Dalla matrice dei dati alla tabella doppia

Operaio	Importo	Livello	Operaio	Importo	Livello	Operaio	Importo	Livello
1	133754	A	41	139637	B	81	156488	A
2	177321	D	42	196198	C	82	191405	A
3	198093	B	43	183375	B	83	117894	F
4	198951	F	44	148518	F	84	161926	A
5	128050	A	45	126191	B	85	102978	B
6	107152	B	46	148488	C	86	171470	A
7	168502	B	47	129230	B	87	131906	A
8	185872	C	48	193780	F	88	179658	C
9	174107	A	49	141154	B	89	146534	A
10	127670	F	50	100256	B	90	137011	B
11	171307	B	51	140573	A	91	112452	D
12	135016	A	52	191271	A	92	117509	A
13	116721	B	53	194093	B	93	185801	C
14	138590	E	54	109994	B	94	172984	A
15	122672	C	55	177444	A	95	103235	B
16	191676	D	56	100239	F	96	195622	B
17	174958	B	57	176015	B	97	127726	D
18	187423	D	58	170692	C	98	121094	A
19	111110	C	59	187677	E	99	193272	B
20	136503	E	60	199348	E	100	148265	B
21	120768	C	61	123781	B			
22	191648	D	62	179708	D			
23	101570	D	63	139825	A			
24	145044	A	64	148948	C			
25	102990	F	65	146901	D			
26	187028	E	66	136471	D			
27	124437	D	67	104697	A			
28	122079	C	68	152657	E			
29	163468	E	69	170503	B			
30	140935	A	70	135280	D			
31	146843	A	71	107743	B			
32	172497	C	72	171517	D			
33	122209	D	73	193946	C			
34	135783	D	74	170884	A			
35	150789	C	75	181407	B			
36	121587	A	76	124571	E			
37	133415	D	77	139906	A			
38	194731	F	78	142344	A			
39	176619	B	79	190776	A			
40	104960	A	80	141811	B			

Su n=100 operai è stato rilevato l'importo dello straordinario settimanale e la classe stipendiale.

In questa forma i dati non sono leggibili;

Organizziamo gli importi in classi:

Excel: Tabella pivot

Count of Operaio	Livello						
Imp.MGL	A	B	C	D	E	F	Grand Total
<120	3	7	1	2	0	3	16
120-140	8	5	3	7	3	1	27
140-160	7	3	3	1	1	1	16
>160	9	12	7	6	4	3	41
Grand Total	27	27	14	16	8	8	100

La tabella rivela che il 41% si colloca nella 4ª classe; che il 12% si trova nella combinazione (4,B) e che il livello "A" fa più straordinari (27%) rispetto a tutti gli altri.

Trattazione generale

Partiamo dalla varaibile doppia: $(X_i, Y_i);\ i = 1,2,\ldots,n$

Supponiamo che siano state organizzate in una tabella con "r" modalità distinte per la variabile sulle righe (X) e "c" modalità per la variabile sulle colonne (Y)

Dove

	Y_1	Y_2	…	Y_c	
X_1	n_{11}	n_{12}		n_{1c}	$n_{1.}$
X_2	n_{21}	n_{22}		n_{2c}	$n_{2.}$
M					
X_r	n_{r1}	n_{r2}		n_{rc}	$n_{r.}$
	$n_{.1}$	$n_{.2}$	…	$n_{.c}$	n

$$n_{i.} = \sum_{j=1}^{c} n_{ij} = n_{i1} + n_{i2} + \ldots + n_{ic} = totale\ \ di\ \ riga$$

$$n_{.j} = \sum_{i=1}^{r} n_{ij} = n_{j1} + n_{j2} + \ldots + n_{rj} = totale\ \ di\ \ colonna$$

$$n = \sum_{i=1}^{r} n_{i.} = \sum_{j=1}^{c} n_{.j} = \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}$$

il punto indica l'indice rispetto a cui si è sommato

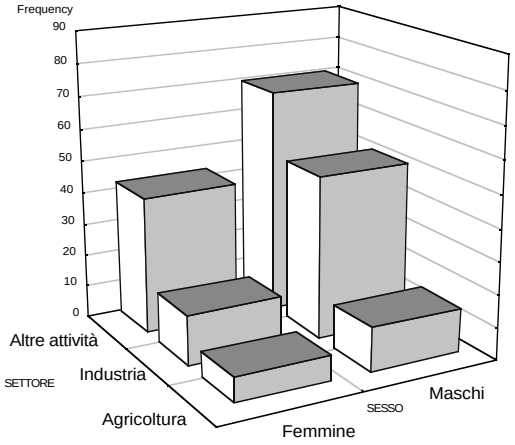
Esempio

Occupati per settori di attività economica (media annua). Dati in migliaia

		Sesso	
Settori	**Maschi**	**Femmine**	**Totale**
Agricoltura	**1.485**	**812**	**2.297**
Industria	**5.270**	**1.626**	**6.896**
Terziario	**7.232**	**4.318**	**11.550**
Totale	**13.987**	**6.756**	**20.743**

r=3; c=2; n=20'743

La diversa struttura delle due componenti è evidente dal grafico



Distribuzione congiunta di due variabili

Anche nella tabella doppia possiamo usare le frequenze relative:

	Y_1	Y_2	…	Y_c	
X_1	f_{11}	f_{12}		f_{1c}	$f_{1.}$
X_2	f_{21}	f_{22}		f_{2c}	$f_{2.}$
:					
X_r	f_{r1}	f_{r2}		f_{rc}	$f_{r.}$
	$f_{.1}$	$f_{.2}$	…	$f_{.c}$	1

$$0 \le f_{ij} \le 1$$

$$f_{i.} = \sum_{j=1}^{c} f_{ij}$$

$$f_{.j} = \sum_{i=1}^{r} f_{ij}$$

$$\sum_{i=1}^{r}\sum_{j=1}^{c} f_{ij} = 1$$

Le f_{ij} sono dette frequenze relative congiunte;

Le "$f_{i.}$" e le "$f_{.j}$" sono le frequenze relative marginali.

L'insieme delle coppie (X_i, Y_i) e delle rispettive frequenze relative f_{ij} costituisce la distribuzione congiunta delle variabili X ed Y;

Essa associa ad ogni combinazione di modalità (X_i,Y_j) un numero in (0,1) e la cui somma è pari ad uno

Distribuzioni marginali

A partire dalla distribuzione congiunta si definiscono le distribuzioni per ciascuna delle variabili a prescindere dall'altra

$$f(X=x_i)=\sum_{j=1}^{c} f(X=x_i,Y=y_j)=\sum_{j=1}^{c} f_{ij}=f_{i.};\ i=1,2,\ldots,r$$

$$f(Y=y_j)=\sum_{i=1}^{r} f(X=x_i,Y=y_j)=\sum_{i=1}^{r} f_{ij}=f_{.j};\ j=1,2,\ldots,c$$

Per ottenere la distribuzione marginale si somma rispetto alla variabile che NON interessa

Valore atteso delle marginali

Le distribuzioni marginali sono delle vere e proprie distribuzioni univariate.

In particolare, ci interessa il loro valore atteso

$$E(X)=\sum_{i=1}^{r} X_i f_{i,}=\mu_x\ ;\qquad E(Y)=\sum_{j=1}^{c} Y_j f_{.j}=\mu_y$$

Esempio

Y/X	2	4	6	
1	$4/20$	$2/20$	$4/20$	$10/20$
3	$4/20$	$1/20$	$5/20$	$10/20$
	$8/20$	$3/20$	$9/20$	1

$$\mu_x=2\left(\frac{8}{20}\right)+4\left(\frac{3}{20}\right)+6\left(\frac{9}{20}\right)=\frac{82}{20}=4.1$$

$$\mu_y=1\left(\frac{10}{20}\right)+3\left(\frac{10}{20}\right)=\frac{40}{20}=2$$

Distribuzioni condizionate

Per studiare il comportamento della "Y" rispetto alla "X" dividiamo la distribuzione Congiunta in tante sottodistribuzioni

La distribuzione della Y CONDIZIONATA (PARZIALE) dal fatto che "X" è ad un dato livello è

$$f(Y=y_j|X=x_i)=\frac{f(X=x_i,Y=y_j)}{f(Y=y_j)};\ j=1,2,\ldots,c$$

cioè un riscalamento pro-quota delle righe della tabella per assicurare la somma unitaria

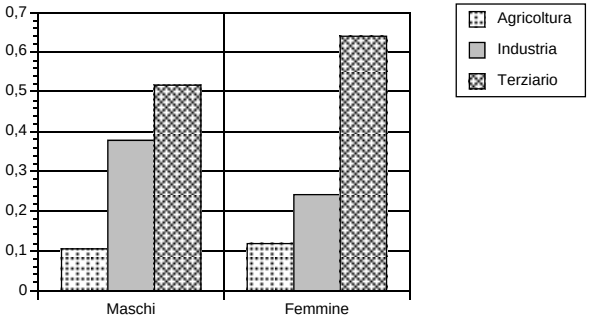
Analogamente, la distribuzione della X dato che Y è ad un livello prefissato è:

$$f(X=x_i|Y=y_j)=\frac{f(X=x_i,Y=y_j)}{f(X=x_i)};\ i=1,2,\ldots,r$$

Esempio

Distribuzione congiunta

		Sesso	
Settori	**Maschi**	**Femmine**	**Totale**
Agricoltura	**7,16%**	**3,91%**	**11,07%**
Industria	**25,41%**	**7,84%**	**33,24%**
Terziario	**34,86%**	**20,82%**	**55,68%**
Totale	**67,43%**	**32,57%**	**100,00%**



Distribuzione marginale Donne

Settori	**Femmine**
Agricoltura	**12,02%**
Industria	**24,07%**
Terziario	**63,91%**
Totale	**100,00%**

Distribuzione marginale maschi

Settori	**Maschi**
Agricoltura	**10,62%**
Industria	**37,68%**
Terziario	**51,71%**
Totale	**100,00%**

Indipendenza di eventi e di variabili

Perché abbia senso lo studio CONGIUNTO esso deve essere più informativo dello studio SEPARATO delle due componenti

Se la "X" assume valori in relazione ad eventi indipendenti da quelli che generano i valori della "Y" non esiste alcun legame statistico interessante

ESEMPIO

Lancio di due dadi di diverso colore

X: punteggio del dado rosso;
Y: punteggio del dado blù;

Sapere che lanciando i due dadi, X= 4 e, contemporaneamente, Y= 3 è come sapere che X=4 (ignorando "Y") e che Y=3 (ignorando "X")

$$P(X=4 \cap Y=3) = P(X=4)P(Y=3)$$

$$ovvero \quad P(X=4|Y=3) = P(X=4)$$

Indipendenza in distribuzione

Se la condizionata di Y|X non cambia al variare di X allora Y è INDIPENDENTE IN DISTRIBUZIONE da X.

$$f(X=x_i,|Y=y_j) = f(X_i); \quad i=1,2,\ldots,r; \quad j=1,2,\ldots,c$$

L'indipendenza è una relazione simmetria: Se X è indipendente da Y anche Y è indipendente da X

Se fra le due variabili c'è indipendenza, le frequenze assolute sono pari al prodotto delle frequenze marginali diviso per il totale frequenze:

$$f(X=x_i,|Y=y_j) = f(X_i) \Rightarrow \frac{n_{ij}}{n_{.j}} = \frac{n_{i.}}{n} \Rightarrow \quad n_{ij} = \frac{n_{i.} * n_{.j}}{n} \quad = \frac{\left(\frac{n_{i.}}{n}\right)\left(\frac{n_{.j}}{n}\right)}{1} = f_{i.} * f_{.j}$$

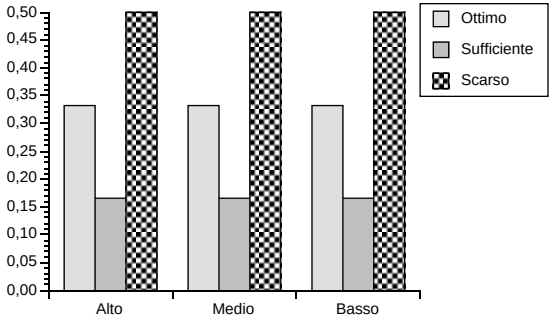
Esempio

Reddito familiare e rendimento scolastico

Rendimento	Alto	Medio	Basso	Totale
Ottimo	16	32	40	88
Sufficiente	8	16	20	44
Scarso	24	48	60	132
Totale	48	96	120	264

	Reddito familiare			
Rendimento	Alto	Medio	Basso	Totale
Ottimo	0,3333	0,3333	0,3333	0,3333
Sufficiente	0,1667	0,1667	0,1667	0,1667
Scarso	0,5000	0,5000	0,5000	0,5000
Totale	1,0000	1,0000	1,0000	1,0000

Le frequenze assolute sono diverse, ma quelle relative coincidono per ogni distribuzione condizionata del rendimento.



Verifica:

$$40 = \frac{88*120}{264}; \; 16 = \frac{44*96}{264}$$

Test del χ^2 (chi quadrato)

L'ipotesi da sottoporre a verifica è

$$H_0: \; \pi_{ij} = \pi_{i.}\pi_{.j} \quad per \; ogni \; "i \; e \; j"$$
$$H_1: \pi_{ij} \neq \pi_{i.}\pi_{.j} \quad per \; almeno \; un \; "i \; e \; j"$$

Si adopera la seguente statistica test:

$$\chi_c^2 = n\left[\sum_{i=1}^{k}\frac{\left(f_{ij}-\pi_{ij}\right)^2}{\pi_{ij}}\right] = \sum_{i=1}^{n}\frac{\left(n_{ij}-n\pi_{ij}\right)^2}{n\pi_{ij}} = \left[\sum_{i=1}^{n}\frac{n_{ij}^2}{n\pi_{ij}}\right] - n$$

che è è nulla se e solo se c'è indipendenza in distribuzione tra le due variabili.

Il χ_c^2 aumenta all'aumentare della differenza tra le frequenze ipotizzate in caso di indipendenza e quelle effettivamente osservate.

Valori elevati di χ_c^2 ridurranno il p-Value dell'ipotesi nulla

Esempio

Produzione di palloni di cuoio. Per il controllo della qualità i prodotti sono classificati rispetto a: X=pressione interna e Y=superficie esterna.

