



Statistica bivariata, correlazione, aggiustamento coi minimi quadrati, regressione

per Farmacia Ospedaliera

AA2008/09



regressione

- supponiamo di avere due variabili (di cui una almeno aleatoria) che campioniamo insieme, la regressione riguarda la relazione fra le due variabili – per es. è una tecnica per descrivere una relazione fra le variabili
- la regressione lineare nei coefficienti può essere usata per trovare i coefficienti della relazione fra le due variabili – caso semplice, in particolare consideriamo l'es. di due coefficienti (una retta)
- un problema analogo è dato dall'adattamento di una legge o curva teorica ai dati (o dal confronto di dati con distribuzioni a priori): si tratta di trovare i parametri della curva che meglio si adattano ai dati [minimi quadrati]
- si ottengono le stesse formule nei due casi

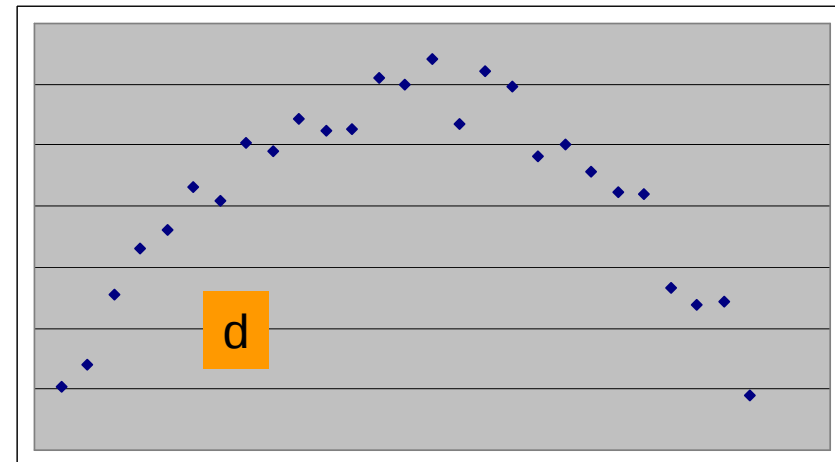
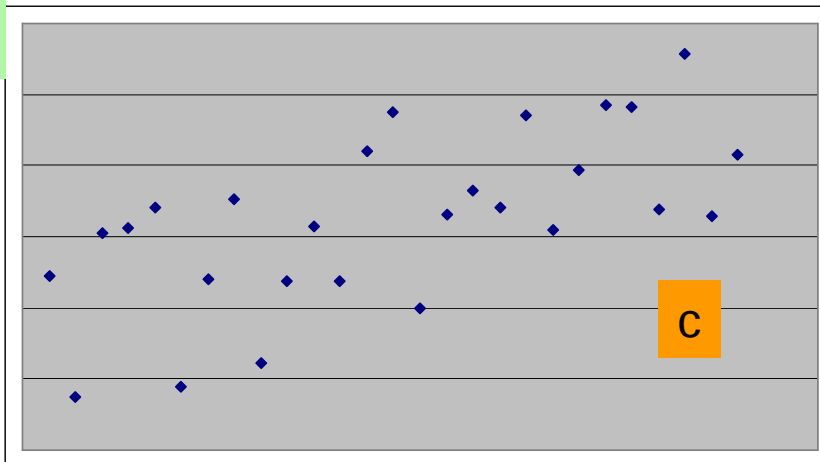
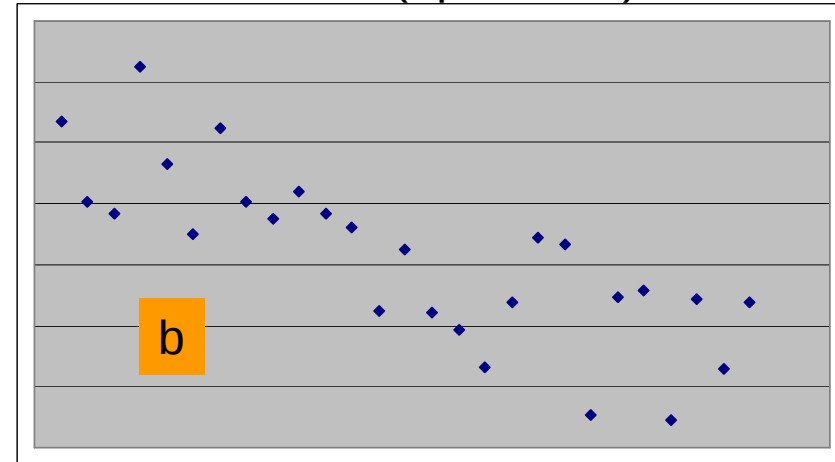
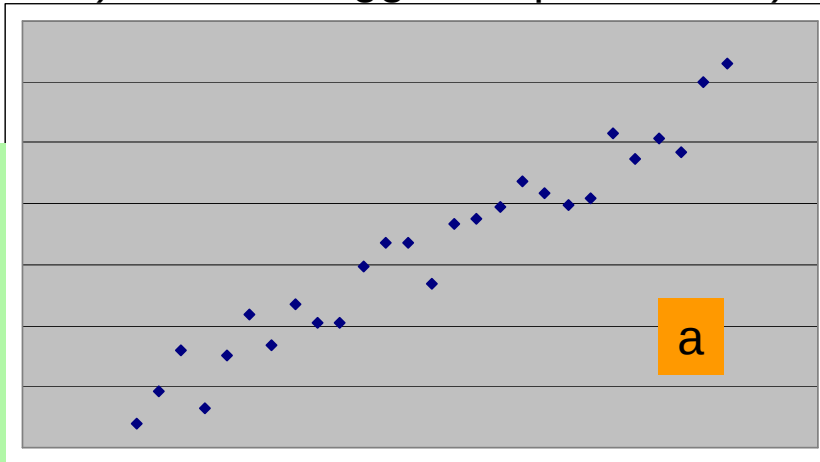


