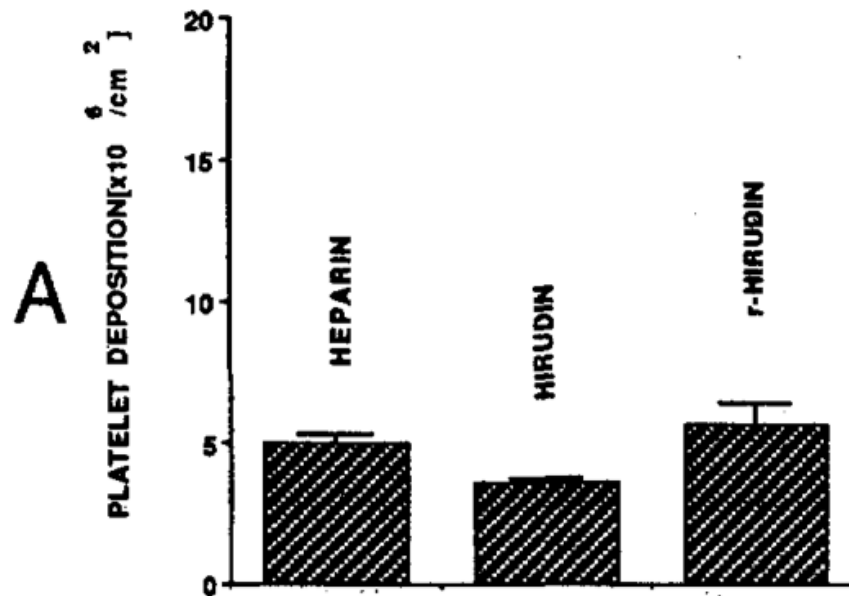


Modelli regressivi parametrici e non parametrici

Prof.ssa Marta Di Nicola

Thrombin Regulation of Platelet Interaction With Damaged Vessel Wall and Isolated Collagen Type I at Arterial Flow Conditions in a Porcine Model: Effects of Hirudins, Heparin, and Calcium Chelation

By L. Badimon, J.J. Badimon, R. Lassila, M. Heras, J.H. Chesebro, and V. Fuster



Analisi della Varianza (ANOVA)

- Confronto simultaneo tra più di due popolazioni,
- L' **analisi della varianza** estende il confronto a **p** gruppi con $p > 2$.
- **Fattore**: variabile utilizzata per differenziare un gruppo da un altro gruppo.
- **Livello (o trattamento)**: uno dei possibili valori/stati/caratteristiche che il fattore può assumere.
- **Variabile risposta**: variabile quantitativa oggetto dello studio.

Il disegno completamente randomizzato

- E' il disegno sperimentale più semplice
- Si utilizza quando si considera un solo fattore sperimentale a più livelli.
- I trattamenti/livelli sono assegnati alle unità sperimentali in modo casuale (randomizzazione).
- Se il numero di repliche è uguale per tutti i trattamenti il disegno è detto bilanciato (preferibile), altrimenti è detto sbilanciato.

In genere i livelli o gruppi possono essere non solo numerici ma anche **qualitativi**.

I fattori che definiscono i gruppi possono essere più di uno.

Con un solo fattore → **analisi della varianza ad un fattore o ad una via**

Con due (o più) fattori → **analisi della varianza a due (o più) fattori o a due vie (o più vie)**

Predisposizione dei dati

Fattore							
repliche	1	2		i		p	
1	y_{11}	y_{21}		y_{i1}	$\bar{y}$	y_{p1}	
2	y_{12}	y_{22}		y_{i2}		y_{p2}	
	...	...		...			
J	y_{1j}	y_{2j}		y_{ij}		y_{pj}	
	...						
n_i	$Y_{1n(i)}$	$Y_{2n(i)}$		$Y_{in(i)}$		$Y_{pn(i)}$	
Medie	$\bar{y}_1$	$\bar{y}_2$		$\bar{y}_i$		$\bar{y}_p$	

Il modello lineare

Il modello può essere rappresentato in questa forma:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