Y/X	Rifiutati	Imperfetti	Standard	
Rifiutati	12	23	89	124
Imperfetti	8	12	62	82
Standard	21	30	119	170
	41	65	270	376

Y/X	R.	I.	S.	
R.	$\frac{14 \cdot 124}{376} * 41$	21	89	124
I.	9	14	59	82
S.	18	30	122	170
	41	65	270	376

$$\chi^2_c = \frac{(12-14)^2}{14} + \frac{(23-21)^2}{21} + \ldots + \frac{(119-122)^2}{122} = 1.599$$

Gli scarti tra frequenze teoriche ed osservato sono dette contingenze

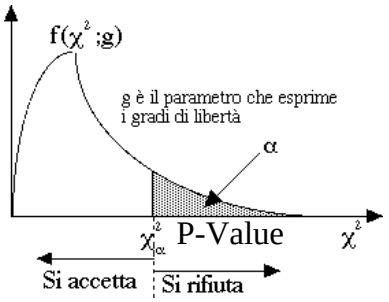
Il valore dell'indice sembra basso, ma è abbastanza basso?

La distribuzione campionaria del χ^2_c, per n grande, è ben approssimata dalla variabile casuale detta del chi-quadrato che quindi è usata per calcolare il p-value del test.

Esempio (continua)

$$\chi^2_c = \frac{(12-14)^2}{14} + \frac{(23-21)^2}{21} + \ldots + \frac{(119-122)^2}{122} = 1.599$$

$$gdl = (3-1)*(3-1) = 4$$



Per i gradi di libertà si tiene conto che frequenze marginali teoriche ed osservate coincidono e che la somma la loro somma deve essere pari ad uno (quest'ultimo totale deve essere contato solo una volta)

gdl=rc-(c+r-1)=(r-1)(c-1)

DISTRIB.CHI(1.599;4)=0.80897

In caso di indipendenza, la probabilità di osservare un valore minore o uguale a quello osservato (1.599) è dell'80%.

Non c'è quindi evidenza di un legame tra la pressione interna e la superficie esterna dei prodotti.

Esercizio (Excel)

Indagine sulla mobilità di voto. Uso dello strumento PivotTable

Soggetto	Ha votato	Voterà			
Adua	Centro	Destra	Iris	Centro	Centro
Aida	Sinistra	Sinistra	Irma	Destra	Destra
Alda	Destra	Destra	Jula	Sinistra	Centro
Alea	Centro	Centro	Kara	Sinistra	Destra
Alfa	Destra	Centro	Lara	Destra	Sinistra
Anna	Sinistra	Sinistra	Leda	Centro	Centro
Asia	Centro	Destra	Lena	Sinistra	Sinistra
Atte	Sinistra	Centro	Lisa	Centro	Centro
Beba	Sinistra	Sinistra	Lory	Sinistra	Centro
Bice	Centro	Destra	Mara	Centro	Destra
Cira	Centro	Sinistra	Mena	Centro	Sinistra
Cleo	Destra	Destra	Mina	Sinistra	Sinistra
Cora	Sinistra	Destra	Mira	Sinistra	Sinistra
Demi	Centro	Destra	Olga	Centro	Destra
Dina	Centro	Centro	Pina	Centro	Centro
Dora	Destra	Destra	Rina	Destra	Centro
Edda	Centro	Destra	Rita	Destra	Destra
Elsa	Destra	Sinistra	Rosa	Sinistra	Sinistra
Emma	Sinistra	Sinistra	Sara	Destra	Destra
Enza	Centro	Destra	Teti	Centro	Destra
Etta	Centro	Centro	Tina	Sinistra	Sinistra
Fede	Destra	Destra	Vega	Sinistra	Sinistra
Gina	Sinistra	Centro	Vera	Centro	Destra
Gisa	Centro	Destra	Zita	Destra	Destra
Ines	Destra	Destra	Zora	Centro	Centro

Count	Voterà			
Ha votato	Centro	Destra	Sinistra	Totale
Centro	8	11	2	21
Destra	2	9	2	13
Sinistra	4	2	10	16
Totale	14	22	14	50

	5.88	9.24	5.88	21 =I3*$F$6/$I$6
	3.64	5.72	3.64	13
	4.48	7.04	4.48	16
	14	22	14	50
	0.764	0.335	2.560	3.660 =(F3-F8)^2/F8
	0.739	1.881	0.739	3.359
	0.051	3.608	6.801	10.461
C hi-quadrato				17.480
Gdl				4 =(3-1)(3-1)
p-Value				0.0016 =Distrib.Chi(I16;I17)

Esercizio

Un'indagine ha classificato i rivenditori di hardware di una regione secondo il tipo di società ed il tipo di collocazione

Determinate il p-Value dell'ipotesi di indipendenza

	Tipologia società			
Negozio	Persone	Cooperativa	Impresa	Totale
Autonomo	34	16	4	54
Supermercato	4	2	3	9
Misto proprio	17	21	32	70
Misto altri	13	5	6	24
Totale	68	44	45	157

Test dell'omogeneità di due campioni

il test del χ^2 può essere utilizzato anche nella seguente situazione:

Si osservano due classificazioni campionarie.

Le ampiezze possono essere diverse, ma le categorie sono identiche.

Ci si chiede se i campioni provengono dalla stessa popolazione

$$H_0:\ \pi_{1i} = \pi_{2i}\ \ per\ \ ogni\ \ "i"$$
$$H_1:\pi_{1i} \neq \pi_{2i}\ \ per\ \ almeno\ \ un\ \ "i"$$

La statistica test assume ora la formula $\chi_c^2 = \sum_{i=1}^{k} \frac{(\pi_{i1} - \pi_{i2})^2}{\pi_{i2}}$

Esempio

Due campioni di misurazioni del tempo precedente la prima disfunzione in due diversi macchinari hanno dato luogo alla tabella

Tempi	C_1	C_2	C_2/risc	$(f1i-f'2i)^2/f'2$
0 - 20	29	56	32.78	0.436
20 - 40	27	41	24.00	0.375
40 - 60	14	32	18.73	1.195
60 - 70	11	19	11.12	0.001
60 - 80	8	13	7.61	0.020
>80	7	3	1.76	15.659
	96	164	96.00	17.686

Le "teoriche" sono ottenute come: $n'_{2i} = n_1 f_{2i}$

La statistica test è χ^2 =17.686 con 6-1=5 gradi di libertà

DISTRIB.CHI(17.686;5)=0.00336 (p-Value)

L'ipotesi si deve rifiutare aldilà di ogni ragionevole dubbio

Esercizio

Numero di incidenti automobilistici lungo una strada a scorrimento veloce. Frequenza per lato di percorrenza

Verso Sud	Settimane	Verso Nord	Settimane
0	39	0	82
1	28	1	60
2	19	2	40
3	13	3	27
4	8	4	18
5	4	5	8
>5	2	>5	4
	113		239

Determinate il p-value dell'ipotesi di omogeneità dei due campioni

Valori attesi nelle distribuzioni doppie

Nel caso di variabili quantitative metriche siamo interessati al …

Valore atteso della somma

$$E(X+Y) = \sum_{i=1}^{r}\sum_{j=1}^{c}(X_i + Y_j) f_{ij}$$

Valore atteso del prodotto

$$E(XY) = \sum_{i=1}^{r}\sum_{j=1}^{c}(X_i Y_j) f_{ij}$$

Valore atteso della somma

$$E(X+Y)=\sum_{i=1}^{r}\sum_{j=1}^{c}X_i f_{ij}+\sum_{j=1}^{c}\sum_{i=1}^{r}Y_j f_{ij}=\sum_{i=1}^{r}X_i\sum_{j=1}^{c}f_{ij}+\sum_{j=1}^{c}Y_j\sum_{i=1}^{r}f_{ij}$$

$$=\sum_{i=1}^{r}X_i f_{i}+\sum_{j=1}^{c}Y_j f_{.j}=\mu_x+\mu_y$$

Esempio

Y/X	2	4	6	
1	$4/20$	$2/20$	$4/20$	$10/20$
3	$4/20$	$1/20$	$5/20$	$10/20$
	$8/20$	$3/20$	$9/20$	1

$$E(X+Y)=(2+1)\left(\frac{4}{20}\right)+(4+1)\left(\frac{2}{20}\right)+(6+1)\left(\frac{4}{20}\right)+$$
$$(2+3)\left(\frac{4}{20}\right)+(4+3)\left(\frac{1}{20}\right)+(6+3)\left(\frac{5}{20}\right)$$
$$=\frac{12+10+28+20+7+45}{20}=\frac{122}{20}=6.1$$

$$\mu_x+\mu_y=4.1+2=6.1$$

Valore atteso del prodotto

$$E(XY)=\sum_{i=1}^{r}\sum_{j=1}^{c}X_i Y_j f_{ij}$$

Esempio

Y/X	2	4	6	
1	$4/20$	$2/20$	$4/20$	$10/20$
3	$4/20$	$1/20$	$5/20$	$10/20$
	$8/20$	$3/20$	$9/20$	1

$$E(XY)=(2)\left(\frac{4}{20}\right)+(4)\left(\frac{2}{20}\right)+(6)\left(\frac{4}{20}\right)+$$
$$(6)\left(\frac{4}{20}\right)+(12)\left(\frac{1}{20}\right)+(18)\left(\frac{5}{20}\right)$$
$$=\frac{8+8+24+24+12+90}{20}=8.3$$

E(XY) in caso di indipendenza

$$E(XY)=\sum_{i=1}^{r}\sum_{j=1}^{c}X_i Y_j f_{i.} f_{.j}=\sum_{i=1}^{r}X_i f_{i.}\sum_{j=1}^{c}Y_j f_{.j}=\mu_x\mu_y$$

In questo caso, la media dei prodotti è pari al prodotto delle medie.

Esempio

Y/X	2	4	6	
1	$3/20$	$3/20$	$6/20$	$12/20$
3	$2/20$	$2/20$	$4/20$	$8/20$
	$5/20$	$5/20$	$10/20$	1

$$E(XY)=(2)\left(\frac{3}{20}\right)+(4)\left(\frac{3}{20}\right)+(6)\left(\frac{6}{20}\right)+$$
$$(6)\left(\frac{2}{20}\right)+(12)\left(\frac{2}{20}\right)+(18)\left(\frac{4}{20}\right)$$
$$=\frac{6+12+36+12+24+72}{20}=8.1$$

$$\mu_x=(2)\left(\frac{5}{20}\right)+(4)\left(\frac{5}{20}\right)+(6)\left(\frac{10}{20}\right)=4.5$$

$$\mu_y=(1)\left(\frac{12}{20}\right)+(3)\left(\frac{8}{20}\right)=1.8;\quad \mu_x\mu_y=4.5*1.8=8.1$$

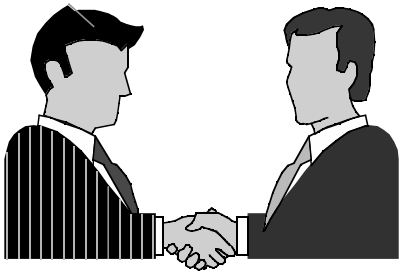
Esercizio

	Variabile Y				
		-1	*0*	*1*	
Var.	*-2*	1/25	2/25	3/25	6/25
X	*-1*	3/25	1/25	2/25	6/25
	0	4/25	3/25	6/25	13/25
		8/25	6/25	11/25	1

- **Calcolare E(X+Y)**
- **Calcolare E(X-Y)**
- **Calcolare E(XY)**

La concordanza

Un aspetto essenziale della dipendenza tra due variabili su scala almeno intervallare è la concordanza, cioè la ricerca della direzione della dipendenza tra Y ed X.



Ci si chiede se valori inferiori (superiori) alla media si accompagnino con valori inferiori (superiori) alla media nell'altra

Per ognuna delle combinazione di possibili valori si può averne una indicazione dagli SCARTI MISTI:

$$S_i = (X_i - \mu_x)(Y_i - \mu_y)$$

Significato della concordanza

Il segno degli scarti è utile per sapere se, per la combinazione dei valori "X_i" e "Y_i" l'andamento delle due variabili è concorde oppure discorde:

CONCORDANZA

$$S_i > 0 \Rightarrow (X_i > \mu_x) \; e \; (Y_i > \mu_y) \quad oppure \quad (X_i < \mu_x) \; e \; (Y_i < \mu_y)$$

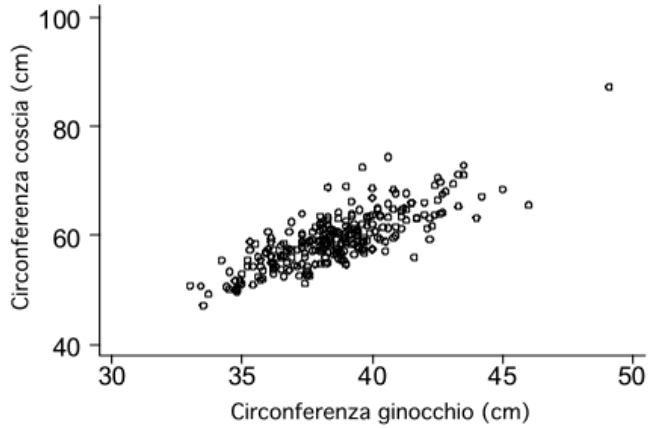
DISCORDANZA

$$S_i > 0 \Rightarrow (X_i > \mu_x) \; e \; (Y_i < \mu_y) \quad oppure \quad (X_i < \mu_x) \; e \; (Y_i > \mu_y)$$

E' difficile cogliere il senso della concordanza analizzando uno per uno TUTTI gli scarti misti.

Scatterplot (diagramma di dispersione)

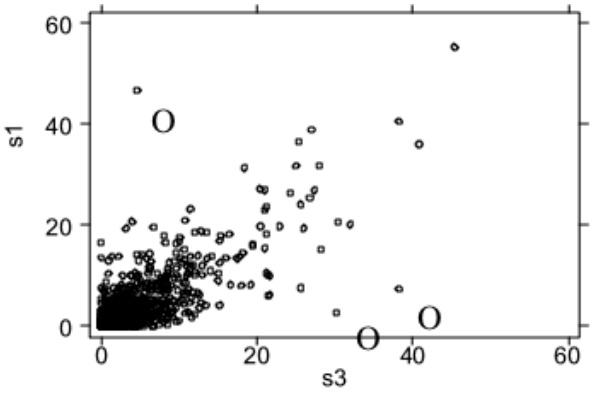
Su due assi coordinati ed in scala opportuna si riportano i valori delle due variabili ed ogni combinazione (X,Y) è rappresentata da un punto.



