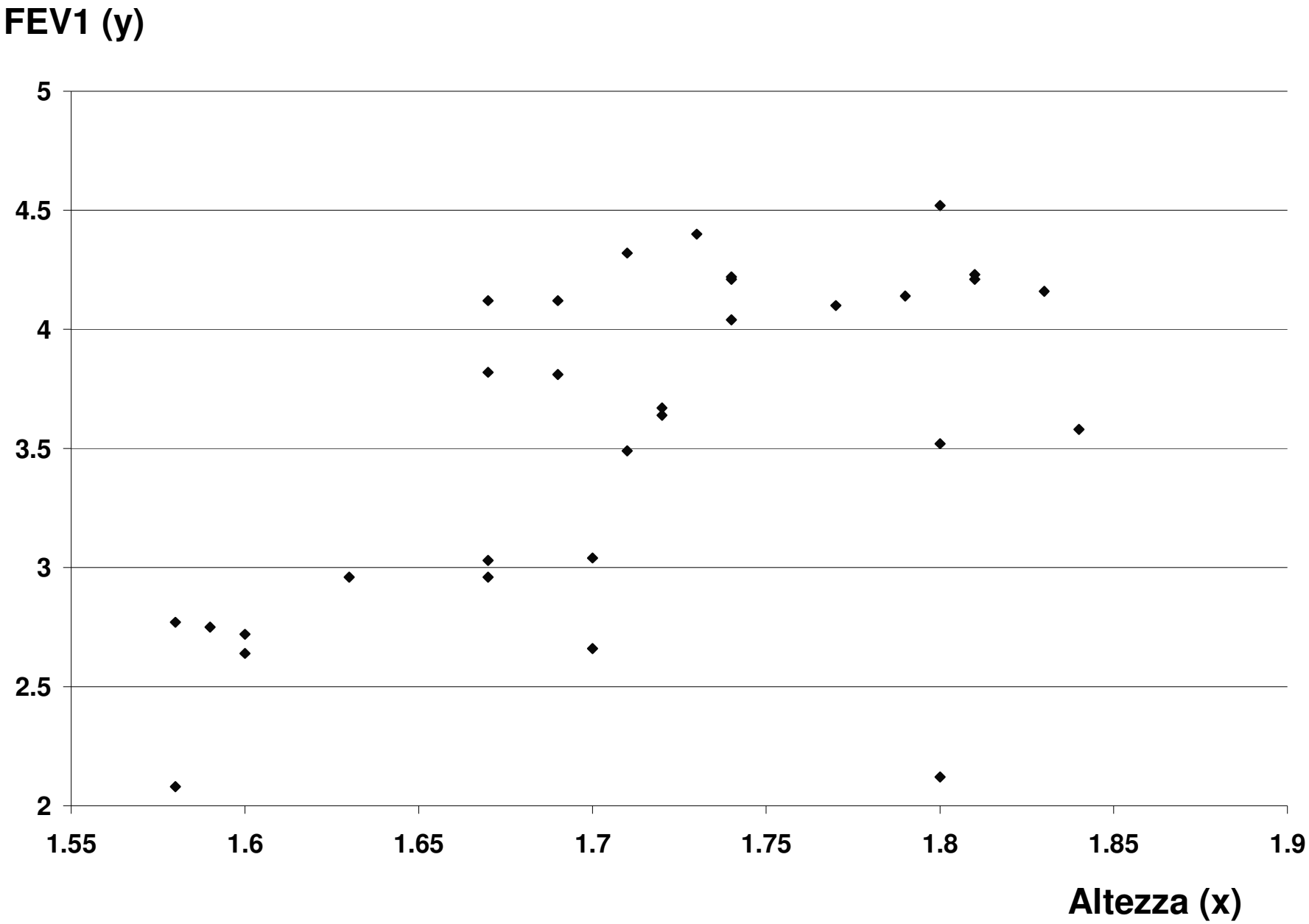
Esercizio

Nella tabella seguente sono riportati i dati relativi ad altezza e FEV1 (*forced expiratory volume in 1 second*) per 30 soggetti (dati ECRHS).