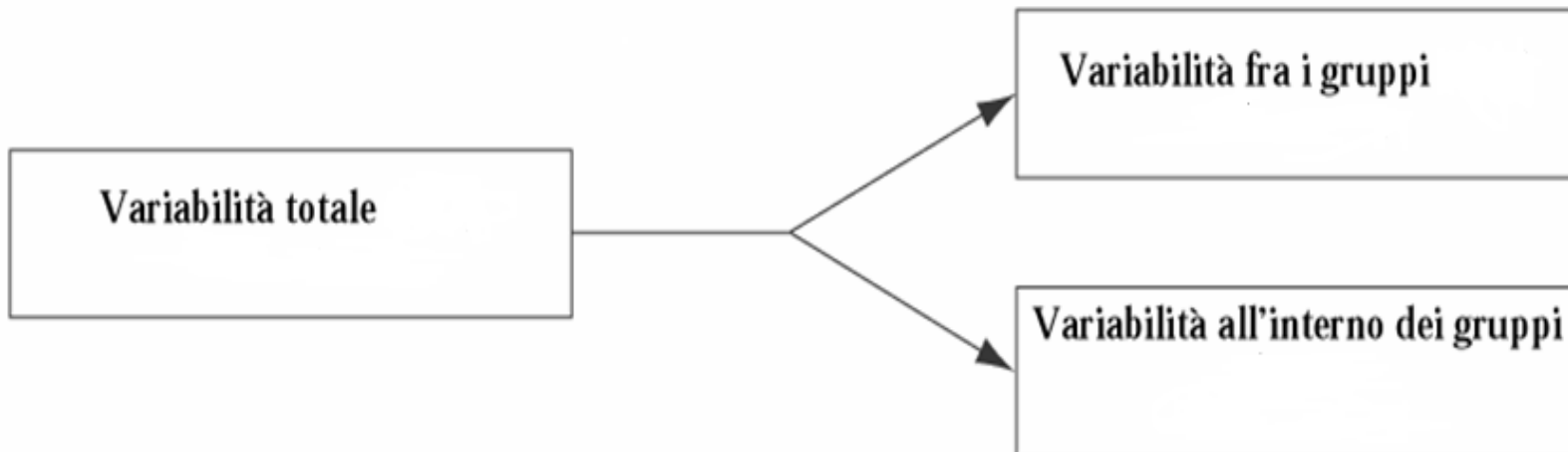
- con μ media di tutte le popolazioni rappresentate nell'esperimento
- $\alpha_i = \mu - \mu_i$ **effetto** dell'i-esimo trattamento/livello

Generalmente si assume:

- $i = 1, \dots, p$ (p numero dei livelli)
- $j = 1, \dots, n_i$ (n_i numero di repliche all'interno di un livello)
- $n_t =$ n.ro totale di osservazioni

Se il disegno è **bilanciato**, $n_1 = n_2 = \dots = n_p$

Scomposizione della variabilità totale



Variabilità all'interno dei gruppi (**MSE**) → **errore sperimentale**

Variabilità tra i gruppi (**MSA**) → **effetti del trattamento/livello**

Si ha che:

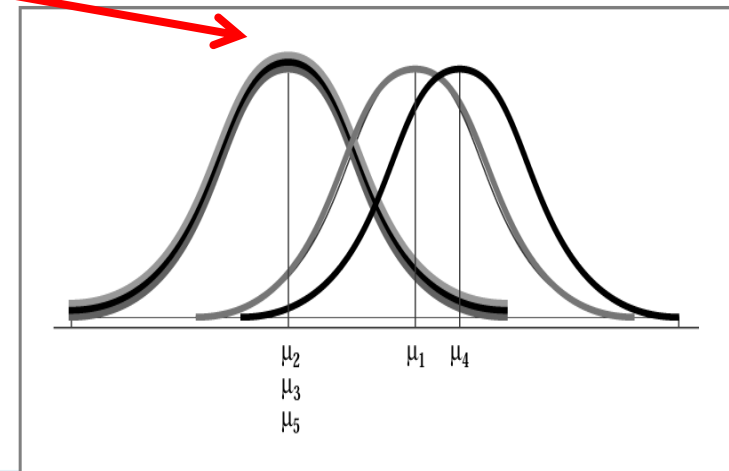
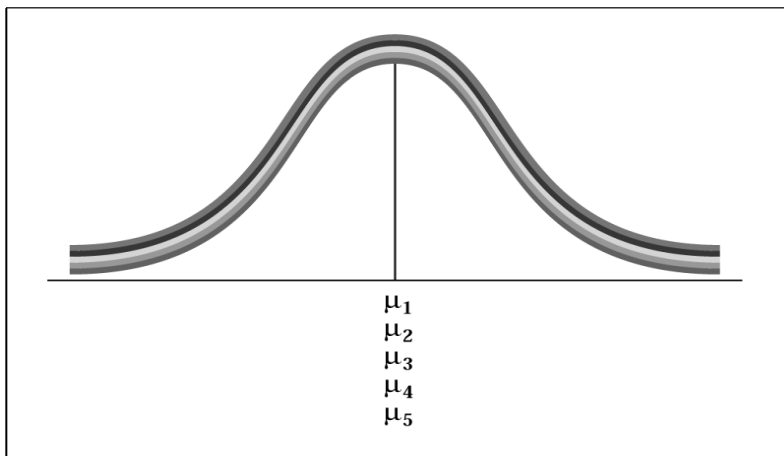
$$MST = MSA + MSE$$

