Due fattori continui

Analisi della regressione

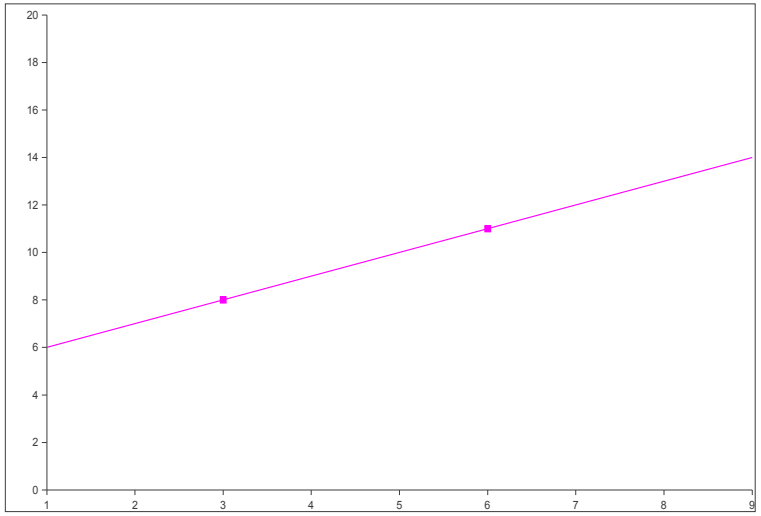
Voglio studiare due fattori dipendenti uno dall'altro

L'esempio classico sono le rese di macellazione:
il peso di un organo aumenta infatti all'aumentare del peso dell'animale (quale relazione li lega?), voglio rappresentare tale andamento con legge biologica (= una funzione matematica) che leghi i due parametri:

- fra le varie funzioni l'equazione della retta (costanza di rapporto) è la più semplice
- fra le varie rappresentazioni geometriche (curve) la retta è la più semplice.

L'analisi della regressione lineare equivale a trovare la relazione matematica (costante) che lega le due variabili

Dalla matematica: per due punti passa una ed una sola retta.



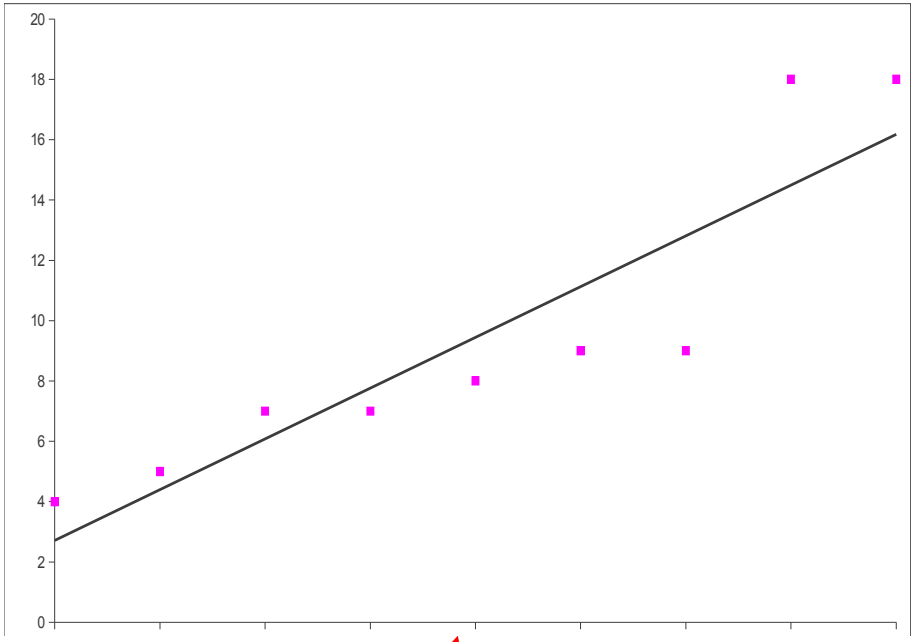
$$y = a + bx$$

equazione generica della retta

Dove **b** = pendenza della retta (cioè la variazione di y corrispondente alla variazione unitaria di x (= angolo con le ascisse oppure tgα)

