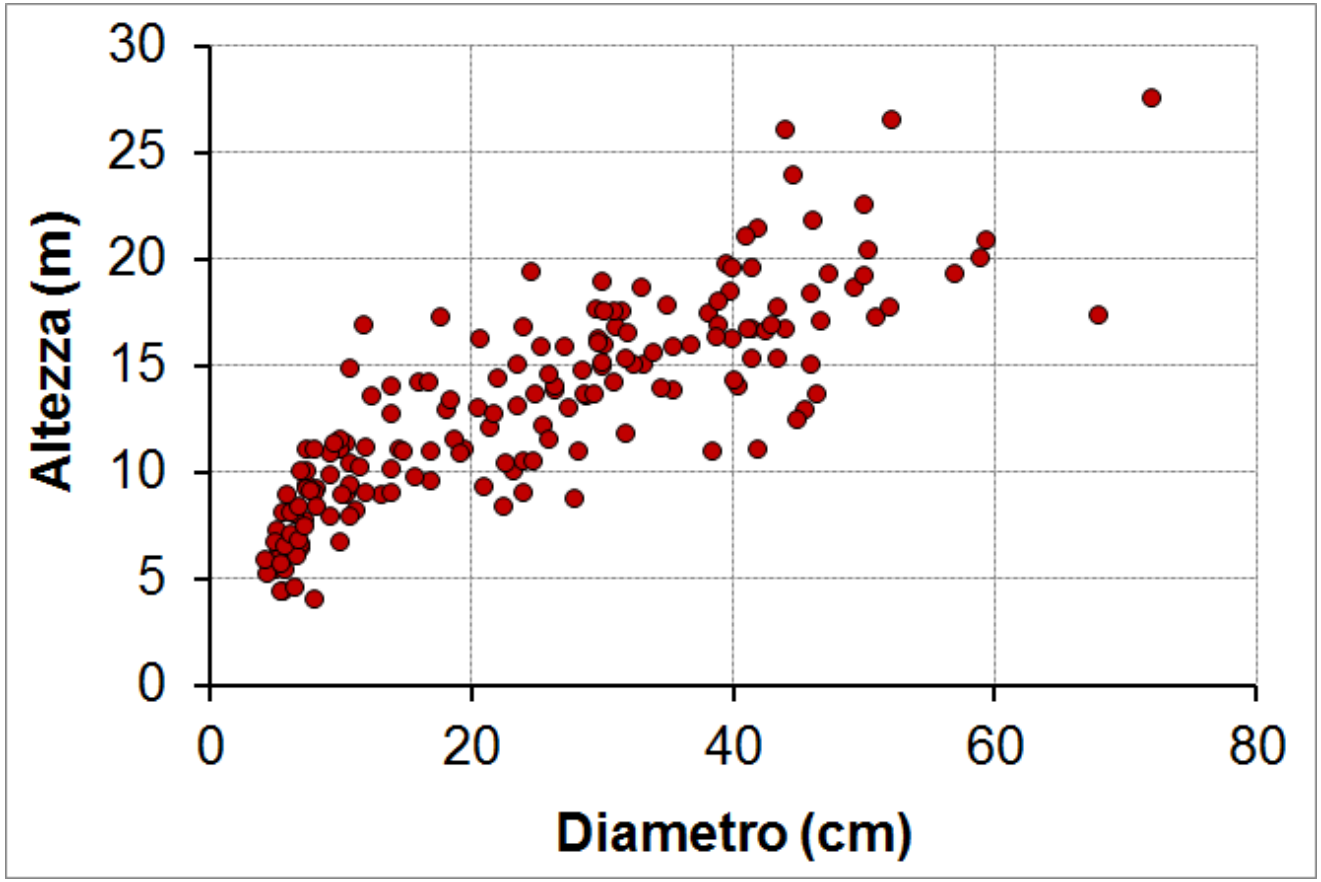
Correlazione

Correlazione

Misure diametri e altezze dei lecci nel Parco Gussone (Prof. A. Saracino). Non definisce un rapporto causa-effetto ma misura il grado di associazione tra due variabili indipendenti senza ottenere un rapporto funzionale



Calcolo della correlazione di Pearson

Definizione della covarianza che è un indice statistico che misura la concordanza o la discordanza tra due variabili quantitative x e y