Come fare inferenza

Assumendo che i p gruppi (popolazioni) da cui vengono estratte casualmente le osservazioni siano distribuiti normalmente e abbiano uguali varianze, l'ipotesi sottoposta a verifica è:

$$H_0: \mu_1 = \mu_2 = \dots = \mu_p \quad \text{oppure} \quad H_0 : \alpha_i = 0$$

H_A : non tutte le μ_i sono uguali



Come costruire il test?

Il test è basato sulle seguenti considerazioni:

- Se è **vera** l'ipotesi **nulla**, i dati differiscono tra loro per il **solo** effetto della variabilità **casuale**.
- Se invece è **vera** l'ipotesi **alternativa** (quindi rifiuto l' H_0), **entrambe** le fonti di variabilità contribuiscono a determinare la variabilità **complessiva**.
- Il test è quindi basato sull'analisi della variabilità complessiva in funzione delle diverse cause (da cui il termine **Analisi della Varianza**).

Test F per ANOVA a un fattore

Per verificare l'ipotesi di uguaglianza delle medie utilizzo il **test F** che confronta MSA e MSE.

$$F = \frac{MSA}{MSE} = \frac{\text{varianza} - \text{tra} - \text{gruppi}}{\text{varianza} - \text{entro} - \text{gruppi}}$$

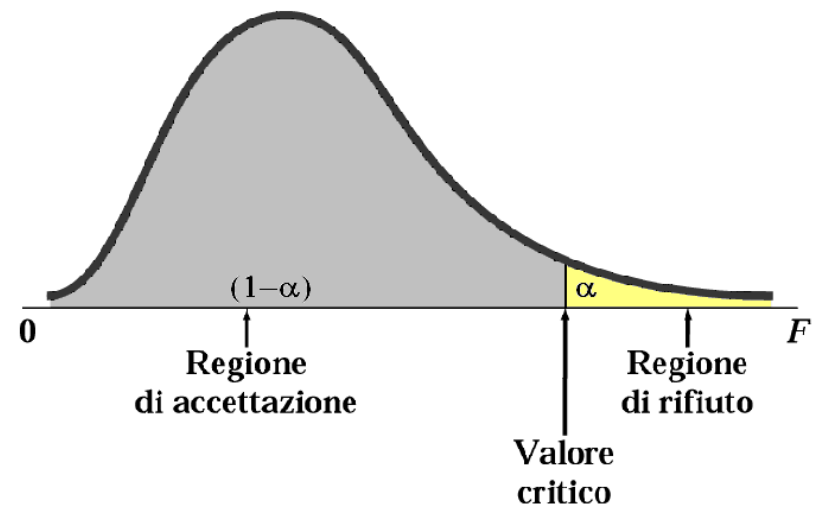
Il **test F** segue una **distribuzione F di Fisher** con $(p-1, n_t-p)$ gradi di libertà.

La regola decisionale è: **Rifiuto H_0 se $F > F_\alpha$**

Test F per la ANOVA a un fattore

Il valore critico della F viene determinato in funzione del livello di significatività α del test.

