Unità 11

Studio di più variabili

Interpolazione

Regressione

Correlazione

Notazione matriciale

INTERPOLAZIONE

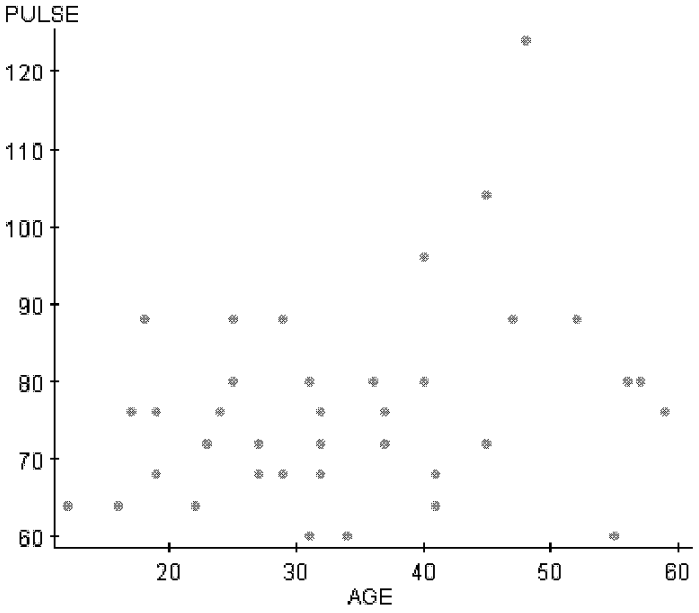
Può spesso capitare di scoprire l'esistenza di una relazione sperimentale tra due (o più) variabili ed è allora naturale ricercare un'espressione matematica ***(equazione)*** che leghi le variabili in questione.

In questa sede ci limiteremo ad analizzare il caso in cui si considerino solo due variabili (***X*** e ***Y***).

Dopo avere raccolto i dati che forniscono i valori corrispondenti delle variabili *X* e *Y*, le osservazioni ottenute possono essere rappresentate graficamente in un sistema di coordinate cartesiane.

La rappresentazione grafica che si ottiene viene detta ***diagramma*** (o ***grafico***) ***a dispersione*** o ***scatter plot***.

La figura sotto mostra un esempio di diagramma a dispersione ottenuto in un campione di donne, in cui nell'asse delle ascisse è riportata l'età (anni) e in quello delle ordinate sono riportate le pulsazioni cardiache (battiti al minuto).



Dall'analisi dello scatter plot è spesso possibile avere un'idea intuitiva dell'esistenza o meno di una possibile relazione fra *X* e *Y* e dell'andamento di una curva che passa *"abbastanza vicino ai dati"*.

Una curva di questo genere è detta **curva interpolatrice**.

Nell'esempio in ***Figura 1*** i dati sembrano ***bene interpolati da una retta***. In questo caso possiamo pensare che tra le due variabili esista una ***relazione lineare***.

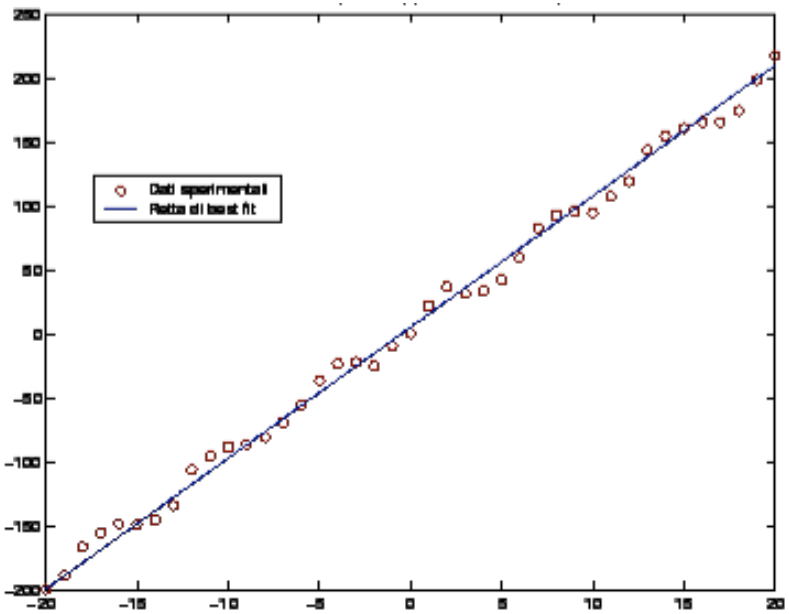


Figura 1

Considerando invece l'esempio in ***Figura 2*** si può ancora ipotizzare che esista una relazione fra *X* e *Y*, ma in questo caso la ***relazione*** è ***non lineare***.

