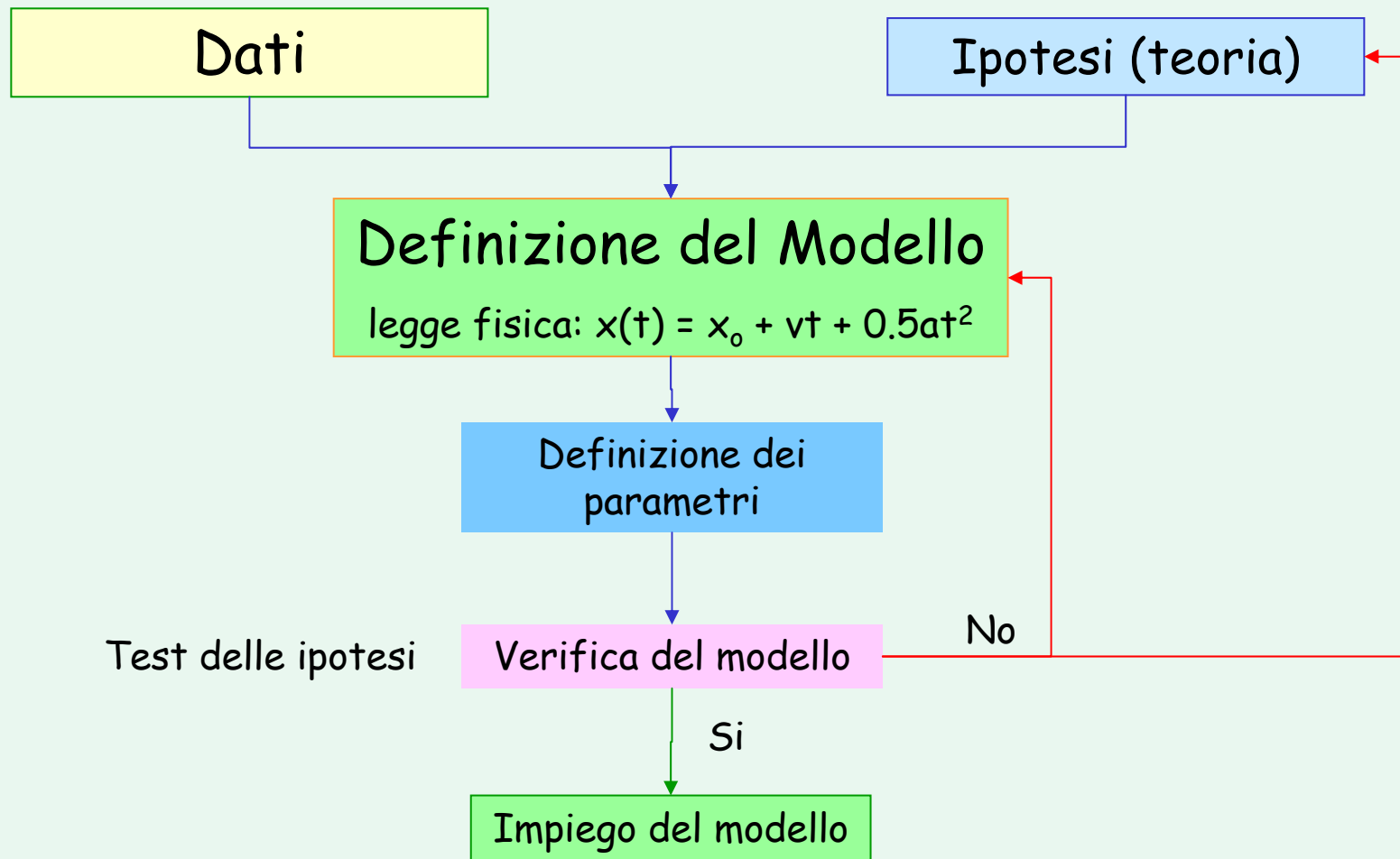


Ulteriori Conoscenze di Informatica e Statistica

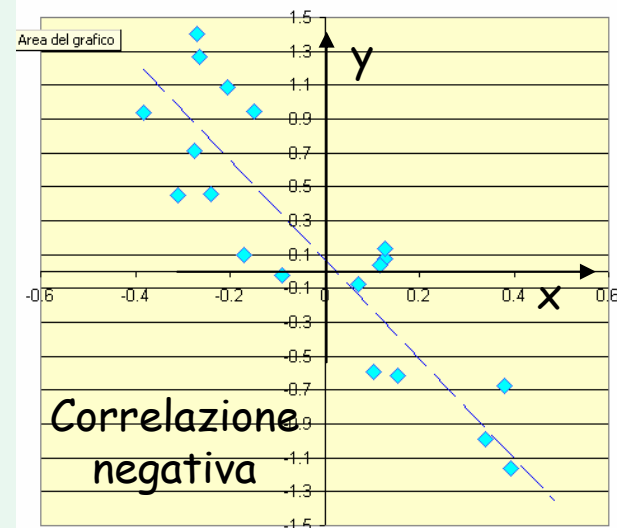
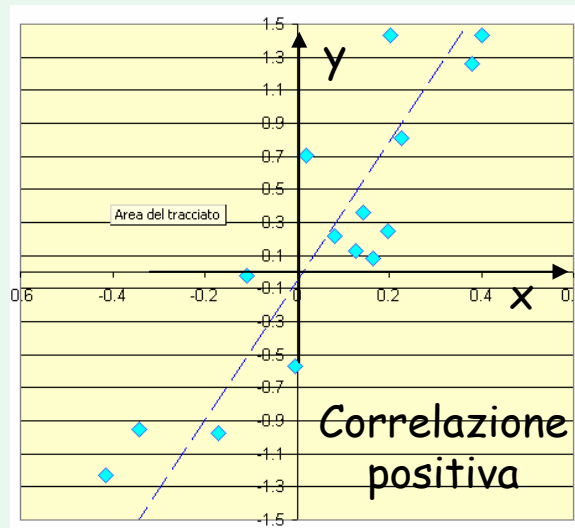
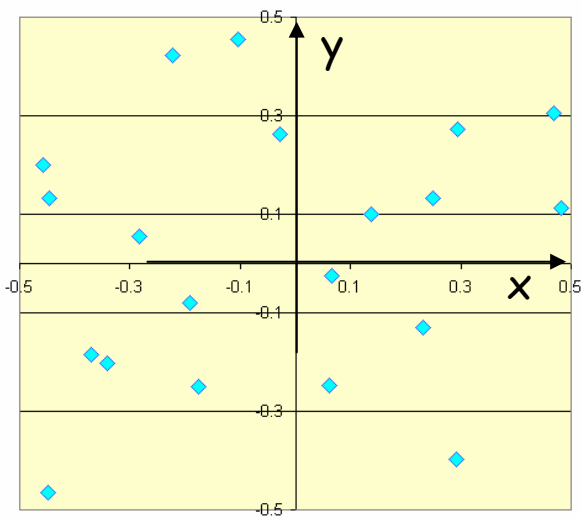
Carlo Meneghini

