

Corso Integrato di Statistica Informatica e Analisi dei dati

prof. Carlo Meneghini

Dip. di Fisica “E. Amaldi”
via della Vasca Navale 84

meneghini@fis.uniroma3.it

<http://webusers.fis.uniroma3.it/~meneghini>

Grafici di dispersione e rette di regressione

Indice

Relazioni funzionali tra variabili

Grafici di dispersione (x,y)

Retta di regressione

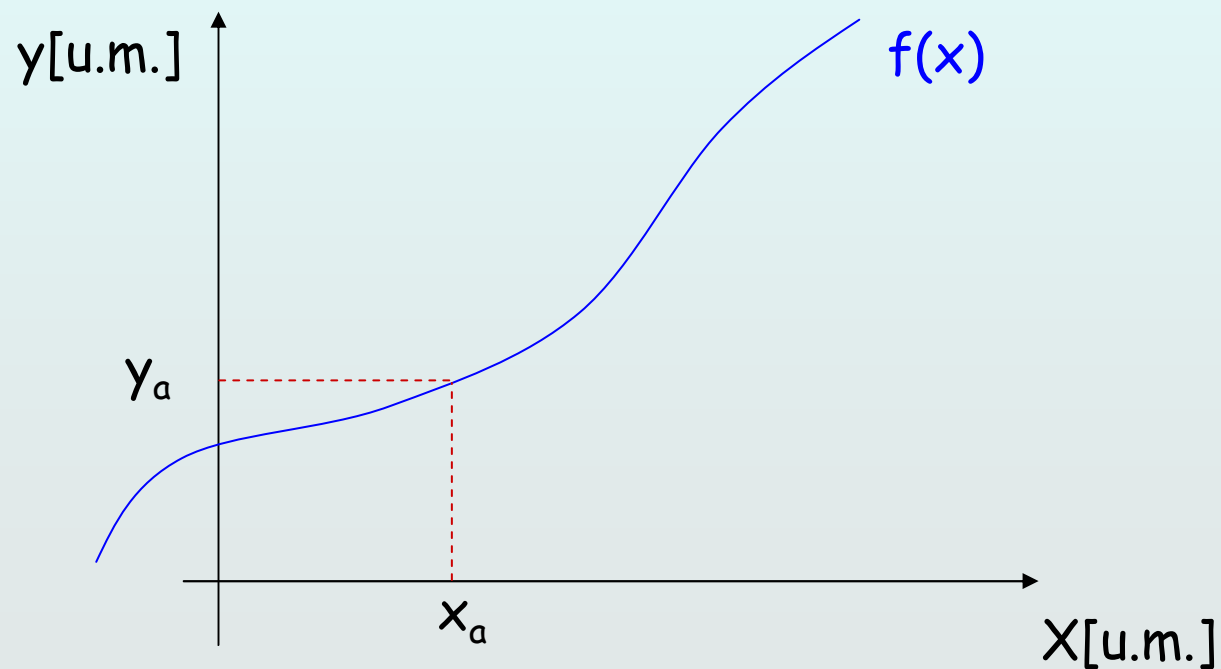
Correlazione

Relazione funzionale tra y e x

$$y = f(x)$$

Y= variabile dipendente

X=variabile indipendente



Relazione statistica tra y e x

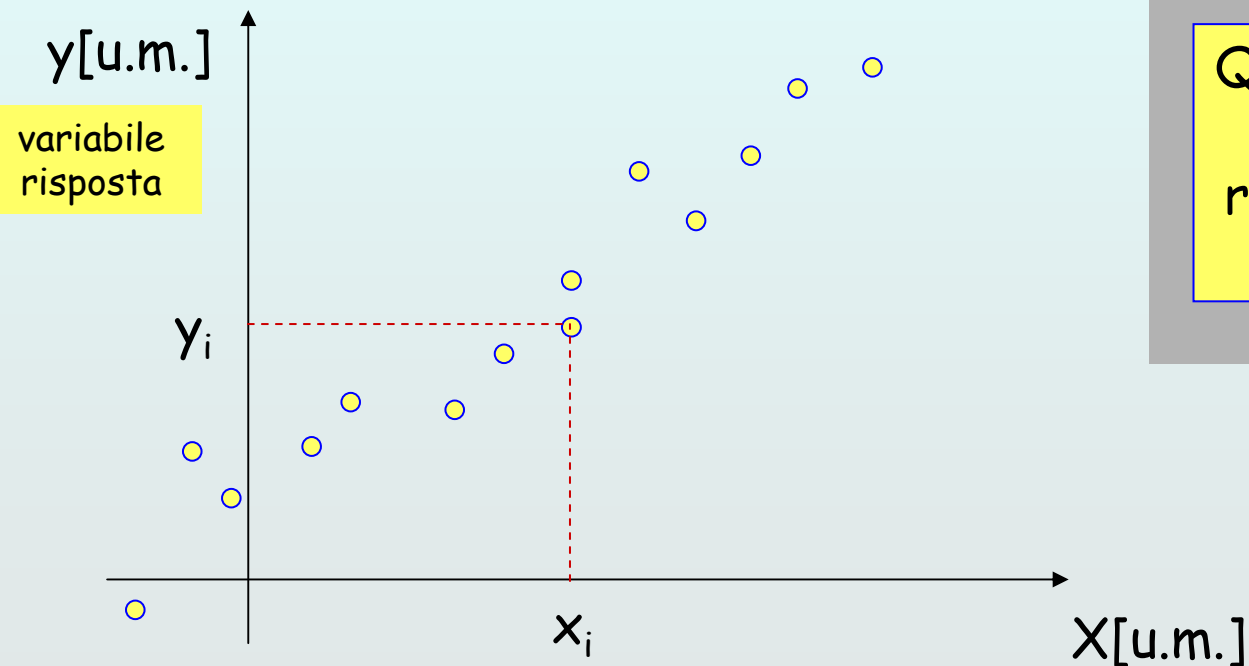
Relazione funzionale

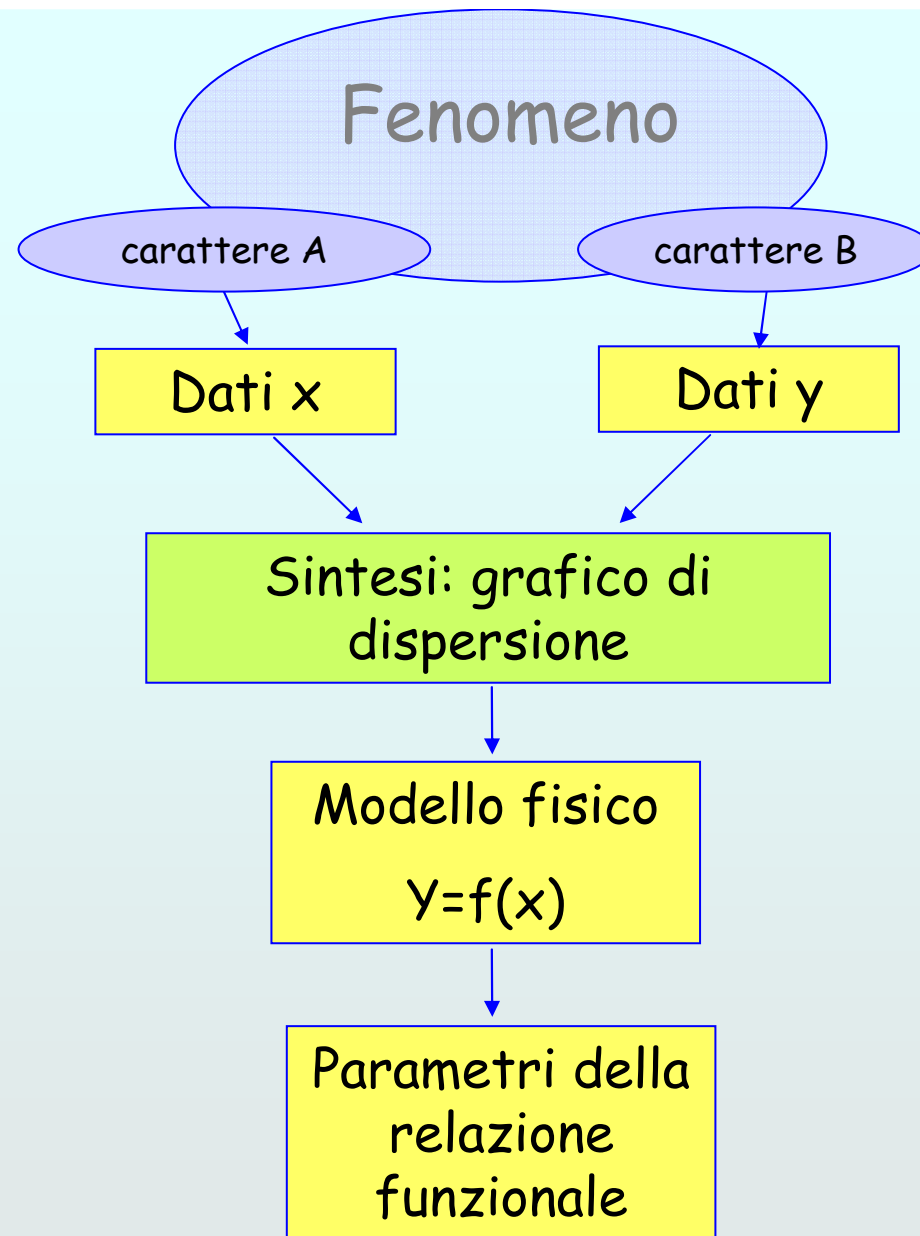
Effetti stocastici