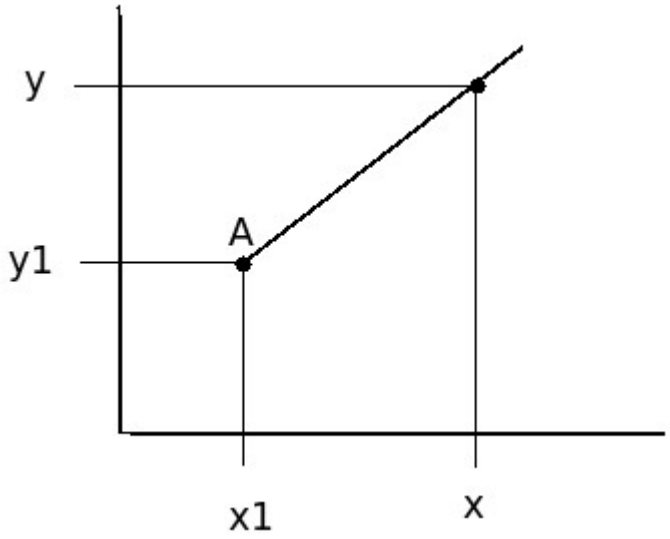
a = intercetta cioè il valore della retta quando x=0 (non ha significato solo per rette parallele a Y)

Dalla statistica: per più punti è possibile trovare una ed una sola retta la cui somma dei quadrati degli scarti delle distanza dei punti dalla retta è la minima possibile.



La retta più vicina ai punti

come calcolare questa retta?



Passando da A (x1, y1) a un punto posto sulla stessa retta (una qualsiasi delle rette che partono da A) è necessario che l'aumento sia proporzionale (altrimenti non è una retta) ovvero che la differenza fra le ordinate sia pari alla differenza fra le ascisse per un valore che deve essere sempre lo stesso

Ricordando dall'analisi matematica:
Indicando le coordinate dei due punti con:
A = (x1, y1) B = (x2, y2)
per trovare l'equazione della retta passante per questi due punti devo fare i seguenti passaggi:
I. calcolo il fascio di rette che passa per il punto A = (x1, y1) pari a $y - y1 = m(x - x1)$
II. tra tutte le rette scelgo quella che passa per il punto B = (x2, y2)
cioè quella che ha come coefficiente angolare $m = \frac{y2 - y1}{x2 - x1}$

III. Sostituisco "m" nella precedente e ottengo l'equazione

$$y - y1 = \frac{y2 - y1}{x2 - x1}(x - x1)$$

IV.Trasporto i termini con y prima dell'uguale e ottengo:

$$\frac{y - y1}{y2 - y1} = \frac{x - x1}{x2 - x1}$$

$$y - y1 = \frac{x\,y2 - x\,y1 - x1y2 + x1y1}{x2 - x1} \qquad y = \frac{x\,y2 - x\,y1 - x1y2 + x1y1}{x2 - x1} + y1$$

Esempio numerico:
A = (x1; y1) (-4, 3) B = (x2; y2) (2, -1)
Sostituendo nella formula

$$\frac{y - 3}{(-1) - 3} = \frac{x - (-4)}{2 - (-4)}$$

$$\frac{y - 3}{-1 - 3} = \frac{x + 4}{2 + 4}$$

$$\frac{y - 3}{-4} = \frac{x + 4}{6}$$

Moltiplicando in croce $6(y - 3) = -4(x + 4)$; $6y - 18 = -4x - 16$; $6y = -4x - 16 + 18$; $6y = -4x + 2$;
Equazione della retta: $y = -4/6\,x + 2/6$

indico con:

SSx^2= somma dei quadrati degli scarti delle x dalla media X
SSy^2 = somma dei quadrati degli scarti delle y dalla media Y
SSxy = somma dei prodotti degli scarti delle y dalla media Y e delle x dalla media X

$$Y = \frac{xy2-xy1-x1y2+x1y1}{x2-x1} + y1$$

Inclinazione della retta

o

Coefficiente angolare $= b = \dfrac{\frac{SSxy}{n-1}}{\frac{SSx^2}{n-1}} = \dfrac{SSxy}{SSx^2}$