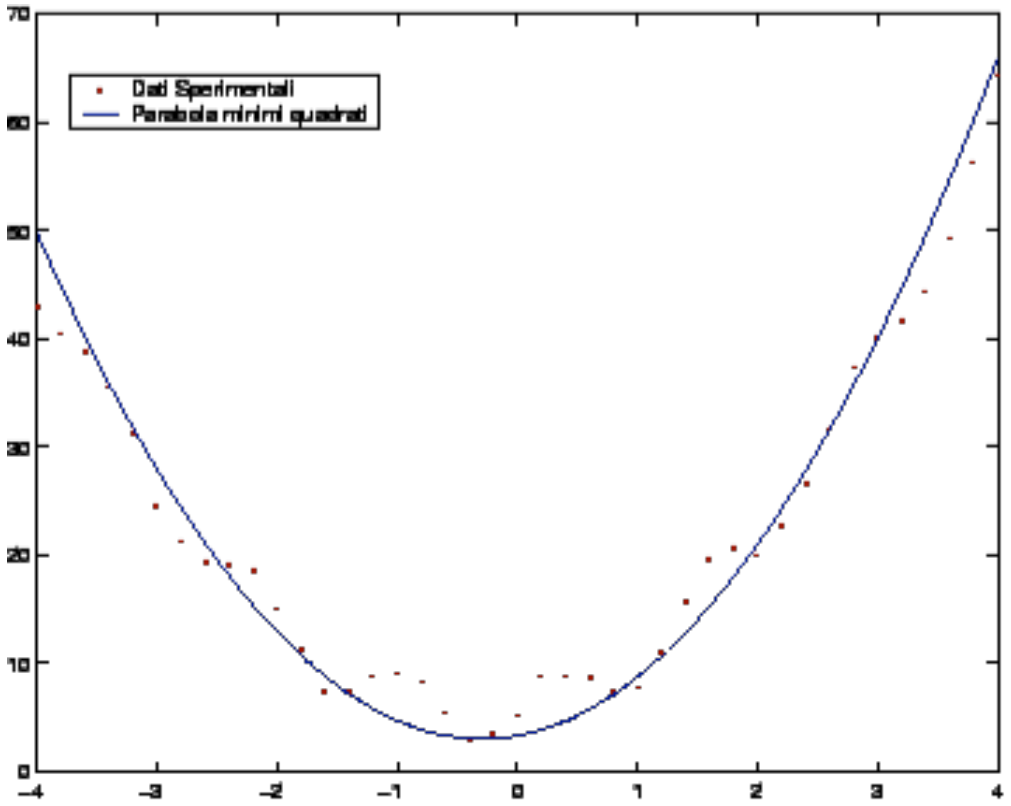


Figura 2

Considerando infine il precedente scatter plot età vs pulsazioni cardiache (***Figura 3***) ***è difficile ipotizzare che esista una relazione*** fra *X* e *Y*, anche se sembra che vi sia una debole tendenza all'aumento delle pulsazioni all'aumentare dell'età.