altezza (m)	FEV1 (l)
X	Y
1.79	4.14
1.8	4.52
1.72	3.64
1.69	4.12
1.72	3.67
1.84	3.58
1.6	2.72
1.7	3.04
1.83	4.16
1.58	2.08
1.74	4.04
1.74	4.22
1.67	3.82
1.71	3.49
1.67	2.96
1.58	2.77

altezza (m)	FEV1 (l)
X	Y
1.71	4.32
1.67	3.03
1.67	4.12
1.73	4.4
1.81	4.21
1.81	4.23
1.8	3.52
1.69	3.81
1.7	2.66
1.74	4.21
1.77	4.1
1.8	2.12
1.6	2.64
1.63	2.96
1.59	2.75

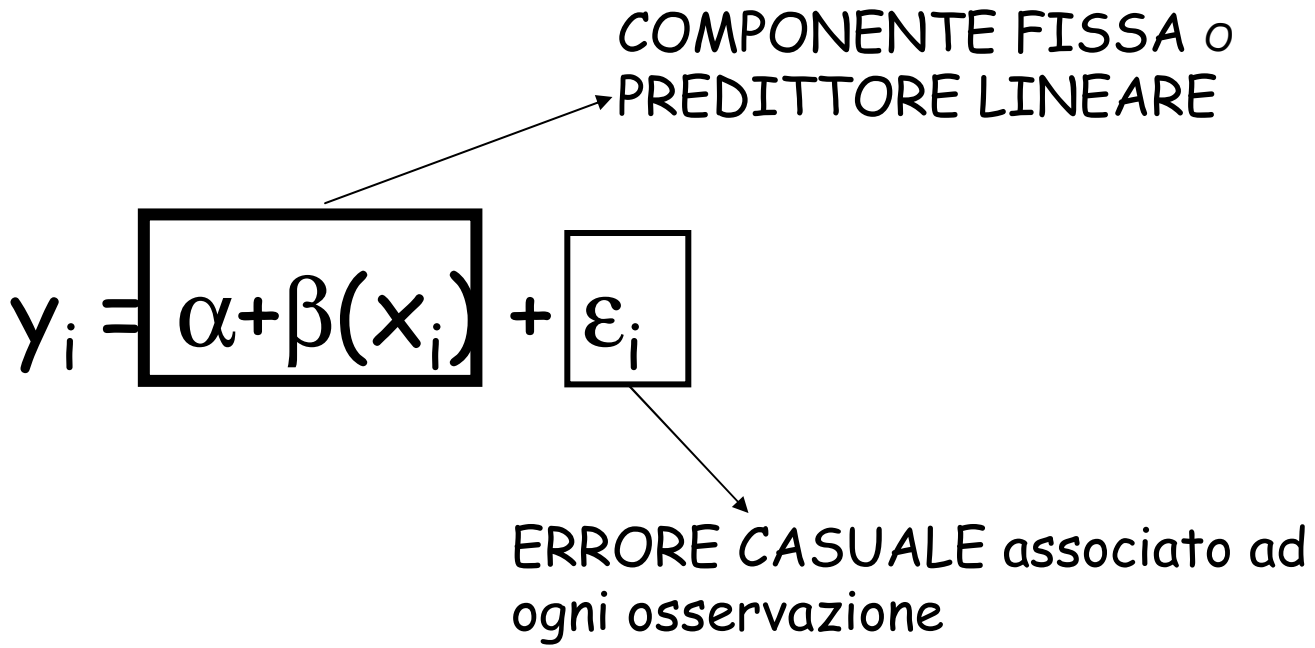
1. Rappresentiamo i dati in un diagramma a dispersione di punti



2. Assumeremo che nella popolazione il legame tra altezza (X) e FEV1 (Y) possa essere espressa da:

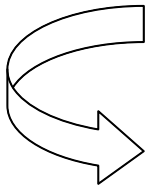
$$E(y)=\alpha+\beta(x)$$

L'osservazione di Y nell'i-mo individuo avrà quindi la seguente struttura:

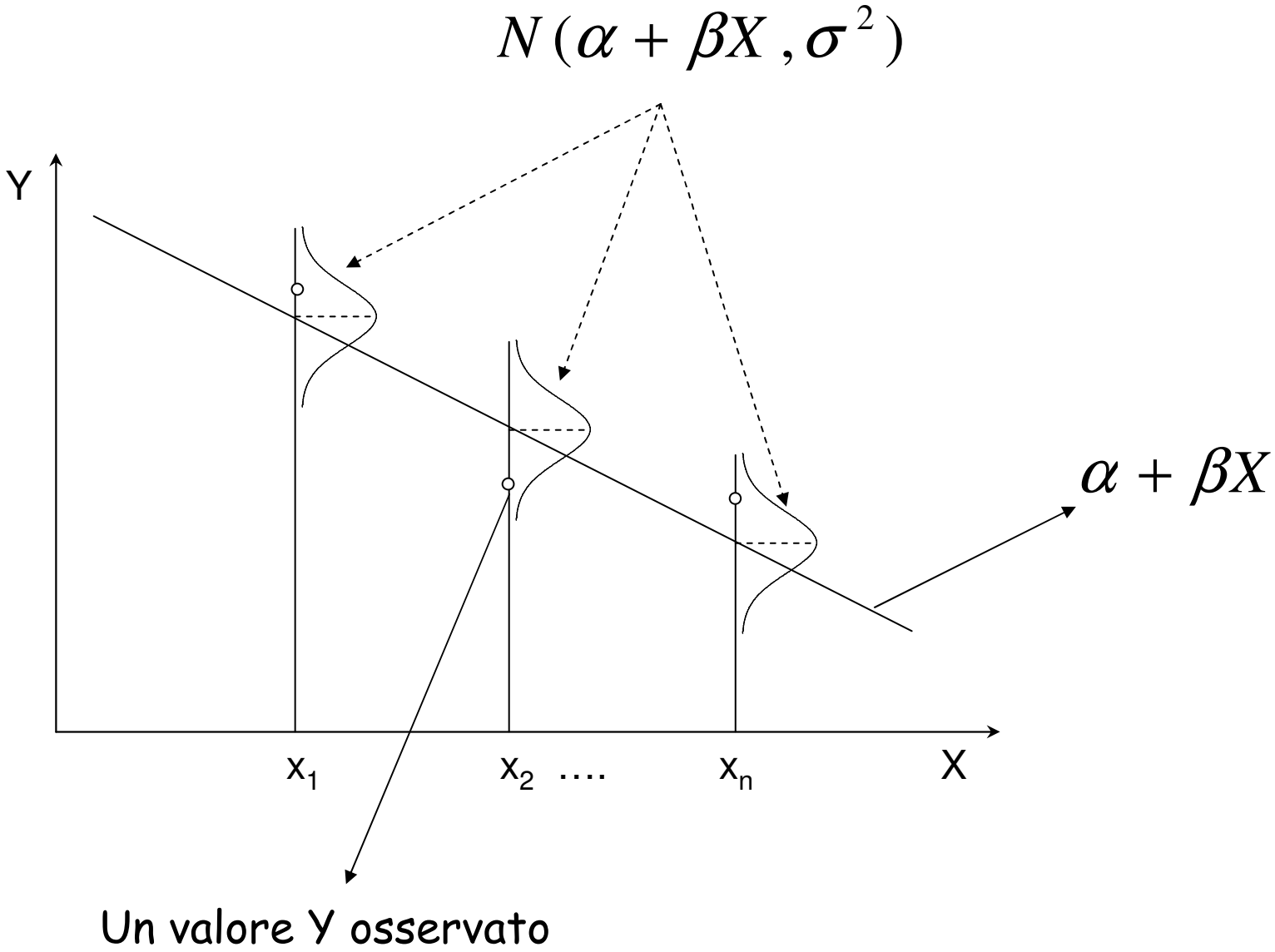


Dove:

- **Y** è la variabile di risposta (o dipendente)
- $\alpha+\beta$ sono parametri ignoti da stimare sulla base dei dati disponibili
- **X** è la variabili esplicativa (indipendente)
- ε_i (errore casuale) ~$N(0,\sigma^2)$