Questo è il grafico più noto ed è di realizzazione e lettura molto semplice evidenziando il prevalere o meno della concordanza.

Scatterplot/2

Lo scatterplot offre una comoda rappresentazione delle possibili relazioni tra due variabili quantitative.

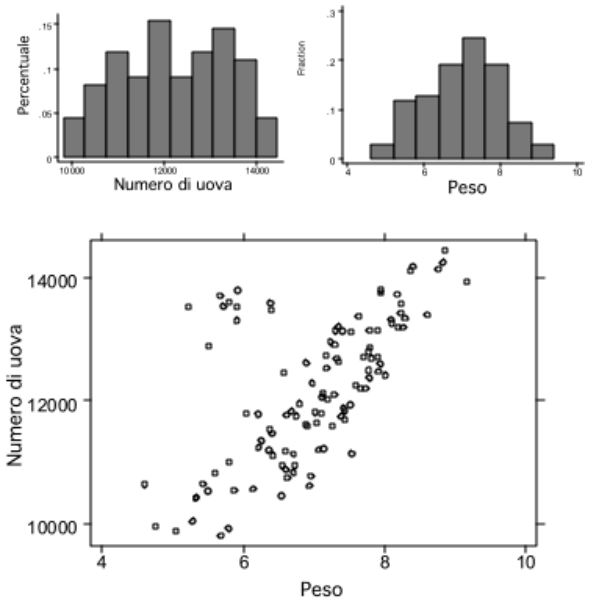


Il grafico evidenzia il gradiente dei dati, l'intensità del legame nonché i possibili valori anomali (Outliers) cioè osservazioni lontane, a prima vista, dal centro della relazione.

Esempio

La presentazione congiunta delle due variabili rivela aspetti che rimangono oscurati nella rappresentazione separata dei due aspetti.

Lo scatterplot indica la presenza di un gruppo di soggetti (in alto a sinistra) diversi dal resto.

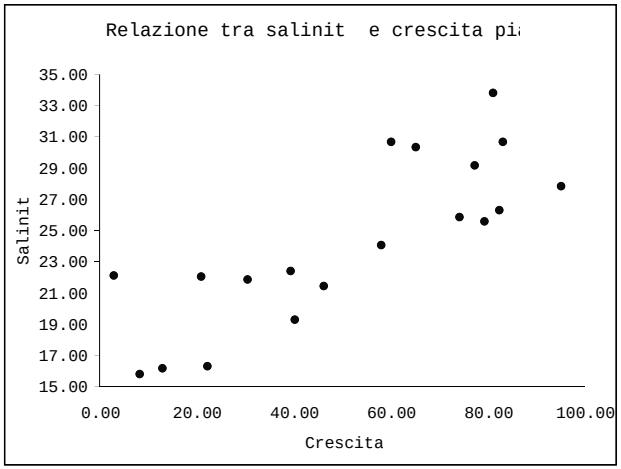


Esercizio

C'e una relazione tra il tasso di crescita delle mangrovie e la salinità del suolo?

Utilizzare il foglio elettronico per ottenere la rappresentazione grafica.

Prelievi	Salinity	Crescita
1	2.90	22.12
2	40.25	19.29
3	60.05	30.69
4	8.24	15.80
5	58.05	24.08
6	95.07	27.85
7	79.31	25.58
8	8.35	14.59
9	12.93	16.17
10	22.21	16.31
11	77.23	29.17
12	74.11	25.87
13	20.91	22.05
14	83.08	30.68
15	81.02	33.82
16	82.31	26.30
17	46.19	21.45
18	65.12	30.34
19	30.46	21.86
20	39.31	22.42



Dominano gli scartj concordi

La covarianza

La sintesi più semplice di tutti gli scarti misti è il loro valore atteso che costituisce la covarianza tra Y ed X

$$Cov(X,Y) = \left(\frac{1}{n}\right)\sum_{i=1}^{n}\left(X_i - \mu_x\right)\left(Y_i - \mu_y\right)$$

Se Cov(Y,X)>0; Predominano gli scarti di segno concorde. Ci si aspetta X e Y tendano a cambiare nella stessa direzione

Se Cov(Y,X)<0; Predominano gli scarti di segno discorde. Ci si aspetta X e Y tendano a cambiare in direzioni opposte

Se Cov(Y,X)=0; le forze di discordanza e di concordanza sono bilanciate

Esempio di calcolo della covarianza

	A	B	C	D	E	F	G
1	Guidatrice	Percorso	Velocità media				
2	Anna	87.05	53.77		94.6790		
3	Gina	76.82	9.11		=COVARIANZA(B2:B21;C2:C21)		
4	Lena	87.56	64.02				
5	Lara	98.04	96.39				
6	Bice	86.28	50.05				
7	Zora	82.97	38.70				
8	Olga	85.38	53.60				
9	Rosa	88.01	71.17				
10	Pina	81.80	39.04				
11	Dina	86.74	51.71				
12	Dora	85.19	56.92				
13	Fede	84.28	53.07				
14	Beba	97.44	97.43				
15	Tina	85.96	55.53				
16	Sara	83.97	48.46				
17	Elsa	89.04	63.51				
18	Enza	94.77	81.92				
19	Edda	92.06	78.58				
20	Mara	88.06	64.08				
21	Emma	85.54	53.99				

Dominano gli scarti concordi

Formula semplificata per la covarianza

Usando le proprietà delle sommatorie si ottiene

$$Cov(X,Y)=\left(\frac{1}{n}\right)\sum_{i=1}^{n}(X_i-\mu_x)(Y_i-\mu_y)=\left(\frac{1}{n}\right)\sum_{i=1}^{n}(X_i-\mu_x)Y_i-(X_i-\mu_x)\mu_y$$
$$=\left(\frac{1}{n}\right)\sum_{i=1}^{n}(X_i-\mu_x)Y_i-\left(\frac{1}{n}\right)\mu_y\sum_{i=1}^{n}(X_i-\mu_x)=\left(\frac{1}{n}\right)\sum_{i=1}^{n}(X_iY_i-\mu_xY_i)-0$$
$$=\left(\frac{1}{n}\right)\sum_{i=1}^{n}X_iY_i-\left(\frac{1}{n}\right)\mu_x\sum_{i=1}^{n}Y_i=\left(\frac{1}{n}\right)\sum_{i=1}^{n}X_iY_i-\left(\frac{1}{n}\right)\mu_x n\mu_y=\left(\frac{1}{n}\right)\sum_{i=1}^{n}X_iY_i-\mu_x\mu_y$$

che semplifica il calcolo e soprattutto l'interpretazione della covarianza che è nulla in caso di indipendenza in distribuzione delle due variabili

Covarianza e trasformazioni lineari

La covarianza, come la varianza, risente di trasformazioni moltiplicative, ma non di quelle additive

$$W_i=a+bX_i;\quad Z_i=c+dY_i$$

$$Cov(X,Y)=\left(\frac{1}{n}\right)\sum_{i=1}^{n}W_iZ_i-\mu_w\mu_z=\left(\frac{1}{n}\right)\sum_{i=1}^{n}(a+bX_i)(c+dY_i)-[a+b\mu_x][c+d\mu_y]$$
$$=\left(\frac{1}{n}\right)\sum_{i=1}^{n}[ac+bcX_i+adY_i+bdX_iY_i]-[ac+ad\mu_y+bc\mu_x+bd\mu_x\mu_y]$$
$$=bd\left(\frac{1}{n}\right)\sum_{i=1}^{n}X_iY_i-bd\mu_x\mu_y=bdCov(X,Y)$$

i parametri additivi "a" e "c" sono scomparsi, quelli moltiplicativi compaiono come fattore

Disuguaglianza Cauchy-Schwartz

Consideriamo la relazione che lega linearmente gli scarti medi di Y agli scarti medi di X

$$\left(\frac{1}{n}\right)\sum_{i=1}^{n}\left[(Y_i-\mu_y)-b(X_i-\mu_x)\right]^2=\left(\frac{1}{n}\right)\left[\sum_{i=1}^{n}(Y_i-\mu_y)^2+b^2(X_i-\mu_x)^2-2b(Y_i-\mu_y)(X_i-\mu_x)\right]\geq 0$$
$$=\left(\frac{1}{n}\right)\left[\sum_{i=1}^{n}(Y_i-\mu_y)^2+b^2\left(\frac{1}{n}\right)\sum_{i=1}^{n}(X_i-\mu_x)^2-2b\left(\frac{1}{n}\right)\sum_{i=1}^{n}(Y_i-\mu_y)(X_i-\mu_x)\right]$$
$$=Var(Y)+b^2Var(X)-2Cov(X,Y)\geq 0$$

Perchè tale disequazione di 2° grado in "b" sia sempre soddisfatta, il discriminante deve essere negativo e cioé:

$$=[2\mathrm{Cov}(x,y)]^2-4\mathrm{var}(x)*\mathrm{var}(y)\leq 0\quad \text{ovvero}\quad [\mathrm{Cov}(x,y)]^2\leq \mathrm{var}(x)*\mathrm{var}(y)$$

La covarianza, al quadrato, è inferiore o uguale al prodotto delle varianze delle distribuzioni marginali

Limiti della Covarianza

La covarianza ha tutti i difetti delle misure di variabilità assoluta (dipendenza dall'unità di misura, mancanza di limiti predefiniti, etc.)

E' però legata alla variabilità dei due caratteri nel senso che non può superare, in valore assoluto, il prodotto degli SQM di Y e X.

Per ottenere un indice normalizzato e standardizzato covarianza è calcolata sulle variabili standardizzate. L'indice risultante si chiama coefficiente di correlazione

$$Cov(X^*,Y^*)=\left(\frac{1}{n}\right)\sum_{i=1}^{n}\left(\frac{X_i-\mu_x}{\sigma_x}\right)\left(\frac{Y_i-\mu_y}{\sigma_y}\right)$$

Coefficiente di correlazione

- **E' compreso tra -1 e +1 perché espresso come rapporto di una quantità (la covarianza) al suo massimo (in valore assoluto)**

$$r(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X) * Var(Y)}}$$

- **E' standardizzato. Se una o entrambe le variabili subiscono una trasformazione lineare il coefficiente rimane lo stesso:**

r(a+bX,c+dY) = r(X,Y)

- **E' simmetrico rispetto alle due variabili: r(Y,X)=r(X,Y)**

- **E' uguale a zero se c'è indipendenza tra le due variabili (il numeratore in questo caso è infatti zero)**

Coefficiente di correlazione/2

Assume i valori estremi solo in caso di relazione lineare esatta

$$Cov(X,a+bX) = \left(\frac{1}{n}\right)\sum_{i=1}^{n} X_i(a+bX_i) - \mu_x(a+b\mu_x) = \left(\frac{1}{n}\right)\left[\sum_{i=1}^{n}(aX_i) + \sum_{i=1}^{n} bX_i^2\right] - a\mu_x - b\mu_x^2$$

$$= a\left(\frac{1}{n}\right)\sum_{i=1}^{n}(aX_i) - a\mu_x + b\left[\left(\frac{1}{n}\right)\sum_{i=1}^{n} X_i^2 - \mu_x^2\right] = a\mu_x - a\mu_x + b[Var(x)]$$

$$= bVar(x)$$

Ne consegue che

$$r(X,a+bX) = \frac{bVar(x)}{\sqrt{Var(x)Var(a+bx)}} = \frac{bVar(x)}{\sqrt{Var(x)b^2Var(x)}} = \frac{b}{|b|} = \begin{cases} -1 \ \ se \ \ b<0 \\ +1 \ \ se \ \ b>0 \end{cases}$$

il coefficiente di correlazione misura, quindi, l'intensità del legame lineare che sussiste tra le due variabili.

Calcolo di r(x,y)

il calcolo è molto semplice purché opportunamente organizzato.

Supponiamo che le due variabili presentino con frequenza 1/6 le coppie di valori inserite nelle prime due colonne.

	X	Y	X2	Y2	XY
	0	0	0	0	0
	1	0	1	0	0
	1	0	1	0	0
	1	1	1	1	1
	1	1	1	1	1
	2	1	4	1	2
Σ=	6	3	8	3	4

$$\mu_x = \frac{6}{6} = 1; \quad \mu_y = \frac{3}{6} = \frac{1}{2};$$

$$\sigma^2(X) = \frac{8}{6} - 1^2 = \frac{1}{3}; \quad \sigma^2(Y) = \frac{3}{6} - \left(\frac{1}{2}\right)^2 = \frac{1}{4};$$

$$Cov(X,Y) = \frac{4}{6} - 1*\frac{1}{2} = \frac{1}{6}; \quad r(X,Y) = 0.5773$$

Le due variabili presentano una correlazione positiva tendendo a presentare insieme i valori più grandi

P-Value per r(x,y)

Una volta calcolato r(x,y) su di un campione di osservazioni cosa si può dire sul coefficiente di correlazione che lega le due variabili nell'intera popolazione?