esempi di relazione fra 2 variabili aleatorie

- a) relazione ~lineare positiva; b) idem negativa, maggiore dispersione;
c) ancora maggior dispersione; d) relazione non lineare (~parabola)

↑
v
a
r
i
a
b
i
l
e

A





minimi quadrati

- consideriamo ad es. la sperimentazione di un farmaco, le due variabili da mettere in relazione possono essere il tempo medio di guarigione, t_i (aleatoria con dev. stand. s_i), ed il dosaggio di un farmaco, x_i (deterministica); $i=1,2,\dots,k$ dove k è il n.o di campioni di pazienti
- supponiamo di ipotizzare una relazione funzionale $t_i \sim f(x_i; p_j)$ dove p_j con $j=1,2,\dots,n$ sono dei parametri a priori incogniti
- per trovare i parametri costruisco una 'distanza' fra variabile aleatoria (dati sperimentali) e modello

$$d^2 = \sum_{i=1}^k (t_i - f(x_i; p_j))^2 / s_i^2$$

- la migliore scelta dei parametri sarà ottenuta quando la 'distanza' fra dati e modello è minima, matematicamente $\partial d^2 / \partial p_j = 0$ dove ∂ indica la derivata rispetto ad un p_j tenendo gli altri fissi (derivata parziale) $\Rightarrow$ n eq. nelle n incognite $\Rightarrow$ semplici da risolvere nel caso di dipendenza lineare dai parametri $\Rightarrow$ ho i p_j



minimi quadrati – relazione lineare

- semplifico assumendo $s_i = s$, costante, per tutti i $\underline{t}_i$ (stesso peso) e semplifico le notazioni ponendo $\underline{T}_i = \underline{t}_i$ e $X_i = \underline{x}_i$ (in gen. anche x_i potrebbe essere aleatoria)
- come modello (hp), in gen. $T_i = T(X_i)$, prendiamo ad es. una relazione lineare

$$T = mX + c$$

$$\begin{aligned} d^2 &= \sum_{i=1}^k [T_i - (mX_i + c)]^2 / s^2 = \\ &= \sum_{i=1}^k [T_i^2 - 2T_i(mX_i + c) + m^2X_i^2 + 2mcX_i + c^2] / s^2 \end{aligned}$$

dove m, c sono i coefficienti da determinare dal sistema formato da $\partial d^2 / \partial m = 0$ e $\partial d^2 / \partial c = 0$

- pongo $S_X = \sum_{i=1}^k X_i$; $S_T = \sum_{i=1}^k T_i$; $S_{TX} = \sum_{i=1}^k T_i X_i$;
 $S_{XX} = \sum_{i=1}^k X_i^2$ (e salto tutti i passaggi)



minimi quadrati - formule

- si ottiene per la pendenza e l'intercetta della retta interpolante

$$m = (kS_{XT} - S_X S_T) / (kS_{XX} - S_X S_X)$$

$$c = (S_{XX} S_T - S_{XT} S_X) / (kS_{XX} - S_X S_X)$$

NB usualmente
la 2^a variabile è
indicata con y

=> m, c ottimali (minima 'distanza')