Dip. di fisica - via della Vasca Navale 84,
st. 83 (I piano) tel.: 06 55 17 72 17

meneghini@fis.uniroma3.it



Regressione (raffinamento, fitting, etc...): relazione funzionale tra variabili ottenuta con metodi statistici. Il metodo della regressione consente di derivare una relazione statistica tra una variabile dipendente (Y) e una o più variabili esplicative ($X_1, X_2...X_n$). Origine: Galton 1886 - regression toward mediocrity



Ipotesi: esiste una relazione lineare tra i valori della x e i valori della y

$$y_i = \beta_o + \beta_1 x_i + \epsilon_i$$

Errori sulle osservazioni

$$\sigma_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \quad \text{Varianza } x$$

$$\sigma_y^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 \quad \text{Varianza } y$$

$$\sigma_{xy}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \quad \text{covarianza } xy$$

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y} \quad \text{coefficiente di correlazione di Pearson}$$

Retta di regressione:

$$Y = \beta_o + \beta_1 X$$

$$\beta_1 = \frac{\sigma_{xy}}{\sigma_x^2} \quad \text{coeff. angolare}$$

$$\beta_o = \bar{y} - \beta_1 \bar{x} \quad \text{intercetta}$$

Varianza degli
errori

$$\sigma_{y_i}^2 = \frac{1}{n} \sum (y_i - (\beta_0 + \beta_1 x_i))^2 = \frac{1}{n} \sum (\epsilon_i)^2$$

$$\sigma_{\beta_0}^2 = \frac{\sigma_{y_i}^2}{n} \left(1 + \frac{n\bar{x}^2}{\sum (x_i - \bar{x})^2} \right)$$

$$\sigma_{\beta_1}^2 = \frac{\sigma_{y_i}^2}{\sum (x_i - \bar{x})^2}$$

Nota:

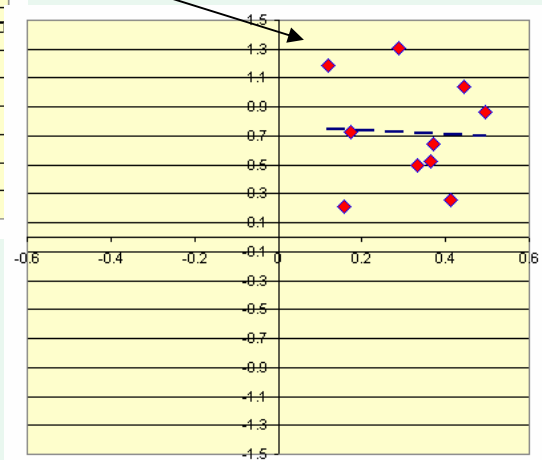
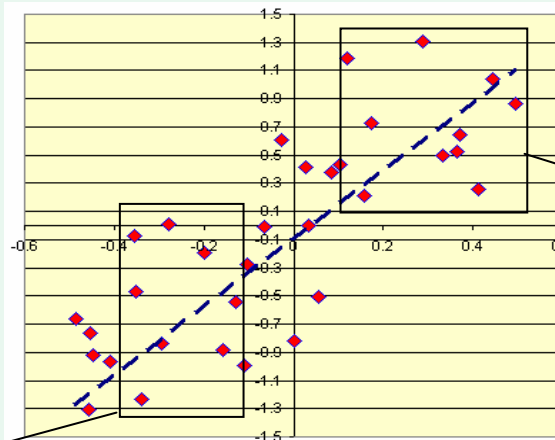
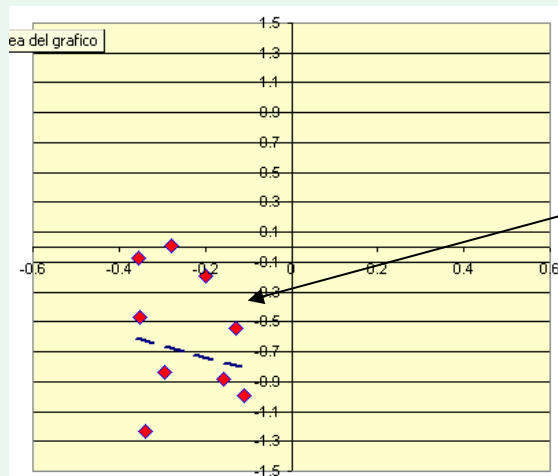
$$\text{Corr}(\beta_0, \beta_1) = - \frac{\bar{x}\sqrt{n}}{\sqrt{\sum (x_i)^2}}$$

Utilizzando

$$x' = x_i - \bar{x}$$

si annulla la
correlazione
e tra β_0 e β_1

L'incertezza sulle stime dei parametri della
regressione decresce aumentando la
varianza campionaria di x