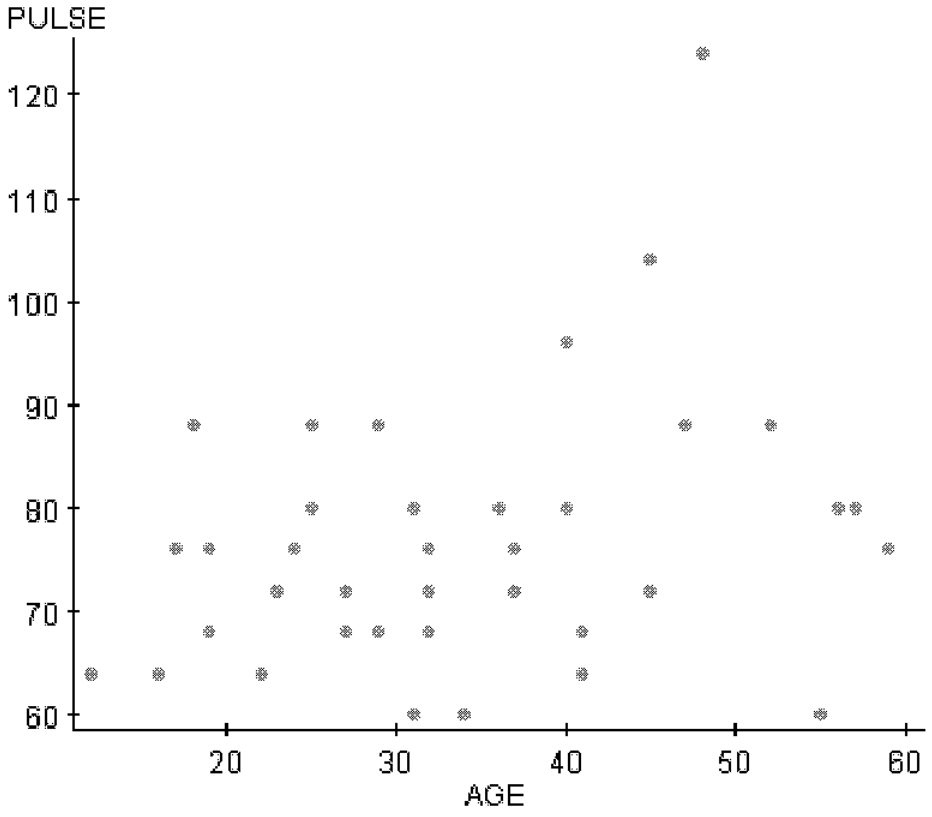


Figura 3

Il problema generale di trovare ***l'equazione*** di una curva che interpoli certi dati è detto ***interpolazione***.

Così per i dati in Figura 1 si potrebbe usare l'equazione di una retta

$$Y = a + b\ X$$

mentre per quelli in Figura 2 si potrebbe usare l'equazione di una parabola

$$Y = a + b\ X + c\ X^2$$

REGRESSIONE

Uno degli scopi principali dell'interpolazione è stimare una delle variabili (***variabile dipendente***) per mezzo dell'altra (***variabile indipendente***).

Il procedimento di stima è detto ***regressione***.

Se, utilizzando un'opportuna equazione, *Y* è stimata a partire da *X*, la relazione matematica impiegata è detta ***equazione di regressione di Y in X***.

La curva corrispondente è detta ***curva di regressione di Y in X***.

METODO DEI MINIMI QUADRATI

Esiste più di una curva di un certo tipo che interpola i dati.

Onde evitare valutazioni personali e soggettive, è necessario definire un metodo per ottenere la *"migliore"* curva interpolante.

A titolo di esempio si consideri la curva in ***Figura 4***, dove i dati sperimentali sono i punti (x_i , y_i) con $i = 1, 2, \ldots, n$

Per ogni x_i ci sarà una differenza fra y_i e il corrispondente valore determinato dalla curva C.

Si indichi con d_i tale differenza, detta ***errore***.

Ovviamente d_i potrà essere maggiore, minore o uguale a 0.

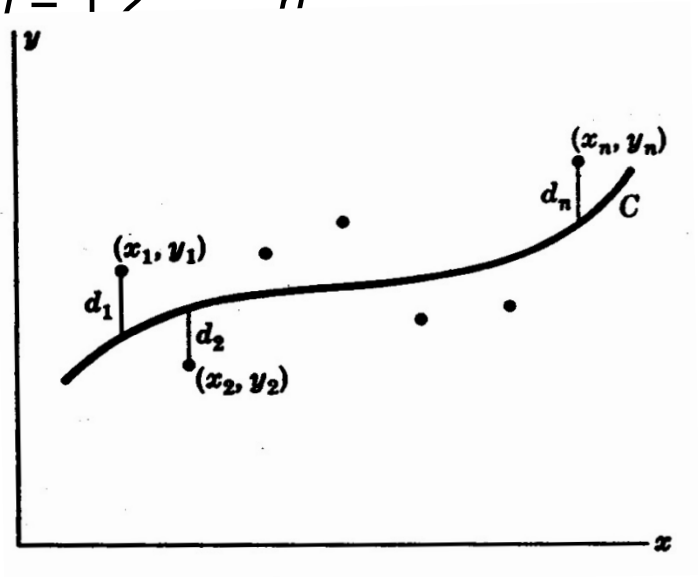


Figura 4

Una misura della bontà dell'adattamento della curva C all'insieme dei dati è rappresentata dalla somma di tutti gli errori elevati al quadrato, cioè da

$$J = d_1^2 + d_2^2 + \ldots\ldots + d_n^2$$

È ovvio che tanto minore è *J*, tanto migliore è l'adattamento.

Definizione. La migliore curva interpolatrice è quella che rende minima la precedente somma *J* (***curva di regressione dei minimi quadrati*** o, semplicemente, ***curva dei minimi quadrati***).

In particolare, se la curva C è una retta, si parlerà di ***retta dei minimi quadrati***.

CORRELAZIONE E REGRESSIONE LINEARE

La correlazione e la regressione sono tecniche per analizzare la relazione fra due o più variabili ***continue***.

La domanda più semplice da porre è: ***C'è un'associazione lineare fra le variabili?*** Ovvero: ***Esiste una relazione del tipo Y = a + bX che lega le variabile X e Y?***

Con la ***correlazione*** si cerca ***un'associazione lineare fra due variabili*** e ***la forza dell'associazione è indicata dal coefficiente di correlazione***.