- e se la relazione non è una retta?
 - relazione lineare nei parametri
 $y = a_0 + a_1 x + a_2 x^2 + \dots + a_n x^n \Rightarrow$ **n eq. in n incognite**
 - **relazioni linearizzabili prendendo il logaritmo**

- es.1 $z = z_0 e^{-mx} \Rightarrow \ln(z) = \ln(z_0) - mx$

pongo $y = \ln(z)$; $c = \ln(z_0)$ $\Rightarrow$ **$y = c - mx$**

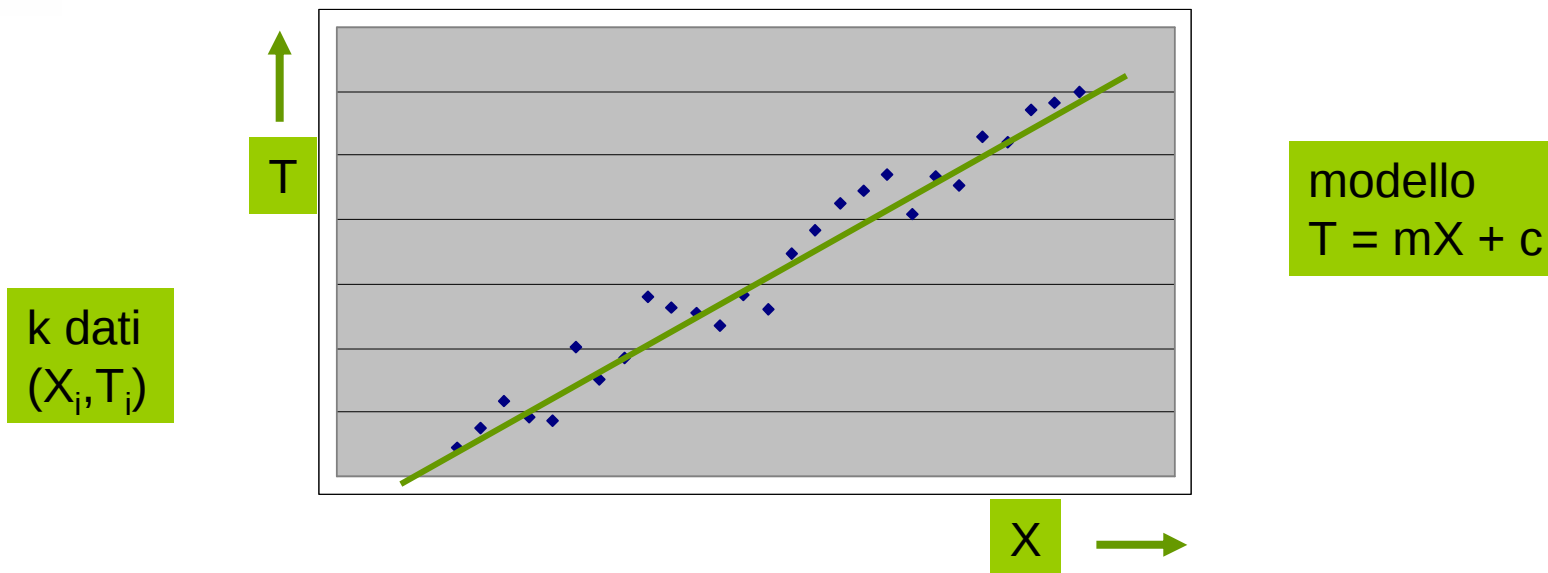
- es.2 $z = z_0 u^a \Rightarrow \ln(z) = \ln(z_0) + a \cdot \ln(u)$

pongo $y = \ln(z)$; $c = \ln(z_0)$; $x = \ln(u)$ $\Rightarrow$ **$y = c + ax$**

in entrambi i casi ottengo l'eq. di una retta per cui posso ottenere pendenza e intercetta ottimali (ricordandomi poi delle posizioni fatte!)



