Retta di regressione

La statistica nella ricerca di leggi sperimentali

Ecco un altro problema per cercare la retta di regressione che non ha più il vincolo di passare per l'origine O(0, 0).

Completa la scheda di lavoro per affrontare il problema.

Che cosa hai trovato

La retta ‘dei minimi quadrati’ che non passa per O

Quesito 1

La Capacità Vitale (CV) *è il volume massimo d’aria contenuto nei polmoni dopo un’ispirazione profonda. Per studiare gli effetti del fumo di sigarette, i medici hanno studiato la relazione fra Capacità Vitale e numero di sigarette fumate al giorno in un gruppo di fumatori.*

I dati sono raccolti nella tabella qui sotto.

Numero di sigarette **X**	2	4	6	7	8	10	12	14	16	20
CV (litri d’aria) **Y**	6,5	6,5	5,9	5,5	5,5	4,8	4,4	4,1	3,8	3,1

1. Spiega perché la retta *s* che meglio raccorda i punti sperimentali qui sopra non può passare per O(0, 0).

Perché non ha senso una capacità vitale 0, con 0 sigarette fumate.

La retta di regressione

Quesito 2

La relazione cercata dai medici porta a cercare i coefficienti m_s e q_s nell'equazione della retta s del tipo

$$Y = mX + q.$$

Con lunghi calcoli si trova:

$$m_s = \frac{\sum_{k=1}^{k=N} (x_k - M_x) \cdot (y_k - M_y)}{\sum_{k=1}^{k=N} (x_k - M_x)^2}$$

$$q_s = M_y - m_s \cdot M_x$$

dove M_x è la media dei dati X e M_y è la media dei dati X

2. Spiega perché la retta s che meglio raccorda i punti sperimentali passa per il punto M $(M_x ; M_y)$

Da $q_s = M_y - m_s \cdot M_x$ esplicito M_y e ottengo $M_y = m_s \cdot M_x + q_s$
Questo vuol dire che le coordinate di M soddisfano l'equazione della retta s, perciò la retta s passa per M.

Pendenza della retta di regressione e covarianza

La formula per ottenere la pendenza m_s della retta *s* di regressione è lunga e complicata da scrivere, ma può essere sintetizzata.

$$m_s = \frac{\sum_{k=1}^{k=N}(x_k - M_x)\cdot(y_k - M_y)}{\sum_{k=1}^{k=N}(x_k - M_x)^2}$$

Introduco la covarianza

Ricordo la varianza

Varianza dei dati X: $\sigma_x^2 = \frac{\sum_{k=1}^{k=N}(x_k - M_x)^2}{N}$

Covarianza: $\sigma_{xy} = \frac{\sum_{k=1}^{k=N}(x_k - M_x)(y_k - M_y)}{N}$

$$m_s = \frac{\sigma_{xy}}{\sigma_x^2}$$

Attenzione al significato dei simboli!

Esamino solo tre coppie di dati per riflettere sul significato dei simboli statistici.

X	Y
0	2
1	9
8	7
$M_x = 3$	$M_y = 6$

Varianza dei dati X: $\sigma_x^2 = \frac{(0-3)^2+(1-3)^2+(8-3)^2}{3} = \frac{38}{3} \cong 12{,}67$