$H_0: \ \rho = 0$ *(assenza di un legame lineare)*

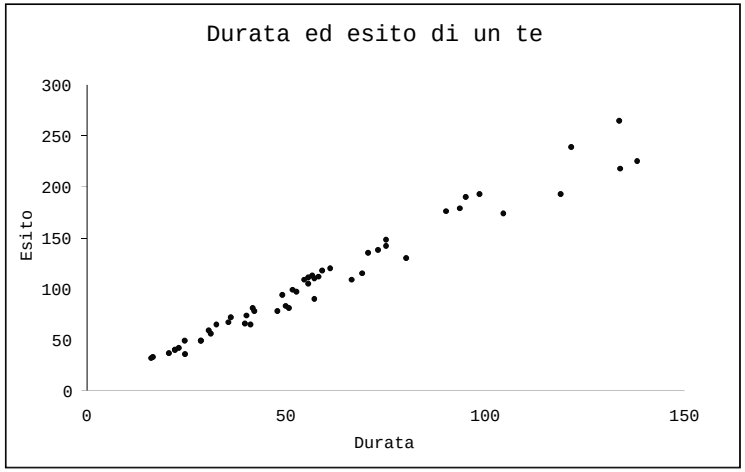
$H_1: \ \rho \neq 0$ *(presenza di un legame lineare)*

La statistica test che si po' usare è $t_c = \sqrt{n-2}\left(\frac{r(y,x)}{\sqrt{1-[r(y,x)]^2}}\right)$

La cui distribuzione è approssimata dalla t-Student con (n-2) gradi di libertà

Esercizio - Excel

Esito test	Durata		
134	265	53	97
122	239	49	94
138	225	57	90
134	218	50	83
119	193	42	81
99	193	51	81
95	190	42	78
94	179	48	78
90	176	40	74
104	174	36	72
75	148	36	67
75	142	40	66
73	138	41	65
71	135	33	65
80	130	31	59
61	120	31	56
59	118	25	49
69	115	29	49
57	113	29	49
58	112	23	42
56	111	22	40
57	110	22	40
66	109	21	37
55	109	25	36
56	105	17	33
52	99	16	32



=CORREL(B2:B53;C2:C53)

$$n = 52 : gld = 50 : t_c = 39.56, \quad p - Value = 0$$

Esercizio

Supponendo che una v.c. doppia presenti i seguenti valori

X	Y
21	40
63	134
32	78
48	103
35	85
66	137
50	119
45	96
61	122
19	41

a) Disegnare lo SCATTERPLOT

b) Calcolare la Correlazione tra X e Y

Significato di r(x,y)

Dalle proprietà di "r(x,y)" si deduce il suo significato:

Quanto più i suoi valori si avvicinano, in modulo, ad uno tanto più i valori delle variabili risultano collegabili con una retta.

D'altra parte, quanto più "r" è vicino a "±1" tanto più la conoscenza di una delle variabili permette, attraverso la relazione lineare, di conoscere l'altra.

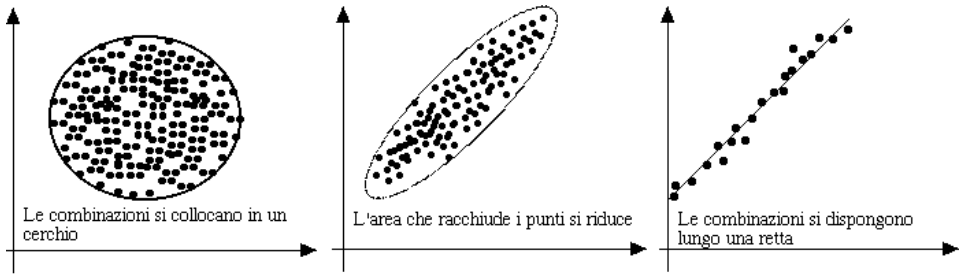
In questo senso "r" è una misura del grado di concordanza tra i valori della variabile doppia (X,Y)

- **INTENSITA' DEL LEGAME LINEARE**
- **PREVEDIBILITA' DI UNA VARIABIULE CONOSCENDO L'ALtra**
- **GRADO DI CONCORDANZA**

Scatterplot e correlazione

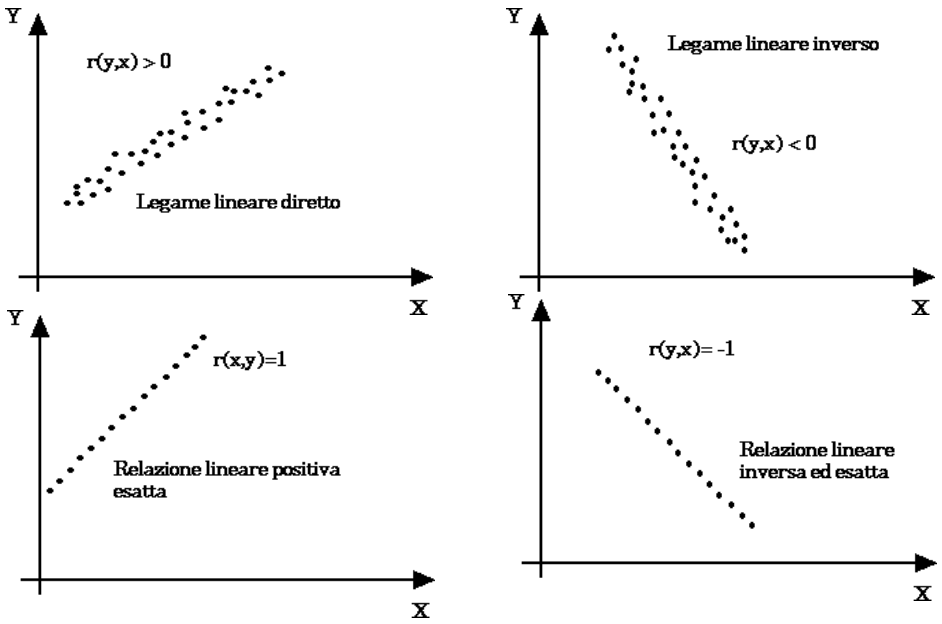
Lo scatterplot fornisce una idea immediata della intensità del legame che intercorre tra le due variabili

Si realizza riportando -in scala opportuna- le combinazioni osservate dei valori

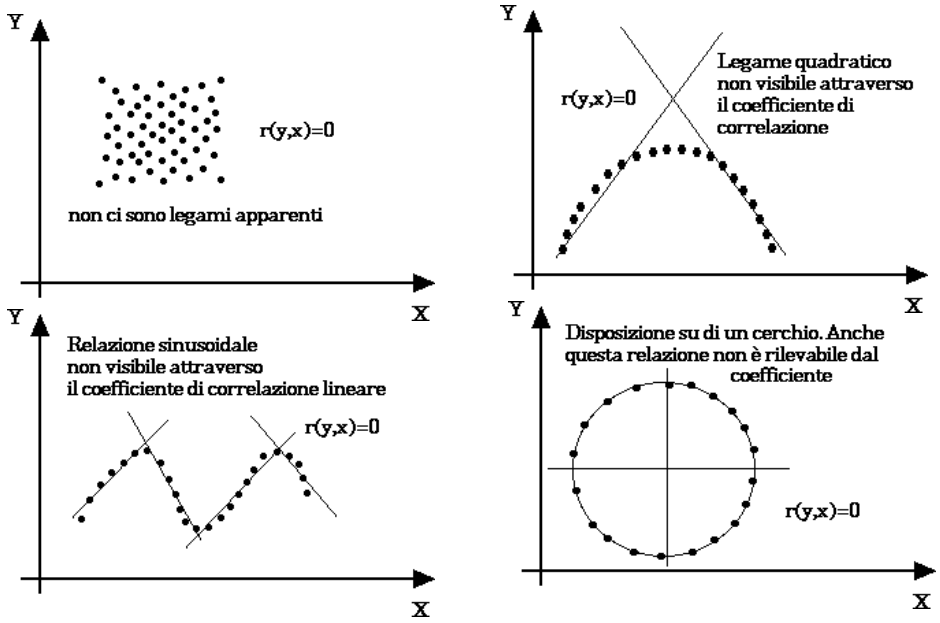


La relazione tra due variabili tende a divenire più stretta ma mano che la nube di punti passa dalla forma circolare, alla ellisse ed alla retta

Scatterplot e correlazione/2



Assenza di legami lineari



Correlazione e causa-effetto

L'esistenza di correlazione, per quanto intensa, non implica una relazione di causa ed effetto.

LEGAME PLAUSIBILE

Il tasso di criminalità è fortemente legato al tasso di disoccupazione.

LEGAME SPURIO

Nei bambini, la misura delle scarpe è molto correlata con la capacità di lettura.

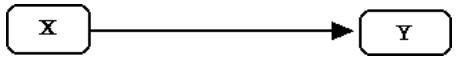
La correlazione indica solo che l'andamento di una variabile tende a disporsi secondo una retta se rappresentato in un diagramma cartesiano insieme all'altra.

I "perchè?" di questa tendenza vanno cercati al di fuori della statistica.

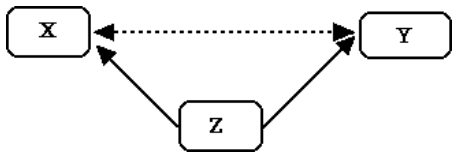
Correlazione spuria

Spesso, il valore di r(y,x) altro non è che l'apparenza di un legame la cui sostanza è invece dovuta a fenomeni esterni.

La situazione di causalità tra X e Y:



Non è distinguibile dal legame spurio che fra di esse si pone a causa della comune dipendenza da una terza variabile Z



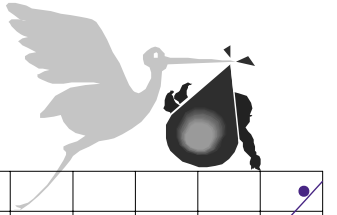
L'apprendimento di nuove parole non rende i piedi più grandi ovvero avere piedi più grandi non aiuta a conoscere nuove parole.
C'un terzo fattore nascosto dietro la correlazione: l'età

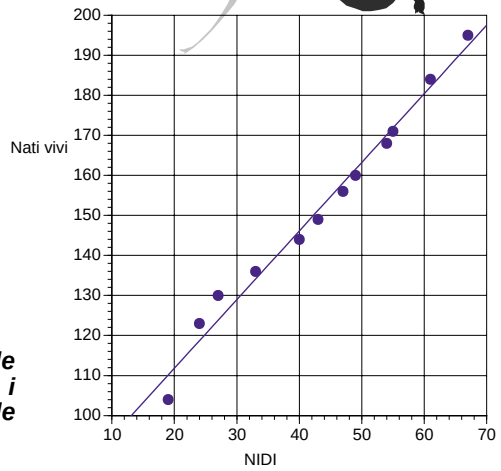
Questo si verifica spesso a causa dell'esistenza di fenomeni tendenziali di lungo periodo che incidono allo stesso modo su variabili diverse

Esempio

In una zona del Nord Europa è stato monitorato il numero di nidi costruiti dalle cicogne ed il numero di nati vivi nel loro periodo di permanenza.

Anno	Nidi di cicogne	Nati vivi
1972	19	104
1973	24	123
1974	27	130
1975	33	136
1976	40	144
1977	43	149
1978	47	156
1979	49	160
1980	54	168
1981	55	171
1982	61	184
1983	67	195





Dal punto di vista della correlazione le ipotesi che siano le cicogne a portare i bambini o che siano i bambini a portare le cicogne sono equivalenti.

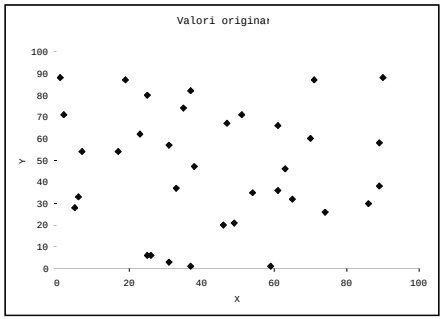
Esercizio

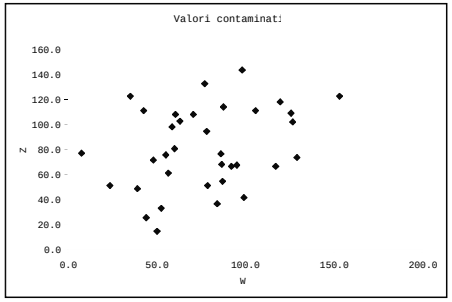
X	Y	W	Z	A
54	35	87.0	68.0	22
63	46	126.0	109.0	42
71	87	78.5	94.5	5
35	74	59.0	98.0	16
1	88	35.5	122.5	23
33	37	57.0	61.0	16
61	66	106.0	111.0	30
49	21	79.0	51.0	20
86	30	129.5	73.5	29
25	6	44.5	25.5	13
65	32	87.5	54.5	15
70	60	86.5	76.5	11
61	36	127.0	102.0	44
38	47	39.5	48.5	1
5	28	48.5	71.5	29
47	67	60.5	80.5	9
19	87	43.0	111.0	16
74	26	84.5	36.5	7
26	6	53.0	33.0	18
89	38	117.5	66.5	19
25	80	77.5	132.5	35
2	71	8.0	77.0	4
7	54	61.0	108.0	36
31	3	95.5	67.5	43
17	54	71.0	108.0	36
31	57	88.0	114.0	38
37	82	98.5	143.5	41
6	33	24.0	51.0	12
59	1	99.5	41.5	27
89	58	153.5	122.5	43
90	88	120.0	118.0	20
46	20	92.5	66.5	31
51	71	55.5	75.5	3
37	1	50.5	14.5	9
23	62	63.5	102.5	27

```
Corr(X,Y)  -0.037
Corr(W,Z)   0.309
```

$$W = X + 1.2A$$

$$Z = Y + 1.2A$$





Dipendenza dei ranghi

Riguarda le variabili riportate in scala quantitativa ordinale.

> Perché non esiste una vera misura, ma solo un punteggio o valutazione

> Perché le misurazioni su sono imprecise o viziate da errore

> Perché sono presenti dei valori remoti

Le modalità sono poste in corrispondenza con dei numeri naturali (ranghi)

Per ogni unità si osserva una coppia di modalità che si trasforma poi in una coppia di ranghi

Esempio

Un gruppo di clienti di una banca classificato per reddito e per importo del prestito

Cliente	Reddito (x)	Rango (x)	Prestito (y)	Rango (y)
A	25 600	6	8 600	5
B	17 800	9	8 800	4
C	167 200	1	500	8
D	44 200	3	6 600	6
E	36 400	4	10 500	3
F	27 400	5	74 400	1
G	83 600	2	0	9
H	18 600	8	6 300	7
I	24 500	7	12 100	2

Correlazione tra ranghi (rho di Spearman)

La misura forse più popolare della dipendenza tra i ranghi è la seguente

$$r_S = \frac{\sum_{i=1}^{n}\left(r_i - \frac{n+1}{2}\right)\left(s_i - \frac{n+1}{2}\right)}{\sqrt{\sum_{i=1}^{n}\left(r_i - \frac{n+1}{2}\right)^2 \sum_{i=1}^{n}\left(s_i - \frac{n+1}{2}\right)^2}}$$

detto rho di Spearman

Caso delle n coppie di valori senza posizioni di parità.