I valori critici si individuano nelle tavole della distribuzione F in base ai gradi di libertà e al livello di significatività scelto



Se H_0 è falsa ci aspettiamo che F assuma valori maggiori rispetto ai valori tabulati nella tavola della $F \rightarrow$ la variabilità totale è dovuta soprattutto all'effetto del trattamento/fattore.

Se H_0 è vera ci aspettiamo che il valore osservato di F sia minore al valore tabulato.

Multiple comparisons - post hoc

- Sono confronti che si fanno **a posteriori**, se l'Anova è significativa e se ci sono più di 2 gruppi in una variabile indipendente;
- La logica è quella di tenere sotto controllo i problemi di significatività legati ai confronti multipli;
- Vi sono diverse procedure di confronti alcuni presumono che le varianze siano uguali: LSD (Least Significant Difference), Bonferroni, Sidak, Scheffé, SNK (Student-Neumann-Kouls), Tukey HSD (Honestly Significant Difference), Duncan, Hochberg, Gabriel, Waller-Duncan, Dunnett;
- altre no: Tamhane, Dunnett, Games-Howell, C di Dunnett

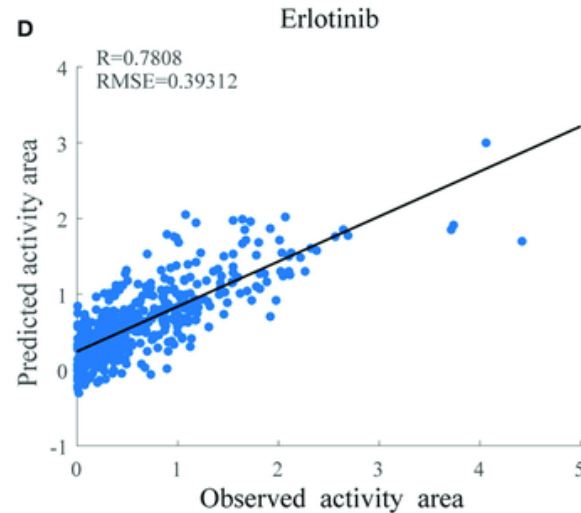
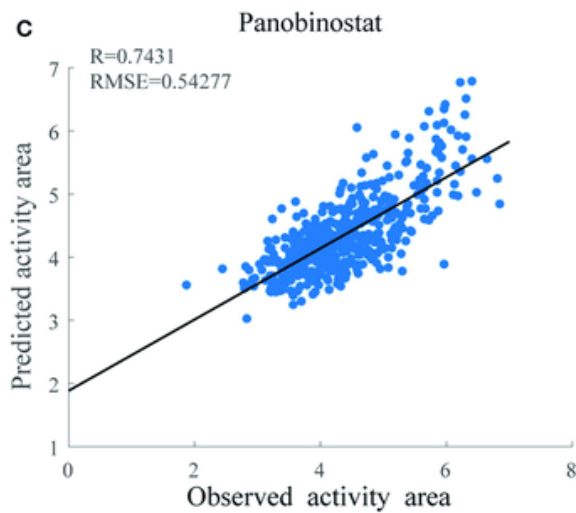
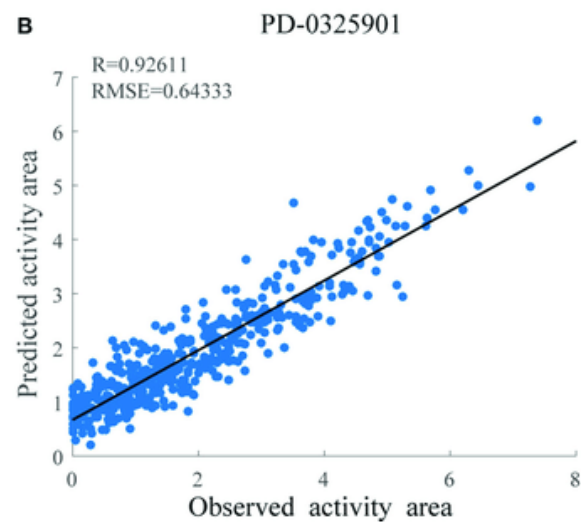
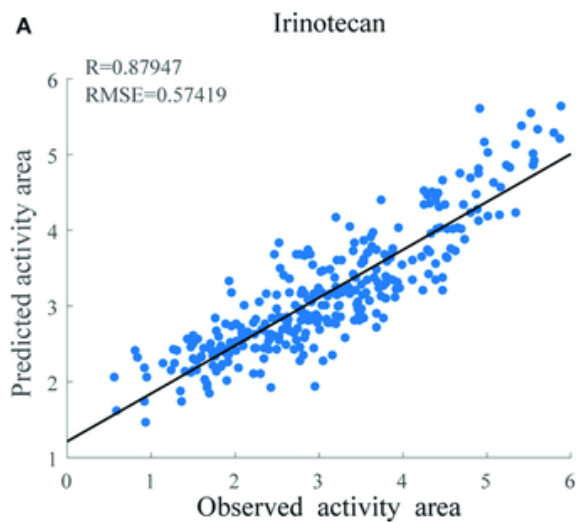


