RELAZIONE TRA VARIABILI QUANTITATIVE

Lezione 7ª

Accade spesso nella ricerca in campo biomedico, così come in altri campi della scienza, di voler studiare come il variare di una o più variabili (variabili indipendenti o esplicative e regressori) modifichi un'altra variabile (variabile dipendente o risposta, etc).Quando le variabili indipendenti sono ≥2 si parla di Regressione multipla, quando è una sola si parla di regressione semplice.

Se entrambe le variabili (o caratteri) sono quantitative si calcola:

1. La REGRESSIONE se è ipotizzabile un rapporto di "causa-effetto" tra le variabili;
2. La CORRELAZIONE se non è ipotizzabile un rapporto di causa-effetto tra le variabili ma queste dipendono, almeno in parte, da cause comuni.

1. LA REGRESSIONE

La regressione studia il tipo e il grado di dipendenza tra due variabili quantitative ossia di "quanto" varia..

L'obiettivo della regressione è quello di trovare l'equazione di una curva che meglio interpreta il meccanismo con il quale una variabile è relazionata ad un'altra.

Usi della regressione

1. DESCRITTIVO: per descrivere la relazione tra due variabili, ossia di quanto Y è associata ad X;
2. INTERPRETATIVO: per interpretare il meccanismo con il quale una variabile è relazionata ad un'altra;

3. Come modello PREVISIVO: per prevedere un "livello" di Y per un nuovo valore di X;
4. Come strumento di RELAZIONE, non sempre di casualita' (Ad esempio, la regressione tra altezza e peso non significa che un incremento di peso causa un aumento di altezza);
5. Per scoprire eventuali dati anomali (DATI ABBERRANTI).

In questo capitolo si tratterà solamente della Regressione lineare semplice o bivariata.

REGRESSIONE SEMPLICE O BIVARIATA

Per analisi bivariata si intende l'applicazione di una serie di metodologie statistiche al fine di individuare e studiare le eventuali relazioni intercorrenti tra due variabili (es. tra età e pressione sistolica, tra età e altezza, tra BMI e pressione diastolica, etc.).

Per regressione lineare si intende una procedura che permette di trovare una funzione di primo grado (lineare) del tipo

$y = a + bx$ che descriva il legame esistente tra una variabile y considerata dipendente (variabile risposta) ed una variabile x considerata indipendente (variabile esplicativa).

La procedura matematica che permette di trovare le formule per calcolare i parametri a e b della funzione è chiamata metodo dei minimi quadrati. Il metodo della regressione, quale tecnica per l'analisi delle relazioni intercorrenti tra due variabili, può essere usato solamente se si conosce quale delle due variabili è quella dipendente e quale quella indipendente.

Facendo riferimento ad una distribuzione doppia (su ogni unità statistica (u.s.) vengono rilevate due variabili statistiche X e Y) del tipo:

u.s.	1	2	3	...	...	i	...	...	N
X	x_1	x_2	x_3	...	...	x_i	...	...	x_N
Y	y_1	y_2	y_3	...	...	y_i	...	...	y_N

Come primo step essa viene graficata mediante un diagramma a dispersione (nuvola di punti).

Ogni punto, individuato dalla coppia di valori (x_i, y_i), rispettivamente ascissa e ordinata, rappresenta una unità statistica.

Se la nuvola di punti presenta caratteristiche di linearità (i punti si dispongono approssimativamente lungo una ipotetica retta) l'obiettivo sarà quello di trovare i valori dei parametri a e b dell'equazione della retta $y = a + bx$ che meglio interpola la nuvola di punti.

Operativamente la procedura che permette di calcolare i parametri a e b è quella dei " metodo dei minimi quadrati" che minimizza la quantità $\sum (y_i - y^*_i)^2$, in cui le y^*_i rappresentano il valore teorico della variabile Y (calcolato mediante la funzione) in corrispondenza del valore empirico x_i. Le formule di tipo statistico che permettono di calcolare a e b sono:

$$b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{CODEVIANZA(xy)}{DEVIANZA(x)} = \frac{\sigma_{xy}}{\sigma_x^2}$$

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

Con $\overline{y}$ $\overline{x}$ la media delle variabili x e y rispettivamente.

Da un punto di vista statistico l'equazione rappresenta la formalizzazione matematica del legame associativo tra le due variabili *x* e *y* e il parametro *b*, detto **coefficiente di regressione**, rappresenta la variazione della variabile dipendente *y* conseguente ad una variazione unitaria della variabile indipendente *x*; il parametro *a* rappresenta, invece, il valore della funzione $y=a+bx$ corrispondente al valore 0 della variabile *x*. Il segno del coefficiente di regressione *b* indica il "verso" della relazione: il segno positivo indica una concordanza tra le variabili (ad un aumento della *x* corrisponde un aumento della *y*), il segno negativo una discordanza (ad un aumento della *x* corrisponde una diminuzione della *y*). Il valore assoluto della *b* indica il "grado" della relazione (quanto più il valore di *b* è grande tanto più la variabile *x* influenza la variabile *y*).