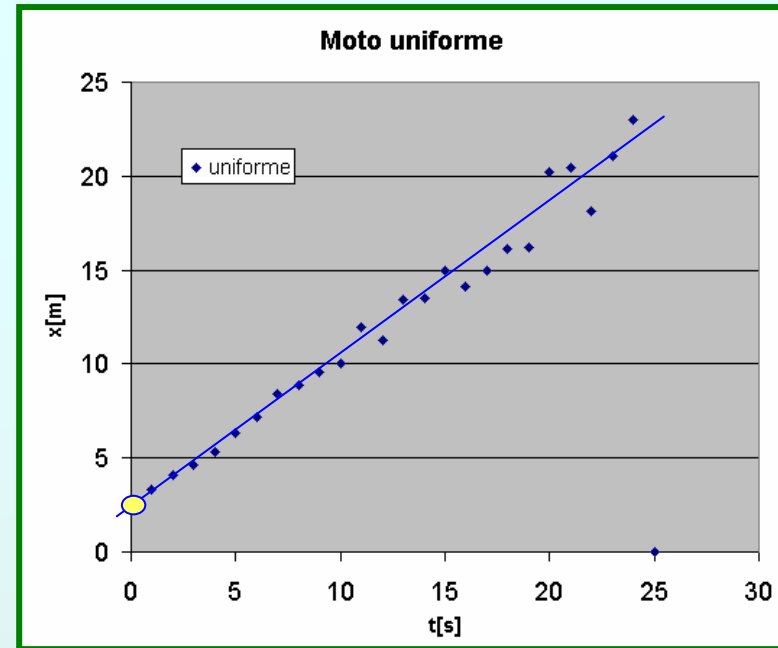
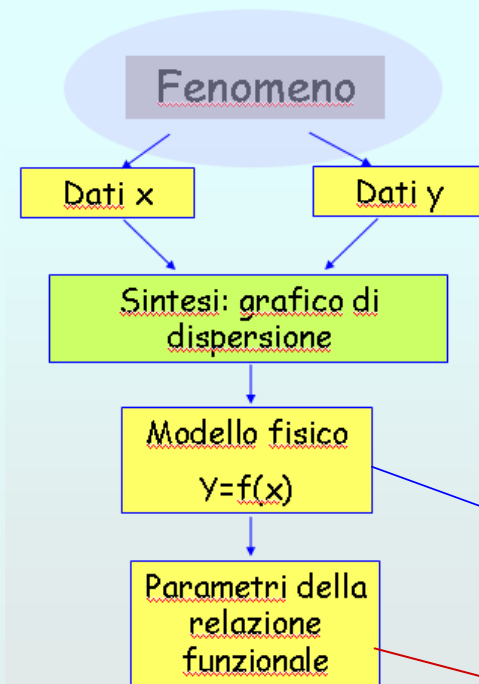
$$y_i = f(x_i) + \epsilon_i$$

Esiste una relazione tra x e y?

Quanto è ragionevole assumere una relazione lineare tra x e y?







Moto rettilineo uniforme

$$x(t) = x_0 + vt$$

Caso semplice: relazione lineare

$$f(x) = mx + c$$

Caso semplice: relazione lineare

$$f(x) = mx + c$$

coefficiente angolare

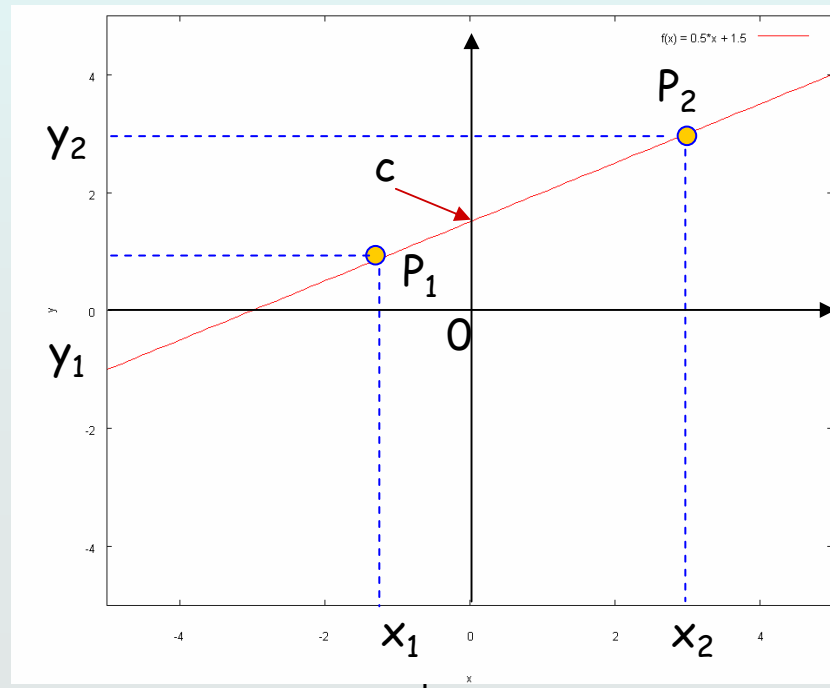
$$m = \frac{y_2 - y_1}{x_2 - x_1}$$

Termine noto

$$c = f(0)$$

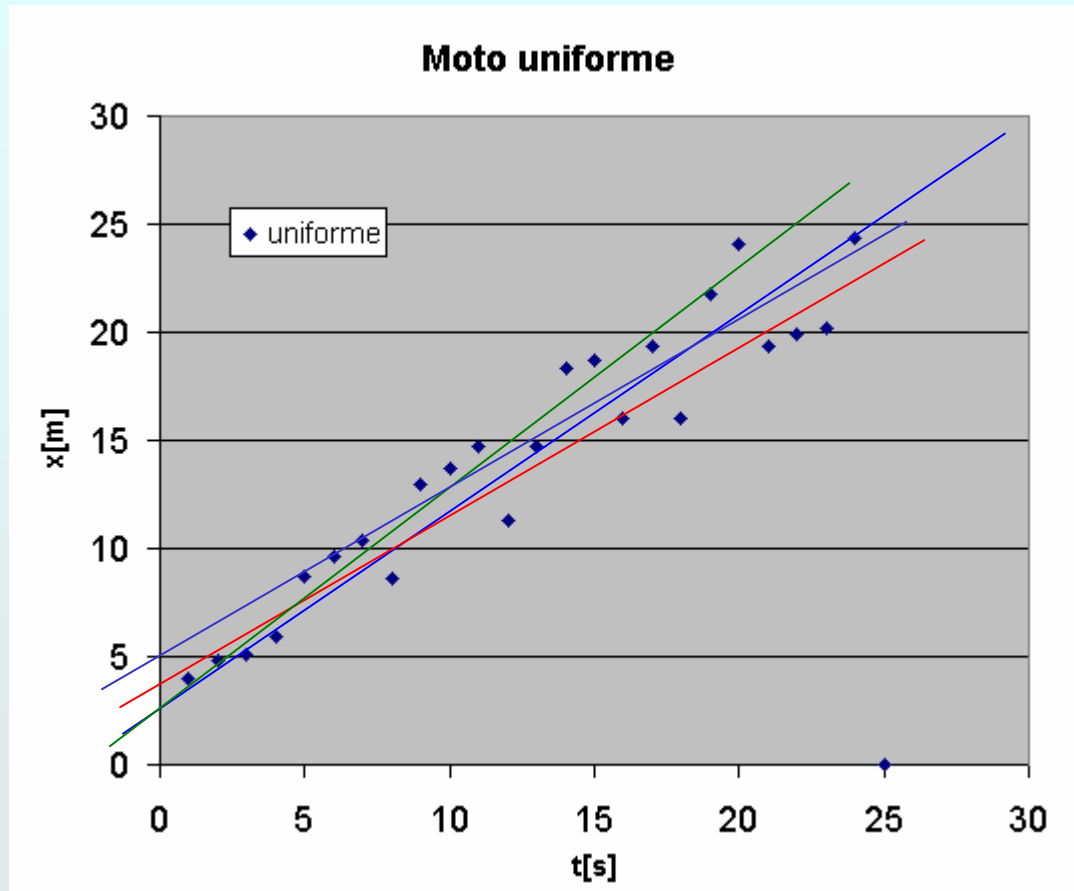
Coordinate
di un punto
nel piano
cartesiano

$$P_i = (x_i, y_i)$$



Caso semplice: relazione lineare

Quale è il modello migliore?