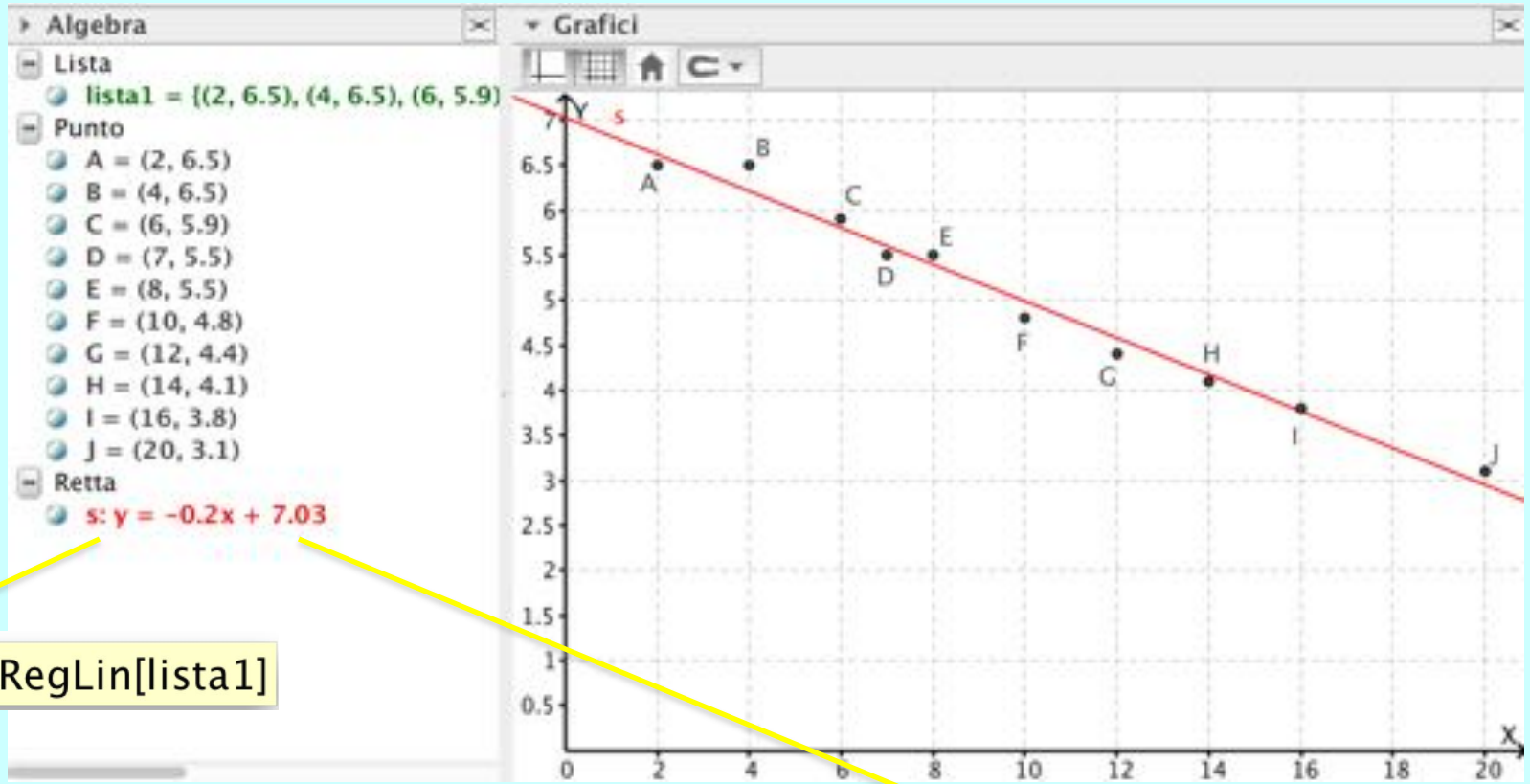
Varianza dei dati Y: $\sigma_y^2 = \frac{(2-6)^2+(9-6)^2+(7-6)^2}{3} = \frac{26}{3} \cong 8{,}67$

Covarianza: $\sigma_{xy} = \frac{(0-3)(2-6)+(1-3)(9-6)+(8-3)(7-6)}{3} = \frac{11}{3} \cong 3{,}67$

Pendenza della retta di regressione: $m_s = \frac{\frac{11}{3}}{\frac{38}{3}} = \frac{11}{38} \cong 0{,}29$

La retta di regressione

Quesiti 4, a, b, c, d



Retta **s**: RegLin[lista1]

c. Scrivi qui l'equazione della retta di regressione. $Y = -0{,}2X + 7{,}03$

d. Quale Capacità Vitale puoi prevedere per chi fuma 18 e 22 sigarette al giorno? 3,43 e 2,63

Scrivi qui sotto il procedimento seguito per rispondere.

