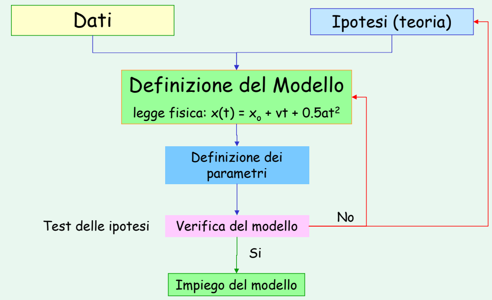
Ulteriori Conoscenze di Informatica e Statistica

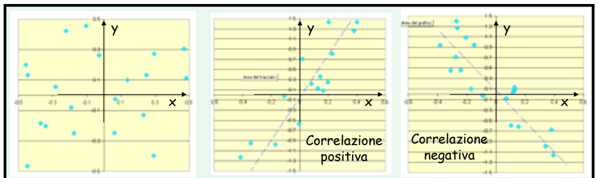
Carlo Meneghini

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

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

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

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

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

Retta di regressione:

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

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

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

Varianza degli errori

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

$$\sigma_{\beta_o}^2 = \frac{\sigma_{y_i}^2}{n}\left(1 + \frac{n\bar{x}^2}{\sum (x_i - \bar{x})^2}\right) \qquad \sigma_{\beta_1}^2 = \frac{\sigma_{y_i}^2}{\sum (x_i - \bar{x})^2}$$

Nota:

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

Utilizzando

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

si annulla la correlazione tra β_o e β_1

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

Stime campionarie

$$s_x^2 = \frac{1}{n-1}\sum (x_i - \bar{x}_i)^2 \qquad s_y^2 = \frac{1}{n-1}\sum (y_i - \bar{y}_i)^2$$

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

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

s: errore standard della regressione: $s = \frac{n}{n-2} s_y^2 (1 - (r_{xy})^2)$

Errori standard sui parametri

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

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

t-test

$t_\nu = \beta_i / es(\beta_i)$

$\nu = n-2$

L'intervalli di confidenza

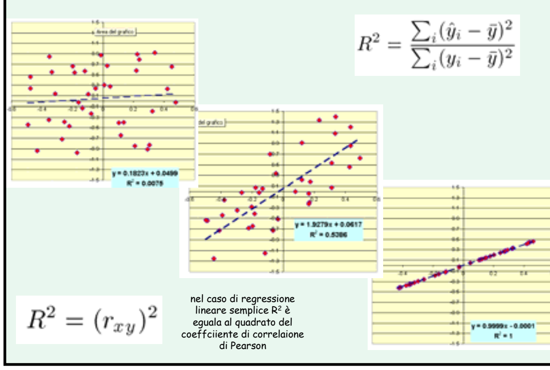
$$\beta_o - es(\beta_o)t_{(\alpha/2;n-2)} \le \hat{\beta}_o \le \beta_o + es(\beta_o)t_{(\alpha/2;n-2)}$$

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

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 eguala al quadrato del coeffciiente di correlaione 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 puó spiegare mediante la curva di regressione

---

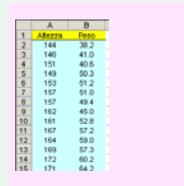
$$r = \frac{C_{xy}}{s_x s_y}$$

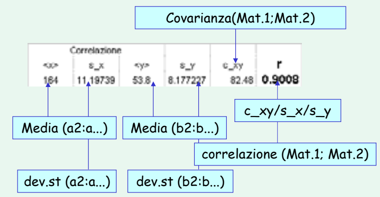
$$C_{xy} = \frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y})$$ Covarianza

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

calcolare:

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





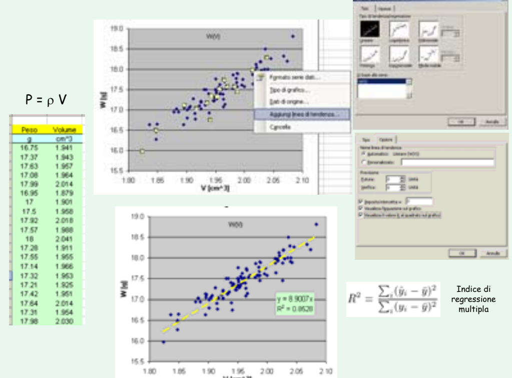
Quanto è significativa la correlazione?