Table 3. Beta estimates and confidence interval for the association between DII and inflammation in the HELENA Study, 2006–2007

DII	Beta estimates and confidence interval for inflammatory markers ^{a)}					
	DII (continuous)	p-Value ^{c)}	Tertile 1	Tertile 2	Tertile 3	p _{trend} ^{b),c)}
CRP mg/L	0.04 (–0.04, 0.11)	0.32	Ref	0.18 (–0.08, 0.44)	0.09 (–0.18, 0.36)	0.41
TNF- α pg/mL	0.04 (0.002, 0.07)	0.03	Ref	–0.01 (–0.13, 0.11)	0.13 (0.007, 0.26)	0.06
IL-6 pg/mL	0.04 (–0.04, 0.13)	0.31	Ref	–0.05 (–0.35, 0.25)	0.09 (–0.22, 0.40)	0.66
IL-1 pg/mL	0.07 (–0.005, 0.15)	0.06	Ref	0.11 (–0.16, 0.38)	0.30 (0.02, 0.58)	0.04
IL-2 pg/mL	0.10 (–0.01, 0.20)	0.05	Ref	–0.14 (–0.50, 0.22)	0.42 (0.04, 0.79)	0.08
IL-4 pg/mL	0.08 (–0.04, 0.20)	0.20	Ref	–0.05 (–0.45, 0.35)	0.17 (–0.25, 0.59)	0.19
IL-10 pg/mL	0.01 (–0.06, 0.08)	0.69	Ref	–0.12 (–0.38, 0.13)	0.09 (–0.17, 0.35)	0.75
INF- γ pg/mL	0.13 (–0.007, 0.27)	0.04	Ref	0.25 (–0.22, 0.72)	0.58 (0.09, 1.06)	0.03
ICAM ng/mL	0.01 (–0.01, 0.04)	0.23	Ref	–0.04 (–0.13, 0.05)	0.02 (–0.08, 0.11)	0.27
VCAM ng/mL	0.03 (0.01, 0.04)	<0.001	Ref	0.03 (–0.03, 0.08)	0.07 (0.01, 0.13)	0.07

a) Adjusted for age, sex, body mass index (BMI) categories, city, smoking, and PA.

b) p_{trend} determined through the median approach.

c) p-Values < 0.005 were deemed significant after Bonferroni approach.

RELAZIONE TRA DUE VARIABILI QUANTITATIVE

Quando si considerano due o più caratteri (variabili) si possono esaminare anche il tipo e l'intensità delle relazioni che sussistono tra loro.

Nel caso in cui per ogni individuo si rilevino congiuntamente due variabili quantitative, è possibile verificare se esse variano simultaneamente e quale relazione “matematica” sussista tra queste variabili.

Si ricorre all'analisi della regressione e a quella della correlazione:

Analisi della regressione: per sviluppare un modello statistico che possa essere usato per prevedere i valori di una variabile, detta dipendente o più raramente predetta ed individuata come l'effetto, sulla base dei valori dell'altra variabile, detta indipendente o esplicativa, individuata come la causa.

Analisi della correlazione: per misurare l'intensità dell'associazione tra due variabili quantitative, di norma non legate direttamente da causa-effetto, facilmente mediate da almeno una terza variabile, ma che comunque variano congiuntamente.