$$f(x) = mx + c$$

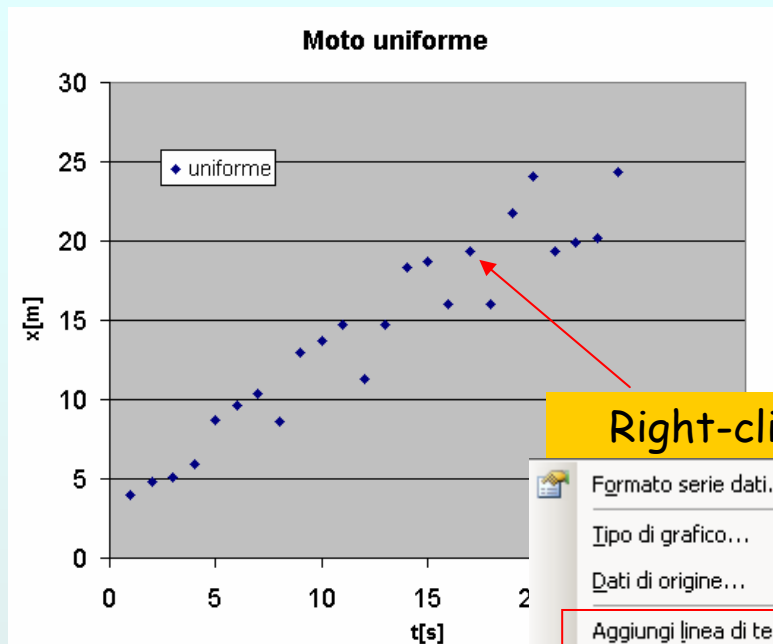
$$m = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^N (x_i - \bar{x})^2}$$
$$c = \bar{y} - m\bar{x}$$

Per il calcolo:

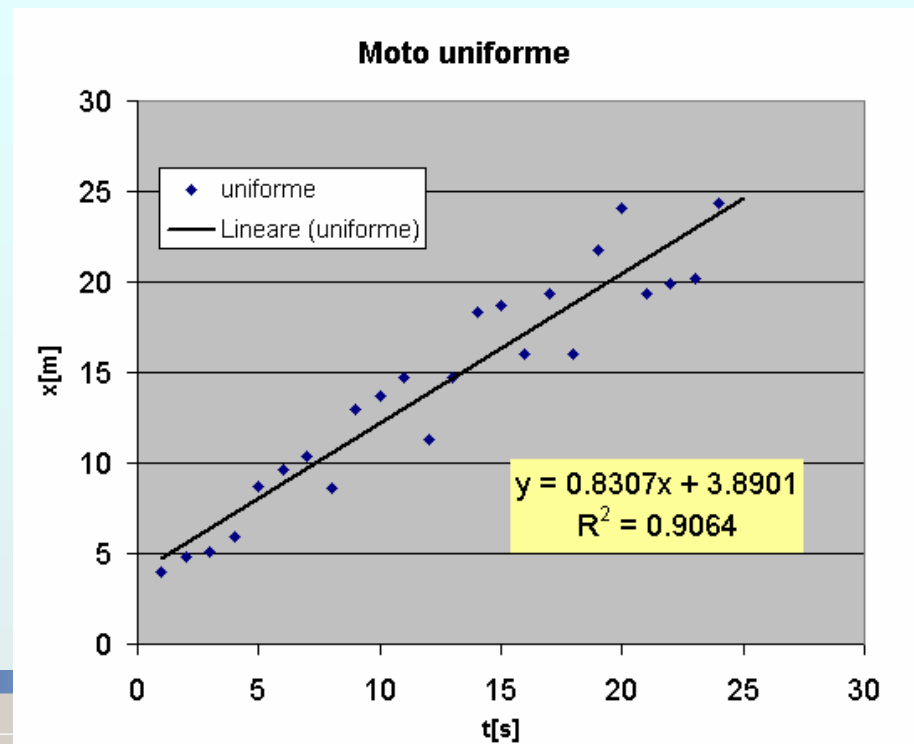
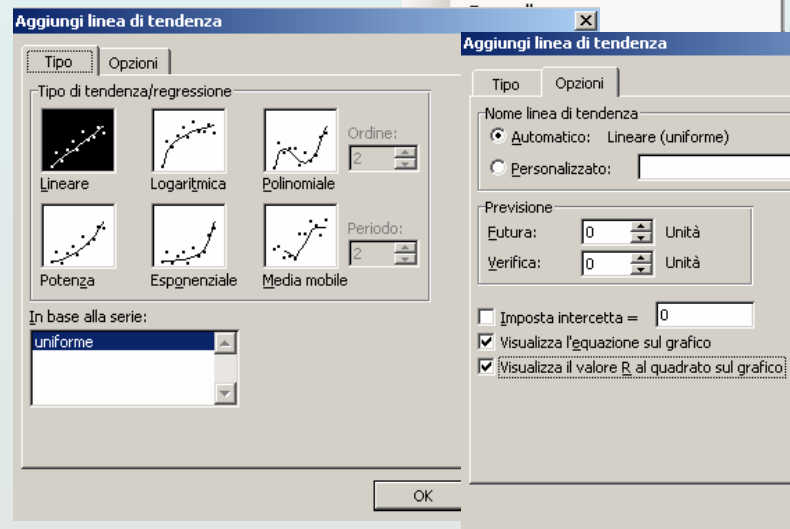
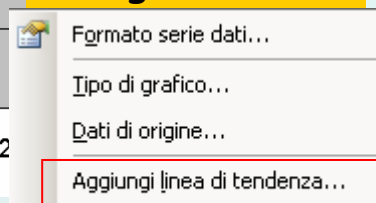
Funzioni
predefinite dei
fogli elettronici

$$y_i = f(x_i) + \epsilon_i$$

Calcolo della retta di regressione 1.



Right-click



si

Calcolo della retta di regressione 2.

	B	C	D	E	F	G
1	uniforme					
2	t[s]	x[m]				
3	1	3.94456	pendenza	0.831	m/s	
4	2	4.856494	intercetta	3.890	m	
5	3	5.114162				

Funzioni predefinite

PENDENZA(Y,X)

INTERCETTA(Y,X)

Argomenti funzione

PENDENZA

Y_nota C3:C26 = {3.94456027476722}

X_nota B3:B26 = {1\2\3\4\5\6\7\8\9\1}

= 0.830731114

Restituisce la pendenza della retta di regressione lineare fra le coordinate note.

X_nota è l'insieme dei valori indipendenti e può essere costituito da numeri, nomi, matrici o riferimenti contenenti numeri.

Risultato formula = 0.830731114

[Guida relativa a questa funzione](#)

OK Annulla

Argomenti funzione

INTERCETTA

Y_nota C3:C26 = {3.94456027476722}

X_nota B3:B26 = {1\2\3\4\5\6\7\8\9\1}

= 3.890082936

Calcola il punto di intersezione della retta con l'asse y tracciando una regressione lineare fra le coordinate note.

X_nota è l'insieme indipendente di osservazioni o dati e può essere costituito da numeri, nomi, matrici o riferimenti contenenti numeri.

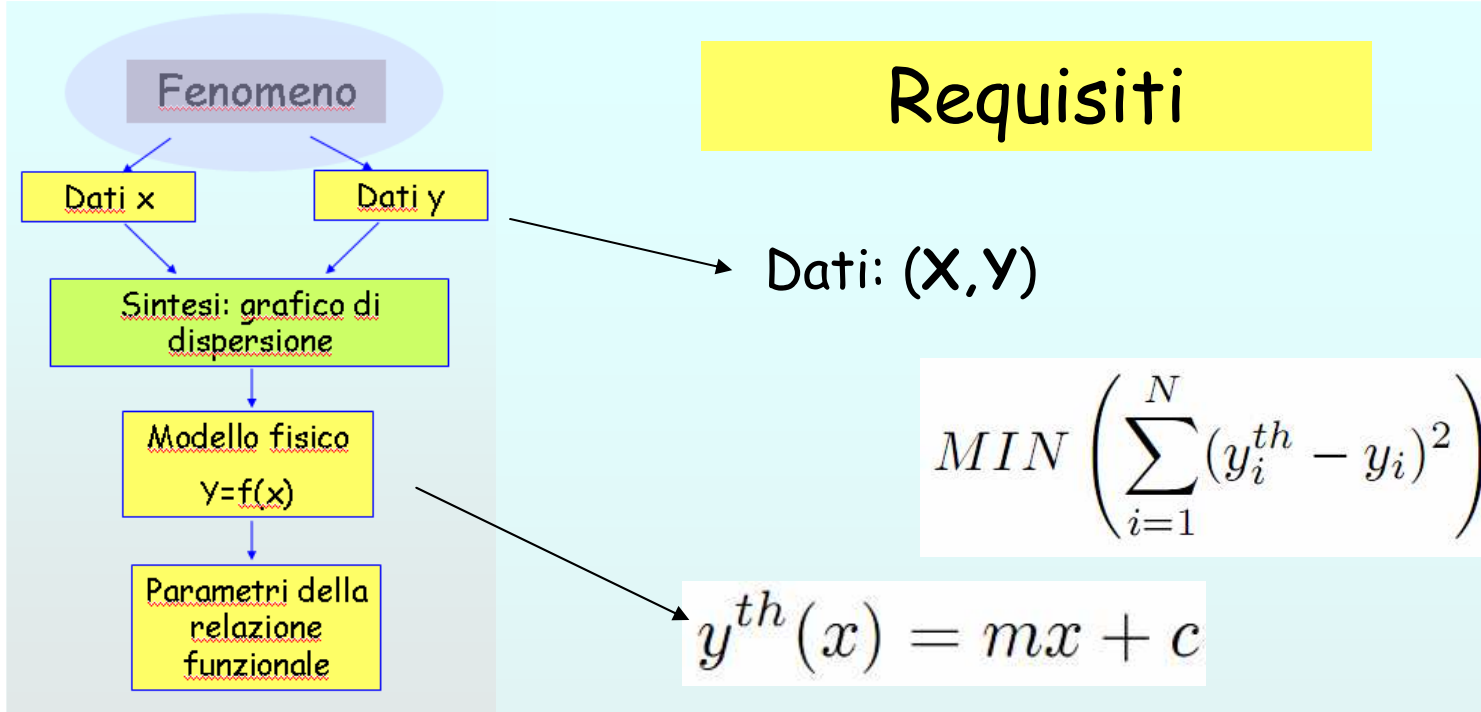
Risultato formula = 3.890082936

[Guida relativa a questa funzione](#)