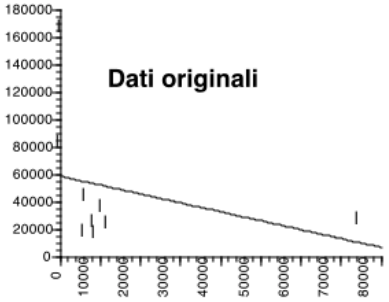
La definizione di r_S è la stessa del coefficiente di correlazione. Comunque il particolare tipo di dati coinvolti consente delle semplificazioni

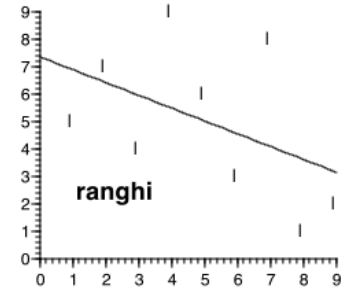
$$r_S = 1 - \frac{6\sum_{i=1}^{n}(r_i - s_i)^2}{n\left(n^2 - 1\right)}$$

Esempio

Cliente	Reddito	r_i	Prestito	s_i	$(r_i - s_i)^2$
A	25600	6	8600	5	1
B	17800	9	8800	4	25
C	167200	1	500	8	49
D	44200	3	6600	6	9
E	36400	4	10500	3	1
F	27400	5	74400	1	16
G	83600	2	0	9	49
H	18600	8	6300	7	1
I	24500	7	12100	2	25
					176

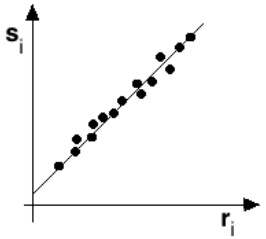
$$r_S = 1 - \frac{6*176}{9*(81-1)} = -0.4667$$

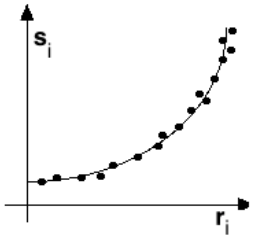




Considerazione sul rho di Spearman

- **Per costruzione l'indice rho varia tra -1 ed 1**
- **Misura la dipendenza monotonica tra le due variabili. Assume il valore massimo (minimo) quando gli ordinamenti sono perfettamente concordi (discordi)**





r_S = ±1 se tra la Y e X esiste una relazione monotona, crescente o decrescente, senza vincolo di linearità

- **Rho mette tutte le osservazioni sullo stesso piano (considera solo l'ordine) e perciò annulla l'effetto dei valori remoti**
- **L'indice rimane invariato se una o entrambe le variabili subiscono una trasformazione monotona, ad esempio Y'=Log(Y)e/o X'=exp(X)**

P-Value per rho

Una volta calcolato rho su di un campione, cosa si può dire sul rho di Spearman che lega le variabili nell'intera popolazione?

$$H_0: \ \rho = 0 \ \ (NON \ \ esiste \ \ una \ \ relazione \ \ monotona)$$
$$H_1: \ \rho \neq 0 \ \ (Esiste \ \ una \ \ relazione \ \ monotona)$$

La statistica test che si puo' usare è la stessa del coefficiente di correlazione

$$t_c = \sqrt{n-2}\left(\frac{\rho}{\sqrt{1-\rho^2}}\right)$$

Per n<11 deve essere usata con prudenza

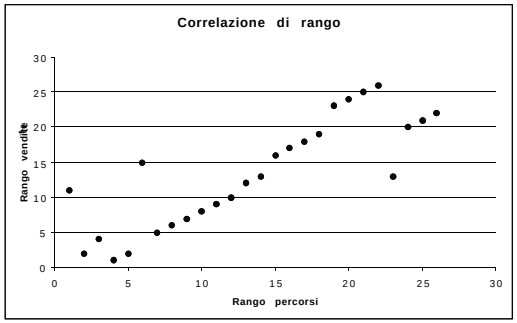
La cui distribuzione è approssimata dalla t-Student con (n-2) gradi di libertà

Esempio

Unit	X Percorsi	Y Vendite	Rank(X)	Rank(Y)
A	121.5	373	21	25
B	151.5	314	25	21
C	146.2	301	24	20
D	106.7	263	16	17
E	98.9	204	11	9
F	95.1	176	9	7
G	90.1	138	4	1
H	115.5	329	19	23
I	71.7	225	1	11
J	111.7	300	18	19
K	93.6	164	7	5
L	109.6	284	17	18
M	105.3	252	15	16
N	125.0	400	22	26
O	91.7	239	6	15
P	88.7	161	3	4
Q	101.9	226	13	12
R	162.3	322	26	22
S	96.4	185	10	8
T	90.7	143	5	2
U	100.0	212	12	10
V	102.6	232	14	13
X	94.5	171	8	6
Y	88.6	143	2	2
W	119.4	358	20	24
Z	142.9	232	23	13

rho= 0.850084703
gdl= 25
tc= 7.907679188
p-Value 2.90161E-08

Venditori porta-a-porta per vendite e Km percorsi



La correlazione è elevata e significativa

Esercizio

Calcolare il rho per le graduatorie dei fattori essenziali per la localizzazione di un nuovo impianto nel Mezzogiorno. Disegnare lo scatterplot

Fattori	Imprese Meridionali	Imprese non Meridionali
Costo del lavoro	1	1
Incentivi	2	2
Trasporti	7	5
Preferenze Personali	13	13
Servizi alle imprese	11	6
Finanziameni	4	12
Materie prime	10	9
Sicurezza	8	3
Tassazione	6	7
Mercato	5	4
Commesse pubbliche	3	10
Collegamenti intern.	12	11
Qualità della vita	9	8

Rilevazione diretta dei ranghi

Un certo insieme di *n* oggetti o situazioni sono ordinate secondo il grado con cui presentano una certa caratteristica X.

Supponiamo …
Che la caratteristica sia un mix di immaterialità che può essere graduato, ma non misurato.

Che le valutazioni siano espresse con voti {1,2,…,n} così ottenendo la permutazione $\{s_1, s_2,\ldots,s_n\}$

Ripetiamo la rilevazione per una *Y* misurata allo stesso modo e che produce una diversa permutazione degli interi: $\{q_1, q_2,\ldots,q_n\}$

Condizione di ansia e stress
Prima e dopo una separazione

Il rho di Spearman cerca di quantificare l'intensità del legame monotono tra i due insiemi di giudizi

Esempio: giudizi degli esperti

Ad un esperto è stato chiesto di pronunciarsi sulla posizione che le 20 squadre di un campionato di calcio occuperanno alla fine: $\{s_1, s_2,\ldots,s_{20}\}$.

Alla fine della stagione i giudizi sono comparati con le posizioni reali: $\{q_1, q_2,\ldots,q_{20}\}$.

Per semplificare il calcolo possiamo disporre le due serie di posizioni secondo l'ordine crescente della prima

Squadra	A	B	C	D	E	F	G	H	I	L	M	N	O	P	Q	R	S	T	U	V	Totale
Prima	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	210
Dopo	9	2	4	7	5	1	3	8	6	11	13	10	14	18	15	12	16	20	17	19	210

ρ=0.87 (p-value 0.000001). L'esperto ha dato un buon giudizio sebbene sembri più in grado di indovinare le squadre che avranno una cattiva stagione rispetto a quelle che l'avranno buona

Esercizio

Ad un campione di consumatori stato chiesto di giudicare la qualit di un servizio con un voto da 0 a 12.

E anche stato chiesto di valutare con un voto da 0 a 12 la reputazione dell azienda che forniva il servizio

Azienda	Rating servizio	Reputazione azienda
Alfa01	8	9
Alfa02	9	12
Alfa03	2	0
Alfa04	5	10
Alfa05	6	6
Alfa06	4	11
Alfa07	7	4
Alfa08	10	3
Alfa09	3	5
Alfa10	1	2
Alfa11	0	1
Alfa12	12	7
Alfa13	11	8

Risulta che ci sia un legame tra le due valutazioni?

Rho-Spearman=	0.4890
Tc=	1.8593
p-value	0.0899

Presenza di valori uguali

Se ci sono degli ex-aequo sorge il problema di assegnare il rango ai valori uguali

In genere si assegna a ciascun elemento di un gruppo di parità, la media dei ranghi che sarebbero loro spettati se fossero stati distinti.

Esempio

intensità	101	108	108	117	117	117	141	142	154	154	154	154	154	173
ranghi teorici	1	2	3	4	5	6	7	8	9	10	11	12	13	14
medie		(2+3)/2=2.5		(4+5+6)/3=5					(9+10+11+12+13)/5 = 11					
ranghi reali	1	2.5	2.5	5	5	5	7	8	11	11	11	11	11	14

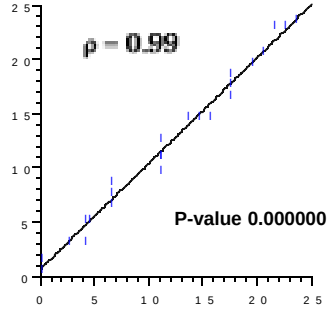
Ai fini del calcolo nulla è cambiato se non la perdita delle semplificazioni possibili solo in caso di assenza di parità

Esempio

E	E-or	rE	B	B-or	rB	$(rE)^2$	$(rB)^2$	rE*rB
10.3	9.2	0.5	-1.3	-1.8	1	0.25	1	0.5
9.9	9.2	0.5	-1.1	-1.6	2	0.25	4	1
10.3	9.3	3.0	1.3	-1.5	3.5	9	12.25	10.5
9.7	9.4	4.5	0.5	-1.5	3.5	20.25	12.25	15.75
9.6	9.4	4.5	1.8	-1.3	5.5	20.25	30.25	24.75
9.5	9.5	5.0	-0.4	-1.3	5.5	25	30.25	27.5
10.8	9.8	7.0	-1.6	-1.1	7	49	49	49
9.7	9.6	7.0	-1.8	-0.8	8	49	64	56
9.4	9.6	7.0	1.0	-0.7	9	49	81	63
10.1	9.7	11.5	-1.5	-0.5	10	132.25	100	115
9.4	9.7	11.5	-0.8	-0.4	11.5	132.25	132.25	132.25
10.4	9.7	11.5	-1.3	-0.4	11.5	132.25	132.25	132.25
9.6	9.7	11.5	-0.7	-0.2	13	132.25	169	149.5
9.7	9.8	14.0	-0.4	0.5	15	196	225	210
10.6	9.8	15.0	0.5	0.5	15	225	225	225
9.3	10.1	16.0	1.3	0.5	15	256	225	240
10.5	10.3	18.0	1.1	0.7	17	324	289	306
10.7	10.3	18.0	0.9	0.8	18	324	324	324
9.2	10.3	18.0	0.5	0.9	19	324	361	342
9.7	10.4	20.0	0.8	1.0	20	400	400	400
9.2	10.5	21.0	-1.5	1.1	21	441	441	441
9.6	10.6	22.0	-0.2	1.3	23.5	484	552.25	517
9.8	10.7	23.0	0.7	1.3	23.5	529	552.25	540.5
10.3	10.8	24.0	-0.5	1.9	24	576	576	576
						4130	4988	4898.5

Accertamento di una relazione d'ordine tra il tasso di interesse effettivo "E" dei BOT trimestrali e l'indice di borsa "B"

$$r_S = \frac{\frac{4898.5}{24} - \frac{(25)^2}{4}}{\sqrt{\left[\frac{4830}{24} - \frac{(25)^2}{4}\right] * \left[\frac{4988}{24} - \frac{(25)^2}{4}\right]}}$$



Esercizio

Voti in due discipline per un camipone di studenti.

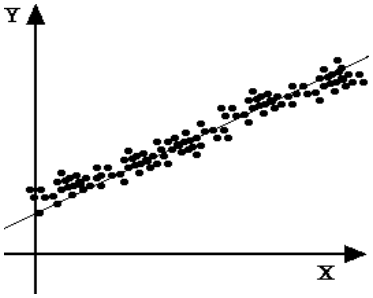
C un legame tra I due voti?

rho-Spearman	0.7836
Tc	7.5669
p-value	0.0000

Matricola	Disciplina A	Disciplina B
50825	18	18
64506	18	18
64289	18	18
31136	18	18
81016	20	19
91817	20	19
42720	20	19
92614	21	20
33491	21	20
31947	21	21
56554	21	21
83355	22	21
95516	22	21
44659	22	22
93637	22	22
70350	22	22
53806	23	24
44509	23	24
92149	23	24
86848	23	24
35750	24	24
95748	24	25
76681	25	25
70776	25	25
43071	26	26
42950	26	26
45653	26	26
56123	28	27
53240	28	27
91805	28	27
69069	28	27
77209	29	27
84099	29	27
55360	30	29
48820	30	29
76747	30	30
92951	30	30
66366	30	30

Relazione tra due variabili

Dopo aver rappresentato graficamente i dati a mezzo dello scatterplot si è interessati a determinare una curva che passi vicino ai punti

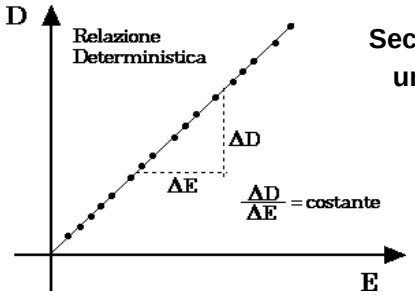


- **Per sostituire uno schema semplice alla nube dei punti**
- **Per sintetizzare le tendenze di fondo**
- **Per ricostruire o determinare il valore della Y noto quello della X o viceversa**

il presupposto è che esiste una variabile (la "X" detta indipendente o esogrena) che è causa o comunque agisce sulla'altra (la "Y" detta dipendente o endogena).

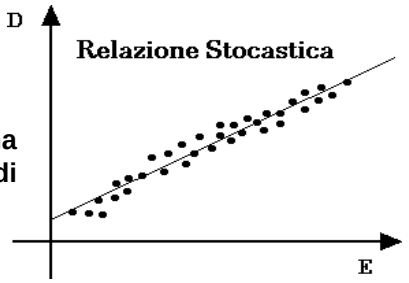
La scelta del ruolo delle due variabili è però esterna alla statistica.