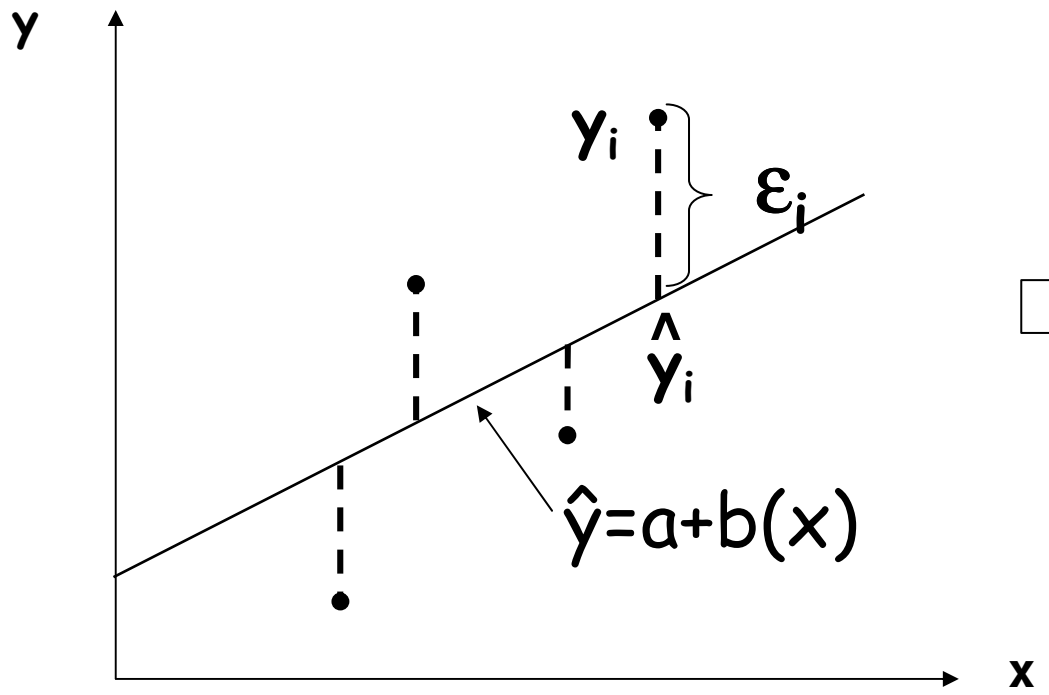
- Y, cioè il FEV1, dipende dall'altezza dell'individuo (X, parte sistematica) e da altre caratteristiche individuali (ε_i, parte probabilistica)


$N(\alpha + \beta X, \sigma^2)$
Y
$\alpha + \beta X$
x_1
x_2 ....
x_n
X
Un valore Y osservato

3. A questo punto, come scegliamo la retta che meglio si adatta ai nostri dati?

⇨ Come stimiamo α e β?

STIMA DEI PARAMETRI CON IL METODO DEI MINIMI QUADRATI



⇨ Cerchiamo la retta che rende minima la distanza tra y e $\hat{y}$, per ogni i

⇨ Cerchiamo a e b (stime di α e β) in modo da minimizzare la seguente quantità:

$$\sum_i \varepsilon_i^2 = \sum_i (y_i - \alpha - \beta x_i)^2 = \sum_i (y_i - \hat{y}_i)^2$$

⇩

$$b = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{\sum_i x_i y_i - \left[\left(\sum_i x_i\right)\right]\left(\sum_i y_i\right)/n}{\sum_i x_i^2 - \left(\sum_i x_i\right)^2/n} = \frac{codev.}{dev} = \frac{Sxy}{Sxx}$$

Stima dei parametri della retta di regressione

$$b = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{Sxy}{Sxx}$$

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

Si noti che il punto di coordinate $(\bar{x}, \bar{y})$ appartiene alla retta di regressione. Infatti:

$$\hat{y} = \bar{y} - b\bar{x} + bx \Rightarrow \hat{y} = \bar{y} + b(x - \bar{x})$$

E per $x = \bar{x} \Rightarrow \hat{y} = \bar{y}$

Quindi nell'esempio:

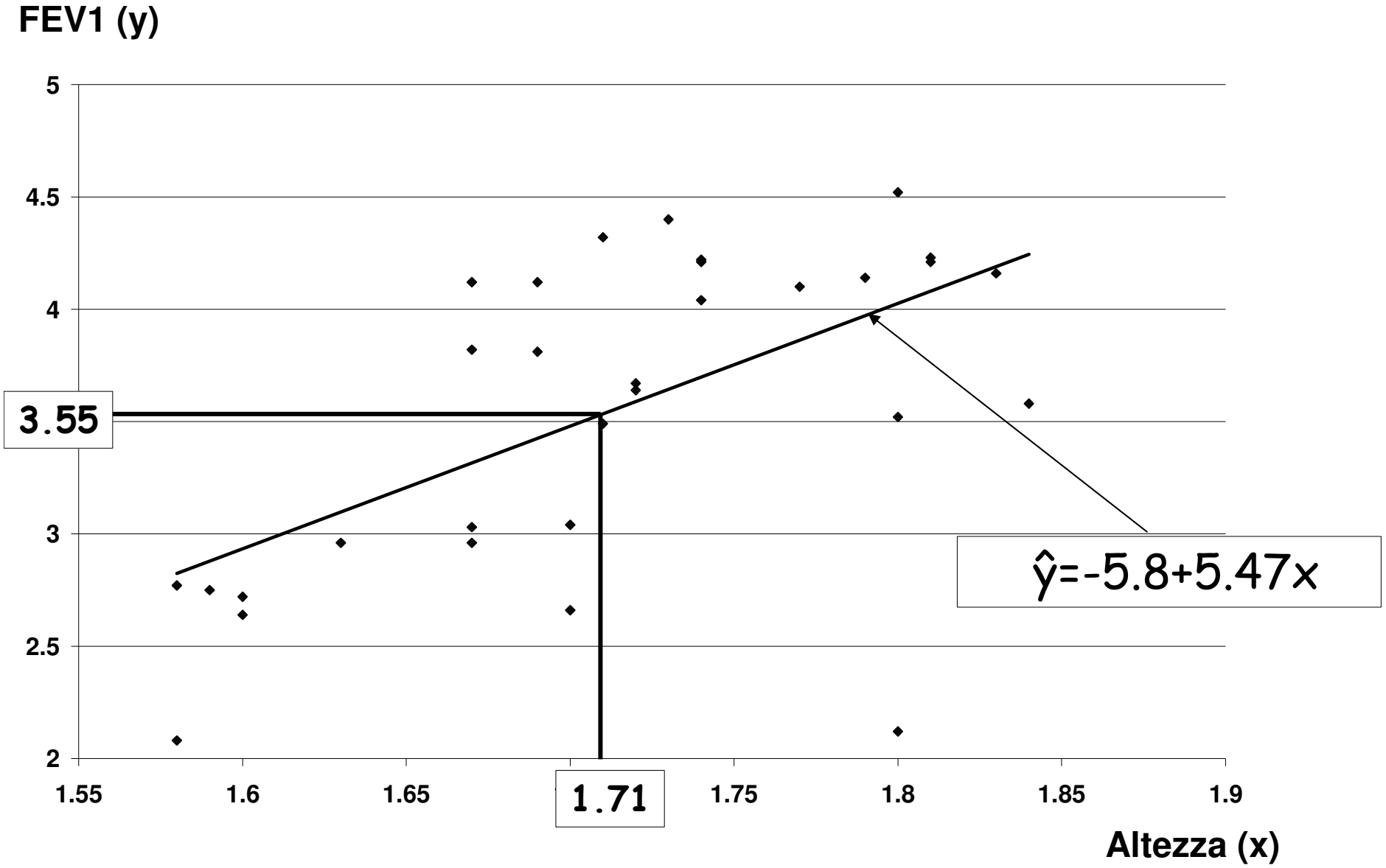
$\bar{x}=1.71, \qquad \bar{y}=3.55$

$Sxx=0.1748,\ Syy=15.1098,\ Sxy=0.9562$

$b=Sxy/Sxx=0.9562/0.1748=5.47$

$a=\bar{y}-b\bar{x}=3.55-5.47\cdot 1.71=-5.8$

$$\hat{y}=-5.8+5.47x$$



$(\overline{x}, \overline{y})$ appartengono alla retta di regressione

5. Stima della varianza residua

Varianza delle osservazioni, Y, intorno al modello di regressione (varianza d'errore)