OK Annulla

Y_nota=ordinate

X_nota=ascisse



Nota: m e c sono effettivamente i parametri del modello che meglio approssima di dati se sono verificate le seguenti condizioni:

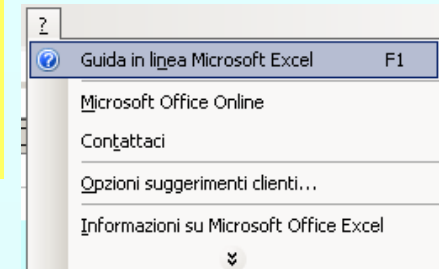
1. i valori misurati x_i sono noti con errore trascurabile
2. il valore atteso di y_i è $\bar{y}_i = mx + c$ (correttezza del modello)
3. i valori y_i seguono una distribuzione normale attorno al valor medio $\bar{y}_i$ con la stessa deviazione standard (omoschedastici)
4. I valori misurati y_i sono tutti indipendenti (indipendenza delle misure)

Calcolo della retta di regressione 3. Avanzato!

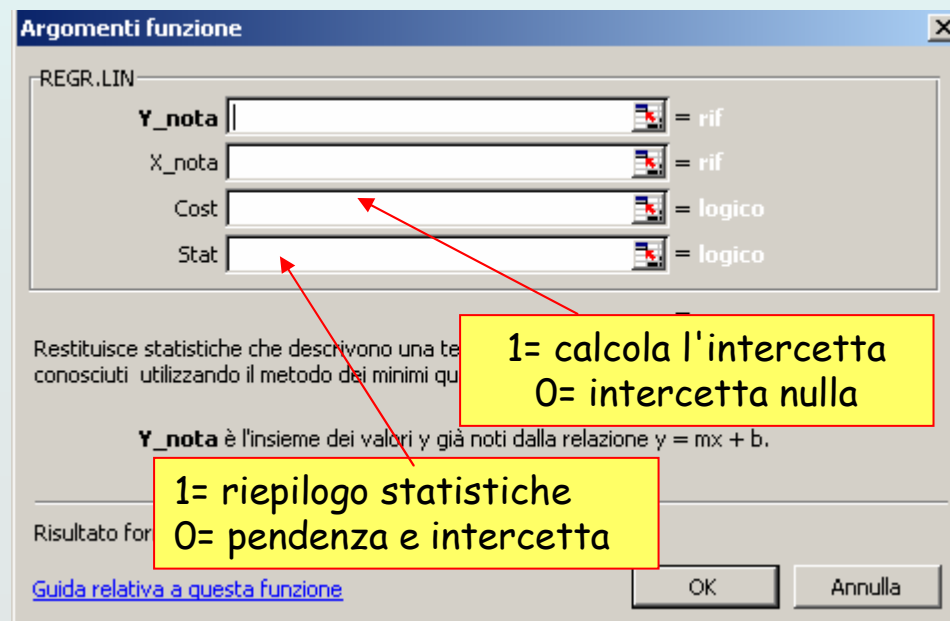
	B	C
1	uniforme	
2	t[s]	x[m]
3	1	3.94456
4	2	4.856494
5	3	5.114162
6	4	5.919289
7	5	8.668887
8	6	9.660259
9	7	10.33091
10	8	8.618057
11	9	12.93659
12	10	13.67318

Regr.lin(Y,X,cost,stat)

pendenza	intercetta
0.831	3.8901
0.057	0.8134
0.906	1.9304
212.976	22
793.631	81.98062



1. Selezionare una regione 2x5 celle
2. Inserire (o scegliere da f_x)
=Regr.lin(Y,X,cost,stat)
3. digitare: **ctrl+shift+enter**



Regr.lin(X,Y, cost, stat)

Y = colonna
delle ordinate

X = colonna
delle ascisse

Nota: una formula in forma di matrice si riferisce ad un intero gruppo di celle e si presenta tra parentesi {} nella barra delle formule

f_x {=REGR.LIN(B1:B17,A1:A17,1,1)}

dei dati sperimentali

Calcolo della retta di regressione 3. Avanzato!

	B	C
1	uniforme	
2	t[s]	x[m]
3	1	3.94456
4	2	4.856494
5	3	5.114162
6	4	5.919289
7	5	8.668887
8	6	9.660259
9	7	10.33091
10	8	8.618057
11	9	12.93659
12	10	13.67318

pendenza	intercetta
0.831	3.8901
0.057	0.8134
0.906	1.9304
212.976	22
793.631	81.98062

Pendenza: m	Intercetta: c
Errore standard sulla pendenza: σ_m	Errore standard sull'intercetta: σ_c
Coefficiente di determinazione R^2	Altre informazioni σ_y (da vedere più avanti) gradi di lib.
F	
SS_{reg}	
	SS_{red}

Nota:

Nel riportare i risultati che si ottengono da una regressione lineare ricordarsi di:

1. indicare le unità di misura,
2. riportare gli errori con 1 o 2 cifre significative
3. riportare il valore di R^2

$$m = 0.831 \pm 0.057 \text{ [u.m.]}$$

$$c = 3.89 \pm 0.81 \text{ [u.m.]}$$

$$R^2 = 0.91$$

Calcolo della retta di regressione 3. Avanzato!

	B	C	D
1	uniforme		
2	t[s]	x[m]	
3	1	3.94456	
4	2	4.856494	
5	3	5.114162	
6	4	5.919289	
7	5	8.668887	
8	6	9.660259	
9	7	10.33091	
10	8	8.618057	
11	9	12.93659	
12	10	13.67318	

pendenza	intercetta
0.831	3.8901
0.057	0.8134
0.906	1.9304
212.976	22
793.631	81.98062

=indice(matrice,riga,colonna)

=indice(Regr.lin(Y,X,cost,stat),riga,colonna)

=indice(Regr.lin(Y,X,cost,stat),1,1)

=indice(Regr.lin(Y,X,cost,stat),4,2)

1,1	1,2
2,1	2,2
3,1	3,2
4,1	4,2
5,1	5,2

Indici di riga e colonna

Pendenza: m	Intercetta: c
Errore standard sulla pendenza: σ_m	Errore standard sull'intercetta: σ_c
Coefficiente di determinazione R²	σ_y
F	gradi di lib.
SS_{reg}	SS_{red}

deviazione standard sulla stima di y

$$y_i = f(x_i) + \epsilon_i$$

$$\epsilon_i = \sigma_i$$

= N-1 se intercetta nulla altrimenti = N-2

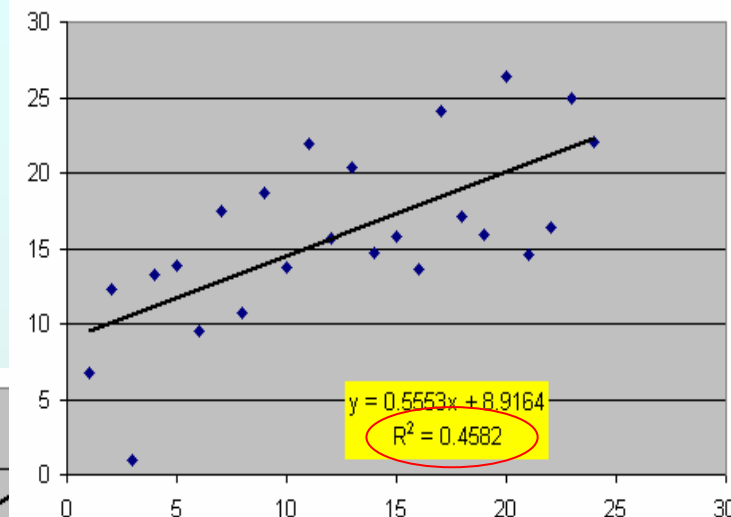
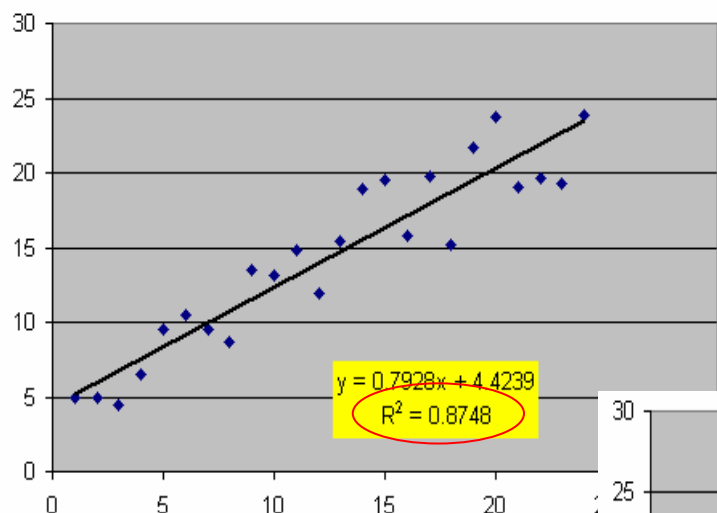
$$\sum_{i=1}^N ((mx_i + c) - \bar{y})^2$$