regressione – coefficiente di correlazione



coefficiente di correlazione

$$r = [\sum (X_i - \underline{X})(T_i - \underline{T})/k] / [\sum (X_i - \underline{X})^2/k \cdot \sum (T_i - \underline{T})^2/k]^{1/2} \quad (\text{si indica anche con } \rho \text{ [pop.]})$$

dove le $\sum$ vanno da 1 a k; il numeratore è la **covarianza** ed il denominatore contiene sotto radice il prodotto delle varianze;
inoltre $\underline{X} = (\sum X_i)/k$; $\underline{T} = (\sum T_i)/k$ sono i valori medi di X e T

$$-1 \leq r \leq +1$$



regressione – coefficienti della retta

- assumendo il modello $T = mX + c$ la regressione al valor medio dà

$$m = [\sum (X_i - \underline{X}) \cdot (T_i - \underline{T}) / k] / [\sum (X_i - \underline{X})^2 / k]$$
$$= [\sum X_i T_i - k \underline{X} \underline{T}] / [\sum X_i^2 - k \underline{X}^2]$$

NB usualmente
la 2^a variabile è
indicata con y

cioè la covarianza divisa per la varianza di X, mentre c può essere ottenuto per sostituzione

$$c = \underline{T} - m \underline{X}$$

- m, c sono uguali a quelli ottenuti dai minimi quadrati con pesi uguali per la retta interpolante i dati (si può dimostrare); forniscono informazioni sul comportamento **in media** di una variabile al variare dell'altra => consentono di interpolare ed estrapolare (ad es. al futuro sulla base dell'andamento passato)



un esempio – n. divorzi vs anno

n. divorzi in Inghilterra e nel Galles

A(nno)	1975	1976	1977	1978	1979	1980
D(ivorzi) x1000	120.5	126.7	129.1	143.7	138.7	148.3

anche in questo caso solo una variabile è aleatoria (D) mentre l'altra (A) è deterministica, un modello ragionevole è una dipendenza lineare di D da A (retta di regressione)

$$D = mA + c \quad \text{con } c, m \text{ coefficienti da determinare}$$

(l'eq. non può essere soddisfatta esattamente per tutte le coppie di dati visto che D è aleatoria)

conviene, per evitare grandi numeri, definire nuove variabili

$$T = D - 120$$

$$X = A - 1975$$

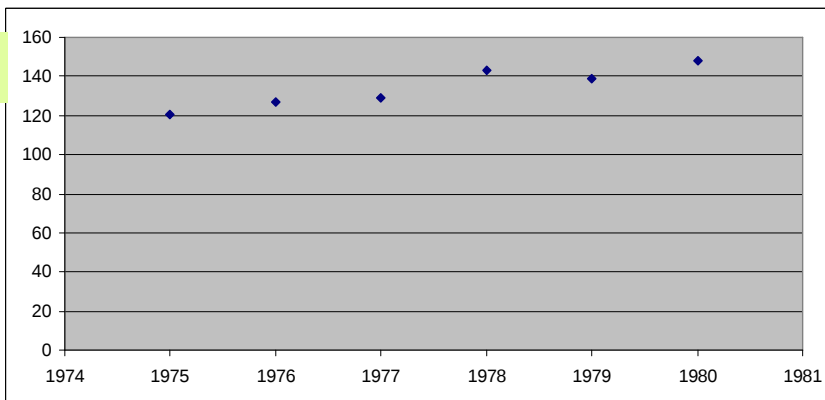
$$T = mX + c'$$

$$\text{dove } c = c' - 1975m + 120$$



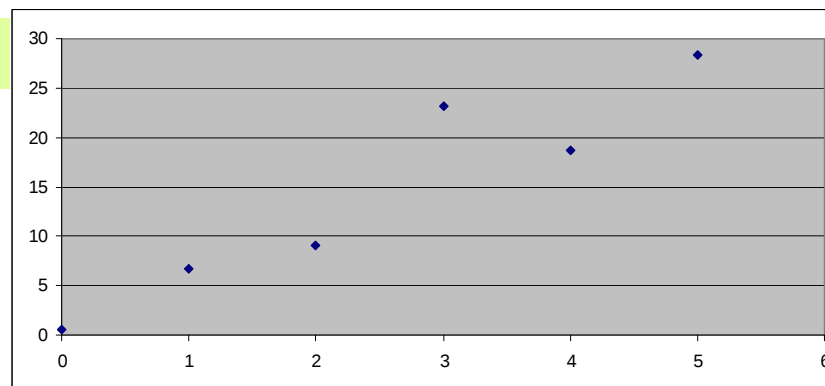
esempio divorzi – minimi quadrati

D



A

T



X

X	T	XX	XT
0	0.5	0	0
1	6.7	1	6.7
2	9.1	4	18.2
3	23.1	9	69.3
4	18.7	16	74.8
5	28.3	25	141.5
Σ	15	55	310.5

$$m = (6 \cdot 310.5 - 15 \cdot 86.4) / (6 \cdot 55 - 15 \cdot 15) = 5.4$$

$$c' = (55 \cdot 86.4 - 310.5 \cdot 15) / (6 \cdot 55 - 15 \cdot 15) = 0.9$$

$$c = 0.9 - 1975 \cdot 5.4 + 120 = -10544.1$$

Domanda D(2006)?