Metodo dei minimi quadrati

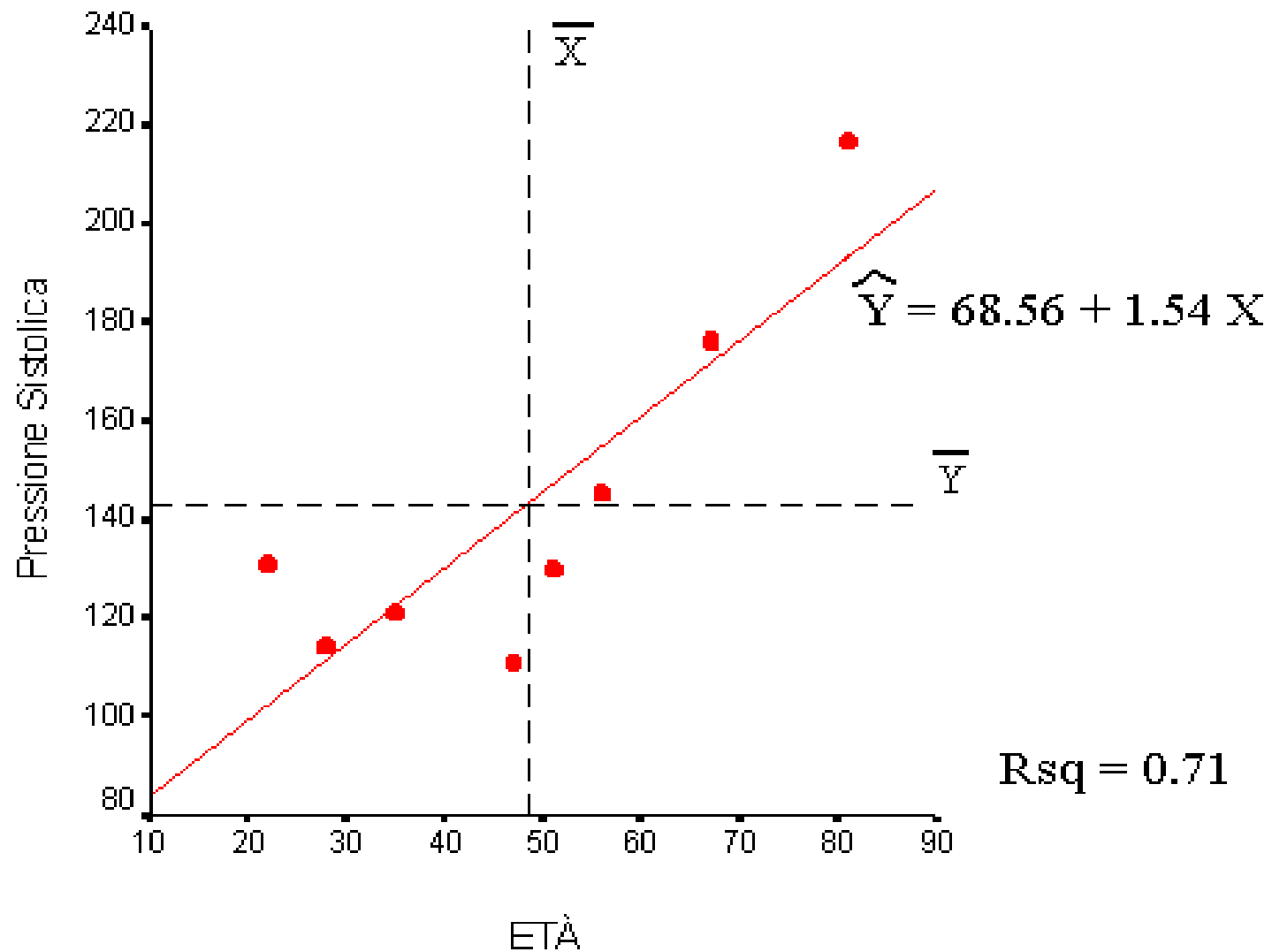
La retta migliore per rappresentare la distribuzione dei punti è quella che **minimizza** la somma:

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

Secondo il principio dei minimi quadrati si stimano matematicamente a e b:

$$b = \frac{\text{Codev}(x,y)}{\text{Dev}(x)}$$

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



ESEMPIO

	ETA' (X)	PAS (Y)	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(y - \bar{y})^2$	$(x - \bar{x})(y - \bar{y})$
1	22	131	-26.4	-12.1	696.96	146.41	+319.44
2	28	114	-20.4	-29.1	416.16	846.81	+593.64
3	35	121	-13.4	-22.1	179.56	488.41	+296.14
4	47	111	-1.4	-32.1	1.96	1030.41	+44.94
5	51	130	+2.6	-13.1	6.76	172.61	-34.06
6	56	145	+7.6	+1.9	57.76	3.61	+14.44
7	67	176	+18.6	+32.9	345.96	1082.41	+611.94
8	81	217	+32.6	+73.9	1062.76	5461.21	+2409.14
	48.4	143.1	0	0	2767.88	9230.88	4255.62
	MEDIA(X)	MEDIA (Y)			DEV(X)	DEV(Y)	CODEV(X,Y)

Si ottiene:

coefficiente di regressione $b = \frac{4255.62}{2767.88} = 1.54$

Intercetta $a = 143.1 - 1.54 * 48.4 = 68.56$

Interpretando i valori dei coefficienti della retta di regressione si può dire:

1. l'aumento medio della pressione è di circa **$b=1.5$ mmHg** per l'aumento di un anno di età.
2. alla nascita il valore della pressione *sarebbe* (!) di **$a=68.56$ mmHg**, ma questa è una indicazione teorica perché non è possibile stimare il valore della pressione arteriosa per età fuori del range considerato (22-81 aa).

L'intercetta è quel valore che assume la variabile dipendente quando quella indipendente è uguale a 0.

Coefficiente di Determinazione

Espresso a volte in percentuale ed indicato in alcuni testi con **R²** o **Rsq**, serve per misurare “quanto” della variabile dipendente Y sia predetto dalla variabile indipendente X e, quindi, per valutare la bontà dell’equazione di regressione ai fini della previsione sui valori della Y .

E' una misura che ha scopi descrittivi dei dati raccolti. Non è legata ad inferenze statistiche, ma a scopi pratici, specifici dell'uso della regressione come metodo per prevedere Y conoscendo X .

Il suo valore, compreso tra 0 e 1, è tanto più elevato quanto più la retta passa vicino ai punti, fino a raggiungere 1 (o 100%) quando tutti i punti sperimentali sono collocati esattamente sulla retta e quindi ogni Y_i può essere predetto con precisione totale dal corrispondente valore di X_i

Nell'esempio con le 8 osservazioni di età e pressione il valore del coefficiente di determinazione è:

$$R^2 = \frac{6543.1}{9230.9} = 0.71$$

Ciò significa che, noto il valore dell'età, quello della pressione è stimato mediante la retta di regressione con una approssimazione di circa il 71%.

Il restante $1-r^2=29\%$ è determinato dalla variabilità individuale di scostamento dalla retta ed indica la parte di variabilità della variabile risposta imputabile eventualmente ad altri fattori diversi dall'età.

La valutazione del valore di r^2 è in stretto rapporto con la disciplina oggetto di studio. Si può ritenere in alcuni ambiti che il modello lineare abbia un **buon fitting** con i valori sperimentali se $r^2 > 0.6$, ma va detto anche che nelle scienze sociali spesso si reputa alto un valore uguale a 0.30 mentre i fisici stimano basso un valore pari a 0.98.

CORRELAZIONE LINEARE SEMPLICE

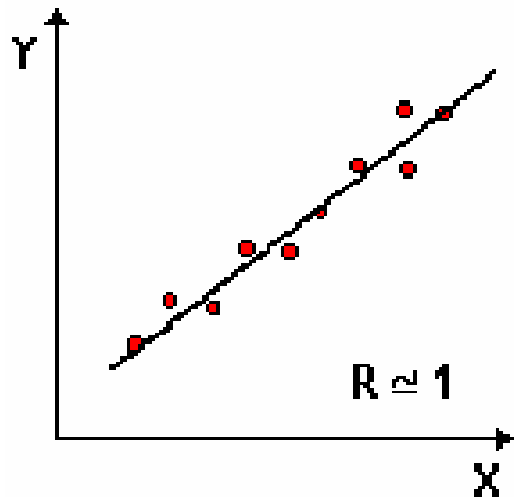
Una misura della bontà del modello lineare può essere ottenuta studiando l'**interdipendenza** tra due caratteri statistici quantitativi X e Y.