$$\sum_{i=1}^N (y_i - (mx_i + c))^2$$

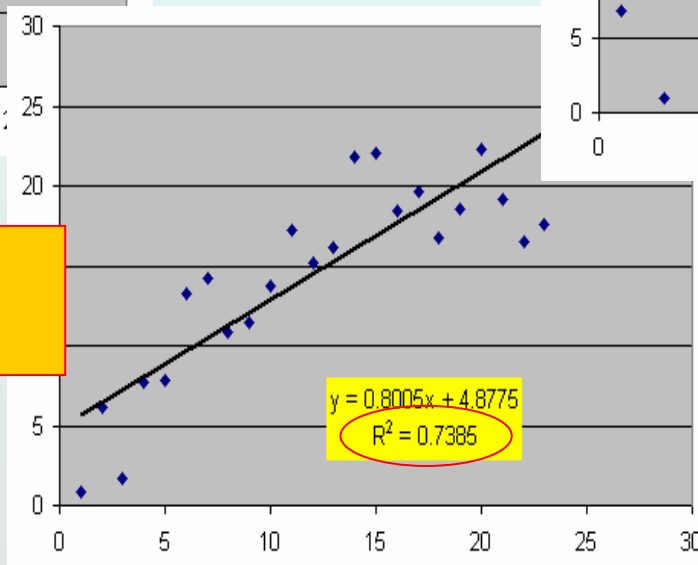
Statistica della funzione F,
corso avanzato

Retta di regressione

Una retta è un buon modello per i dati?



R^2 = coefficiente di determinazione



$$0 \leq R^2 \leq 1$$

☹️ ☺️

Una retta è un buon modello per i dati?

Correlazione lineare $R^2 = r$

Coefficiente di correlazione di Pearson: r

$$r = \frac{COV_{x,y}}{s_x s_y}$$

covarianza

$$cov_{x,y} = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})$$

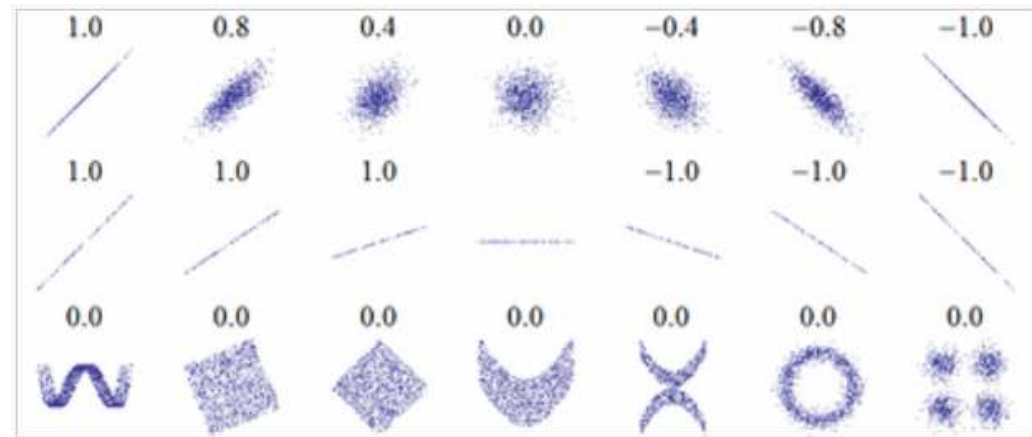
Varianza dei valori x e y

$$s_x^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2$$

$$s_y^2 = \frac{1}{N-1} \sum_{i=1}^N (y_i - \bar{y})^2$$

$$r = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left(\sum_{i=1}^N (x_i - \bar{x})^2\right) \cdot \left(\sum_{i=1}^N (y_i - \bar{y})^2\right)}}$$

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



Wikipedia: correlation coefficient

$r=0$ non significa che non c'è una correlazione tra i dati ma solo che la relazione lineare non è un buon modello

Covarianza e correlazione lineare

COVARIANZA(matrice1,matrice2)

$$cov_{x,y} = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})$$

CORRELAZIONE(matrice1,matrice2)

PEARSON(matrice1,matrice2)

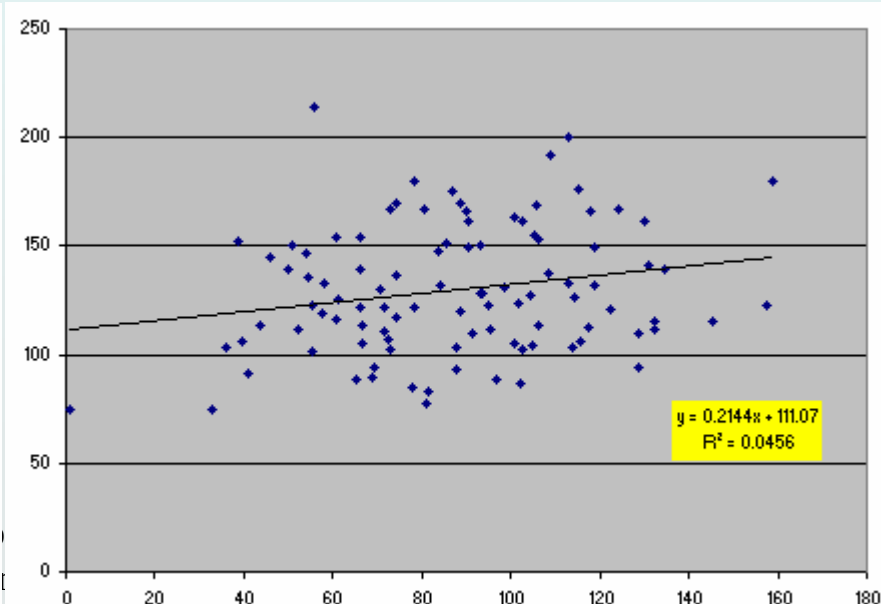
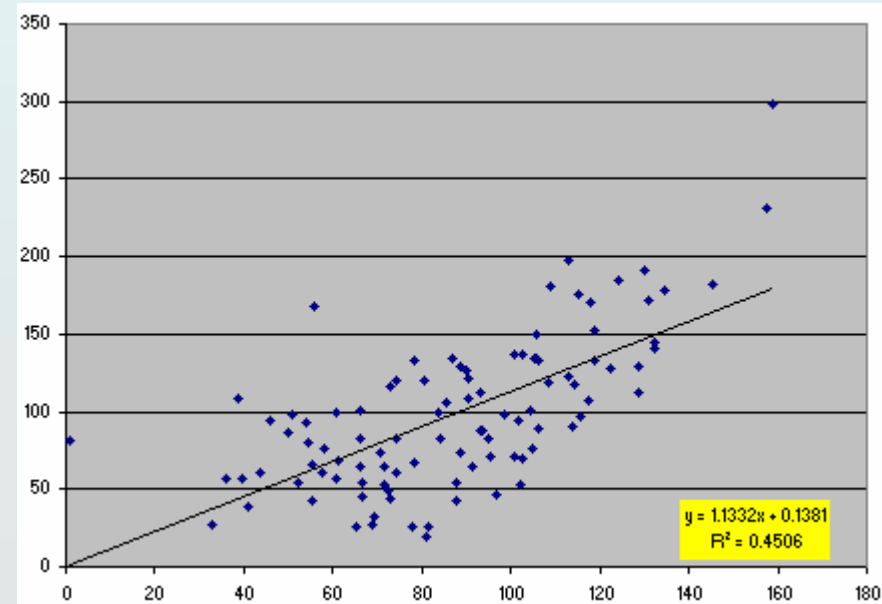
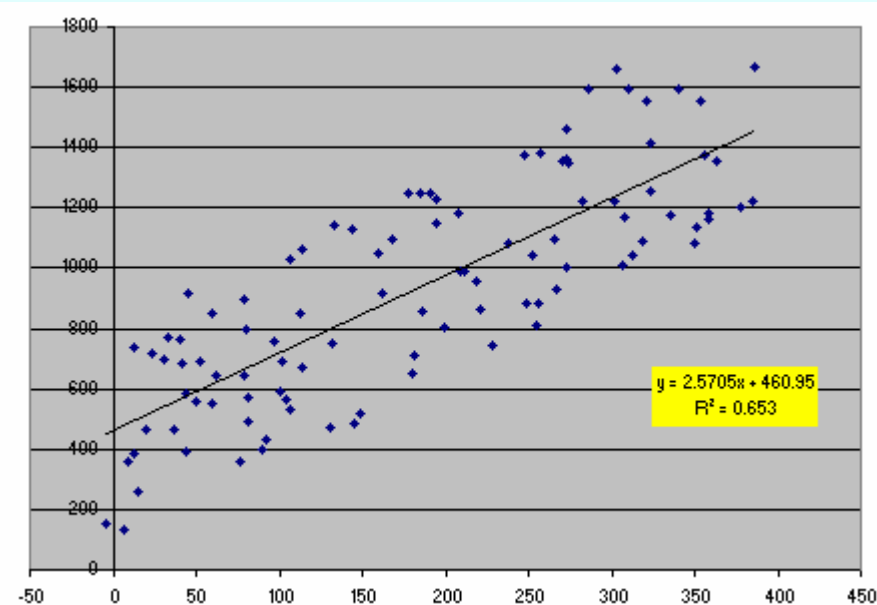
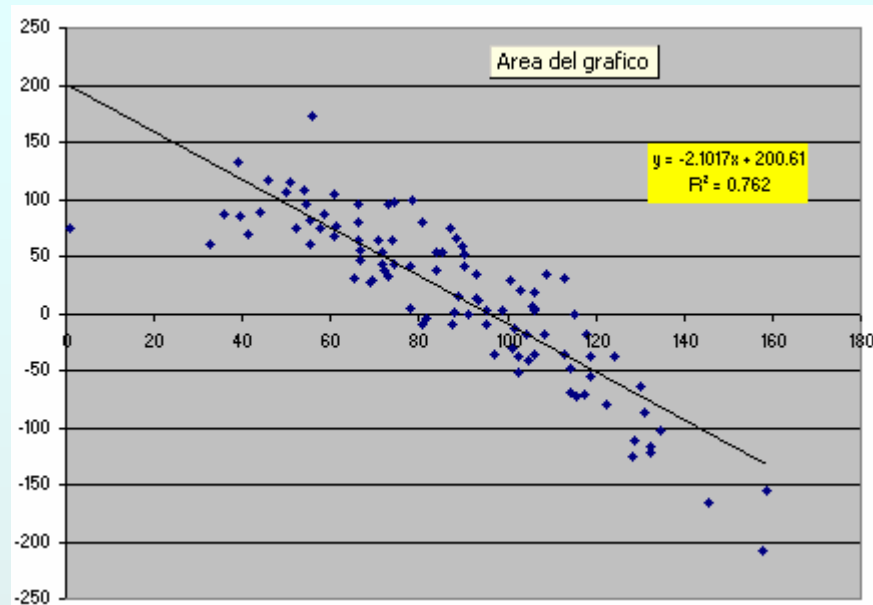
$$r = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left(\sum_{i=1}^N (x_i - \bar{x})^2\right) \cdot \left(\sum_{i=1}^N (y_i - \bar{y})^2\right)}}$$