$$D(2006) = 5.4 \cdot 2006 - 10544.1 = 228.3 \quad \&tc.$$



esempio divorzi – regressione

- nelle variabili trasformate $\underline{X} = 15/6 = 2.5$; $\underline{T} = 86.4/6 = 14.4$

$$\begin{aligned} m &= [\sum (X_i - \underline{X}) \cdot (T_i - \underline{T}) / k] / [\sum (X_i - \underline{X})^2 / k] \\ &= [\sum X_i T_i - k \underline{X} \underline{T}] / [\sum X_i^2 - k \underline{X}^2] = (310.5 - 6 \cdot 2.5 \cdot 14.4) / (55 - 6 \cdot 6.25) \\ &= 5.4 \text{ k divorzi/anno} \end{aligned}$$

come trovato coi minimi quadrati (il passaggio si ha semplificando k e ricordando che $\sum X_i = k \underline{X}$)

$$m = [\sum (A_i - \underline{A}) \cdot (D_i - \underline{D}) / k] / [\sum (A_i - \underline{A})^2 / k] \quad \text{è assolutamente inutile rifare i conti perchè } A_i - \underline{A} = X_i - \underline{X} \text{ e } D_i - \underline{D} = T_i - \underline{T} !!$$

nelle variab. non trasformate è immediato ricavare c

$$c = \underline{D} - m \underline{A} = 134.4 - 5.4 \cdot 1977.5 = -10544.1 \text{ divorzi}$$

che corrisponde al valore all'anno 0



Retta di regressione – pressione vs altitudine(*)

dati Aviazione Civile, $p = p_0 e^{mz}$ - dipende esponenzialmente da z $\Rightarrow$ prima si deve linearizzare $\ln p = \ln p_0 + mz = c + mz$

X		T		XT	XX
z(m)		p(hPa)		z*ln p	z*z
0		1013		0	0
2000		795		13356.68	4000000
4000		616		25692.99	16000000
6000		472		36941.87	36000000
Σ		12000		26.1792	75991.55

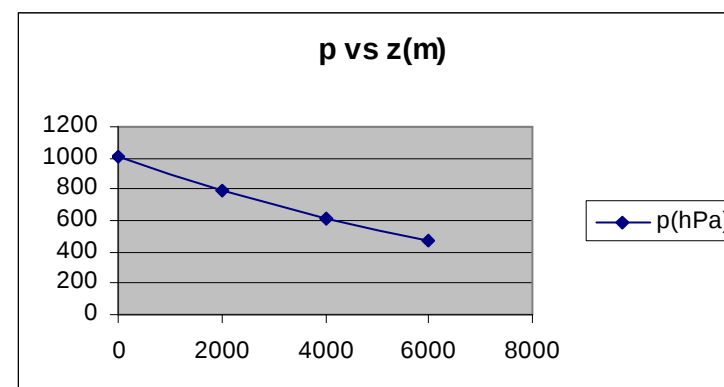
valore dalla regressione	
z	lineare
1000	904.20
5000	543.80

$$m = -0.0001273$$

$$c = 6.9267358 \quad p_0 = 1019.162$$

$$m = (kS_{XT} - S_X S_T) / (kS_{XX} - S_X^2)$$

$$c = (S_{XT} - S_X S_T / S_X) / (kS_{XX} - S_X^2)$$



(*) facoltativo

FLN nov 2009

12



Retta di regressione – altitudine vs pressione(**)

- ottenuti m , $c=\ln(p_0)$ abbiamo
$$p(z) = e^c e^{mz} = 1019.2e^{-0.0001273z}$$

NB $1019.2 \neq 1013 = p(0)$ nella tabella, a causa dell'aggiustamento che tiene conto degli altri 3 dati e della loro dispersione
 - con i parametri *posso(*)* invertire e ottenere $z(p)$
$$z(p) = (6.927 - \ln(p))/0.0001273$$

e calcolare per es.
 $z(750\text{hPa}) = 2410 \text{ m}$
&tc. ed una misura della pressione (differenza di p) può essere usata per misurare l'altitudine (la differenza di altitudine)
- (*) **N.B.** in generale la regressione di z su p non dà gli stessi coefficienti dell'inversione della regressione di p su z (a meno che il coefficiente di correlazione non sia $\sim \pm 1$, qui $r = -0.9998$)