Stime campionarie

$$s_x^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$$

$$s_y^2 = \frac{1}{n-1} \sum (y_i - \bar{y})^2$$

Nell'ipotesi che la varianza sia la stessa per tutti i valori osservati y_i , la stima campionaria della varianza sugli errori è

$$s_{y_i}^2 = \frac{1}{n-2} \sum (y_i - \hat{y}_i)^2 = s^2 \quad \text{con} \quad \hat{y}_i = \beta_0 + \beta_1 x_i$$

S: errore standard della regressione:

$$s = \frac{n}{n-2} s_y^2 (1 - (r_{xy})^2)$$

Errori standard sui parametri

$$es(\beta_0) = \frac{s}{\sqrt{n}} \left(1 + \frac{n\bar{x}^2}{\sum (x_i - \bar{x})^2} \right)^{\frac{1}{2}}$$

$$es(\beta_1) = \frac{s}{s_x \sqrt{n}}$$

t-test

$$t_v = \beta_i / es(\beta_i)$$

$$v = n-2$$

L'intervalli di confidenza

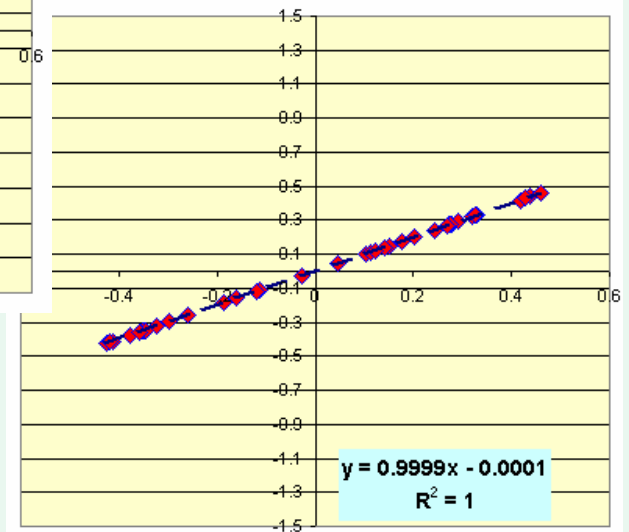
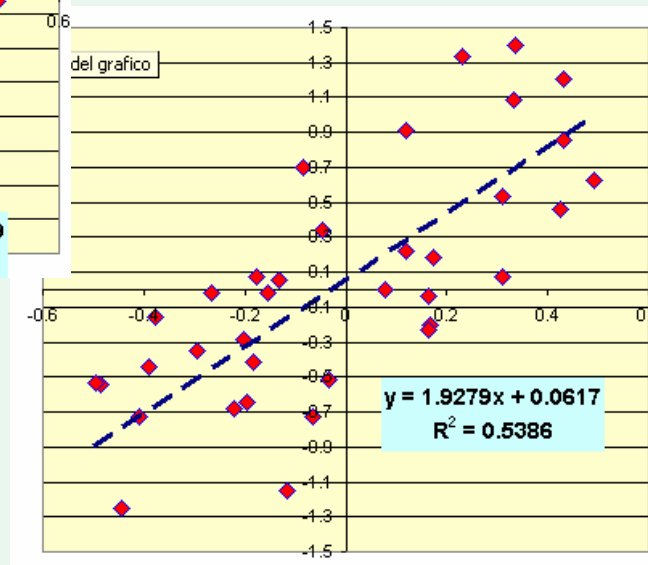
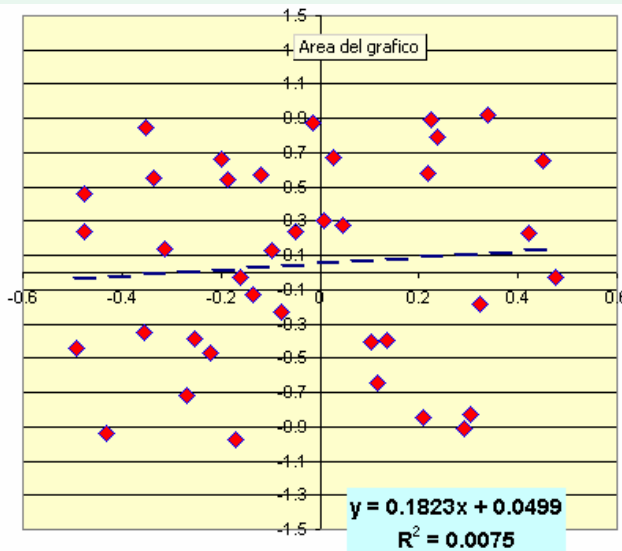
$$\beta_0 - es(\beta_0) t_{(\alpha/2; n-2)} \leq \hat{\beta}_0 \leq \beta_0 + es(\beta_0) t_{(\alpha/2; n-2)}$$

$$\beta_1 - es(\beta_1) t_{(\alpha/2; n-2)} \leq \hat{\beta}_1 \leq \beta_1 + es(\beta_1) t_{(\alpha/2; n-2)}$$

Indice (o coefficiente) di determinazione multipla R^2

è un indice della "bontà della retta di regressione nello spiegare la variabilità di Y mediante X

$$R^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$$



$$R^2 = (r_{xy})^2$$

nel caso di regressione lineare semplice R^2 è eguale al quadrato del coefficiente di correlazione di Pearson

$$\sum_i (y_i^{exp} - \bar{y})^2 = \sum_i (y_i^{th} - \bar{y})^2 + \sum_i (y_i^{exp} - y_i^{th})^2$$

$$SST = SSR + SSE$$

SST= somma dei quadrati della differenza tra dati e valor medio

SSR= somma dei quadrati degli scarti tra punti teorici e valor medio (sulla retta di regressione)

SSE= somma dei quadrati degli scarti tra valori sperimentali e teorici

$$R^2 = \frac{SSR}{SST} = \frac{\sum_i (y_i^{th} - \bar{y})^2}{\sum_i (y_i^{exp} - \bar{y})^2}$$