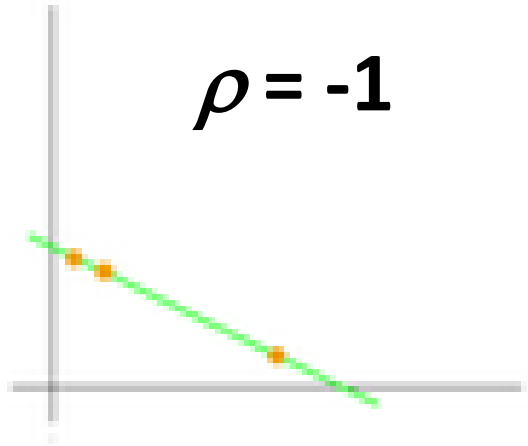
$$Covarianza(x,y) = \frac{1}{N-1}\sum_{i=1}^{N}[(x_i - \bar{x})(y_i - \bar{y})]$$

La covarianza ha il difetto di dipendere dall'unità di misura delle variabili x e y, quindi normalizziamo la covarianza trovando la correlazione…

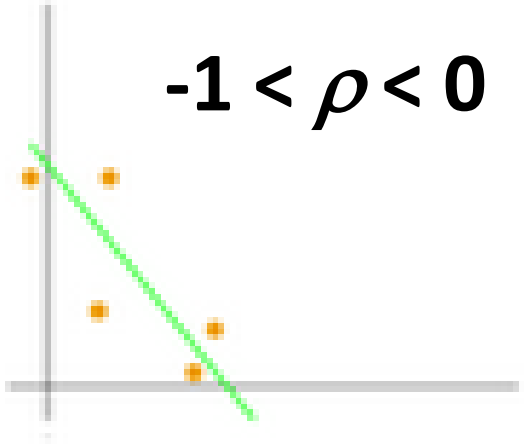
Calcolo della correlazione di Pearson

$$\rho = \frac{Covarianza(x, y)}{\sigma_x \sigma_y}$$

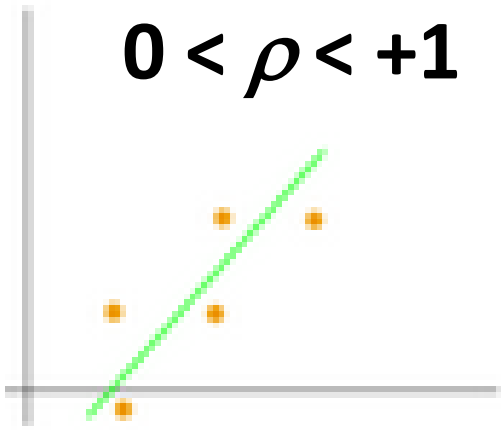
ρ = -1



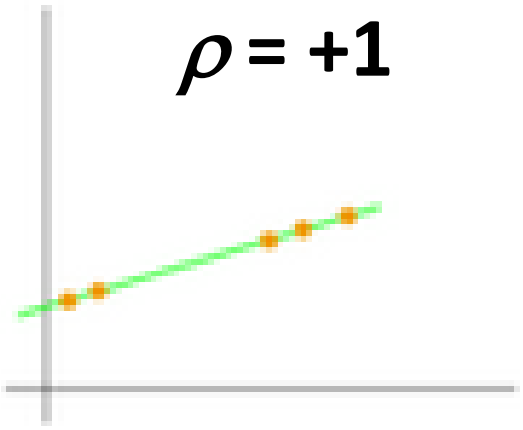
-1 < ρ < 0



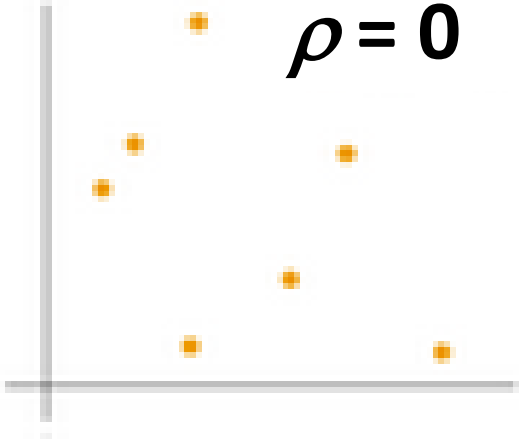
0 < ρ < +1



ρ = +1



ρ = 0



Calcolo della correlazione di Spearman

Indice di correlazione per ranghi. Il rango ordina un elenco di numeri a seconda di un criterio prestabilito (ad esempio dal più piccolo al più grande; vedi funzione RANGO su Excel)

$$\rho_s = 1 - \frac{6 \sum_i D_i^2}{N(N^2 - 1)}$$

dove

$$D_i = r_i - s_i$$

è la differenza dei ranghi (essendo r_i e s_i rispettivamente il rango della prima variabile e della seconda variabile della *i*-esima osservazione. N è il numero complessivo di osservazioni