$Y_{18} = -0{,}2 \cdot 18 + 7{,}03 = 3{,}43$ $\quad$ $Y_{22} = -0{,}2 \cdot 22 + 7{,}03 = 2{,}63$

Daniela Valenti, 2020

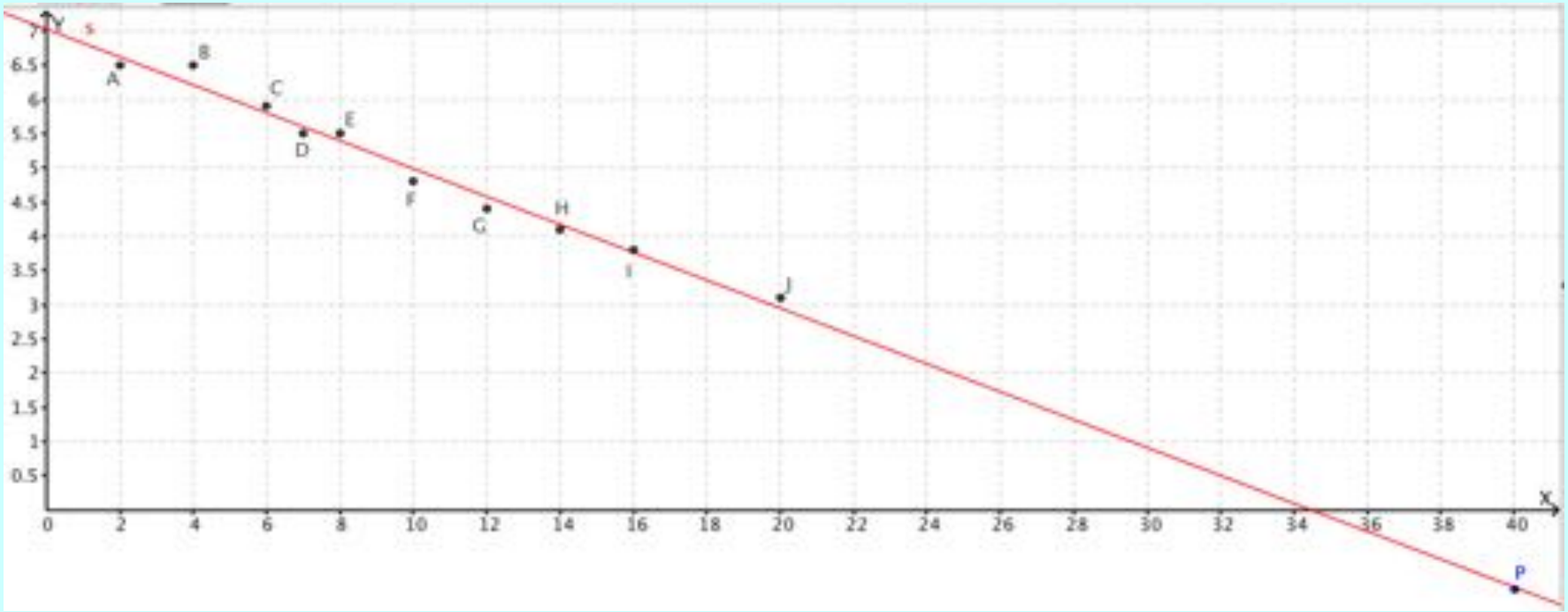
Una riflessione importante

Provo a prevedere la CV per chi fuma 40 sigarette.

$$Y_{40} = -0{,}2 \cdot 40 + 7{,}03 = -0{,}97$$

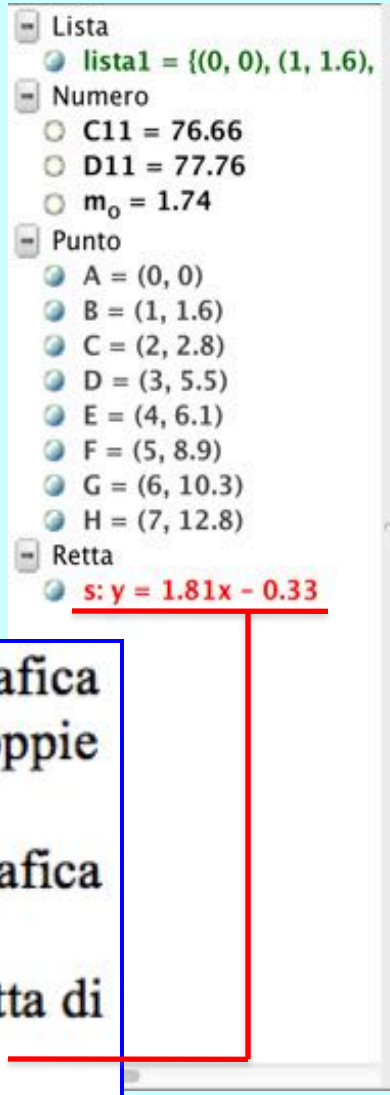
Inaccettabile CV negativa!