Relazioni stocastiche e deterministiche



Secondo questo grafico ed ogni età corrisponde un ben determinato diametro del tronco

il diametro del tronco aumenta con l'età, ma l'incremento non è certo: talvolta aumenta di più, altre volte di meno

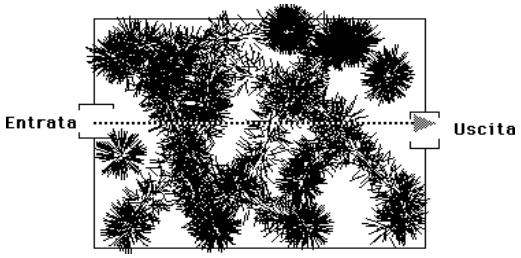


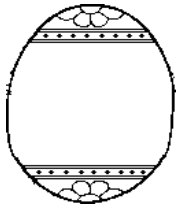
Proposta del modello lineare

La semplicità è una convenzione: giudichiamo semplice ciò che è regolare o prevedibile e che appare com'è a chiunque possa e voglia vedere

il rasoio di Occam

Se è necessario dare una soluzione ad un problema di cui si sa poco, la risposta più semplice comporta meno rischi in caso di errore ed è spesso quella giusta





L uovo di Colombo

Principio di semplicità di Galilei

La natura procede per vie semplici ed offre così la sicura scelta tra le varie spiegazioni possibili dei suoi fenomeni

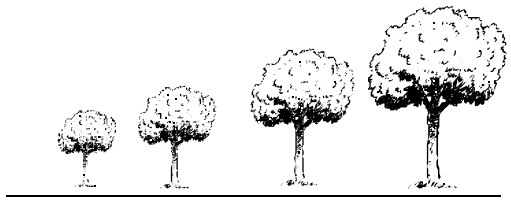
Esempio

La teoria di un fenomeno può spesso essere sintetizzata da un modello espresso da una equazione.

Sia "D" l'ampiezza in cm del diametro alla base del tronco di una data specie arborea e sia "E" l'età .

L'idea che il diametro sia più grande secondo l'età può essere espressa dalla relazione funzionale:

D=f(E)

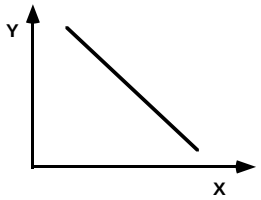


queste variazioni assicurano all'albero adeguata resistenza e flessibilità.

Proposta del modello lineare/2

il legame più semplice tra due variabili è quello lineare

$$D = \beta_0 + \beta_1 E + u$$



Ipotizziamo che l'ordinata "Y" sia dovuta alla combinazione ADDITIVA di due valori: la parte deterministica (lineare) ed un errore

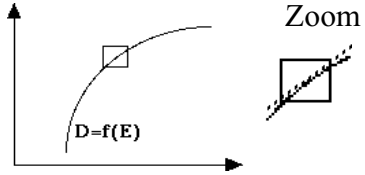
il termine "u" è il risultato di:

- **Errori e carenze nella misurazione e nella rilevazione di "D" e di "E"**
- **Insufficienza del solo fattore E a "spiegare" da solo la D**
- **Inadeguatezza della "semplice" relazione lineare**

Una ragione di più

Teorema di Taylor.

Se la funzione "f" che lega "D" ad "E'" ha derivate prime e seconde continue in un intorno del punto E0, in tale intorno la "f" è ben approssimata dalla retta



In generale si può dire che la scelta del modello lineare è motivata da

- **Ragioni di semplicità**
- **Esigenze di sintesi**
- **Approssimazione funzionale**

Formulazione matematica

Supponiamo di disporre di "n" coppie di osservazioni

X	Y
X_1	Y_1
X_2	Y_2
---	---
X_{n-1}	Y_{n-1}
X_n	Y_n

Il modello di regressione lineare semplice è

$$Y_i = \beta_0 + \beta_1 X_i + e_i \quad \text{per } i = 1,2,\ldots,n$$

"i" indice che denota l'ordine

COMPONENTE DETERMINISTICA

COMPONENTE STOCASTICA (Non osservabile)

β_0= intercetta.

Rappresenta il valore a cui tende a stabilizzarsi la Y quando la X è zero.

β_1 = Coefficiente angolare.

Rappresenta la variazione che si realizza nella Y per ogni aumento unitario in X

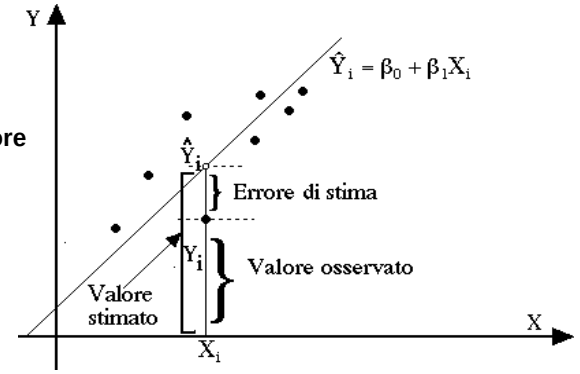
La terminologia

- ***Modello.*** **Perché è un insieme di ipotesi rispetto al legame esistente tra la variabile esogena ed endogena. Le ipotesi in genere danno luogo ad una equazione, lineare nel nostro caso**
- ***Regressione*** **. E' una etichetta storica dovuta agli studi di Francis Galton (1889) sull'effetto di regressione: la tendenza a prevalere dei valori medi.**
- ***Lineare.*** **Perché i parametri incogniti vi compaiono con potenza 1**
- ***Semplice.*** **Perchè c'è una sola variabile esplicativa (REGRESSORE) in contrapposizione a *multipla* termine usato quando vi sono più variabili esplicative.**

Calcolo dei parametri

Se per due punti passa una sola retta fra più di due punti non allineati ne passano infinite.

Ogni scelta determina degli errori dovuti alla sostituzione di un valore presunto o teorico ad un valore osservato



$\hat{y}_i =$ **valore stimato (ottenuto in base al modello)**

Occorre stabilire un criterio che ci permetta di scegliere quella che passa più vicino ai punti ovvero si adatta bene allo scatterplot

Soluzione dei minimi quadrati

Partiamo dall'errore i-esimo:

$$(y_i - \hat{y}_i) = y_i - \beta_0 - \beta_1 X_i = y_i - \beta_0 - \beta_1 X_i \pm \bar{y} \pm \beta_1 \bar{x}$$
$$= (y_i - \bar{y}) + (\bar{y} - \beta_0 - \beta_1 \bar{x}) - \beta_1 (X_i - \bar{x})$$

che evidenzia il ruolo del punto $(\bar{x}, \bar{y})$ **baricentro fisico dello scatterplot**

Elevando al quadrato e sviluppando si ottiene:

$$(y_i - \hat{y}_i)^2 = [(y_i - \bar{y}) + (\bar{y} - \beta_0 - \beta_1 \bar{x}) - \beta_1 (x_i - \bar{x})]^2 = (y_i - \bar{y})^2 + (\bar{y} - \beta_0 - \beta_1 \bar{x})^2 + \beta_1^2 (x_i - \bar{x})^2 +$$
$$+ 2(y_i - \bar{y})(\bar{y} - \beta_0 - \beta_1 \bar{x}) - 2\beta_1 (y_i - \bar{y})(x_i - \bar{x}) - 2\beta_1 (\bar{y} - \beta_0 - \beta_1 \bar{x})(x_i - \bar{x})$$

Considerando la somma di tutti gli "n" termini e ricordando che la somma degli scarti dalla media aritmetica e nulla si arriva a:

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 + \sum_{i=1}^{n} (\bar{y} - \beta_0 - \beta_1 \bar{x})^2 + \beta_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 - 2\beta_1 \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})$$

Soluzione dei minimi quadrati/2

Definiamo: $S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$; $S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$; $S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$;

Tali quantità sono note come devianze e codevianze

Ne consegue:

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = S_{yy} + n(\bar{y} - \beta_0 - \beta_1 \bar{x})^2 + \beta_1^2 S_{xx} - 2\beta_1 S_{xy}$$
$$= S_{yy} + n(\bar{y} - \beta_0 - \beta_1 \bar{x})^2 + \beta_1^2 S_{xx} - 2\beta_1 S_{xy} + \frac{S_{xy}^2}{S_{xx}} - \frac{S_{xy}^2}{S_{xx}}$$
$$= S_{yy} + n(\bar{y} - \beta_0 - \beta_1 \bar{x})^2 + S_{xx}\left[\beta_1^2 - 2\beta 11\frac{S_{xy}}{S_{xx}} + \frac{S_{xy}^2}{S_{xx}}\right] - \frac{S_{xy}^2}{S_{xx}}$$
$$= S_{yy} + n(\bar{y} - \beta_0 - \beta_1 \bar{x})^2 + S_{xx}\left[\beta_1 - \frac{S_{xy}}{S_{xx}}\right]^2 - \frac{S_{xy}^2}{S_{xx}}$$

La somma degli errori dipende dalle incognite solo attraverso dei termini al quadrato per cui il minimo si ottiene azzerando quei termini e cioè

$$\hat{\beta}_1 = \frac{S_{yx}}{S_{xx}}; \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

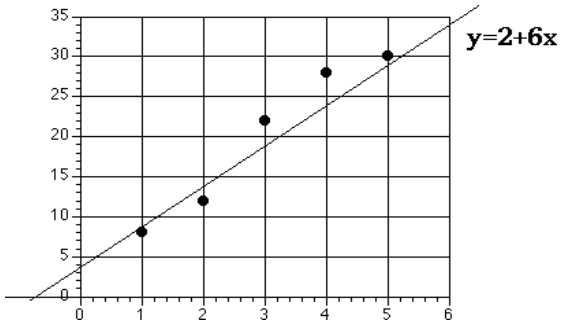
Esempio

X	Y	$(x - \bar{x})$	$(y - \bar{y})$	$(x - \bar{x})(y - \bar{y})$	$(x - \bar{x})^2$
1	8	-2	-12	24	4
2	12	-1	-8	8	1
3	22	0	2	0	0
4	28	1	8	8	1
5	30	2	10	20	4
15	100			60	10

$$\bar{y} = \frac{100}{5} = 20; \quad \bar{x} = \frac{15}{5} = 3$$

$$\hat{\beta}_1 = \frac{60/5}{10/5} = \frac{60}{10} = 6$$

$$\hat{\beta}_0 = 20 - 6 * 3 = 20 - 18 = 2$$



Altro esempio

Il prezzo dell'usato per una particolare auto è stato rilevato in alcuni esemplari

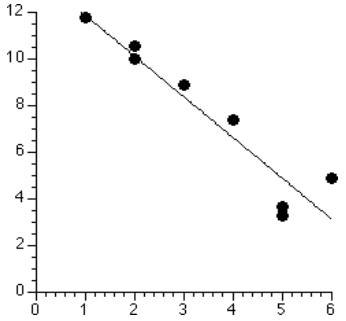
X	Y	X^2	XY
2	10.5	4	21.0
5	3.2	25	16.0
1	11.7	1	11.7
6	4.8	36	28.8
4	7.3	16	29.2
5	3.6	25	18.0
3	8.8	9	26.4
2	9.9	4	19.8
28	59.8	120	170.9

$$\begin{cases} \hat{\beta}_0 = \dfrac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} x_i y_i\right)}{n*\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2} \\ \hat{\beta}_1 = \dfrac{n*\left(\sum_{i=1}^{n} x_i y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n*\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2} \end{cases}$$

1) Calcolare le stime dei parametri
2) Disegnare la retta teorica

$$\hat{\beta}_1 = \frac{8*170.9 - 28*59.8}{8*120 - 28^2} = -1.7454$$

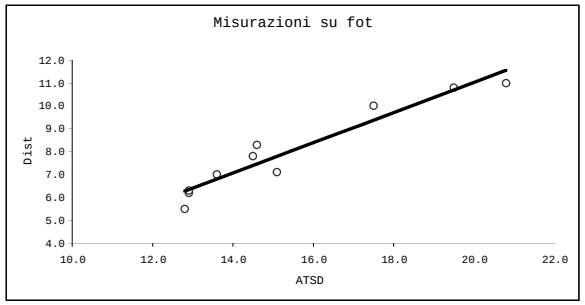
$$\hat{\beta}_0 = \frac{28}{8} + 1.7454*\frac{59.8}{8} = 16.5469$$



Esercizio con l'Excel

Di seguito si riportano dei dati relativi ad X=ampiezza totale della sede stradale e Y= distanza tra un ciclista e un auto (ottenuta con misurazioni su foto)

ATSD	DIST
12.8	5.5
12.9	6.2
12.9	6.3
13.6	7.0
14.5	7.8
14.6	8.3
15.1	7.1
17.5	10.0
19.5	10.8
20.8	11.0



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.9607
R Square	0.9229
Adjusted R Square	0.9133
Standard Error	0.5821
Observations	10

Strumenti-->Analisi dei dati-->Regressione

	Coefficients	Standard Error	t Stat	P-value
Intercept	-2.1825	1.0567	-2.0654	0.0727
X Variable 1	0.6603	0.0675	9.7858	0.0000

Esercizio

Analisi della relazione tra i rapporti Terra/Lavoro (T/L) e Redditività del Lavoro (RL/L) in alcuni settori colturali calabresi

Settore	T/L X	RL/L Y
1	1.9	3035
2	5.4	6332
3	4.0	4068
4	4.9	4943
5	5.1	4215
6	6.3	3146
7	2.0	2686
8	2.7	6775
9	5.0	4214
10	3.7	6086

1) Calcolare i parametri della regressione di Y su X
2) Disegnare la retta teorica e lo scatterplot

Proprietà della retta di regressione

La retta di regressione passa per il punto di coordinate $(\bar{x}, \bar{y})$

La retta stimata può essere scritta come:

$$y = \bar{y} + \hat{\beta}_1(x - \bar{x}) \;\Rightarrow\; \bar{y} + \hat{\beta}_1(\bar{x} - \bar{x}) = \bar{y}$$

La somma degli scarti tra osservate e teoriche è nulla:

$$\sum_{i=1}^{n} y_i - \hat{y}_i = \sum_{i=1}^{n} y_i - \bar{y} - \hat{\beta}_1(x - \bar{x}) \;\Rightarrow\; \sum_{i=1}^{n}(y_i - \bar{y}) - \hat{\beta}_1 \sum_{i=1}^{n}(x - \bar{x}) = 0 - \hat{\beta}_1 * 0 = 0$$

Ciò implica che Media osservate = Media teoriche

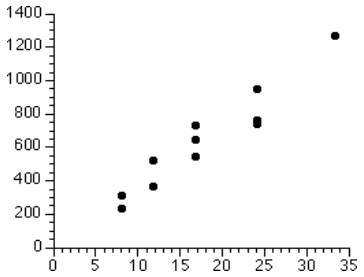
$$\frac{\sum_{i=1}^{n} \hat{y}_i}{n} = \frac{\sum_{i=1}^{n} \bar{y} + \hat{\beta}_1(x - \bar{x})}{n} = \frac{n\bar{y} + \hat{\beta}_1 \sum_{i=1}^{n}(\bar{x} - \bar{x})}{n} = \frac{n\bar{y} + 0}{n} = \bar{y}$$

Esercizio

L'urbanista Palmira Morrone sta investigando la relazione tra il flusso di traffico X (in termini migliaia di auto ogni 24 ore) ed il contenuto di piombo Y nella cortegga degli alberi che fiancheggiano una superstrada (peso a secco in µg/g)

i	1	2	3	4	5	6	7	8	9	10	11
x_i	8.3	8.3	12.1	12.1	17.0	17.0	17.0	24.3	24.3	24.3	33.6
y_i	227	312	362	521	640	539	728	945	738	759	1263

a) disegnare lo scatterplot; b) Stimare i parametri; c) Calcolare i valori teorici d) Verificare che le proprietà indicate nel precedente lucido.