R^2 indica la frazione della varianza dei dati che si può spiegare mediante la curva di regressione

$$r = \frac{C_{xy}}{s_x s_y}$$

$$C_{xy} = \frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y})$$

Covarianza

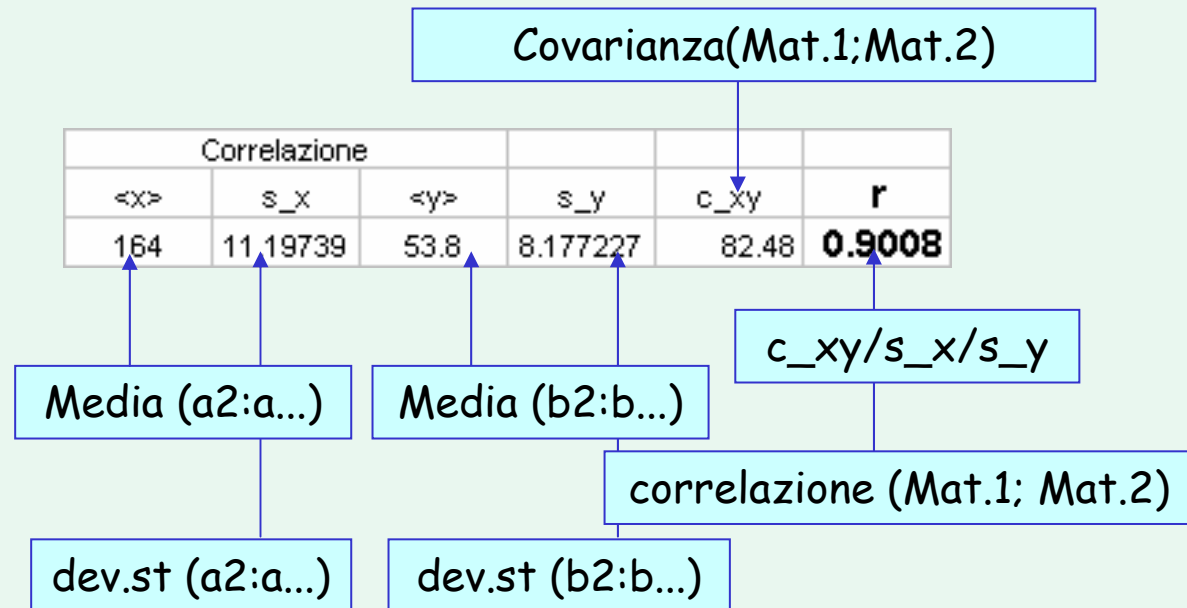
$$s_x = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2}$$

Deviazione
standard

calcolare:

$\bar{x}$, $\bar{y}$, s_x , s_y , C_{xy}

	A	B
1	Altezza	Peso
2	144	38.2
3	146	41.0
4	151	40.6
5	149	50.3
6	153	51.2
7	157	51.0
8	157	49.4
9	162	45.0
10	161	52.8
11	167	57.2
12	164	59.0
13	169	57.3
14	172	60.2
15	171	64.2



Quanto è significativa la correlazione?

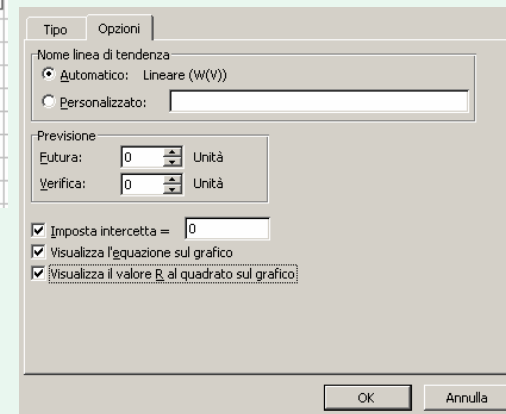
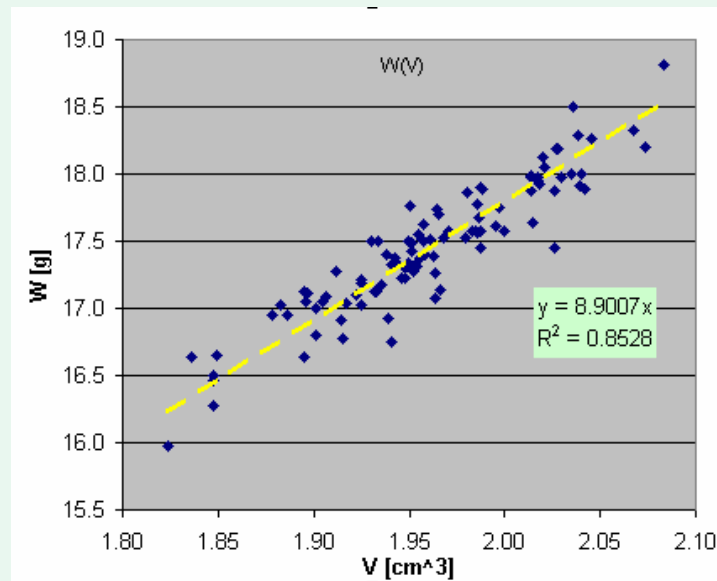
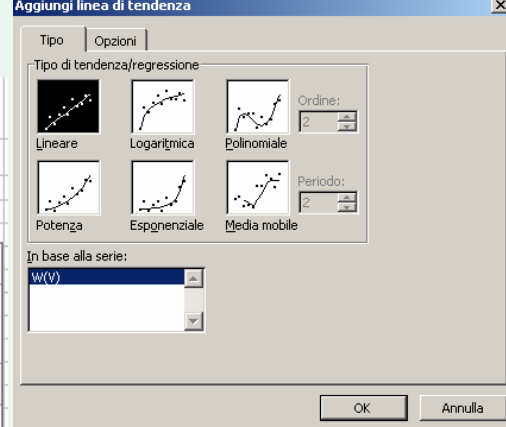
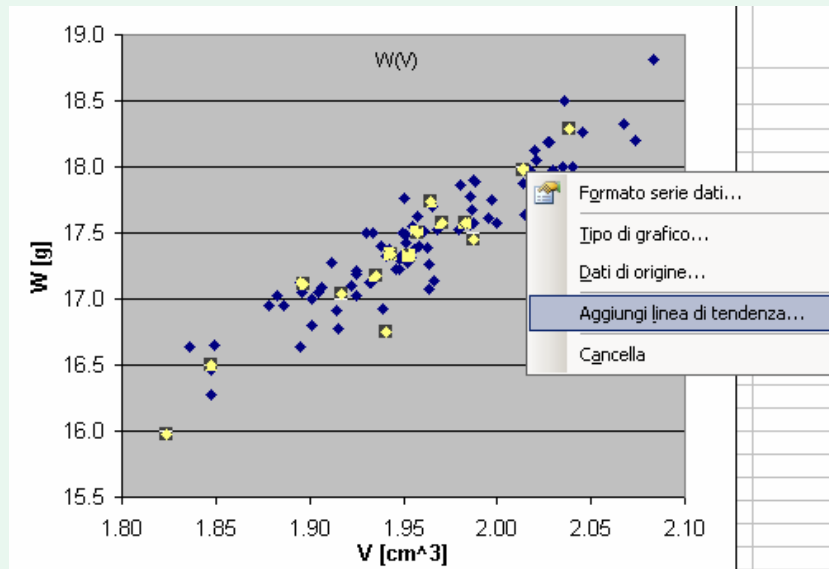
T-test sulla correlazione