RQ(matrice1,matrice2)

$$R^2 = r^2$$

	serie 1
sx	114.10
sy	362.93
covarianza	33128.47
correlazione - r	0.81
R^2	0.65
pendenza	0.25
intercetta	-53.67

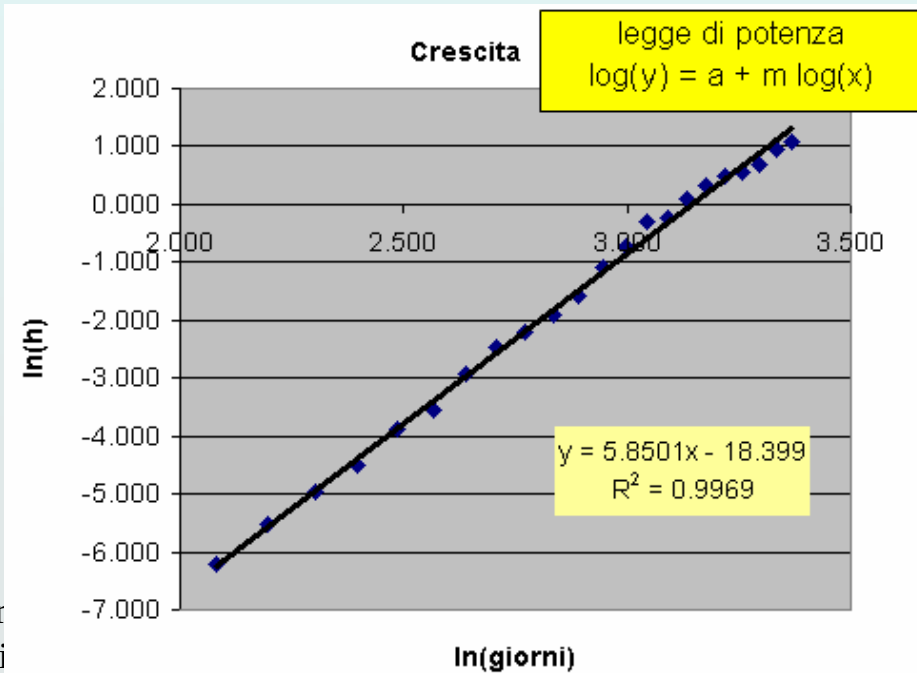
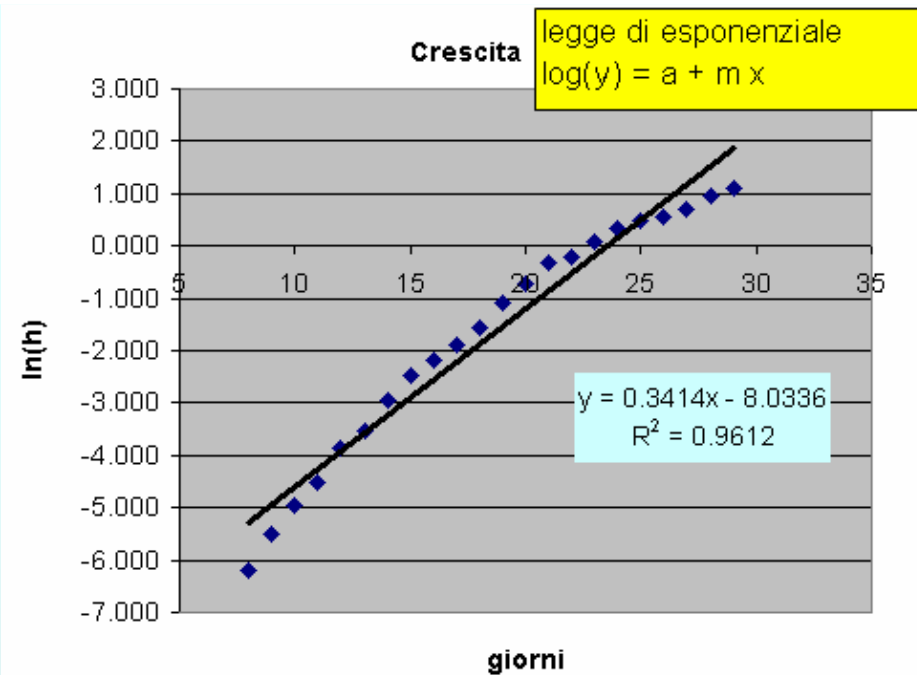
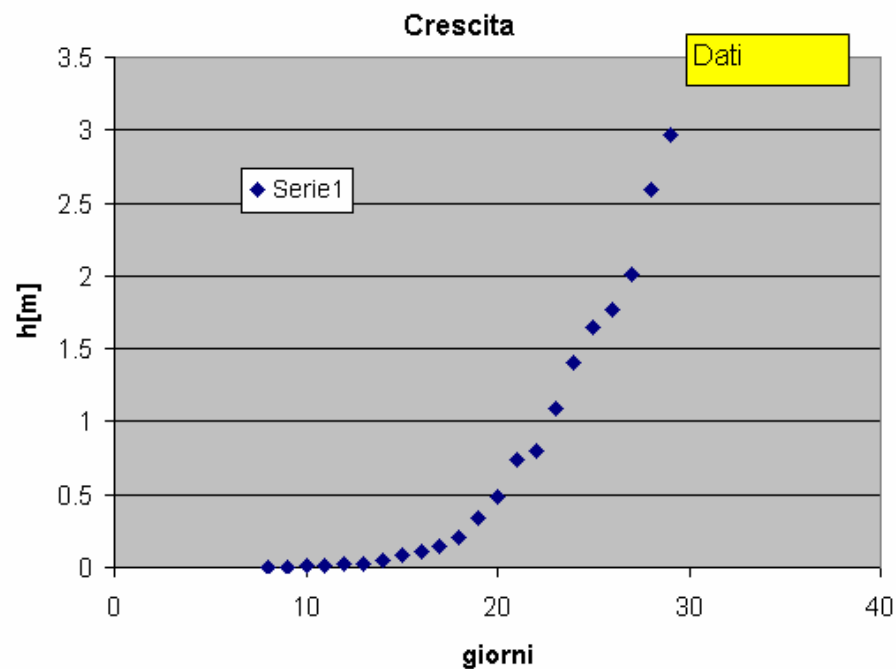
Una retta non è un buon modello per i dati?



Una retta non è un buon modello per i dati!