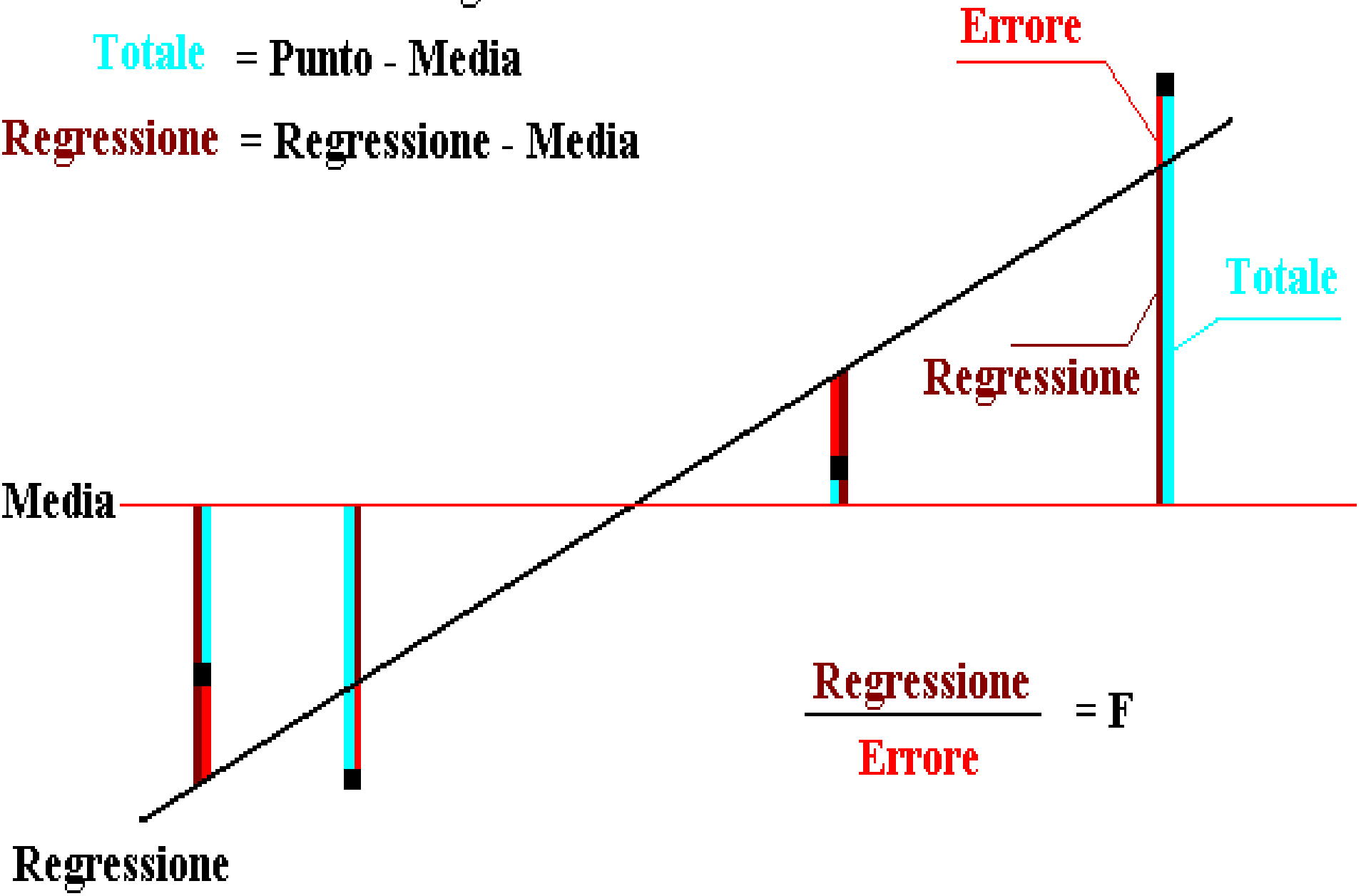
b = SSxy/SSx^2
a = mediaY - b*mediaX
Y = a + bx