$$t_{\nu} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

Intervallo di confidenza per la correlazione

$$P = \rho V$$

Peso	Volume
g	cm ³
16.75	1.941
17.37	1.943
17.63	1.957
17.08	1.964
17.99	2.014
16.95	1.879
17	1.901
17.5	1.958
17.92	2.018
17.57	1.988
18	2.041
17.28	1.911
17.55	1.955
17.14	1.966
17.32	1.953
17.21	1.925
17.42	1.951
17.64	2.014
17.31	1.954
17.98	2.030



$$R^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$$

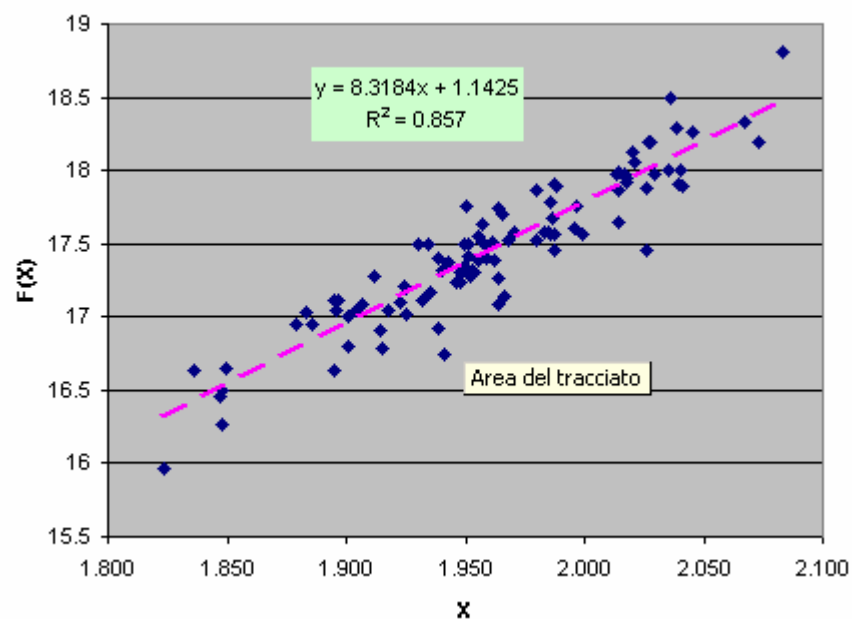
Indice di
regressione
multipla

F(X)	X		
17.63	1.957	COVARIANZA	0.02520
17.08	1.964	ERR.ST.XY	0.18727
16.75	1.941	CORRELAZIONE	0.92970
17.37	1.943	PEARSON	0.92970
17.99	2.014	INTERCETTA	1.14246
16.95	1.879	PENDENZA	8.31838
17	1.901		

$$C_{xy} = \frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y})$$

$$s = \frac{n}{n-2} s_y^2 (1 - (r_{xy})^2)$$

$$r = \frac{C_{xy}}{s_x s_y}$$

 β_0
 β_1


Analisi dati

Strumenti di analisi

Covarianza
 Statistica descrittiva
 Smorzamento esponenziale
 Test F a due campioni per varianze
 Analisi di Fourier
 Istogramma
 Media mobile
 Generazione di un numero casuale
 Rango e percentile
 Regressione

Regressione

Input

Intervallo di input Y:

Intervallo di input X:

☐ Etichette

☐ Passa per l'origine

☐ Livello di confidenza: %

Opzioni di output

☒ Intervallo di output:

☐ Nuovo foglio di lavoro:

☐ Nuova cartella di lavoro

Residui

☒ Residui

☒ Residui standardizzati

☒ Tracciati dei residui

☒ Tracciati delle approssimazioni

Probabilità normale

☒ Tracciati delle probabilità normali

OK

Annulla

?