Uno degli indici molto noto per una tale misura è il **Coefficiente di Correlazione Lineare r**:

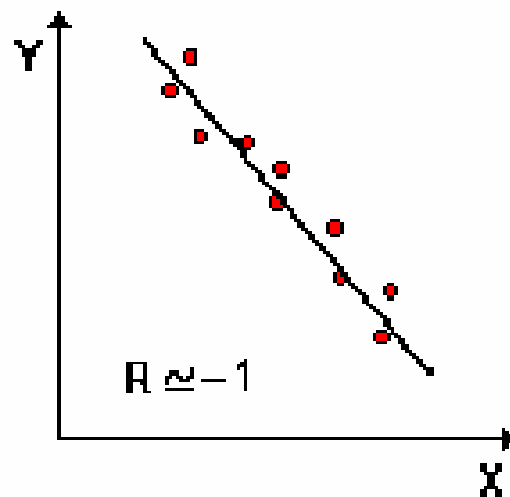
$$r = \frac{\text{CODEV}(X, Y)}{\sqrt{\text{DEV}(X) \cdot \text{DEV}(Y)}}$$

Tale quantità, indicata anche con **R**, varia tra -1 e 1 .

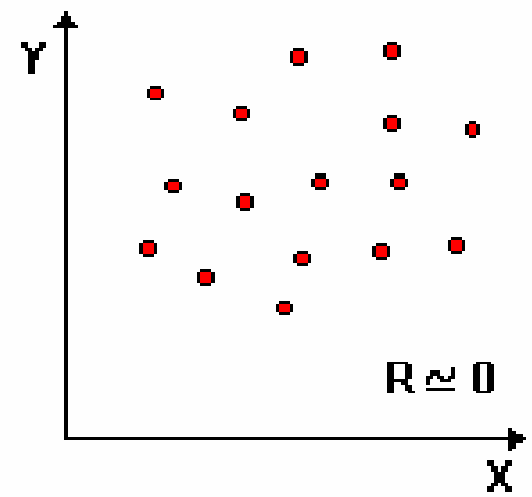
- Un valore di **r** vicino a 1 indica una associazione stretta o molto stretta tra le due variabili; si parla in tal caso di *correlazione lineare positiva* tra X e Y : all'aumentare di una variabile aumenta anche l'altra.
- Un valore di **r** vicino a -1 denota un'alta o molto alta *correlazione lineare negativa (discordanza)* tra X e Y : all'aumentare di una di esse l'altra diminuisce.
- Un valore di **r** = 0 o prossimo a 0 indica *indifferenza (indipendenza)* tra le variabili.



**Correlazione lineare
positiva tra X e Y**



**Correlazione lineare
negativa tra X e Y**



**Variabili X e Y
incorrelate**

Diagrammi di dispersione in alcuni casi di correlazione tra due variabili

L'analisi della correlazione misura **solo** il grado di associazione spazio-temporale di due fenomeni; il coefficiente **r** è semplicemente una misura dell'intensità dell'associazione tra due variabili.

Nell'esempio, utilizzando i calcoli della Tabella, si ha:

$$r = \frac{+4255.62}{\sqrt{2767.88 \cdot 9230.88}} = +0.842$$

e si registra, quindi, un apprezzabile grado di correlazione lineare positiva tra l'età e la pressione sistolica *per i dati presi in esame*.

Valori di r intorno all'80% o superiori possono, in teoria, far ritenere buona l'associazione lineare: ma va tenuto conto dell'ambito disciplinare e della numerosità dei dati.

CENNI REGRESSIONE LINEARE MULTIPLA



Soggetto	Sesso	Età	PAS	PAD	Fumo
1	M	22	131	70	5
2	F	28	114	75	8
3	M	35	121	80	30
4	F	47	111	75	20
5	F	51	130	70	15
6	M	56	145	80	0
7	M	67	176	85	25
8	M	81	217	90	10

$$\text{PAS} = 69 + 1.53 \text{ Et\`a}$$

$$\text{PAS} = 75 + 1.55 \text{ Et\`a} - 0.54 \text{ Fumo}$$