La regressione multipla studia l'influenza di due o più variabili esplicative su una variabile dipendente. Ossia come quest'ultima è determinata da almeno altre due variabili.

Il modello sarà del tipo:

$y = b_0 + b_1X_1 + b_2X_2 + \ldots + b_nX_n + \varepsilon$

dove bi = sono i coefficienti di regressione parziale e misurano il contributo di ciascuna Xi al variare di y e ε una variabile casuale che rappresenta, in estrema sintesi,

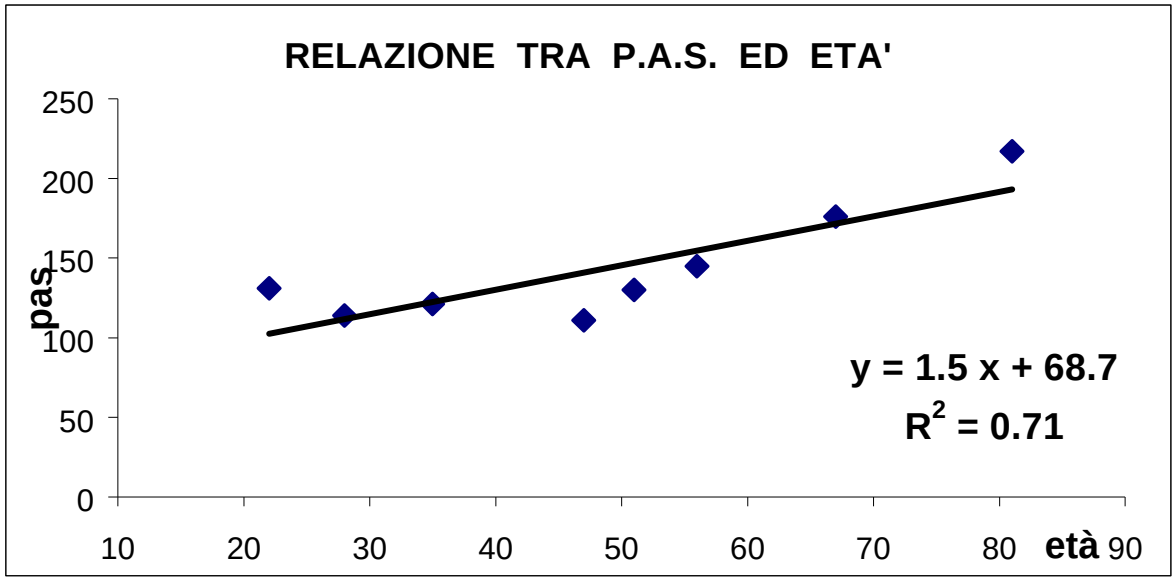
lo scarto tra le y_i osservate ed i corrispondenti valori medi teorici della y risultanti dalla funzione.

Esempio

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

Rappresentato graficamente si ha:



L'andamento dei punti nel grafico a dispersione (i punti si distribuiscono approssimativamente lungo una retta) suggerisce l'esistenza tra le due variabili di una relazione di tipo lineare. Sapendo che tra i due caratteri l'età rappresenta la variabile

indipendente e la PAS quella dipendente la relazione può essere descritta mediante la retta di regressione y=a+bx.

Pts	x	y	$(xi-\overline{x})$	$(yi-\overline{y})$	$(xi-\overline{x})(yi-\overline{y})$	$(xi-\overline{x})^2$	$(yi-\overline{y})^2$
1	22	131	-26.4	-12.1	+319.44	696.96	146.41
2	28	114	-20.4	-29.1	+593.64	416.16	846.81
3	35	121	-13.4	-22.1	+296.14	179.56	488.41
4	47	111	-1.4	-32.1	+44.94	1.96	1030.41
5	51	130	+2.6	-13.1	-34.06	6.76	172.61
6	56	145	+7.6	+1.9	+14.44	57.76	3.61
7	67	176	+18.6	+32.9	+611.94	345.96	1082.41
8	81	217	+32.6	+73.9	+2409.14	1062.76	5461.21
Σ	387	1145			4255.62	2767.88	9230.88

Media Aritmetica $\overline{X}$= 387/8=48.4

$\overline{Y}$= 1145/8=143.1

Codevianza (XY)=4255.62

Devianza (X)=2767.88

b= 4255.62/2767.88=1.54 **a**= 143.1-(1.5*48.4)=68.6

y = 68.6+1.54x

Interpretazione statistica dei parametri: il coefficiente di regressione (b=1.54) sta ad indicare che, teoricamente, la PAS aumenta, mediamente, di 1.54 mm Hg per ogni anno di età, il valore di a (68.6) rappresenta il valore teorico della PAS corrispondente all'età 0.