F(X)	X
17.63	1.957
17.08	1.964
16.75	1.941
17.37	1.943
17.99	2.014
16.95	1.879
17	1.901
17.5	1.958
17.92	2.018
17.57	1.988
18	2.041

F(X)	X
17.63	1.957
17.08	1.964
16.75	1.941
17.37	1.943
17.99	2.014
16.95	1.879
17	1.901

OUTPUT RIEPILOGO	
Statistica della regressione	
R multiplo	0.926
R al quadrato	0.857
R al quadrato corr.	0.856
Errore standard	0.188
Osservazioni	100

ANALISI VARIANZA	
	gdl
Regressione	1
Residuo	98
Totale	99

	Coefficienti	Errore standard	Stat t	Valore di significatività	Inf. 95%	Sup. 95%	Inf. 90.0%	Sup. 90.0%
Intercetta β_0	1.1425	0.6733	1.6969	0.0929	-0.193612165	2.47852	0.02447	2.26044
Variabile X 1 β_1	8.3184	0.3433	24.2320	3.58E-43	7.637145145	8.99961	7.74834	8.88841

COVARIANZA	0.02520
ERR.ST.XY	0.18727
CORRELAZIONE	0.92970
PEARSON	0.92970
INTERCETTA	1.14246
PENDENZA	8.31838

Dati

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

ANOVA

$$\hat{y}_i = \beta_0 + \beta_1 x_i$$

$$\sigma_y^2 = \frac{1}{N} \sum_i (\hat{y}_i - \bar{\hat{y}})^2 = \bar{\hat{y}}^2 - \bar{\hat{y}}^2$$

$$\bar{\hat{y}}^2 = \frac{1}{N} \sum_i (\beta_0 + x_i \beta_1)^2 = \frac{1}{N} \sum_i (\beta_0^2 + x_i \beta_1^2 + 2x_i \beta_0 \beta_1) = \beta_0^2 + \beta_1^2 \bar{x^2} + 2\beta_0 \beta_1 \bar{x}$$

$$\bar{\hat{y}}^2 = \left(\frac{1}{N} \sum_i (\beta_0 + x_i \beta_1) \right)^2 = (\beta_0 + \bar{x} \beta_1)^2 = \beta_0^2 + \bar{x}^2 \beta_1^2 + 2\beta_0 \beta_1 \bar{x}$$

$$\sigma_y^2 = \beta_1^2 (\bar{x^2} - \bar{x}^2) = \beta_1^2 \sigma_x^2$$

Varianza dele medie

$$N \sigma_{\hat{y}}^2 = \beta_1^2 N \sigma_x^2$$

Varianza tra misure

$$F(1, \nu) = \frac{N \sigma_{\hat{y}}^2}{s_{xy}^2}$$

$$R_{corr}^2 = 1 - \frac{N-1}{N-p-1} (1 - R^2)$$

$$N \sigma_{\hat{y}}^2 = \beta_1^2 N \sigma_x^2 = dev.Q(X) * \beta_1^2$$

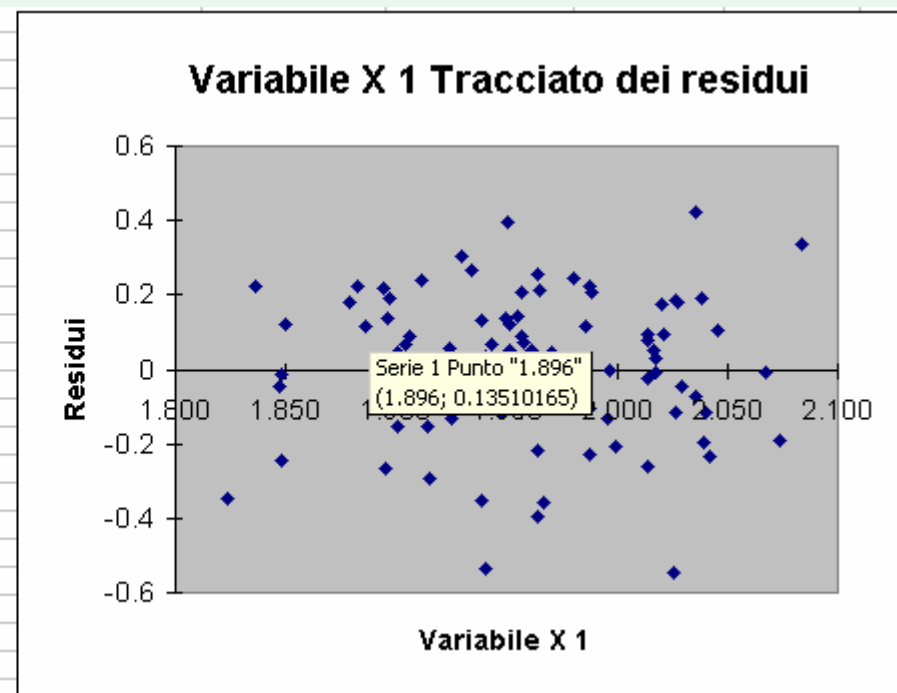
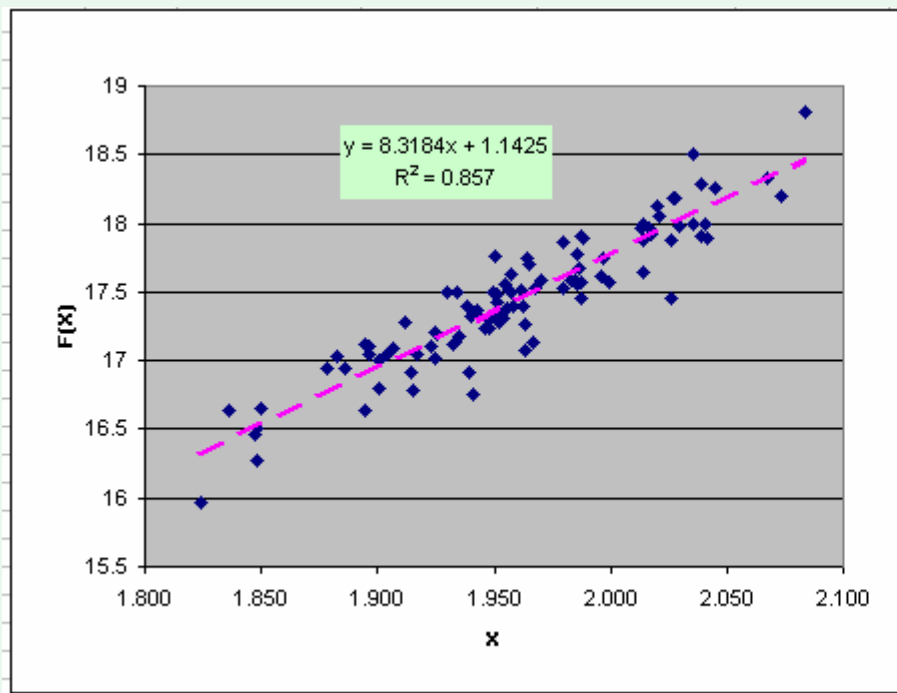
$$N \sigma_{\hat{y}}^2 = \sum_i (\hat{y}_i - \bar{\hat{y}})^2$$