Linearizzazione

Funzione $y = f(x)$	Forma linearizzata $Y = AX + B$	Cambiamenti di variabili e costanti
$y = C \cdot x^A$	$\ln y = A \ln x + \ln C$	$X = \ln x \quad Y = \ln y$ $C = e^B$
$y = C \cdot e^{Ax}$	$\ln y = A \cdot x + \ln C$	$X = x \quad Y = \ln y$ $C = e^B$
$y = A \ln x + B$	$y = A \ln x + B$	$X = \ln x \quad Y = y$
$y = \frac{A}{x} + B$	$y = A \frac{1}{x} + B$	$X = \frac{1}{x} \quad Y = y$
$y = \frac{1}{Ax + B}$	$\frac{1}{y} = Ax + B$	$X = x \quad Y = \frac{1}{y}$
$y = \frac{x}{A + Bx}$	$\frac{1}{y} = A \frac{1}{x} + B$	$X = \frac{1}{x} \quad Y = \frac{1}{y}$
$y = \frac{D}{x + C}$	$y = -\frac{1}{C}(xy) + \frac{D}{C}$	$X = xy \quad Y = y$ $C = -\frac{1}{A} \quad D = -\frac{B}{A}$
$y = \frac{L}{1 + Ce^{Ax}}$	$\ln\left(\frac{L}{y} - 1\right) = Ax + \ln C$	$X = x \quad Y = \ln\left(\frac{L}{y} - 1\right)$ $C = e^B \quad L = \text{costante assegnata}$
$y = \frac{1}{B + Ae^{-x}}$	$\frac{1}{y} = Ae^{-x} + B$	$X = e^{-x} \quad Y = \frac{1}{y}$



C.I. - statistica, informazioni
 dei dati sperimentali



Prepariamo alcuni fogli di alluminio, quadrati, di lato L

L=2cm



L=5cm



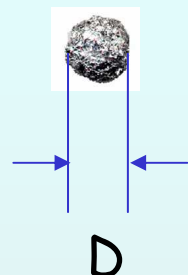
L=10 cm



$$V_f = L^2 \cdot \delta$$

Volume del foglio
di alluminio di lato
L e spessore δ

Da ogni foglio si
fa una pallina di
diametro D



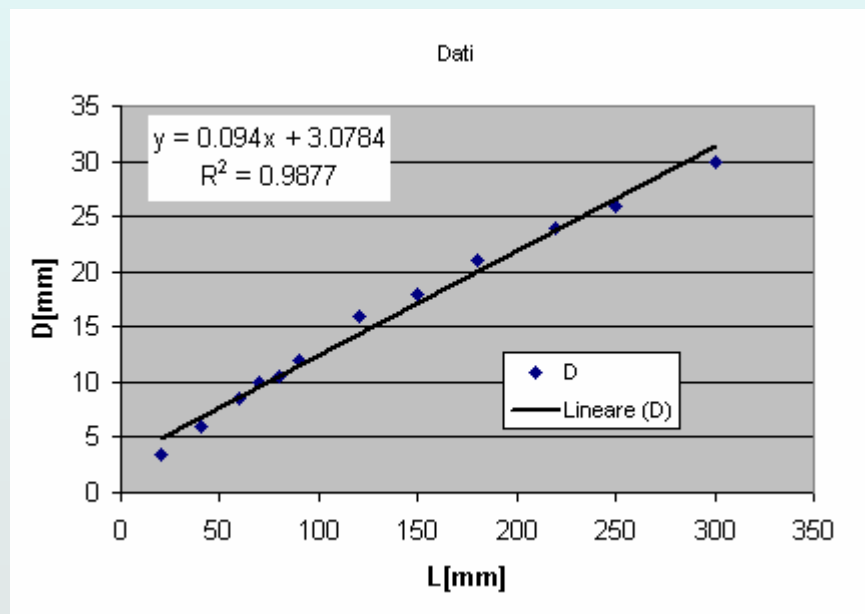
$$V_p = \frac{4}{3}\pi R^3 = \frac{\pi}{6}D^3$$

Volume della
pallina di
diametro D

$$\bar{D} = \left(\frac{6\delta}{\pi} \right)^{1/3} L^m$$

$$Y = \ln(\bar{D}) = \ln \alpha + m \ln L = c + mX$$

Una relazione lineare lega il log.
del diametro medio al log. del lato
dei foglietti



$$\bar{D} = \left(\frac{6\delta}{\pi} \right)^{1/3} L^m$$

$$Y = \ln(\bar{D}) = \ln \alpha + m \ln L = c + mX$$

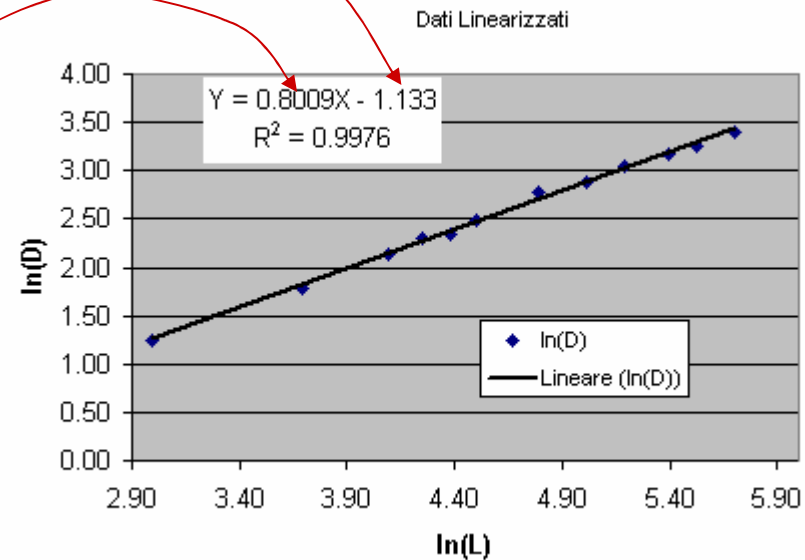
$$c = \ln(\alpha)$$

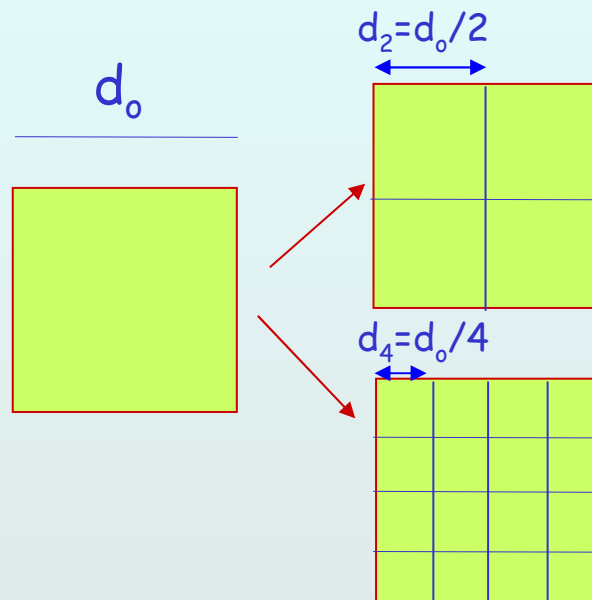
$$\exp(c) = \alpha$$

$$\alpha = \left(\frac{6\delta}{\pi} \right)^{1/3}$$

$$\exp(c) = (6\delta/\pi)^{1/3}$$

$$\delta = \pi (\exp(c))^3 / 6 = 0.0175 \text{ mm}$$





dividendo per 2 l'unità di misura devo utilizzare $N_r = 4$ repliche per costruire il quadrato originale

$$u.m. = d_0/2 \rightarrow N_u = 2, N_r = 4$$

$$D = \frac{\ln 4}{\ln 2} = 2$$

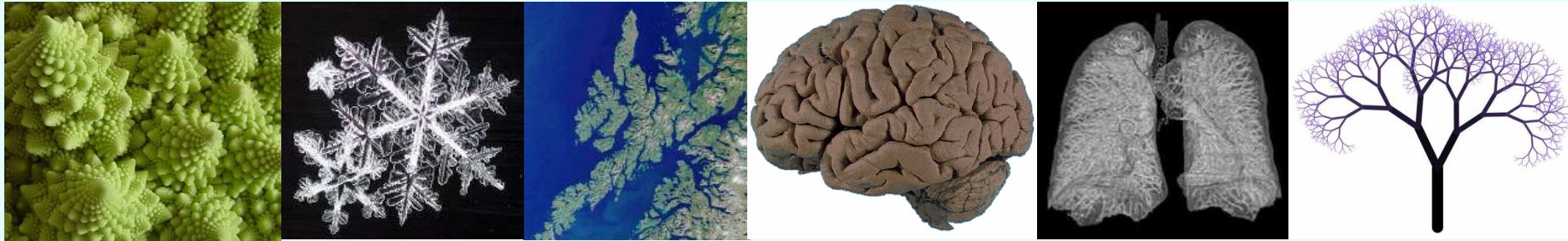
dividendo per 4 l'unità di misura devo utilizzare $N_r = 16$ repliche per costruire il quadrato originale

$$u.m. = d_0/4 \rightarrow N_u = 4, N_r = 16$$

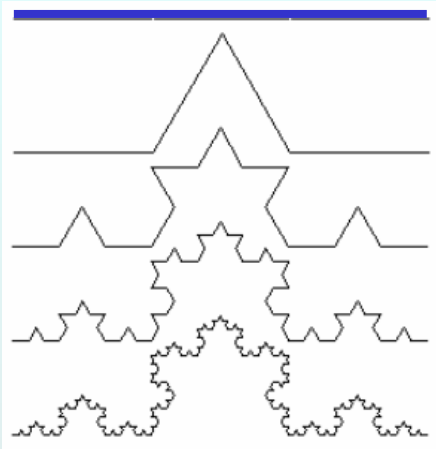
Dimensione di Hausdorff

$$D = \frac{\ln N_r}{\ln N_u} = \log_{N_u} N_r$$

Proprietà dei
logaritmi!