Quando il coefficiente di correlazione è basato su osservazioni di valori originali è noto come ***coefficiente di correlazione di Pearson***.

Quando invece è calcolato dopo avere ordinato i dati è noto come ***coefficiente di correlazione per dati ordinati di Spearman***.

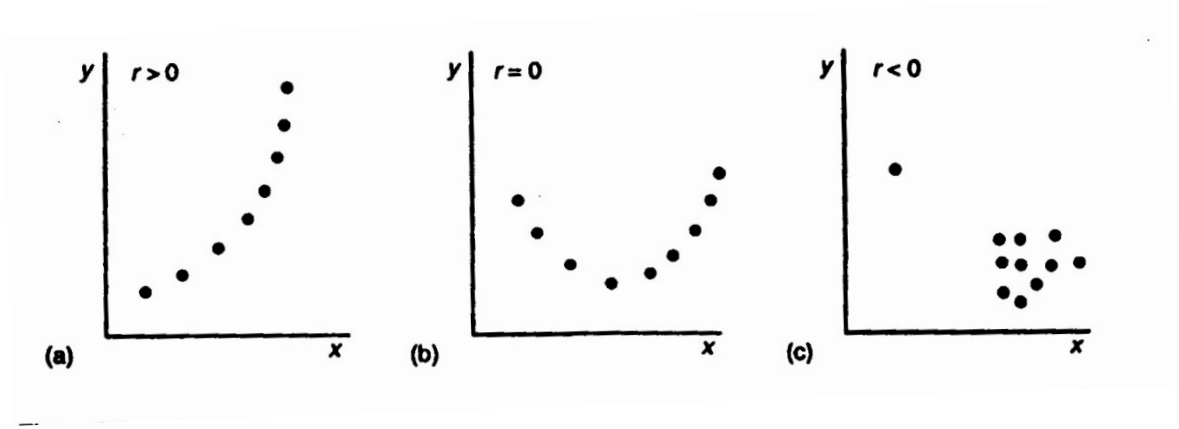
I casi in cui può non essere appropriato utilizzare il coefficiente di correlazione sono:

1) il coefficiente di correlazione non deve essere utilizzato se la relazione è non lineare [Figure a e b];

2) il coefficiente di correlazione deve essere utilizzato con prudenza in presenza di uno più punti estremi (molto distanti dagli altri) [Figura c];

3) il coefficiente di correlazione deve essere utilizzato con cautela quando le variabili sono misurate da più di un gruppo distinto, ad esempio pazienti affetti da una malattia e controlli sani;

4) il coefficiente di correlazione non deve essere usato in quelle situazioni in cui una delle variabili è fissata a priori, ad esempio se si vuole analizzare la risposta a dosi diverse di un farmaco.



VARIANZA E COVARIANZA

Si considerino due variabili casuali continue X ed Y aventi una certa funzione di densità di probabilità $p(x,y)$. Indicati con μ_x e μ_y i valori medi di X e di Y, le loro ***varianze*** sono rispettivamente definite come

$$\sigma_x^2 = E[(X - \mu_x)^2] \qquad \sigma_y^2 = E[(Y - \mu_y)^2]$$

In questo caso si può definire anche un'ulteriore quantità detta ***covarianza*** ed indicata con il simbolo σ_{xy}

$$\sigma_{xy} = E[(X - \mu_x)(Y - \mu_y)]$$

- Quando ***X* e *Y* sono indipendenti** allora $\sigma_{xy} = 0$;
- quando **fra *X* e *Y* esiste una completa dipendenza di tipo lineare** allora $\sigma_{xy} = \pm\sigma_x \sigma_y$.
- **vale sempre la relazione** $|\sigma_{xy}| \le \sigma_x \sigma_y$.

COEFFICIENTE DI CORRELAZIONE

Da quanto detto segue che è possibile introdurre una misura di un'eventuale dipendenza lineare fra *X* e *Y* come

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

che è una quantità adimensionale detta ***coefficiente di correlazione***.

Se c'è una completa dipendenza di tipo lineare fra *X* e *Y*, allora ρ assume il valore +1 oppure –1.

Quando $\rho = 0$ allora $\sigma_{xy} = 0$. In questo caso si dirà che *X* e *Y* sono incorrelate (le variabili sono indipendenti o siamo in presenza di particolari tipi di dipendenza non lineare).

In tutti gli altri casi è immediato verificare che $|\rho| \leq 1$, ovvero $-1 \leq \rho \leq 1$.

ρ è positivo quando al crescere di *X* cresce anche *Y*, mentre è negativo quando *Y* decresce al crescere di *X*.