(SSxy)^2/Sx^2 =	deviazione dovuta alla regressione
SSy^2 - (SSxy)^2/SSx^2 =	deviazione dalla regressione
SSy^2 =	deviazione TOTALE

Errore = Punto - Regressione

Totale = Punto - Media

Regressione = Regressione - Media



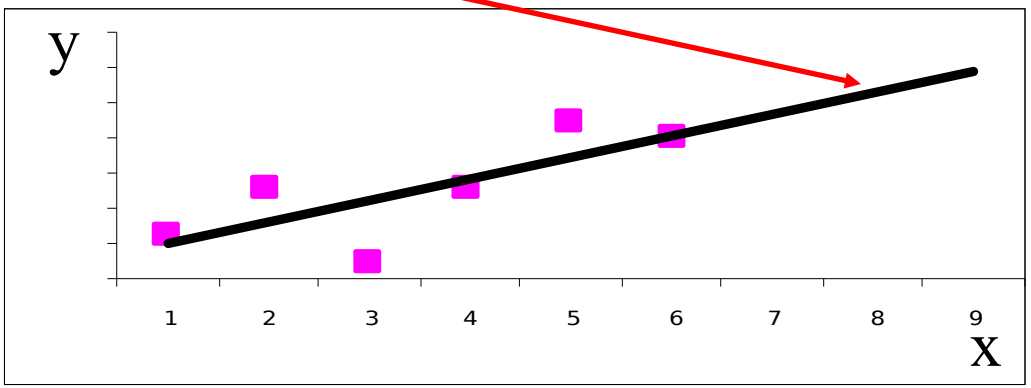
$$\frac{\text{Regressione}}{\text{Errore}} = F$$

Esempio numerico

	x	y
1	11	12
2	26	26
3	3	5
4	18	26
5	48	44
6	34	40

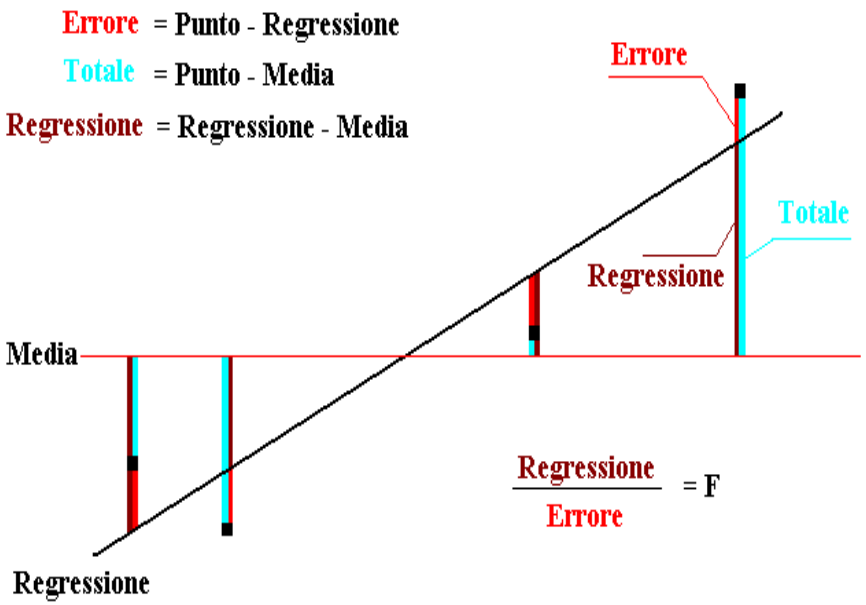
Procedo calcolando le somme dei quadrati degli scarti dalle mediae OVVIAMENTE **con la tecnica del TC (termine di correzione).**