$\hat{\beta}_1 = 36.18385$
$\hat{\beta}_0 = -12.84159$

	y_i	$\hat{y}_i$	$\hat{e}_i$
	227	287.48	-60.48
	312	287.48	24.52
	362	424.98	-62.98
	521	424.98	96.02
	640	602.28	37.72
	539	602.28	-63.28
	728	602.28	125.72
	945	866.43	78.57
	738	866.43	-128.43
	759	866.43	-107.43
	1263	1202.94	60.06
MEDIE	639.4545	639.4545	0.0000

Proprietà della retta di regressione/2

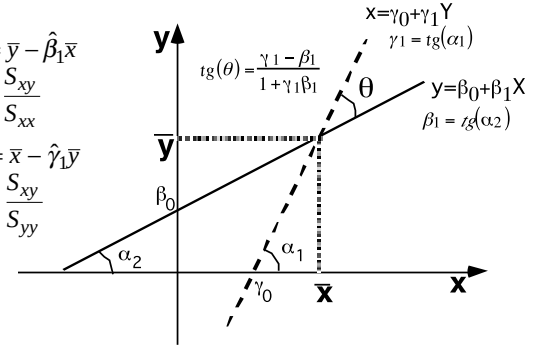
$$tg(\theta) = \frac{r^2 - 1}{r}$$

il ruolo di esogena ed endogena può essere scambiato:

$$y_i = \beta_0 + \beta_1 x_i + e_i \Rightarrow \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \quad dove \begin{cases} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 = \dfrac{S_{xy}}{S_{xx}} \end{cases}$$

$$x_i = \gamma_0 + \gamma_1 x_i + e_i' \Rightarrow \hat{x}_i = \hat{\gamma}_0 + \hat{\gamma}_1 x_i \quad dove \begin{cases} \hat{\gamma}_0 = \bar{x} - \hat{\gamma}_1 \bar{y} \\ \hat{\gamma}_1 = \dfrac{S_{xy}}{S_{yy}} \end{cases}$$

$\bar{x} = \bar{y}$



Le due rette interpolanti sono legate:

$$\hat{\gamma}_1 = \frac{S_{xy}}{S_{yy}} = \frac{\sqrt{S_{xx}}}{\sqrt{S_{yy}}} * \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} = \frac{\sqrt{S_{xx}}}{\sqrt{S_{yy}}} * r$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sqrt{S_{yy}}}{\sqrt{S_{xx}}} * \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} = \frac{\sqrt{S_{yy}}}{\sqrt{S_{xx}}} * r$$

$$\hat{\gamma}_1 * \hat{\beta}_1 = r^2$$

i coefficienti angolari hanno sempre lo stesso segno per cui le due rette non sono mai perpendicolari

Le due rette sono parallele (e coincidenti) se e solo se Y=X dato che tg(θ)=0

Se poi allora le due rette coincidono

Stima della varianza degli errori

Oltre ai due parametri della retta esiste un'altro parametro incognito: σ^2

Le assunzioni del modello indicano gli errori "e" come non osservabili, possiamo però stimarne i valori con gli errori osservati

$$\hat{e}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \quad \text{per} \quad i = 1,2,\ldots,n$$

Per ottenerne un valore approssimato si usa la seguente quantità

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

Che possiede alcune utili proprietà statistiche

Ai fini del calcolo è facile mostrare che

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\mu}_y)^2 - \hat{\beta}_1 * \sum_{i=1}^{n} (y_i - \hat{\mu}_y)(x_i - \hat{\mu}_x)$$

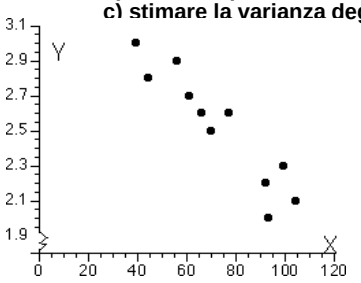
Per cui il calcolo di s_e^2 usa quantità già pronte

Esercizio

L'aziendalista Costantina Tenuta fa parte di una commissione chiamata a valutare una serie di progetti per l'idoneità al finanziamento. Per controllarne la congruità pone in relazione il numero X dei progetti per area e il tempo medio di completamento Y.

Settori di destinazione	Proget.	TMC
Edilizia demaniale	105	2.1
Altri	100	2.3
Opere straddali extraurb	94	2.0
Disinquinamento	93	2.2
Ferrovie	78	2.6
Edilizia sanitaria	71	2.5
Edilizia scolastica	67	2.6
Porti commerciali	62	2.7
Infrastrutture urbane	57	2.9
Energia	45	2.8
Smaltimento R.S.U.	40	3.0
Ferrovie metropol.	36	3.4
Archivi, Biblioteche	30	32.2
Ferrocie in concessione	12	3.3

a) Disegnare lo scatterplot;
b) Stimare i parametri;
c) stimare la varianza degli errori



$\sum y_i = 37.6; \quad \sum y_i^2 = 103.54; \quad \sum x_i = 890; \quad \sum x_i^2 = 67.182;$
$\sum x_i y_i = 2234.30$

$\hat{\beta}_1 = -0.0147109; \quad \hat{\beta}_0 = 3.6209072$

$$s_e^2 = \frac{103.54 - 3.6209072 * 37.6 - (-0.0147109) * 2234.30}{14-2} = \frac{0.2624532}{12} = 0.0219$$

Uso della retta di regressione

INTERPOLAZIONE

Lo scopo è trovare i valori della dipendente o di sostituirne i valori particolarmente anomali, per valori noti della indipendente.

ESTRAPOLAZIONE

Determinazione del valore della dipendente che corrisponde ad un valore della indipendente non necessariamente osservato.

CONTROLLO

Determinazione del valore della indipendente idoneo a determinare un fissato livello della dipendente

In ogni caso si ottengono dei VALORI TEORICI costituenti la stima dei VALORI VERI che rimangono comunque sconosciuti

Esempio

Unità di lavoro part-time e aumento di produzione

X	y	X²	Y²	XY
1	4	1	16	4
2	6	4	36	12
3	10	9	100	30
4	10	16	100	40
5	15	25	225	75
6	15	36	225	90
7	16	49	256	112
8	20	64	400	160
36	96	204	1358	523

$$\hat{\beta}_1 = \frac{8*523 - 36*96}{8*204 - 36^2} = 2.167;$$

$$\hat{\beta}_0 = \frac{204*96 - 36*523}{8*204 - 36^2} = 2.25;$$

$$\text{Semilavorati} = 2.25 + 2.167*\text{ULPT}$$

Ogni unità di lavoro part-time addizionale è responsabile di 2.167 tonn. di produzione.

Se il lavoro part-time non fosse impiegato la produzione media sarebbe a 2.25 tonn.

Supponiamo che si decida di impiegare più di 8 ULPT, diciamo 10, quale sarà l'incremento di produzione?

$$\hat{y}_{10} = 2.25 + 2.167*10 = 2.25 + 21.67 = 23.92$$

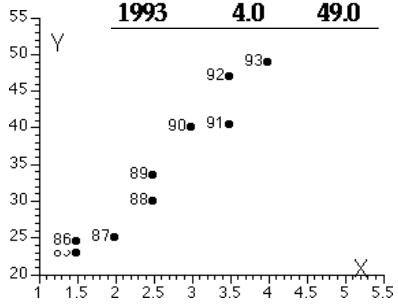
Se invece si volesse stabilire quante unità di lavoro part-time impiegare per ottenere 16 semilavorati allora

$$16 = 2.25 + 2.167*\hat{X} \Rightarrow \hat{X} = \frac{(16 - 2.25)}{2.167} = 6.345$$

Esercizio

La dott.ssa Sarina Bonofiglio, analista finanziario, sta studiando la relazione tra X= Tasso medio sui prestiti nel sistema interbancario e Y=importo della cedola semestrale di un titolo obbligazionario.

Anni	TMPSI	CeSe.
1985	1.5	23.0
1986	1.5	24.5
1987	2.0	25.0
1988	2.5	30.0
1989	2.5	33.5
1990	3.0	40.0
1991	3.5	40.5
1992	3.5	47.0
1993	4.0	49.0



a) Disegnare lo scatterplot

b) Stimare β_0, β_1 e σ^2

c) Supponendo che il dato del 93 sia non affidabile perché affetto dalla crisi nello SME calcolare il valore interpolato.

d) Quale sarà la cedola semestrale se nel '94 il TMPSI arriva a 5.5?

$$\hat{\beta}_0 = 6.448718;\ \hat{\beta}_1 = 10.602564;\ s_e^2 = 6.4803$$

$$\hat{y}_{93} = 6.448718 + 10.602564*4.0 = 48.89$$

$$\hat{y}_{94} = 6.448718 + 10.602564*5.5 = 64.76$$

L'effetto di regressione

I principi del ritorno alla media lo si ritrova in varie occasioni

Un docente che loda gli studenti per il buon risultato raggiunto in una prova vedrà un esito peggiore nella prova successiva.

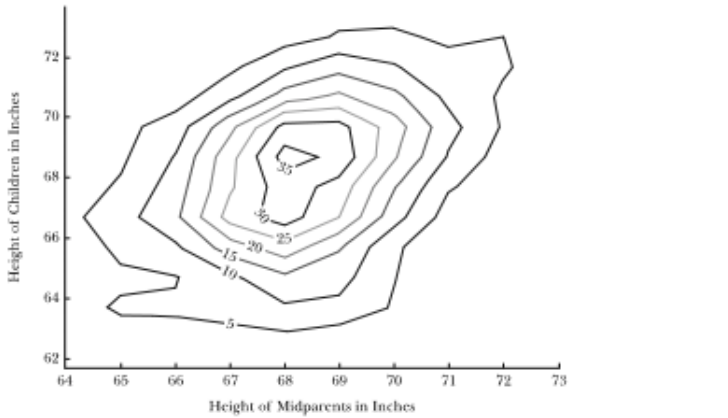
il docente che sgrida gli studenti per la pessima riuscita di un test otterrà risultati molto migliori nella seguente prova.

Un buon governo sarà seguito da una amministrazione inefficace e ad un premier inadeguato succederà un brillante primo ministro.

Nelle competizioni articolate su due fasi è frequente notare il ribaltamento degli esiti tra la prima e seconda prova: i migliori che peggiorano ed i peggiori che migliorano.

Alle origine del concetto di regressione

Figure 4
Galton's Original Smoother±Contour Plots: Contours after Galton Smoother

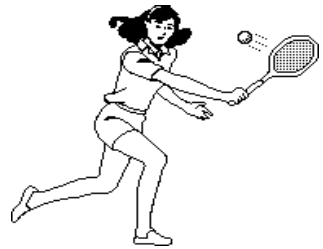


Sir Francis Galton notò che i figli di padri alti erano più alti della media, ma meno di quanto non eccedessero dalla media i loro padri. I figli di padri bassi erano in media bassi, ma meno bassi della media generale di quanto non lo fossero i padri

L'effetto di regressione/2

L'elemento comune è spiegabile così:

Per ottenere un buon risultato in un'impresa difficile concorrono due fattori:



- **Talento/Genio**
- **Sorte**

il successo in una prova ardua implica che entrambi i fattori hanno agito a favore.

Nella seconda prova il talento/genio magari migliorano o agiscono con la stessa intensità

La Sorte è capricciosa e imprevedibile e non si ripete.

Ed ecco l'effetto di regressione alla media in cui gli scarti si annullano tutti.

Misura dell'adattamento

I minimi quadrati ci garantiscono il miglior adattamento possibile, ma questo potrebbe non essere abbastanza.

Dobbiamo trovare misure standardizzate e normalizzate che siano in grado di quantificare il grado di scostamento tra valori stimati e valori osservati

SCARTO QUADRATICO MEDIO DEGLI ERRORI (errore standard della stima)

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

con i minimi quadrati

$$s_e^2 = \frac{S_{yy}}{n-2}\left(1-r^2\right)$$

E' nullo solo in caso di perfetta relazione lineare (r=1). Non varia però entro limiti predefiniti.

Possiamo solo dire che un adattamento è peggiore di un altro, ma non se un dato adattamento è buono o no

Risente anche delle unità di misura della dipendente.