Dato un insieme di n osservazioni appaiate (x_1,y_1), (x_2,y_2), …, (x_n,y_n), le varianze di X e Y, la covarianza ed il coefficiente di correlazione di Pearson sono calcolate come

$$s_x^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1} \qquad s_y^2 = \sum_{i=1}^{n} \frac{(y_i - \bar{y})^2}{n-1}$$

$$s_{xy} = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1} \qquad r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

dove $\bar{x}$ e $\bar{y}$ indicano rispettivamente i valori medi delle variabili X e Y, calcolati dal campione.

Il calcolo di r è quindi semplice perché richiede di determinare solo gli scarti di X e Y rispetto ai valori medi campionari.

Il coefficiente di correlazione *r* del campione permette non solo di riassumere la forza della relazione lineare, ma anche di verificare l'ipotesi che il coefficiente di correlazione ρ di popolazione sia 0.

In altre parole r permette di valutare se l'apparente associazione tra le variabili possa essere dovuta al caso.

Per fare ciò è necessario calcolare l'errore standard ES(*r*) come

$$\mathrm{ES}(r) = \sqrt{[(1 - r^2)/(n - 2)]}$$

e quindi il corrispondente valore di *t* come

$$t = \frac{r}{\mathrm{ES}(r)}$$

Il valore di *t* così ottenuto va confrontato con i valori critici della distribuzione *t* di Student con *n*–2 gradi di libertà.

Esempio

Nella Tabella sotto sono riportati (in litri) i valori di volume espiratorio massimo nel 1° secondo (FEV_1 o VEMS) e di capacità vitale forzata (FVC) misurati in un gruppo di pazienti.

	$FEV_1(x)$	$FVC(y)$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
	1,5	2,0	0,2448	0,1296	0,4624
	1,7	3,0	−0,0512	0,0256	0,1024
	2,1	2,9	0,0528	0,0576	0,0484
	1,6	2,5	0,0468	0,0676	0,0324
	2,4	3,0	0,1728	0,2916	0,1024
Totale	9,3	13,4	0,4660	0,5720	0,7480

Dalla Tabella, $n = 5$, $\bar{x} = 1{,}86$, $\bar{y} = 2{,}68$.

$$r = \frac{0{,}4660}{\sqrt{0{,}5720 \times 0{,}748}} = \frac{0{,}4660}{0{,}6541} = 0{,}71.$$

Quindi $ES(r) = \sqrt{\{(1 - 0{,}71^2)/3\}} = 0{,}41$ e $t = 0{,}71/0{,}41 = 1{,}73$.

Confrontando il valore calcolato di *t* (= 1,73) con quelli riportati nella tabella dei valori critici del *t* di Student (*per un test bilaterale*) con 5 – 2 = 3 gradi di libertà, si vede che il *t* calcolato è maggiore di quello corrispondente ad $\alpha = 0{,}20$ ($t = 1{,}64$), ma è minore del *t* corrispondente ad $\alpha = 0{,}10$ ($t = 2{,}35$).

Quindi **$0{,}10 < p < 0{,}20$**.

In altre parole, anche avendo trovato un valore di *r* uguale a 0,71, non si può rifiutare l'ipotesi nulla che ***l'apparente associazione tra le variabili possa essere dovuta al caso***, se abbiamo scelto $\alpha = 0{,}05$.

Tavola 2: *valori critici per test bilaterali e unilaterali basati sulla distribuzione t di Student*

Gradi di libertà	Livello di significatività				
	20% (bilaterale) 10% (unilaterale)	10% (bilaterale) 5% (unilaterale)	5% (bilaterale) 2,5% (unilaterale)	2% (bilaterale) 1% (unilaterale)	1% (bilaterale) 0,5% (unilaterale)
1	3,08	6,31	12,71	31,82	63,66
2	1,89	2,92	4,30	6,96	9,92
3	1,64	2,35	3,18	4,54	5,84
4	1,53	2,13	2,78	3,75	4,60
5	1,48	2,02	2,57	3,36	4,03
6	1,44	1,94	2,45	3,14	3,71
7	1,41	1,89	2,36	3,00	3,50
8	1,40	1,86	2,31	2,90	3,36
9	1,38	1,83	2,26	2,82	3,25
10	1,37	1,81	2,23	2,76	3,17
11	1,36	1,80	2,20	2,72	3,11
12	1,36	1,78	2,18	2,68	3,05
13	1,35	1,77	2,16	2,65	3,01
14	1,35	1,76	2,14	2,62	2,98
15	1,34	1,75	2,13	2,60	2,95
16	1,34	1,75	2,12	2,58	2,92
17	1,33	1,74	2,11	2,57	2,90
18	1,33	1,73	2,10	2,55	2,88
19	1,33	1,73	2,09	2,54	2,86
20	1,33	1,72	2,09	2,53	2,85
21	1,32	1,72	2,08	2,52	2,83
22	1,32	1,72	2,07	2,51	2,82
23	1,32	1,71	2,07	2,50	2,81
24	1,32	1,71	2,06	2,49	2,80
25	1,32	1,71	2,06	2,49	2,79
26	1,32	1,71	2,06	2,48	2,78
27	1,31	1,70	2,05	2,47	2,77
28	1,31	1,70	2,05	2,47	2,76
29	1,31	1,70	2,05	2,46	2,76
30	1,31	1,70	2,04	2,46	2,75
60	1,30	1,67	2,00	2,39	2,66
90	1,29	1,66	1,99	2,37	2,63
120	1,29	1,66	1,98	2,36	2,62
∞	**1,28**	**1,64**	**1,96**	**2,33**	**2,58**