	x	quadrati	prodotti	y	quadrati				
	11	121	132	12	144				
	26	676	676	26	676	SSx^2 =	sx^2-(sx)^2/n		SSx^2= somma dei quadrati degli scarti delle x dalla media X
	3	9	15	5	25				
	18	324	468	26	676	SSy^2 =	sy^2-(sy)^2/n		SSy^2 = somma dei quadrati degli scarti delle y dalla media Y
	48	2304	2112	44	1936				
	34	1156	1360	40	1600	SSxy =	sxy - (sx)*(sy)/n		SSxy = somma dei prodotti degli scarti delle y dalla media Y e delle x dalla media X
					0				
					0				
					0	b =	SSxy/SSx^2		TCx=(Sx)²/n(coppie)
					0	a =	$\overline{Y} - b\,\overline{X}$		TCxy=(Sx)*(Sy)/n(coppie)
					0				TCx=(Sy)²/n(coppie)
sum	140	4590	4763	153	5057		(SSxy)^2/Sx^2 =		deviazione dovuta alla regressione
n	6			6			SSy^2 - '(SSxy)^2/SSx^2 =		deviazione dalla regressione
media	23,33333			25,5			SSy^2 =		deviazione TOTALE
	SS²	-TC				(SSxy)^2/SSx^2 =	1075,5030227		deviazione dovuta alla regressione
SSx^2 =	4590	3266,67	=	1323,333			79,99697733		deviazione dalla regressione
SSxy =	4763	3570	=	1193,00		SSy^2 =	1155,5		deviazione TOTALE
SSy^2 =	5057	3901,5	=	1155,50	**y =**	**4,4647**	**se x =**	**0**	
b =	0,901511				**y =**	**5,3662**	**se x =**	**1**	
a =	4,464736				**x =**	**-4,9525**	**se y =**	**0**	
	Y =	4,46474	+0,901511	x	**y =**	**8,9723**	**se x =**	**5**	



T.C.

Quanta variabilità è spiegata dalla retta e quanto "vale" l'errore?



Totale =1.155,5

Regressione=$1.193^2/1.323$=1.075,5

Errore= 1.155,5-1.075,5=80

b = 1.193/1.323=0,902

a = 25,5 – 23,3 * 0,90 = 4,5

Totale =somma quadrati scarti di y = **SSy^2**

Regressione=somma dei prodotti degli scarti sulla somma dei quadrati degli scarti della X = **(SSxy)2/sx^2**

Errore= Totale meno **Regressione**

Angolo della retta = b = SSxy/SSx^2

Costante= a = Valore di Y per X=0 cioè a=media di y meno media di x per b cioè a=y-bx (dalla retta generica y=a+bx)

Sorgenti di variazione	Somme dei quadrati degli scarti SS	gradi di libertà gl o df	Varianze MS	Rapporti F
deviazione dovuta alla regressione $(SSxy)^2/SSx^2$	1075,50302267	1	1075,50302267	53,777183017
Deviazione dalla regressione $SSy^2-(SSxy)^2/SSx^2$	79,99697733	4	19,9992443325	
SS TOTALE di Y cioè SSy^2	1155,5	5	231,1	93,08%

**

Può essere utile, in quanto dà una dimensione diretta: esprime la variazione spiegata dalla curva (in questo caso una retta) rispetto alla variazione totale dei dati espressa in percentuale.

1.155,50-
80,00=
1.075,50

Tavola F (0,05)

g.l.	1	2	3	4	5	6	7	8	9	10	12	14	16	18	20	25	30	40
1	161,4	199,5	215,7	224,6	230,2	234	236,8	238,9	240,5	241,9	243,9	245,4	246,5	247,3	248	249,3	250,1	251,1
2	18,51	19	19,16	19,25	19,3	19,33	19,35	19,37	19,39	19,4	19,41	19,42	19,43	19,44	19,45	19,46	19,46	19,47
3	10,13	9,552	9,277	9,117	9,013	8,941	8,887	8,845	8,812	8,785	8,745	8,715	8,692	8,675	8,66	8,634	8,617	8,594
4	7,709	6,944	6,591	6,388	6,256	6,163	6,094	6,041	5,999	5,964	5,912	5,873	5,844	5,821	5,803	5,769	5,746	5,717
5	6,608	5,786	5,409	5,192	5,05	4,95	4,876	4,818	4,772	4,735	4,678	4,636	4,604	4,579	4,558	4,521	4,496	4,464
6	5,987	5,143	4,757	4,534	4,387	4,284	4,207	4,147	4,099	4,06	4	3,956	3,922	3,896	3,874	3,835	3,808	3,774