$$(Err.std.xy)^2$$

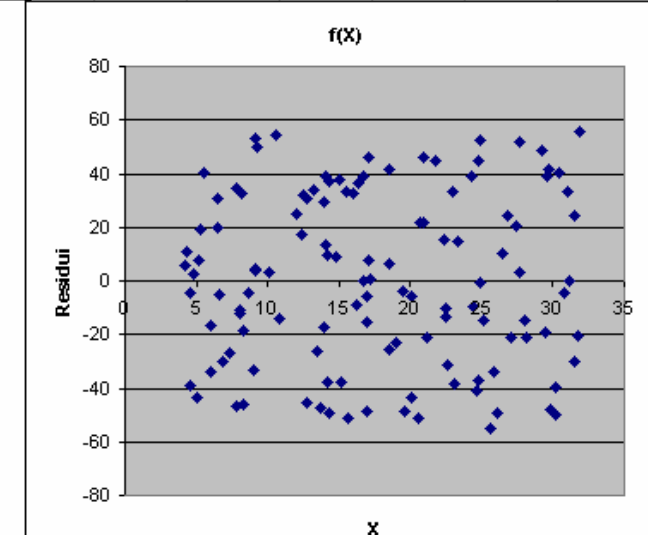
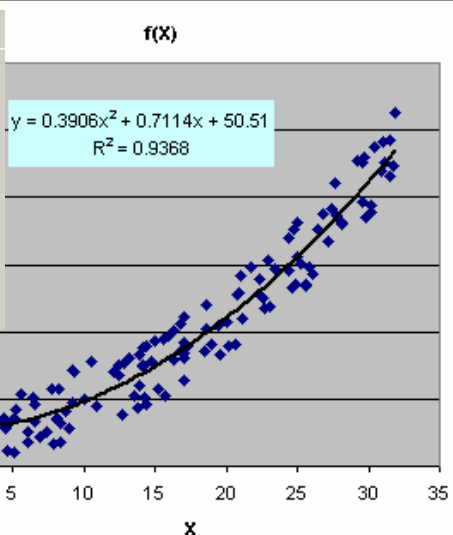
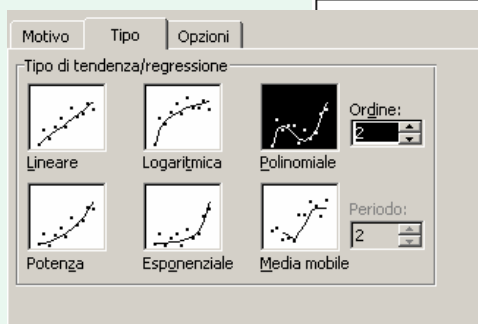
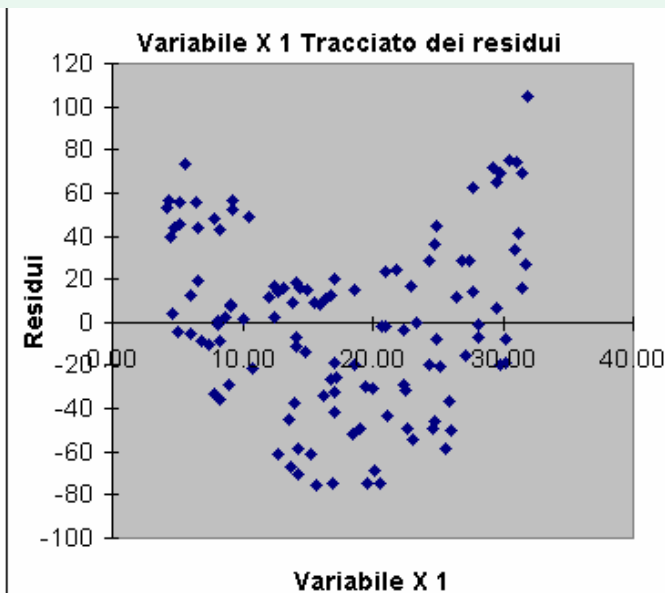
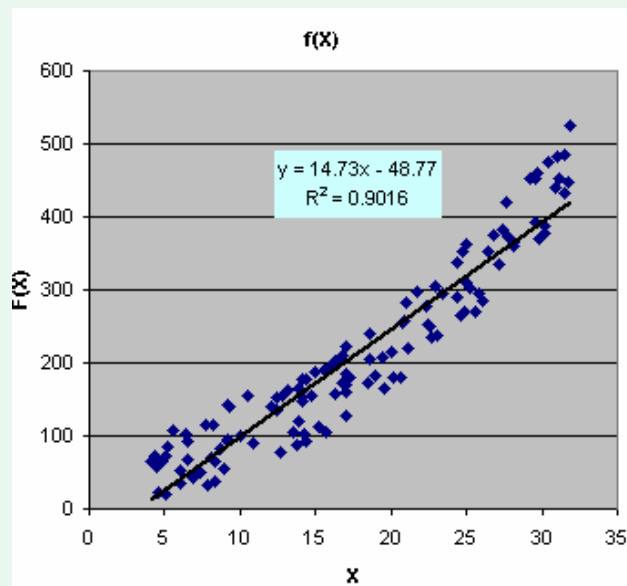
$$F(1, \nu) = \frac{N \sigma_{\hat{y}}^2}{s_{xy}^2}$$