d_0



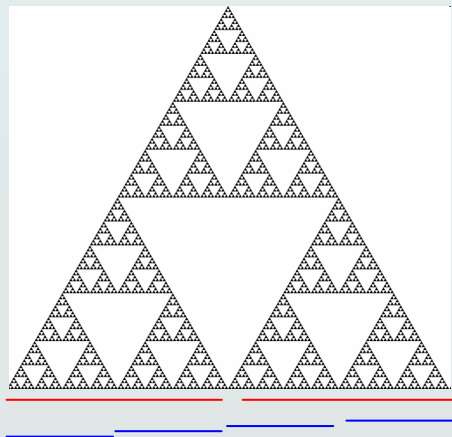
$$D = \frac{\ln N_r}{\ln N_u} = \log_{N_u} N_r$$

$$u_{,m.} = d_0/3 \rightarrow Nu = 3, Nr = 4$$

$$u_{,m.} = d_0/9 \rightarrow Nu = 9, Nr = 12$$

$$u_{,m.} = d_0/27 \rightarrow Nu = 27, Nr = 48$$

$$D = \frac{\ln 4}{\ln 3} = 1.26$$

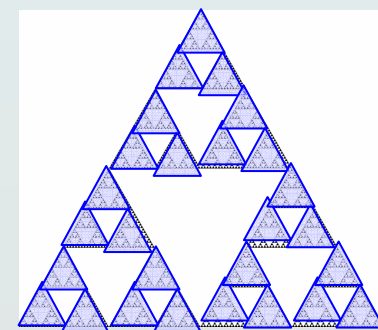
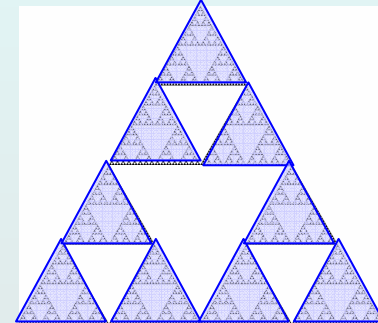
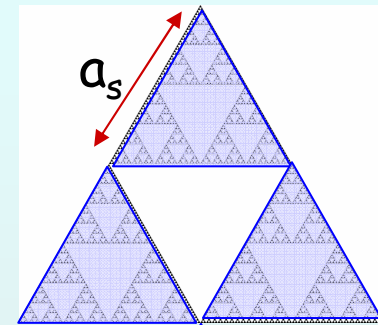
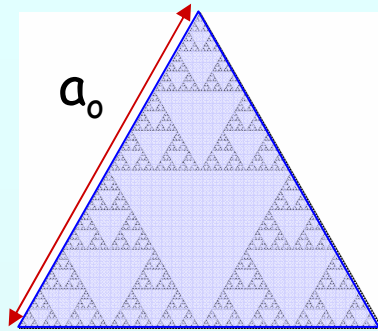


$$u_{,m.} = d_0/2 \rightarrow Nu = 2, Nr = 3$$

$$u_{,m.} = d_0/4 \rightarrow Nu = 4, Nr = 9$$

$$D = \frac{\ln 3}{\ln 2} = 1.58$$

Box counting o ricoprimento



Si sceglie una "forma geometrica" caratterizzata da una dimensione a_0 . Ad esempio un triangolo di lato a_0 , un cerchio di raggio a_0 , una sfera di raggio a_0 , etc... in grado di "ricoprire" il mio oggetto.

Si scelgono sottomultipli dell'unità di misura ($a_s = a_0/s$) e prende una figura, simile all'originale, in scala $1/s$. Si conta quindi il numero di repliche N_s della figura in scala, necessario per ricoprire l'oggetto.

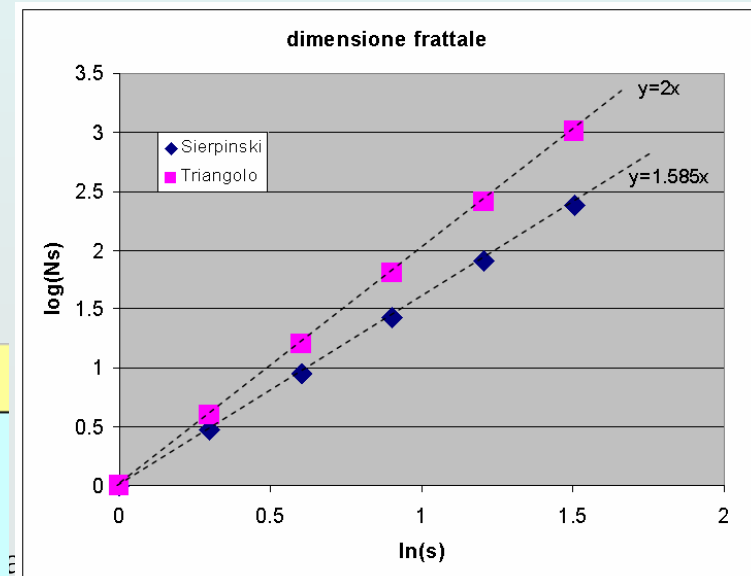
La legge di scala mi dice che si deve avere $N_s \sim s^D$ dove D è la *dimensione* dell'oggetto, quindi:

$$D \sim \ln(N_s) / \ln(s)$$

Riportando su un grafico i valori $X=\ln(s)$, $Y=\ln(N_s)$, il coefficiente angolare della retta è la dimensione frattale dell'oggetto

Sierpinski Triangolo			
s	a_s	N_s	M_s
1	1	1	1
2	0.5	3	4
4	0.25	9	16
8	0.125	27	64
16	0.0625	81	256
32	0.03125	243	1024

Sierpinski Triangolo		
$\ln(s)$	$\ln(N_s)$	$\ln(M_s)$
0	0	0
0.301	0.477	0.602
0.602	0.954	1.204
0.903	1.431	1.806
1.204	1.908	2.408
1.505	2.386	3.010

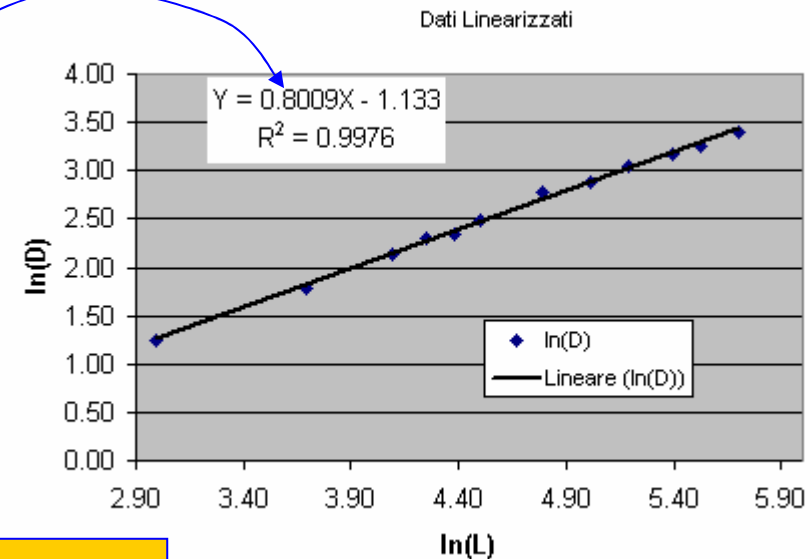


$$\bar{D} = \left(\frac{6\delta}{\pi} \right)^{1/3} L^m$$

$$Y = \ln(\bar{D}) = \ln \alpha + m \ln L = c + mX$$

$$\alpha = \left(\frac{6\delta}{\pi} \right)^{1/3}$$

$$\delta = \pi (\exp(c))^3 / 6 = 0.0175 \text{ mm}$$



$$m = 0.8$$

$$D_{sfere} \approx \alpha L^{0.8}$$

$$Vol_{sfere} \approx \beta (L^{0.8})^3 = \beta L^{2.4}$$

Dim. Frattale della superficie corticale del cervello umano: ~2.4

Risultato: la dimensione frattale delle palline è simile a quella della sup. corticale del cervello.

Nota: patologie e disfunzioni neurologiche possono essere collegate a variazioni della dim. frattale della superficie corticale del cervello (Search google: cortical brain fractal dimension)

Esempi di sistemi biologici a struttura frattale (autosimilare) nei mammiferi

Apparato respiratorio.

- Cornetti nasali.
- Al fondo delle cavità nasali, esistono altre lamine ossee dette turbinati distinte in ecto ed endo turbinati.
- Anelli tracheali.
- Albero bronchiale.

Apparato digerente.

- Microtubuli scavati nella dentina.
- Papille linguali (fungiformi, filiformi, coniche, lenticolari, vallate)
- Creste trasversali del palato duro
- Pliche della mucosa nella superficie interna del ruminale
- Nell'abomaso (stomaco ghiandolare dei ruminanti),
- Villi intestinali e microvilli intestinali
- Muscolatura liscia dell'intestino tenue
- Pliche semilunari dell'intestino crasso
- Fegato. La struttura microscopica epatica

Sistema circolatorio.

- Muscolatura cardiaca.
- La conformazione spaziale arboriforme del sistema circolatorio
- Globuli rossi.

Pelle ed annessi.

- Tessuto podovilloso e podofiloso dello zoccolo equino.

Muscoli ed ossa.

- Tessuto muscolare scheletrico.

Ghiandole endocrine ed esocrine.

Sistema nervoso centrale.

Embriogenesi.

Dimensione frattale della sup. corticale e volume del cervello in soggetti patologici e di controllo sani.

Gruppi	controlli sani	OCD	Schizofrenici
Dimensione frattale			
Volume cranico	2.418	2.392	2.367
emisfero sinistro	2.413	2.383	2.354
emisfero destro	2.423	2.401	2.379