I valori mostrati si riferiscono ai valori critici per ipotesi alternative bilaterali (≠) e unilaterali (>). Il valore critico per i test unilaterali (<) sono il valore negativo dei test unilaterali (>) mostrati nella tavola. Ad esempio, 2,13 è il valore critico per un test bilaterale con livello di significatività 5% utilizzando una distribuzione t Student con 15 gradi di libertà.

RETTA DI REGRESSIONE

Quando si analizza la correlazione fra due variabili *X* e *Y* si può non essere interessati a valutare come *X* predica *Y* o viceversa.

Quando si calcola la regressione si parte dalla premessa che un cambiamento di *X* porterà direttamente ad un cambiamento di *Y*. In questo caso si può essere interessati a predire il valore di *Y* corrispondente ad un dato valore di *X*, anche se ***non si è autorizzati a credere che ci sia un reale rapporto di causa-effetto***.

Convenzionalmente i valori della variabile *X* (variabile indipendente) sono riportati sull'asse orizzontale e quelli della *Y* (variabile dipendente) in quello verticale.

L'equazione

$$Y = \alpha + \beta X$$

è detta ***retta di regressione***.

α* è l'*intercetta* e *β* è il *coefficiente di regressione.

N.B. Nell'equazione precedente si sono impiegate lettere greche per ricordare che si tratta di parametri di popolazione.

Data una serie di *n* coppie di osservazioni (x_1,y_1), (x_2,y_2), …, (x_n,y_n) come si calcolano *α* e *β*?

Pensando ad ***α*** e ***β*** come a parametri caratteristici di popolazione, **si vuole ottenere una loro stima** (***a*** e ***b***, rispettivamente) **a partire da un campione di quella popolazione**.

Utilizzando il metodo dei minimi quadrati le stime *b* e *a* sono date da

$$b = \frac{s_{xy}}{s_x^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad a = \bar{y} - b\,\bar{x}$$

È importante verificare che ***b*** sia significativamente diverso da 0. Per fare ciò si ricorre ancora alla statistica *t*, calcolando

$$t = \frac{b}{\mathrm{ES}(b)} \qquad \textbf{dove} \qquad \mathrm{ES}(b) = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 \; - \; b^2 \sum_{i=1}^{n}(x_i - \bar{x})^2}{(n-2)\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Il valore di *t* così ottenuto viene confrontato con quello riportato nella tabella *t* di Student con ***n – 2*** **gradi di libertà**.

Pertanto l'intervallo di confidenza al 95% per l'inclinazione è dato da

$$[b - t_{0,05}\,\mathrm{ES}(b)\,;\;\; b + t_{0,05}\,\mathrm{ES}(b)]$$

dove $t_{0,05}$ è il valore di *t* riportato nella tabella (*per un test bilaterale*) per il livello di significatività di 0,05 (5%) e $n-2$ gradi di libertà.

Tavola 2: *valori critici per test bilaterali e unilaterali basati sulla distribuzione t di Student*