$$s_{xy}^2$$

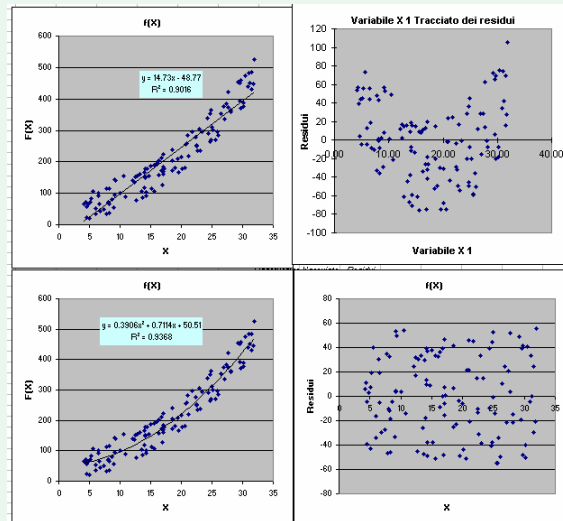
F(X)	X
17.63	1.957
17.08	1.964
16.75	1.941
17.37	1.943
17.99	2.014
16.95	1.879
17	1.901
17.5	1.958
17.92	2.018
17.57	1.988
18	2.041



X_1	f(X)	FIT	RESIDUI
15.57	189.71	156.22	33.49
15.20	113.73	151.53	-37.79
4.79	65.54	62.87	2.67
20.73	254.69	233.15	21.54
14.24	102.32	139.89	-37.57
21.16	219.46	240.49	-21.03
25.57	269.17	324.01	-54.84
16.33	202.83	166.30	36.53
31.77	446.51	467.29	-20.78
9.29	140.78	90.84	49.95
30.14	376.63	426.84	-50.20
18.55	239.37	198.07	41.30
12.78	153.99	123.36	30.63
24.81	271.14	308.57	-37.43
7.38	49.90	77.02	-27.12
9.18	143.26	89.98	53.29
12.70	77.03	122.57	-45.54
24.93	363.51	310.93	52.58
9.15	93.65	89.70	3.95
20.10	178.99	222.63	-43.64



Quale modello è migliore?



		sigma^err.		1026.03			
		media (res)^2	1570.58		media (res)^2	1026.03	
		v	117.00	chi^2	187.66	v	116.00
			prob.Chi	3.69E-05			prob.Chi
		SSQ	186899.16		SSQ	122097.60	
Dati		Lineare				Quadratico	
X_1	f(X)	fit	residui	res.^2		FIT	RESIDUI
15.57	189.71	180.5158	9.19	84.54	0.08	156.22	33.49
15.20	113.73	175.0782	-61.35	3763.26	3.67	151.53	-37.79
4.79	65.54	21.74268	43.80	1918.20	1.87	62.87	2.67
20.73	254.69	256.6171	-1.93	3.72	0.00	233.15	21.54
14.24	102.32	161.0436	58.73	3448.68	3.36	139.89	37.67

$$F(n_{pc} - n_{ps}, N - n_{pc} - 1) = \left(\frac{SSQ_s - SSQ_c}{n_{pc} - n_{ps}} \right) \cdot \left(\frac{SSQ_c}{N - n_{pc} - 1} \right)^{-1}$$

F = 61.56535 2.33287E-12

$$F(n_{pc} - n_{ps}, N - n_{pc} - 1) = \left(\frac{\chi_s^2 - \chi_c^2}{n_{pc} - n_{ps}} \right) \cdot \left(\frac{\chi_c^2}{N - n_{pc} - 1} \right)^{-1}$$

AIC (Akaike information criteria)

$$AIC = N \log \left(\frac{SSQ_{res}}{N} \right) + 2n_p$$

N = numero di dati
SSQ_{res} = somma dei quadrati dei residui
n_p = numero di parametri

Akaike, Hirotugu (1974). "A new look at the statistical model identification". *IEEE Transactions on Automatic Control* **19** (6): 716-723.

Criterio: si preferisce il modello con il valore minore di AIC

Correzioni per piccoli N

McQuarrie definition

$$AIC_u = \log \left(\frac{SSQ_{res}}{N - n_p} \right) + \frac{N + n_p}{N - n_p - 2}$$

McQuarrie, A. D. R., and Tsai, C.-L., 1998. *Regression and Time Series Model Selection*. World Scientific.

Burnham definition

$$AIC_c = AIC + \frac{2n_p(n_p + 1)}{N - n_p - 1}$$

Burnham, K. P., and D. R. Anderson, 2002. *Model Selection and Multimodel Inference: A Practical-Theoretic Approach*, 2nd ed. Springer-Verlag.

Nota: AIC è un criterio, non un test, non fornisce una valutazione del "rischio" statistico associato ad una scelta piuttosto che ad un'altra.

Fit (regressione) non lineare: χ^2

Ho una serie di osservazioni sperimentali x_i, y_i

Voglio cercare di riprodurre i dati sperimentali mediante una curva teorica:

$$f^{th}(X, p_1, p_2, \dots, p_m)$$

funzione degli x_i e di m parametri p_m i cui valori devono essere determinati.
Cerco il valori dei parametri come quelli che minimizzano la funzione:

$$\chi^2_\nu = \sum_{i=1}^N \frac{(f_i^{th} - y_i)^2}{\sigma_i^2}$$

Se la funzione $f(x,p)$ scelta è corretta si ha: $y_i = f(x_i, p_1, p_2, \dots, p_m) + \epsilon_i$

quindi:

$$\chi^2_\nu = \sum_{i=1}^N \frac{(\epsilon_i)^2}{\sigma_i^2} \sim \nu$$

Gradi di libertà: $\nu = m - p - 1$