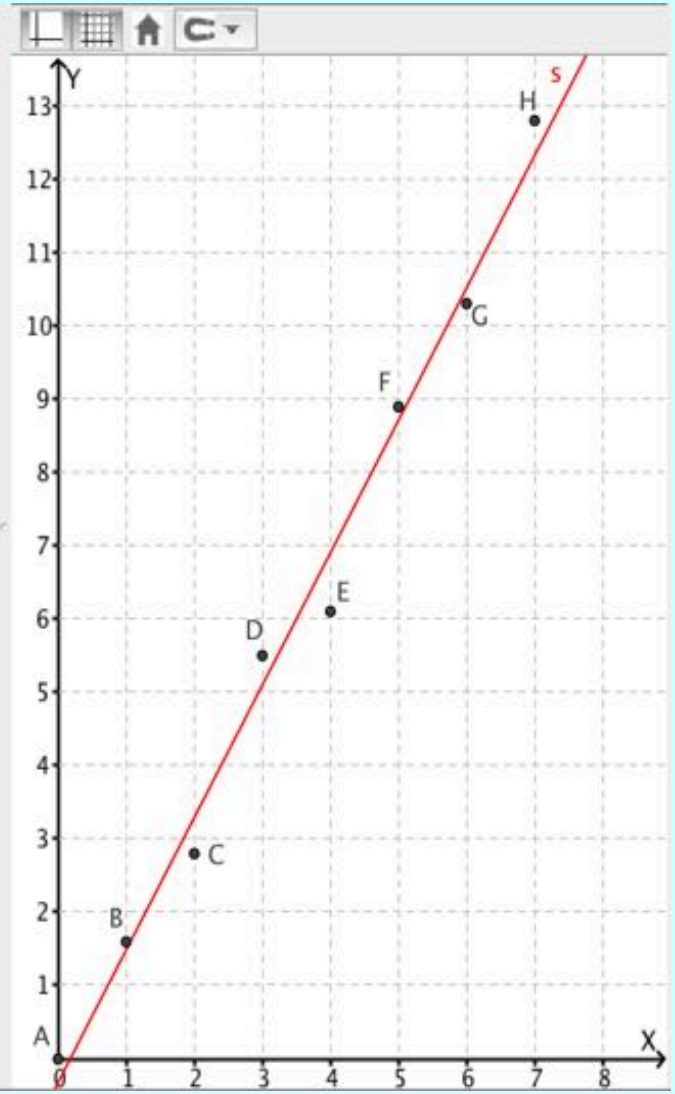
Il modello statistico non fornisce previsioni affidabili su casi lontani dai punti sperimentali.



Retta di regressione per i dati sulla deformazione della trave

Quesiti 5 a, b, c





a. Fai comparire sulla finestra grafica i punti che rappresentano le coppie di dati assegnati.

b. Fai comparire nella finestra grafica la retta di regressione.