T-test sulla correlazione

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

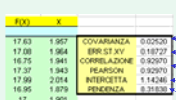
Intervallo di confidenza per la correlazione

P = ρ V



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

---



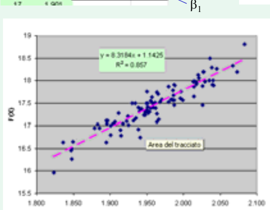
$$C_{xy} = \frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y})$$

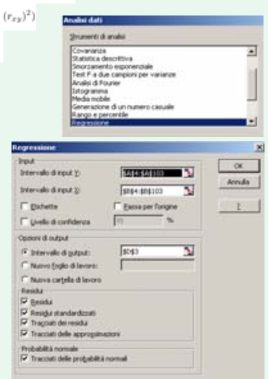
$$s = \frac{n}{n-2} s_y^2 (1 - (r_{xy})^2)$$

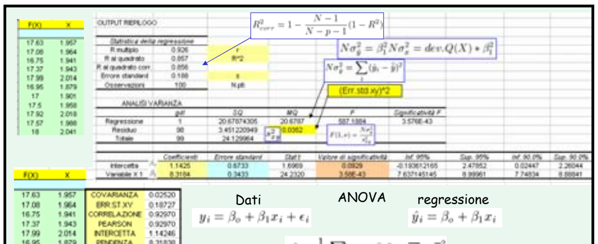
$$r = \frac{C_{xy}}{s_x s_y}$$

β_o

β_1







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

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

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

Dati

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

ANOVA

regressione

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

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

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

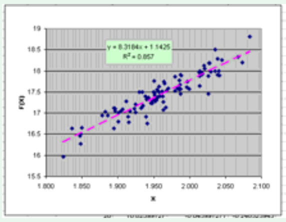
$$\bar{\hat{y}}^2 = \left(\frac{1}{N} \sum_i (\beta_o + x_i \beta_1)\right)^2 = (\beta_o + \bar{x}\beta_1)^2 = \beta_o^2 + \bar{x}^2 \beta_1^2 + 2\beta_o \beta_1 \bar{x}$$

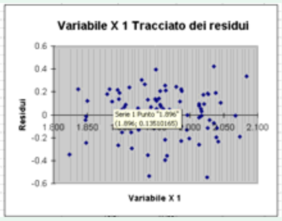
$$\sigma_{\hat{y}}^2 = \beta_1^2 (\overline{x^2} - \bar{x}^2) = \beta_1^2 \sigma_x^2$$ Varianza dele medie

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

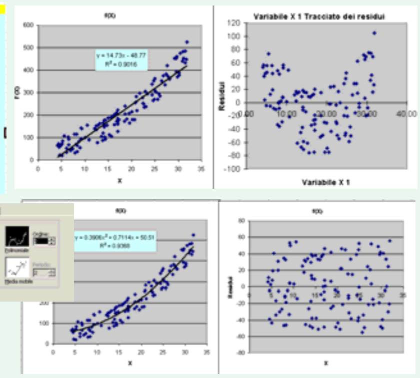
$$F(1, \nu) = \frac{N\sigma_{\hat{y}}^2}{s_{xy}^2}$$

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

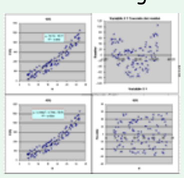




X_1	f(X)	FIT	RESIDUI
15.57	189.71	156.22	33.49
15.20	113.73	151.53	-37.79
4.79	65.54	62.87	2.67
20.73	254.69	233.15	21.54
14.24	102.32	139.89	-37.57
21.16	219.46	240.49	-21.03
25.57	269.17	324.01	-54.84
16.33	202.83	166.30	36.53
31.77	446.51	467.29	-20.78
9.29	140.78	90.84	49.95
30.14	376.63	426.84	-50.20
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	χ2	187.66	v	116.00	χ2	120.60
				prob.Chi	3.69E-05			prob.Chi	0.366
			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	1121.29	1.09
15.20	113.73	175.0782	-61.35	3763.26	3.67	151.53	-37.79	1428.38	1.39
4.79	65.54	21.74268	43.80	1918.20	1.87	62.87	2.67	7.15	0.01
20.73	254.69	256.6171	-1.93	3.72	0.00	233.15	21.54	463.84	0.45

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

F = 61.56535 2.33267E-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, Hirotsugu (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:

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

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:

$$\chi_\nu^2 = \sum_{i=1}^{N} \frac{(f^{th} - y_i)^2}{\sigma_i^2}$$

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

quindi:

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

Gradi di libertà: ν = m - p -1