Retta di regressione – esempio 3

X	T	XT	XX		
mese	indice di inquinamento	m*idi	m^2	mese di riferimento	valore calcolato con regressione lineare
1	254	254	1	18	57.30
2	241	482	4		
3	233	699	9		
4	212	848	16		
5	201	1005	25		
6	195	1170	36		
7	184	1288	49		
8	172	1376	64		
9	159	1431	81		
10	147	1470	100		
11	139	1529	121		
12	128	1536	144		
Σ	78	2265	13088		

ottenuto con Excel usando
FORECAST(18;idi;mesi)

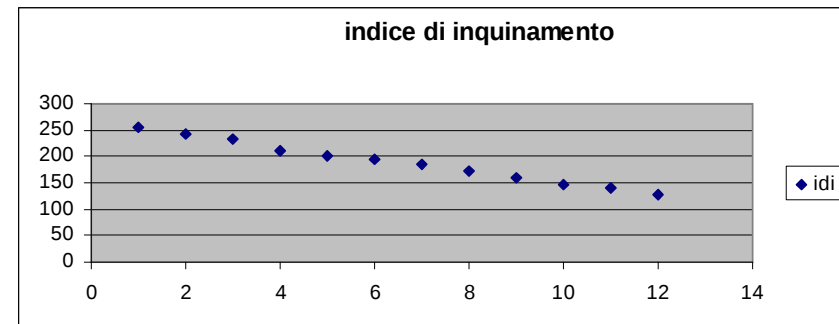
Indice di inquinamento di un lago in funzione del **tempo**, si vuole estrapolare la retta di regressione per trovare il momento in cui l'inquinamento è inferiore ad un certo valore per iniziare il ripopolamento del lago

m = -11.430

idi(18) = 57.30

c = 263.045

$idi(t) = c + mt$ dove c, m sono ottenuti dalla retta di regressione, invertendo
 $t = (idi - c)/m$ e se ad es. voglio $idi \leq 50$ avrò
 $t(50) = (50 - 263.05)/(-11.430) = 18m19g$





uso della calcolatrice e del foglio elettronico

- in pratica in genere i calcoli coi minimi quadrati o la stima della retta di regressione non sono laboriosi se si usa una calcolatrice adatta (molte calc. scientifiche hanno statistica bidimensionale e regressione lineare)
- un foglio elettronico ha alcune funzioni che permettono di far immediatamente il calcolo, ad es. in Excel
FORECAST(x;y-noti;x-noti) permette di calcolare il valore della retta nel punto x nuovo (estrapolato o interpolato) usando i valori noti delle coppie x_i, y_i (X_i, T_i nella ns. notazione), le coppie di numeri possono essere date singolarmente oppure ci si può riferire ad un gruppo contiguo di celle (vettore) ; altre funzioni calcolano m [SLOPE(y-noti,x-noti)] e c [INTERCEPT(y-noti,x-noti)]; il coefficiente di correlazione è calcolato con CORREL(vett1;vett2) che da la correlazione fra due insiemi di dati di uguale ampiezza



varianza dei parametri della retta

- (non fatto a lezione) m , c sono anch'essi aleatori visto che almeno una variabile lo è: se le varianze delle T_i sono costanti per tutti i dati (non dipendono da i), si può mostrare
$$\sigma_m^2 = \sum (T_i - \bar{T})^2 \cdot (1 - r^2) / [(k - 2) \cdot \sum (X_i - \bar{X})^2]$$
- (non fatto a lezione) le formule per σ_m^2 σ_c^2 sono ancora più noiose da calcolare con la calcolatrice. Excel ad es. ha la funzione `LINEST(y_i; x_i; const; stats)` che permette di fare i calcoli – `const` e `stats` sono due quantità logiche – `const=false` permette di fissare a 0 l'intercetta durante il calcolo mentre `const=true` ha $c \neq 0$, con `const=stats=true` si ottiene un quadro (array) di informazioni estraibile con un'altra funzione: `INDEX(LINEST(y_i; x_i; true; true); 2; 1)` dà ad es. σ_m &tc. (si veda l'help di Excel per ulteriori informazioni)
- attenzione quello che noi chiamiamo “ c ” in Excel vers. Inglese è chiamato “ b ” &tc. e questo è il meno!