c. Scrivi qui l'equazione della retta di regressione. $Y = 1{,}81X - 0{,}33$

Daniela Valenti, 2020

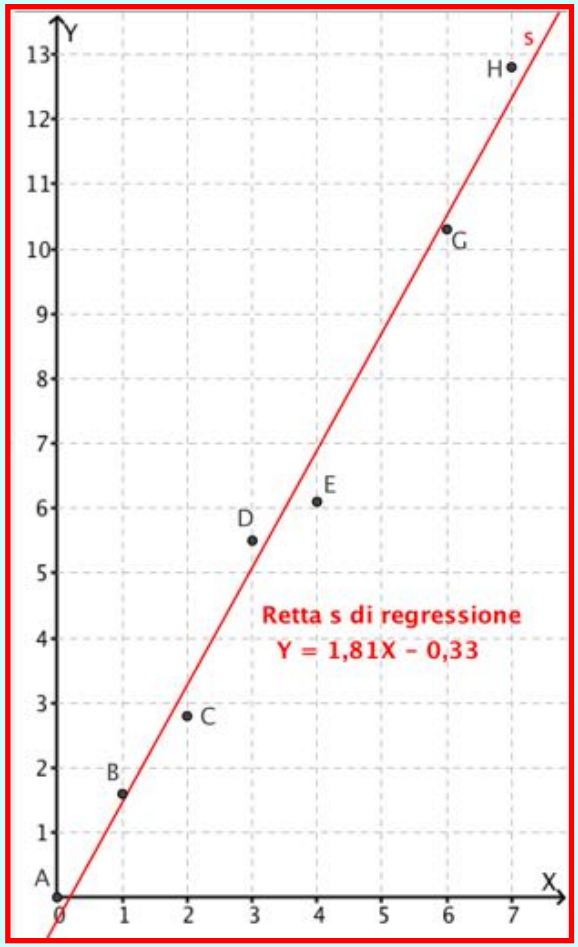
Confronto fra due rette

Quesito 5d

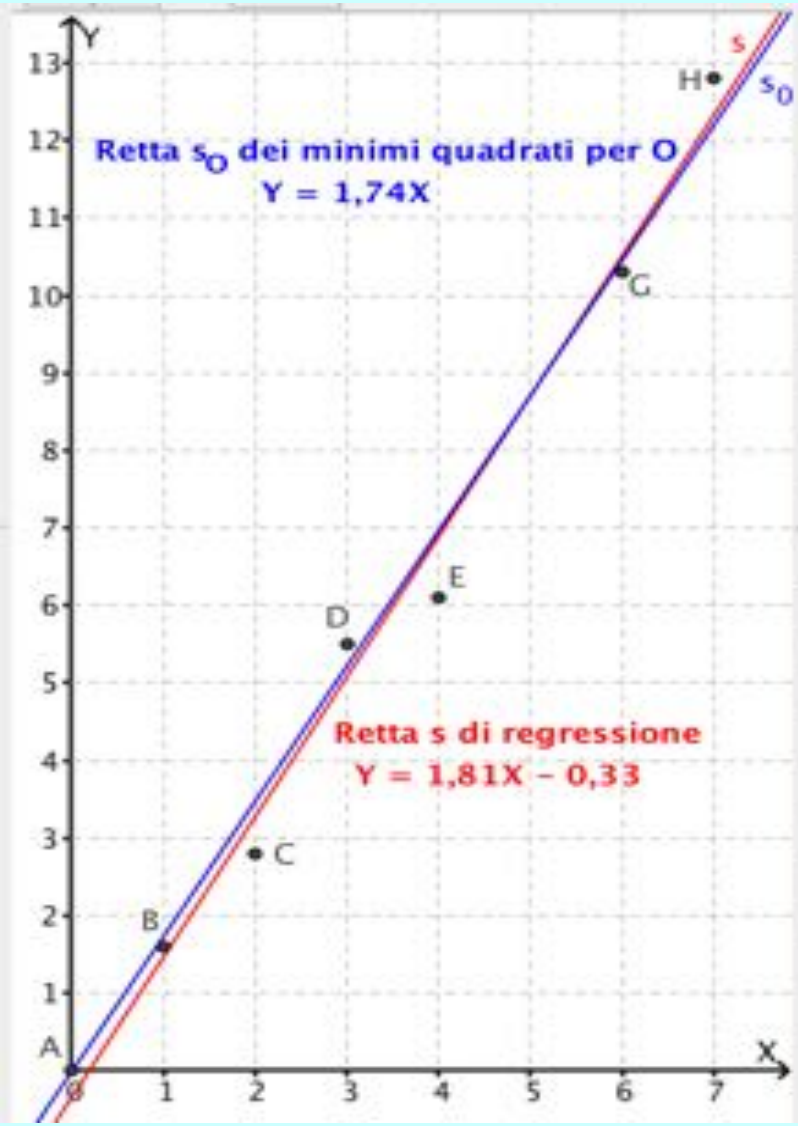
d. Confronta la retta di regressione con la retta passante per O, ottenuta quando hai risolto il quesito sulla deformazione della trave, e scrivi qui sotto le tue osservazioni.