Tavola F (0,01)

g.l.	1	2	3	4	5	6	7	8	9	10	12	14	16	18	20	25	30
1	4052	4999	5404	5624	5764	5859	5928	5981	6022	6056	6107	6143	6170	6191	6209	6240	6260
2	98,5	99	99,16	99,25	99,3	99,33	99,36	99,38	99,39	99,4	99,42	99,43	99,44	99,44	99,45	99,46	99,47
3	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,34	27,23	27,05	26,92	26,83	26,75	26,69	26,58	26,5
4	21,2	18	16,69	15,98	15,52	15,21	14,98	14,8	14,66	14,55	14,37	14,25	14,15	14,08	14,02	13,91	13,84
5	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05	9,888	9,77	9,68	9,609	9,553	9,449	9,379
6	13,75	10,92	9,78	9,148	8,746	8,466	8,26	8,102	7,976	7,874	7,718	7,605	7,519	7,451	7,396	7,296	7,229
7	12,25	9,547	8,451	7,847	7,46	7,191	6,993	6,84	6,719	6,62	6,469	6,359	6,275	6,209	6,155	6,058	5,992

Il rapporto fra la somma dei quadrati degli scarti dovuti alla regressione (SS=MS) e la somma dei quadrati degli scarti totali (SStot=SSregressione+SSerrore) è chiamato coefficiente di determinazione e viene indicato come **R^2**.

93,08% = R2

La radice quadrata di **R^2** (con l'aggiunta del segno che "porta" b) è chiamato coefficiente di correlazione e viene indicato come **r**.

0,96 = r

$$r = \sqrt{\frac{(SSxy)^2}{SSx^2}} \qquad r = \sqrt{\frac{(SSxy)^2}{SSx^2 * SSy^2}} \qquad r = \frac{(SSxy)}{\sqrt{SSx^2 * SSy^2}}$$

Viene calcolato più correttamente anche un coefficiente di determinazione, detto aggiustato in quanto **calcolato per differenza dall'unità**. In questo caso invece di utilizzare il rapporto fra diretto fra le SS si usa il complemento ad uno del rapporto fra le MS.

L'**R^2** aggiustato è pari a 1 meno il rapporto fra variazione residua (cioè dell'errore = MS dalla regressione) e la variazione totale (SS-tot/g.s=MS-tot) e viene indicato come **R^2_{adj}**. (È indispensabile per comparare modelli con numero diverso di regressori),

$$1 - \frac{19{,}999244332}{231{,}1} \quad 1 - \quad 0{,}0865 \quad = \quad 0{,}91$$

Il risultato finale si deve esprimere quindi:

$Y = 4{,}46 + 0{,}902x$; $R^2 = 93{,}08\%$ **

oppure..... meglio!

Sapendo che:

			Varianza	d.s.=e.s.
d.s. di b =	MS errore	19,999244332	0,015112779	0,122934044
	Sx^2	1323,3333333		

$Y = 4{,}46 + 0{,}902(e.s.0.1229)x$; $R^2 = 93{,}08\%$ **

NOTA: b è "un solo numero" (g.l. di b = 1) quindi poiché l'e.s. è pari alla d.s. diviso la radice quadrata di 1, le due quantità sono uguali

Purtroppo le leggi biologiche molto spesso vengono “descritte” da equazioni più complesse di quella lineare (quadratiche cubiche esponenziali ecc) ed inoltre più variabili possono intervenire contemporaneamente (regressioni multiple).

Tuttavia porzioni limitate di curve (equazioni) anche complesse che descrivono un fenomeno possono sempre essere approssimate con porzioni lineari (magari trasformando preventivamente i dati (es trasformazione logaritmica).

Lo studio delle leggi biologiche non lineari è ovviamente più complesso e non può essere trattato in un corso per futuri laureati in scienze applicate.