STATISTICA MEDICA

Dott.ssa Marta Di Nicola
N.P.D. 3° Blocco 2° piano
0871-3554007
m.dinicola@unich.it
http://www.biostatistica.unich.it

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.

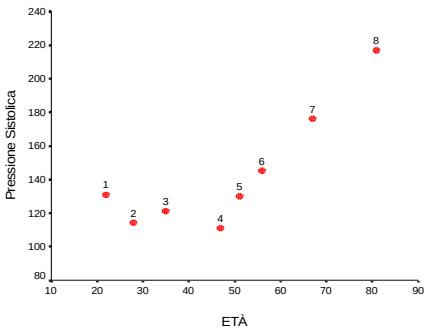
Analisi della correlazione: per misurare l'intensità dell'associazione tra due variabili quantitative, di norma non legate direttamente da causa-effetto, facilmente mediate da almeno una terza variabile, ma che comunque variano congiuntamente.

ESEMPIO

In Tabella sono riportati i valori assunti dai due caratteri quantitativi età (ETA') e pressione sistolica (PAS) misurati in un campione di 8 soggetti:

soggetto	ETA' (anni)	PAS (mm Hg)
1	22	131
2	28	114
3	35	121
4	47	111
5	51	130
6	56	145
7	67	176
8	81	217

Diagramma di Dispersione (Scatter)



Domande

•Di quanto varia la pressione sistolica all'aumentare dell'età ?

•La relazione tra le due variabili è tendenzialmente lineare?

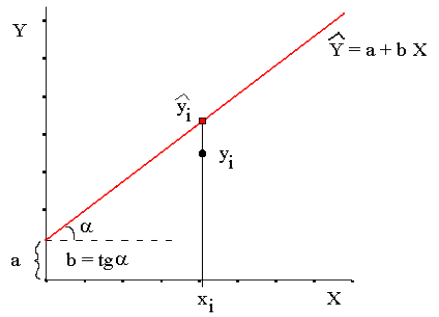
REGRESSIONE LINEARE SEMPLICE

La "forma" di relazione matematica più semplice tra due variabili è la regressione lineare semplice, rappresentata dalla retta di regressione:

$$\hat{Y}=a+b\cdot X$$

dove :

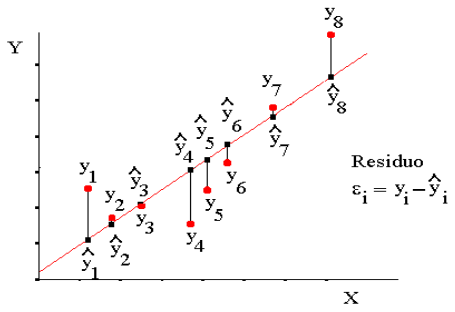
- $\hat{Y}$ valore stimato di Y attraverso il modello regressivo
- X valore empirico di X
- a intercetta della retta di regressione
- b coefficiente di regressione (coefficiente angolare della retta)



Per stimare la retta che meglio approssima la distribuzione dei punti, si può partire considerando che ogni punto osservato Yi si discosta dalla retta di una certa quantità detta errore o **residuo**

Ogni valore di residuo può essere positivo o negativo:

- positivo quando il punto Y sperimentale è sopra la retta
- negativo quando il punto Y sperimentale è sotto la retta



Metodo dei minimi quadrati

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

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

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

$$b=\frac{\text{Codev}(x,y)}{\text{Dev}(x)} \qquad a=\bar{y}-b\cdot\bar{x}$$

$$\text{Codev}(x,y)=\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

$$\text{Dev}(x)=\sum_{i=1}^{n}(x_i-\bar{x})^2$$

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

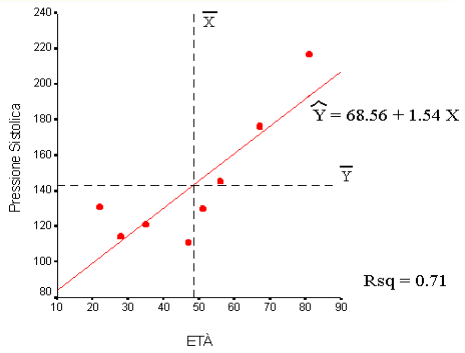
ESEMPIO

	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	-43.06
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
	$\bar{x}$=48.4	$\bar{y}$=143.1	**0**	**0**	**2767.88**	**9230.88**	**4255.62**
					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



Supposto "accettabile" il modello regressivo lineare, affrontiamo le seguenti domande:

1. di quanto aumenta ***mediamente*** la pressione sistolica all'aumentare di un anno di età?
2. che valore ha la PAS alla nascita?

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.

Valore predittivo dell'analisi della regressione

La semplice rappresentazione grafica dei valori osservati e della retta di regressione fornisce alcune indicazioni importanti per l'interpretazione delle relazioni esistenti tra le due variabili.

Il valore del coefficiente angolare indica quanto aumenta in media la variabile dipendente Y all'aumento di una unità della variabile indipendente X. Se si cambia la scala della variabile indipendente o predittiva X (per esempio l'altezza misurata in mm o in m e non più in cm) lasciando invariata quella della variabile dipendente o predetta Y, muta proporzionalmente anche il valore del coefficiente angolare b.

Nell'analisi della regressione:

• è frequente, specialmente negli utilizzi predittivi, il ricorso al tempo come variabile indipendente;
• viene spesso dimenticato che qualsiasi previsione o stima di Y derivata dalla retta è valida solo entro il campo di variazione della variabile indipendente X;
• non è dimostrato che la relazione esistente tra le due variabili sia dello stesso tipo anche per valori minori o maggiori di quelli sperimentali rilevati.

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 statistiche quantitativi X e Y.

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

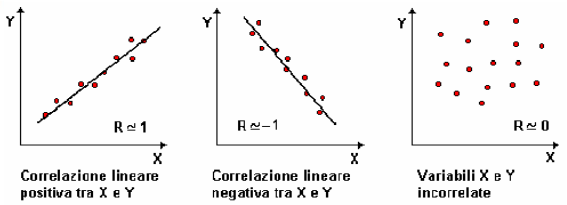
$$R = \frac{\text{CODEV}(X,Y)}{\sqrt{\text{DEV}(X)\cdot\text{DEV}(Y)}}$$

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.



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

PAS = 69 + 1.53 Età

PAS = 75 + 1.55 Età – 0.54 Fumo