Il confronto suggerisce varie osservazioni; ecco qualche esempio.

- La retta d'equazione $Y = 1{,}81X - 0{,}33$ NON passa per O.
- La retta s per $X = 0$ fornisce il risultato $-0{,}33$, che non ha significato rispetto ai dati.
- Il punto O è considerato come tutti gli altri punti sperimentali: la retta di regressione rende minima la somma dei quadrati degli scarti, ma non è 'obbligata' a passare per uno dei punti.



Confronto fra due rette con foglio di calcolo



	A	B
1	X	Y
2	0	0
3	1	1.6
4	2	2.8
5	3	5.5
6	4	6.1
7	5	8.9
8	6	10.3
9	7	12.8
10	m_O	
11	1.74	
12		
13		
14		
15		
16		
17		
18		
19		
20		
21		
22		
23		

Come ottenere l'equazione della retta dei minimi quadrati per O.

Statistica

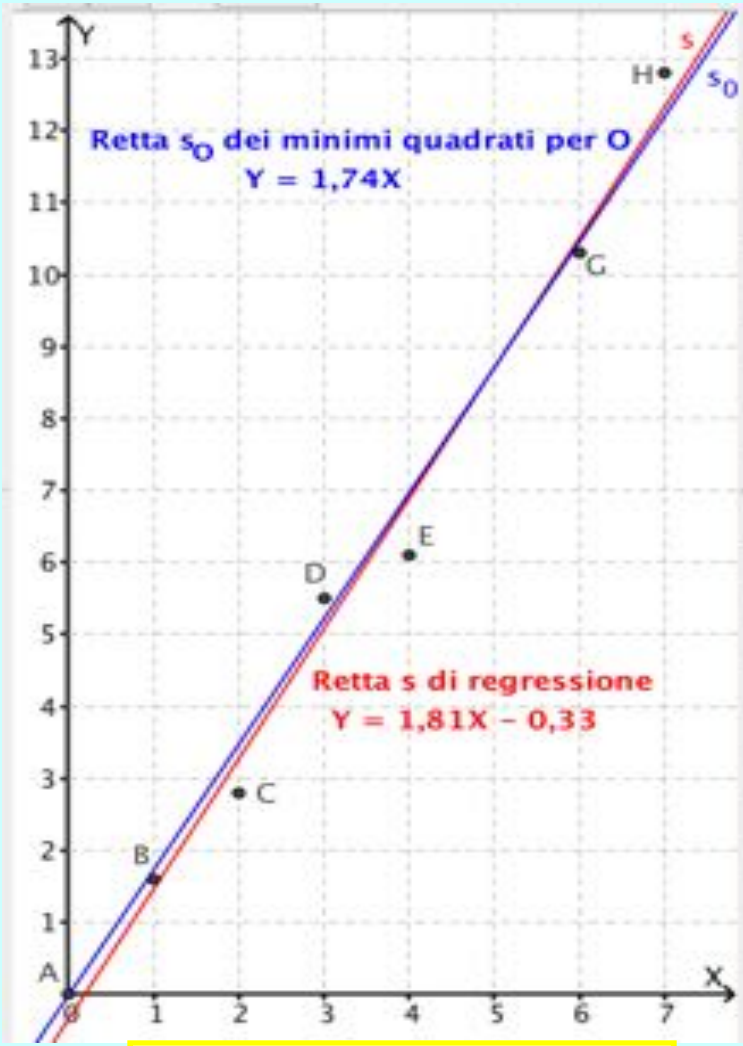
$$m_O = \frac{\sum_{k=1}^{k=N} x_k \cdot y_k}{\sum_{k=1}^{k=N} x_k^2}$$

Foglio di calcolo

SigmaXY[lista1] / SigmaXX[lista1]