Correlazione teoriche-osservate

Una possibilità di valutare l'adattamento potrebbe basarsi su:

$$\frac{Cov(y_i,\hat{y}_i)}{\sigma(y_i)\sigma(\hat{y}_i)} = \frac{\sum_{i=1}^{n}(y_i-\bar{y})\left(\bar{y}-\hat{\beta}_1(x_i-\bar{x})-\bar{y}\right)}{\sqrt{S_{yy}\sum_{i=1}^{n}\left(\bar{y}-\hat{\beta}_1(x_i-\bar{x})-\bar{y}\right)^2}} = \frac{\hat{\beta}_1 S_{xy}}{\sqrt{S_{yy}\hat{\beta}_1^2 S_{xx}}} =$$

$$= \frac{\hat{\beta}_1}{|\hat{\beta}_1|} r$$

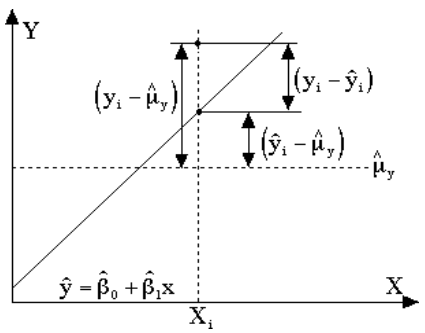
Ne consegue che l'adattamento è misurabile da:

$$\left|\frac{Cov(y_i,\hat{y}_i)}{\sigma(y_i)\sigma(\hat{y}_i)}\right| = \left|\frac{\hat{\beta}_1}{|\hat{\beta}_1|} r\right| = |r|$$

cioè dal valore assoluto del coefficiente di correlazione tra osservate e stimate che coincide con il valore assoluto del coefficiente di correlazione "r" tra X ed Y.

Coefficiente di determinazione (R^2)

La variabilità della "Y" può essere scomposta in due parti distinte. Infatti, l'identità



$$\sum_{i=1}^{n}(y_i - \bar{y}) = \sum_{i=1}^{n}\left[(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})\right]$$

rimane anche quando si considerano i quadrati (se la retta è quella dei minimi quadrati)

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

Dividendo per "n" si ha la seguente relazione:

Varianza totale=Varianza errori+Varianza stime

La varianza delle stime è la parte di variabilità (attitudine a presentare modalità diverse) che il nostro modello riesce a spiegare, quella degli errori è la parte che rimane ignota.

Varianza totale=Varianza NON spiegata+Varianza spiegata

Formula dell' R^2

Dalla formula della scomposizione abbiamo: $\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$

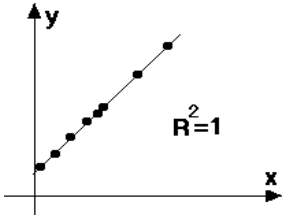
Dividendo i membri per la devianza totale si ha $= \dfrac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} + \dfrac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$

il 1° addendo è il rapporto tra varianza non spiegata e varianza totale, il 2° è il rapporto tra varianza spiegata e varianza totale.

Questo rapporto è usato come indice della bontà di adattamento ed è noto come il COEFFICIENTE DI DETERMINAZIONE

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n}\hat{e}_i^2}{\sum_{i=1}^{n}y_i^2 - n\bar{y}^2}$$

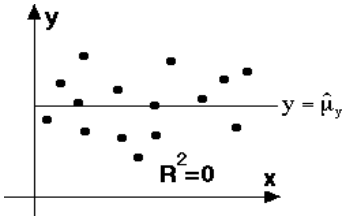
Casi estremi



$\hat{y}_i = \bar{y} + \hat{\beta}_1(x_i - \bar{x})$

Se tutte le osservate sono allineate su di una retta, teoriche ed osservate coincidono e quindi

$$Se\ \ y_i = \hat{y}_i\ \ per\ \ ogni\ \ "i"\ \ \Rightarrow R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1;$$



Se la retta di regressione è piatta (coefficiente angolare nullo) allora le teoriche sono tutte pari alla media e quindi

$$Se\ \ \hat{y}_i = \bar{y}\ \ per\ \ ogni\ \ "i"\ \ \Rightarrow R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - 1 = 0$$

Esempio

Studio della relazione tra il massimo del battito cardiaco sotto stress ed età

X	Y	$(x-M_x)$	$(y-M_y)$	$(x-M_x)(y-M_y)$	$(x-M_x)^2$	$(y-M_y)^2$	y_i	$(y-y_i)^2$
10	210	-25	25	-625	625	625	212.2058	4.8656
20	200	-15	15	-225	225	225	201.3234	1.7514
25	195	-10	10	-100	100	100	195.8822	0.7783
35	190	0	5	0	0	25	184.9998	25.0020
40	185	5	0	0	25	0	179.5586	29.6088
45	175	10	-10	-100	100	100	174.1174	0.7790
50	165	15	-20	-300	225	400	168.6762	13.5144
55	160	20	-25	-500	400	625	163.2350	10.4652
35	185			-1850	1700	2100		86.7647

$$s_e = \sqrt{\frac{86.7647}{6}} = 3.8027$$

$$r(y,\hat{y}) = \left|\frac{-1850}{\sqrt{1700 * 2100}}\right| = 0.9791$$

$\beta_0 = 223.0812;\quad \beta_1 = 1.0882;$

$$R^2 = 1 - \frac{86.7647}{2100} = 0.9572$$

$$R^2 = (-0.9791)^2 = 0.9587$$

Test sul coefficiente angolare

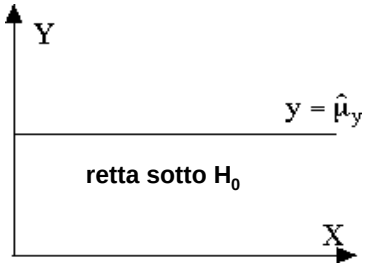
La prima è più importante verifica riguarda l'esistenza o meno di una relazione tra la "Y" e la "X"

ESISTE O NON ESISTE UNA RELAZIONE TRA Y ed X?

Questo si traduce nella verifica dell'ipotesi $\begin{cases} H_0: \beta_1 = 0 \\ H_1: \beta_1 \neq 0 \end{cases}$

infatti, β_1 **esprime la variazione media nella endogena a partire da una variazione unitaria nella esogena**

Se H_0 non potesse essere rifiutata la retta di regressione si presenterebbe come parallela all'asse delle X e non in grado di spiegare la "Y"



Continua test su β1

La statistica test necessaria per la verifica si ottiene come per i test sulla media. In particolare si adopera la t-Student

$$t_{n-2}(\beta_1) = \frac{\hat{\beta}_1}{\frac{s_e}{\sqrt{S_{xx}}}} \quad dove \quad \begin{cases} s_e = \sqrt{\dfrac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}} \\ S_{xx} = \sum_{i=1}^{n}(x_i - \mu_x)^2 \end{cases}$$

Se il campione è abbastanza grande si cambierà la t-Student con la Normale

Esempio

Si ritiene che il consumo di energia elettrica sia determinato da un modello lineare nella temperatura media del giorno

giorno	T.M.	C.E.L.
1	95	214
2	82	152
3	90	156
4	81	129
5	99	254
6	100	266
7	93	210
8	95	204
9	93	213
10	87	150

$$SS_{xy} = \sum_{i=1}^{n}(x_i - \hat{\mu}_x)(y_i - \hat{\mu}_y) = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n} = 180709 - \frac{915 * 1948}{10} = 2556$$

$$SS_{xx} = \sum_{i=1}^{n}(x_i - \hat{\mu}_x)^2 = \sum_{i=1}^{n} x_i - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} = 84103 - \frac{(915)^2}{10} = 380.5$$

$$\hat{\mu}_y = 194.8; \quad \hat{\mu}_x = 91.5; \quad \hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{2556}{380.5} = 6.7175; \qquad R^2 = 0.99$$

$$\hat{\beta}_0 = \hat{\mu}_y - \hat{\beta}_1 \hat{\mu}_x = 194.8 - (6.7175) * 91.5 = -419.85$$

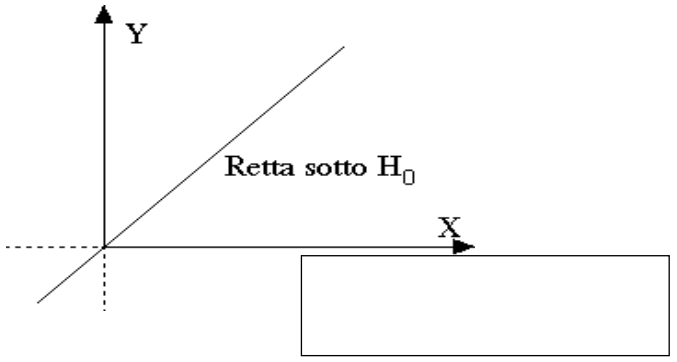
$$s_e^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} = \frac{2093.43}{8} = 261.68 \Rightarrow s_e = \sqrt{261.68} = 16.18 \qquad t_c = \frac{6.7175}{16.18/\sqrt{380.5}} = 8.10$$

Il p-value del test è TDIST(8.1;8;2)= `0.000040`

E' evidente l'elevata significatività del parametro

Test sull'intercetta

La verifica dell'intercetta è poco interessante dato che non ha incidenza sulla bontà di adattamento. In genere si sottopone a verifica l'ipotesi che sia uguale a zero



Statistica test

$$t_c = \frac{\hat{\beta}_0}{s_e \sqrt{\left(\dfrac{\sum_{i=1}^{n} x_i^2}{S_{xx}}\right)}}$$

Nel caso di temperature e consumo di energia si ha: $t_c = \dfrac{-419.85}{16.18\sqrt{\dfrac{84103}{380.5}}} = -1.745$

L'intercetta non risulta significativa: p-Value=DISTRIB.T(1.745;8;2)=11.9%

Esercizio

Reddito	Superficie
22	106
26	117
45	128
37	132
28	80
50	95
56	168
34	65
60	150
40	120
45	110
36	45
80	205
120	230

SUMMARY OUTPUT

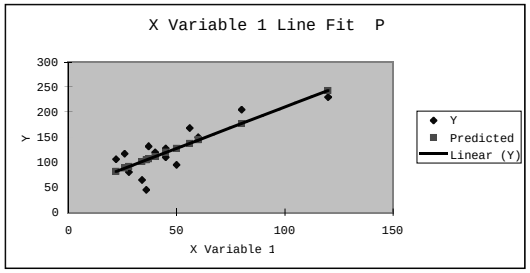
Regression Statistics	
Multiple R	0.832286982
R Square	0.692701621
Adjusted R S	0.667093423
Standard Err	29.31437326
Observations	14

	Beta	SE	t Stat	P-value
Intercept	44.9592	17.2813	2.6016	0.0232
X Variable 1	1.6518	0.3176	5.2010	0.0002

Si supponga che la proprietaria di un'agenzia immobiliare voglia stabilire la relazione tra reddito familiare e superficie dell'appartamento.

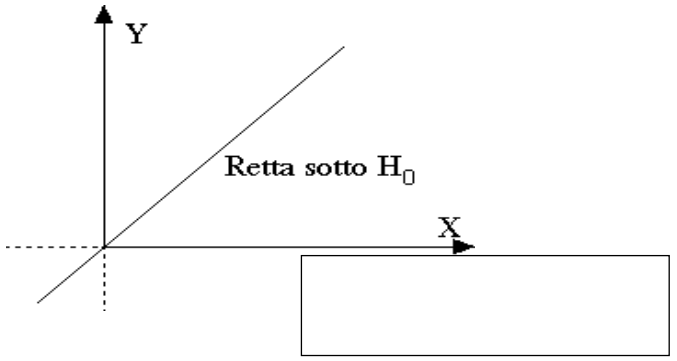
RESIDUAL OUTPUT

Observation	Predicted Y	Residuals
1	81.2988	24.7012
2	87.9060	29.0940
3	119.2901	8.7099
4	106.0757	25.9243
5	91.2096	-11.2096
6	127.5491	-32.5491
7	137.4599	30.5401
8	101.1203	-36.1203
9	144.0671	5.9329
10	111.0311	8.9689
11	119.2901	-9.2901
12	104.4239	-59.4239
13	177.1031	27.8969
14	243.1750438	-13.17504381



Test sull'intercetta

La verifica dell'intercetta è poco interessante dato che non ha incidenza sulla bontà di adattamento. In genere si sottopone a verifica l'ipotesi che sia uguale a zero



Statistica test

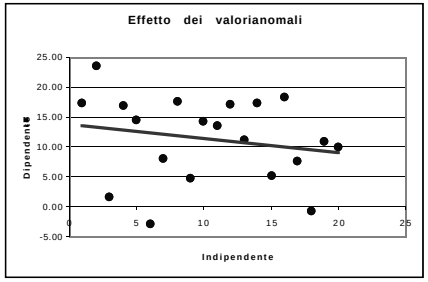
$$t_c = \frac{\hat{\beta}_0}{s_e \sqrt{\left(\frac{\sum_{i=1}^{n} x_i^2}{S_{xx}} \right)}}$$

Nel caso di temperature e consumo di energia si ha: $t_c = \frac{-419.85}{16.18\sqrt{\frac{84103}{380.5}}} = -1.745$

L'intercetta non risulta significativa: p-Value=DISTRIB.T(1.745;8;2)=11.9%

Effetto dei valori anomali

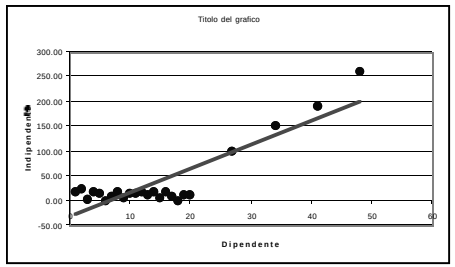
X	Y
1	17.27
2	23.49
3	1.72
4	16.87
5	14.58
6	-2.76
7	8.04
8	17.65
9	4.75
10	14.39
11	13.61
12	17.22
13	11.22
14	17.39
15	5.31
16	18.39
17	7.64
18	-0.73
19	10.86
20	9.89
27	**100.00**
34	**150.00**
41	**190.00**
48	**260.00**



L'impatto della "X" può essere fuorviato dalla presenza di alcuni valori remoti.

R multiplo	0.2064
R al quadrato	0.0426
R al quadrato	-0.0106
Errore standa	7.0803
Osservazioni	20

	Beta	*Stat t*	*p-value*
Intercetta	13.9197	4.2322	0.0005
Variabile X 1	-0.2457	-0.8950	0.3826



Il metodo dei minimi quadrati trascura il blocco di 20 dati tra i quali non c'è relazione significativa (o è negativa). Invece, pone attenzione ai quattro punti tra i quali c'è una relazione positiva

R multiplo	0.8795
R al quadrato	0.7735
R al quadrato	0.7632
Errore standa	32.7075
Osservazioni	24

	beta	*Stat t*	*p-value*
Intercetta	-34.9737	-3.2383	0.0038
Variabile X 1	4.9060	8.6687	0.0000