Congiuntamente alla retta di regressione si è soliti calcolare l'indice R^2 (coefficiente di determinazione) che valuta la bontà di adattamento della retta alla nuvola di punti. Questo indice che assume valori compresi tra 0 ed 1 (quanto più assume valori vicini ad 1 tanto più la retta approssima bene i punti). In sostanza misura l'attendibilità della relazione formalizzata mediante la funzione lineare (come si vedrà nel prossimo paragrafo) .

La retta di regressione del precedente esempio descrive abbastanza bene la relazione in quanto si riscontra un $R^2 = 0.71$.

2. LA CORRELAZIONE

Nel caso in cui tra le due variabili non si può ipotizzare un legame di dipendenza del tipo "causa-effetto", per lo studio dell'eventuale relazione (chiamata in questo caso interdipendenza), viene usato il metodo della correlazione lineare. Questo metodo consiste nel calcolare e nell'interpretare poi, l'indice R (coefficiente di correlazione lineare del Pearson) che quantifica il "verso" (concorde o discorde) ed il grado della relazione tra le variabili.

L'indice R può essere calcolato mediante la formula:

$$R = \pm \sqrt{b \cdot b'} = \frac{COVARIANZA(xy)}{\sqrt{DEV(x)\,DEV(y)}}$$

R è un numero che può assumere valori compresi tra -1 e +1. Il segno di *R* indica il verso della relazione, il valore assoluto il grado. Quanto più il valore assoluto si avvicina ad 1 tanto più vi è interdipendenza (di tipo lineare) tra le variabili. Viceversa,

quanto più si avvicina a 0 tanto più le variabili possono essere considerate linearmente indipendente.

Esempio

Si sia effettuato un certo esperimento riguardante i rapporti tra il peso di 5 cavie e la loro resistenza alla tossina difterica (misurata come tempo di morte). Si sono ottenuti i seguenti valori:

CAVIE		1	2	3	4	5	Σ
Peso (in gr)	X	220	300	210	350	270	1350
Tempo di morte (ore)	Y	32	38	27	50	25	172

PESO MEDIO=1350/5=270

TEMPO MEDIO=172/5=34.4

Tabella per il calcolo dei coefficienti b e R:

Pts	x	y	$(xi- \overline{x})$	$(yi-\overline{y})$	$(xi- \overline{x})(yi-\overline{y})$	$(xi- \overline{x})^2$	$(yi-\overline{y})^2$
1	220	32	-50	-2.4	120	2500	5.72
2	300	38	+30	+3.6	108	900	12.96
3	210	27	-60	-7.4	444	3600	54.76
4	350	50	+80	+15.6	1248	6400	243.36
5	270	25	0	-9.4	0	0	88.36
Σ	1350	172			1920	13400	405.2

Quindi b=1920/13400=0.14

$$R = \frac{\sum (xi - \bar{x})(yi - \bar{y})}{\sqrt{\sum (xi - \bar{x})^2 (yi - \bar{y})^2}} = \frac{1920}{\sqrt{13400 * 405.2}} = 0.82$$

R^2=67%

In definitiva esiste una concordanza positiva (R=0.82) tra peso della cavia e tempo di sopravvivenza.

All'aumentare di 10 grammi di peso delle cavie s'incrementa mediamente di circa 1.4 ore il tempo di sopravvivenza.

Solamente il 67% della variabilità del tempo di sopravvivenza è spiegabile mediante il peso delle cavie.

MISURE DI ASSOCIAZIONE

Studio l'associazione tra due o più variabili rilevate simultaneamente sulle unità statistiche per definire il loro reciproco comportamento.

1. MISURE DI CORRELAZIONE

Studia l'interdipendenza (reciproche relazioni) tra due o più variabili mediante:

- la codevianza
- il coefficiente di correlazione

2. MISURE DELLA DIPENDENZA

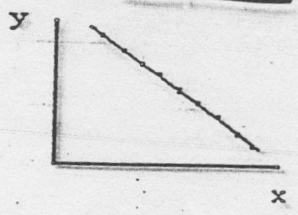
Ricerca l'esistenza di una relazione di dipendenza tra due variabili:

- il coefficiente di regressione (se quantitative)
- ρ di Spearman (se ordinali)
- il Chi-quadrato (se qualitative)

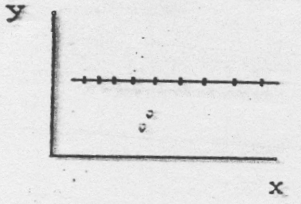
COEFFICIENTE DI CORRELAZIONE LINEARE

r

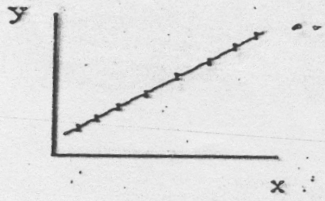
-1 — i due caratteri sono perfettamente correlati: all'aumentare di una variabile l'altra diminuisce



0 — i due fenomeni non sono correlati (sono indipendenti)



+1 — i due caratteri sono perfettamente correlati: all'aumentare di una variabile aumenta anche l'altra



COEFFICIENTE DI CORRELAZIONE LINEARE

-1

correlazione inversa

0

correlazione diretta

+1