Gradi di libertà	Livello di significatività				
	20% (bilaterale) 10% (unilaterale)	10% (bilaterale) 5% (unilaterale)	5% (bilaterale) 2,5% (unilaterale)	2% (bilaterale) 1% (unilaterale)	1% (bilaterale) 0,5% (unilaterale)
1	3,08	6,31	12,71	31,82	63,66
2	1,89	2,92	4,30	6,96	9,92
3	1,64	2,35	3,18	4,54	5,84
4	1,53	2,13	2,78	3,75	4,60
5	1,48	2,02	2,57	3,36	4,03
6	1,44	1,94	2,45	3,14	3,71
7	1,41	1,89	2,36	3,00	3,50
8	1,40	1,86	2,31	2,90	3,36
9	1,38	1,83	2,26	2,82	3,25
10	1,37	1,81	2,23	2,76	3,17
11	1,36	1,80	2,20	2,72	3,11
12	1,36	1,78	2,18	2,68	3,05
13	1,35	1,77	2,16	2,65	3,01
14	1,35	1,76	2,14	2,62	2,98
15	1,34	1,75	2,13	2,60	2,95
16	1,34	1,75	2,12	2,58	2,92
17	1,33	1,74	2,11	2,57	2,90
18	1,33	1,73	2,10	2,55	2,88
19	1,33	1,73	2,09	2,54	2,86
20	1,33	1,72	2,09	2,53	2,85
21	1,32	1,72	2,08	2,52	2,83
22	1,32	1,72	2,07	2,51	2,82
23	1,32	1,71	2,07	2,50	2,81
24	1,32	1,71	2,06	2,49	2,80
25	1,32	1,71	2,06	2,49	2,79
26	1,32	1,71	2,06	2,48	2,78
27	1,31	1,70	2,05	2,47	2,77
28	1,31	1,70	2,05	2,47	2,76
29	1,31	1,70	2,05	2,46	2,76
30	1,31	1,70	2,04	2,46	2,75
60	1,30	1,67	2,00	2,39	2,66
90	1,29	1,66	1,99	2,37	2,63
120	1,29	1,66	1,98	2,36	2,62
∞	**1,28**	**1,64**	**1,96**	**2,33**	**2,58**

I valori mostrati si riferiscono ai valori critici per ipotesi alternative bilaterali (≠) e unilaterali (>). Il valore critico per i test unilaterali (<) sono il valore negativo dei test unilaterali (>) mostrati nella tavola. Ad esempio, 2,13 è il valore critico per un test bilaterale con livello di significatività 5% utilizzando una distribuzione t Student con 15 gradi di libertà.

Per comodità, a lato viene di nuovo mostrata la tabella dei valori critici del *t* di Sudent per un test bilaterale o unilaterale.

ESEMPIO

Si vuole calcolare la retta di regressione fra altezza (cm) e FEV_1 (litri) a partire dai dati riportati nella tabella sotto, che mostra i valori di 5 osservazioni ottenute in pazienti asmatici.

	x cm (statura)	*y* litri (FEV_1)	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
	160	1,5	-8,6	-0,36	73,96	0,1296	3,096
	170	1,7	1,4	-0,16	1,96	0,0256	-0,224
	173	2,1	4,4	0,24	19,36	0,0576	1,056
	165	1,6	-3,6	-0,26	12,96	0,0676	0,936
	175	2,4	6,4	0,54	40,96	0,2916	3,456
Totale	843 $\bar{x}=168,6$	9,3 $\bar{y}=1,86$	0	0	149,2	0,572	8,32

Utilizzando le precedenti relazioni si ottiene

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{8{,}32}{149{,}2} = 0{,}05576 \ \text{(litri/cm)}$$

$$a = \bar{y} - b\,\bar{x} = 1{,}86 - 0{,}05576 \times 168{,}6 = -7{,}542 \ \text{(litri)}$$

La retta di regressione è quindi

$$FEV_1\,(\text{litri}) = -7{,}542 + 0{,}05576 \times \text{altezza}\,(\text{cm})$$

Fissato il livello di significatività al 5%, si effettui ora il **test di significatività su *b***.

$$\mathrm{ES}(b) = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 - b^2 \sum_{i=1}^{n}(x_i - \bar{x})^2}{(n-2)\sum_{i=1}^{n}(x_i - \bar{x})^2}} =$$

$$= \sqrt{\frac{0{,}572 - 0{,}00311 \times 149{,}2}{3 \times 149{,}2}} = 0{,}0155 \text{ (l/cm)}$$

e quindi

$$t = \frac{0{,}05576}{0{,}0155} = 3{,}59$$

Confrontando il valore ottenuto di *t* con i valori critici riportati in tabella con 3 (= 5 – 2) gradi di libertà si ottiene $p < 0{,}05$ e quindi, avendo fissato il livello di significatività al 5%, si può rifiutare l'ipotesi nulla e quindi ***b* è significativamente diverso da 0**.