$$s_e^2 = \frac{\sum_i (y_i - \hat{y})^2}{n-2} =$$

g.l=n°osservazioni-n° parametri

$$s_e^2 = (S_{yy} - \frac{S_{xy}^2}{S_{xx}}) / (n-2)$$

Dimostrazione:

$$\sum_i (y_i - \hat{y})^2 = \sum_i (y_i - a - bx_i)^2$$

$$= \sum_i (y_i - \bar{y} + b\bar{x} - bx_i)^2 = \sum_i \{(y_i - \bar{y}) - b(x_i - \bar{x})\}^2 =$$

$$= \sum_i (y_i - \bar{y})^2 + b^2 \sum (x_i - \bar{x})^2 - 2b \sum (y_i - \bar{y})(x_i - \bar{x})$$

- ($b = S_{XY}/S_{XX}$)

$$= S_{yy} + \frac{S_{xy}^2}{S_{xx}^2} S_{xx} - 2\frac{S_{xy}}{S_{xx}} S_{xy} = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

6. Errore standard di b e test per il modello di regressione

- Si può dimostrare che: $$ES(b) = \frac{s_e}{\sqrt{\sum \left(x_i - \bar{x}\right)^2}} = \frac{s_e}{\sqrt{Sxx}}$$

- La validità del modello viene valutata mediante il seguente sistema d'ipotesi

$$\begin{cases} H_0: \beta_0 = \beta = 0 \\ H_1: \beta \neq 0 \end{cases}$$

$$t = \frac{b - \beta_0}{ES(b)} \sim t_{n-2}$$

- intervallo di confidenza (95%) $b \pm t_{n-2,\, \alpha/2} \cdot ES(b)$

Nell'esempio;

$$s_e^2 = \frac{\sum_i (y_i - \hat{y})^2}{n-2} = 0.34$$

g.l.=31-2=29

$$ES(b) = \frac{s_e}{\sqrt{\sum \left(x_i - \bar{x}\right)^2}} = \frac{\sqrt{0.34}}{\sqrt{0.17}} = 1.396$$

$$t = \frac{b - \beta_0}{ES(b)} = \frac{5.47 - 0}{1.396} = 3.92$$

intervallo di confidenza (95%)

$$b \pm t_{0.025,29} \cdot ES(b) \Rightarrow 5.47 \pm 2.364 \cdot 1.396$$

$$5.47 \pm 3.30$$

$$(2.17;8.77)$$

7. Intervallo di confidenza della retta di regressione e utilizzo nelle previsioni

- Il modello di regressione ci permette di prevedere E(Y|X)

- Tali previsioni saranno affette da un margine di imprecisione, quantificato dall'intervallo di confidenza

- Ad esempio: si supponga di dover prevedere y per $x=x_0$

$\hat{y}=a+bx_0=\bar{y}+b(x_0-\bar{x})$

$Var(\hat{y})=Var(\bar{y})+(x_0-\bar{x})^2 var(b)=\frac{\sigma^2}{n}+\frac{(x_0-\bar{x})^2}{\sum(x_i-\bar{x})^2}.se^2$, stimata da:

$$Var(\hat{y})=s_e\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{Sxx}}$$

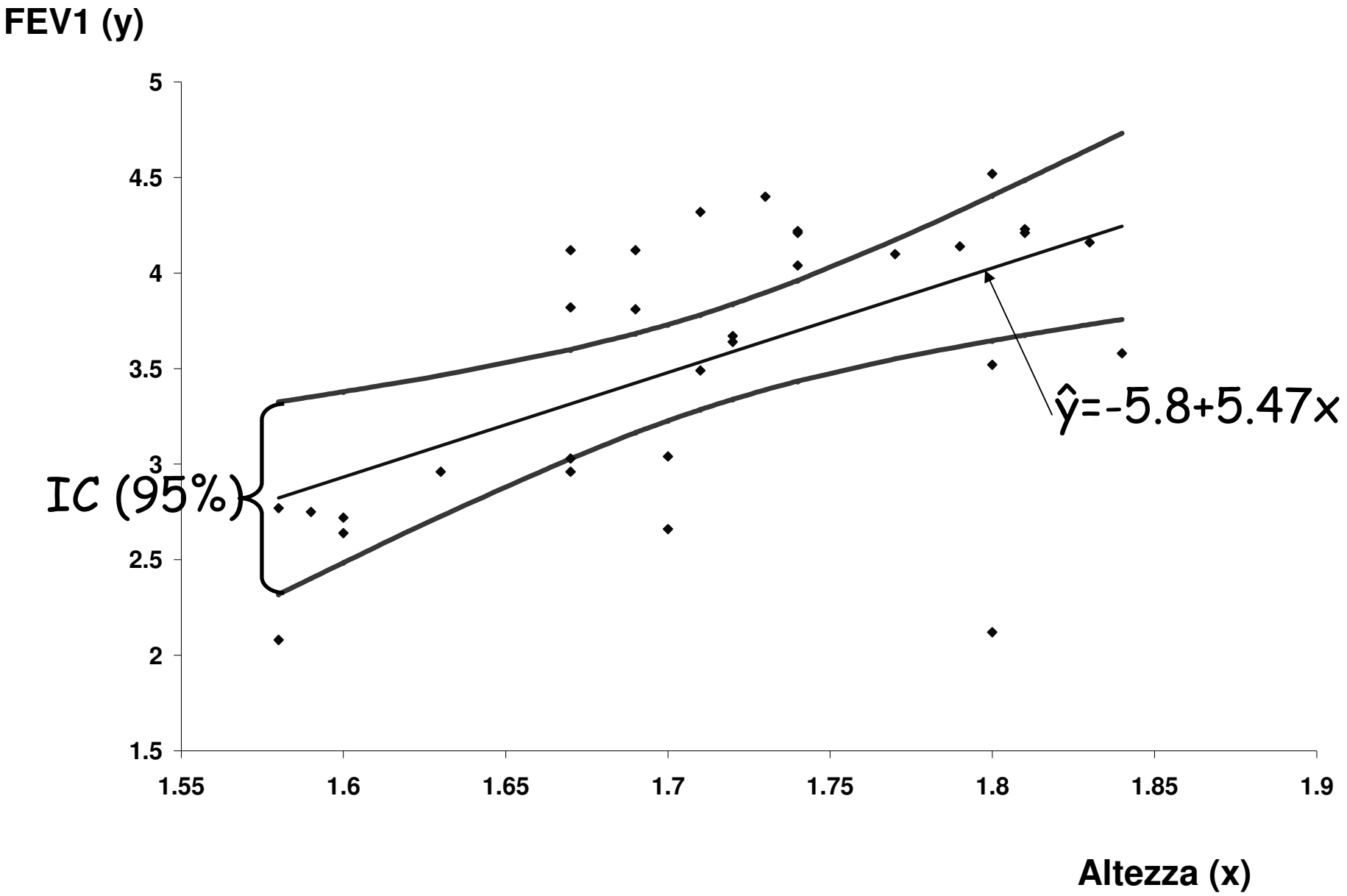
- L'intervallo di confidenza di $\hat{y}$ sarà:

$$\hat{y} \pm t_{n-2,\alpha/2}.ES(\hat{y})$$

Con $ES(\hat{y}) = \sqrt{Var(\hat{y})}$

- L'intervallo di confidenza cresce all'aumentare di $(\bar{x} - x_0)$
- È minimo quando $\bar{x} = x_0$

Retta di regressione e intervallo di confidenza (95%)



SCHEMA RIASSUNTIVO

UTILIZZO DEL MODELLO DI REGRESSIONE

Date n unità sperimentali in cui vengono misurate 2 variabili, X e Y, lo studio della relazione tra X e Y segue i seguenti punti:

•Plot descrittivo dei dati per valutare l'esistenza di una relazione lineare

•Si ipotizza che nell'universo esista una legge funzionale del tipo

$y=\alpha+\beta x$, con $\alpha+\beta$ parametri incogniti

- Si stimano con il metodo dei minimi quadrati i parametri ignoti

- Si procede all'inferenza sui parametri per valutare se i dati "supportano" il modello proposto

- Inferenza e previsione sui valori della variabile Y

- Analisi dei residui per ulteriore conferma del modello