Daniela Valenti, 2020

Confronto fra due rette con foglio di calcolo



	A	B	C	D	E	F
1	X	Y	Y_0	Y_s	$(Y - Y_0)^2$	$(Y - Y_s)^2$
2	0	0	0	-0.33	0	0.11
3	1	1.6	1.74	1.48	0.02	0.02
4	2	2.8	3.49	3.29	0.47	0.24
5	3	5.5	5.23	5.1	0.07	0.16
6	4	6.1	6.97	6.9	0.76	0.65
7	5	8.9	8.71	8.71	0.03	0.03
8	6	10.3	10.46	10.52	0.02	0.05
9	7	12.8	12.2	12.33	0.36	0.22
10	m_0				$Somma_0$	$Somma_s$
11	1.74				1.74	1.48
12						
13						
14						
15						
16						
17						
18						
19						
20						
21						
22						
23						

Punti su s

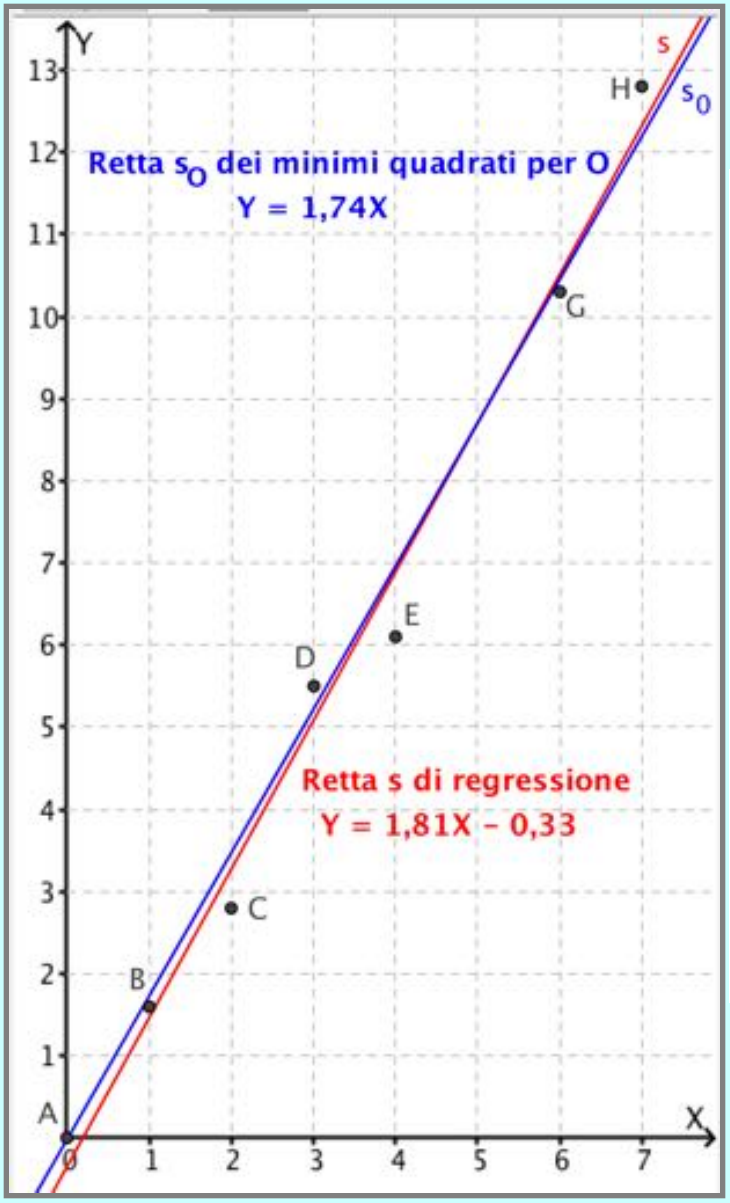
Punti su s_0

Punti sperimentali

Le due rette sono molto vicine, ma sono diverse.

Somma dei quadrati degli scarti più piccola per i punti della retta s

Confronto fra due rette con foglio di calcolo



Riflessioni

- La retta *s* di regressione è quella 'più vicina' ai punti sperimentali, ma non passa per O, perché tratta O come tutti gli altri punti sperimentali.
- È opportuno stabilire prima di tutto se è importante trovare una retta che passa per O(0; 0), per scegliere i procedimenti statistici più adatti ai dati da esaminare.