	Livello di significatività				
Gradi di libertà	20% (bilaterale) 10% (unilaterale)	10% (bilaterale) 5% (unilaterale)	5% (bilaterale) 2,5% (unilaterale)	2% (bilaterale) 1% (unilaterale)	1% (bilaterale) 0,5% (unilaterale)
1	3,08	6,31	12,71	31,82	63,66
2	1,89	2,92	4,30	6,96	9,92
3	1,64	2,35	3,18	4,54	5,84
4	1.53	2.13	2.78	3.75	4.60

Infine, **l'intervallo di confidenza al 95% per l'inclinazione della retta di regressione è dato da**

$$[0{,}05576 - 3{,}182 \times 0{,}0155 \; ; \;\; 0{,}05576 + 3{,}182 \times 0{,}0155 \;\; \text{(litri/cm)}]$$

ovvero $\;\; [0{,}007; \;\; 0{,}105 \;\; \text{(litri/cm)}]$

Variabili multidimensionali: notazione matriciale

VETTORE DELLE MEDIE

È possibile generalizzare al **caso multidimensionale** il concetto di distribuzione di probabilità e, in particolare, nel caso di più variabili casuali continue, quello di densità di probabilità. Le considerazioni che potremmo fare in questo caso sono del tutto simili a quelle fatte nel caso monodimensionale.

In questo modo le definizioni che abbiamo introdotto precedentemente per una variabile possono essere estese a due o più variabili.

Ad esempio, **nel caso di una variabile bidimensionale (*X*,*Y*) possiamo definire i valori medi di *X* e di *Y***

$$E(X) = \mu_x \qquad E(Y) = \mu_y$$

Il valore medio ***M*** di una variabile *n*-dimensionale viene di solito riportato in una colonna composta di ***n*** elementi. Ad esempio, nel caso bidimensionale ***M*** è scritto come

$$M = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}$$

Questo tipo di rappresentazione utilizza la cosiddetta ***notazione matriciale***, dove ***M*** è detto ***vettore delle medie***.

In generale ***un vettore è una stringa composta di più numeri***.

Se la stringa è messa su una colonna si parla di ***vettore colonna***, mentre, se è messa su una riga, prende il nome di ***vettore riga***.

MATRICE DI COVARIANZA

Quando si considerano più variabili casuali, le varianze e le covarianze possono essere messe in una tabella, che prende il nome di ***matrice di covarianza***. Avendo, ad esempio, 3 variabili casuali *X*, *Y* e *Z,* si definisce la seguente matrice di covarianza

$$\Sigma = \begin{pmatrix} \sigma_x^2 & \sigma_{xy} & \sigma_{xz} \\ \sigma_{yx} & \sigma_y^2 & \sigma_{yz} \\ \sigma_{zx} & \sigma_{zy} & \sigma_z^2 \end{pmatrix}$$

Una matrice è una tabella ordinata di elementi numerici avente *n* righe e *m* colonne.

La tabella a lato è una matrice con 3 righe e 4 colonne (si dice semplicemente 3x4).

N.B. Un vettore è una particolare matrice avente una sola riga (vettore riga) o una sola colonna (vettore colonna).

$$A = \begin{pmatrix} -1 & 0 & 2 & 8 \\ 5 & -7 & 3 & -6 \\ 1 & -5 & -4 & 9 \end{pmatrix}$$

Si noti che la matrice di covarianza ha un uguale numero di righe e di colonne, ovvero è una ***matrice quadrata***.

Inoltre, per come è stata definita la covarianza è ovvio che $\sigma_{xy} = \sigma_{yx}$, $\sigma_{xz} = \sigma_{zx}$ ed anche $\sigma_{yz} = \sigma_{zy}$.

Ciò significa che gli elementi al di fuori della diagonale che parte dal punto alto a sinistra e termina nel punto basso a destra della matrice quadrata (detta diagonale principale) sono simmetricamente uguali fra loro.

Una tale matrice è detta simmetrica.

MATRICE DI CORRELAZIONE

Date più variabili casuali, anche i coefficienti di correlazione possono essere messi in una tabella che prende il nome di ***matrice di correlazione***.

Nel caso bidimensionale essa assume la forma

$$\rho = \begin{pmatrix} \rho_{xx} & \rho_{xy} \\ \rho_{yx} & \rho_{yy} \end{pmatrix} = \begin{pmatrix} 1 & \rho_{xy} \\ \rho_{yx} & 1 \end{pmatrix}$$

Si noti che anche ***la matrice di correlazione è una matrice quadrata e simmetrica***.

VANTAGGI DELLA NOTAZIONE MATRICIALE

Impiegare la notazione matriciale è di estrema utilità in quanto:

- ***permette di rappresentare i dati in modo organizzato e particolarmente adatto all'impiego di un calcolatore elettronico;***
- ***mediante l'algebra delle matrici, che specifica le regole per l'uso delle matrici, è possibile estendere anche a queste le principali operazioni matematiche;***
- ***sarebbe estremamente difficoltoso sviluppare tecniche di analisi statistica multivariata facendo a meno delle matrici;***
- ***con i moderni calcolatori è facile e rapido effettuare operazioni matematiche sulle matrici e, quindi, implementare tecniche (anche complesse) di analisi multivariata.***