Regressione lineare

- Regressione semplice
- Regressione multipla

Regressione semplice

Funzione statistica che descrive la relazione tra una variabile dipendente y (effetto) ed un'altra indipendente x (causa)
Questa funzione permette di interpolare ed estrapolare dati (y) in funzione dei dati osservati x:

$$y = f(x) + e$$

dove

$f(x)$ è la componente sistematica (deterministica)

e è la componente casuale (stocastica)

Ipotesi da soddisfare:

1) media attesa degli errori è nulla (scarti positivi e negativi si compensano)
2) la varianza degli errori è costante al variare delle osservazioni (omoschedasticità)
3) la covarianza degli errori è nulla, gli errori si assumono incorrelati e quindi indipendenti fra loro
4) la variabile indipendente (x) non è affetta da errori di misura

Regressione lineare semplice

$y_i = \beta_0 + \beta_1 x_i + e_i \qquad i=1,2,\ldots,N.$

dove
x è la variabile indipendente
y è la variabile dipendente
e è l'errore (quantità stocastica non spiegata dal modello)

β_0 e β_1 sono i coefficienti «veri» della retta della popolazione:

β_0 è l'intercetta

β_1 è il coefficiente angolare della retta di regressione

Stima dei parametri

Problema: non conosco β_0 e β_1 della popolazione, quindi li stimo da uno o più campioni

$\hat{y}_i = \boldsymbol{a} + \boldsymbol{b} x_i$

dove

x_i sono i valori misurati della variabile indipendente

a e **b** sono le stime dei parametri β_0 e β_1

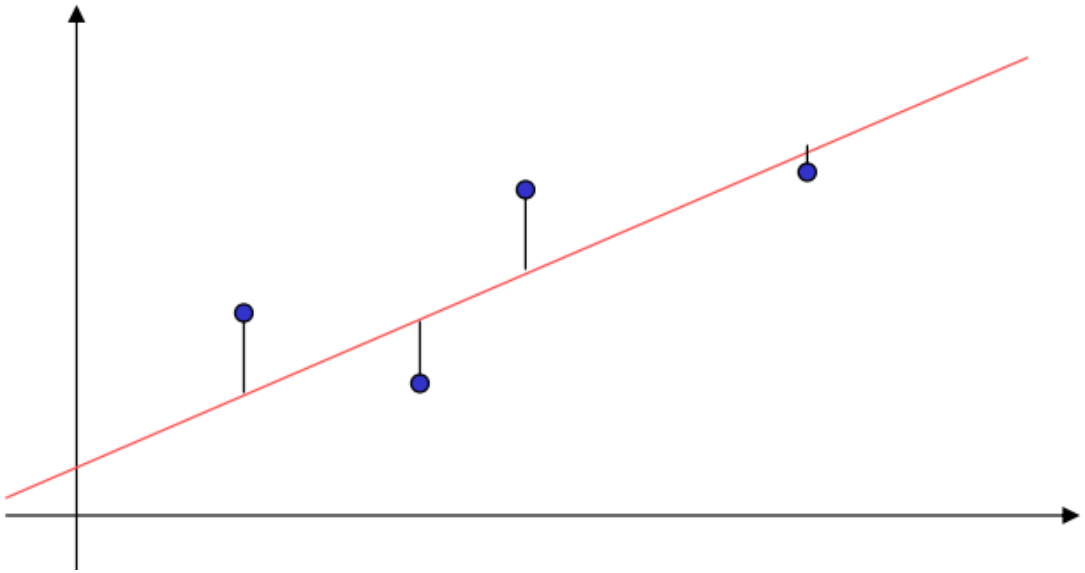
$\hat{y}_i$ sono i valori di y stimati dalla regressione semplice

Obiettivo: trovare quella coppia di **a** e **b** che minimizzino gli errori tra y_i e $\hat{y}_i$ o in altre parole:

$$\min\sum(y_i-\hat{y}_i)^2 = \min\sum[y_i-(\mathbf{a}+\mathbf{b}x_i)]^2$$

Metodo dei minimi quadrati

così minimizzo il contributo della componente stocastica della regressione, concepita come residuo o deviazione rispetto alla relazione tra x e y



Stima dei parametri

dopo alcuni passaggi matematici ***b*** si calcola con la seguente formula:

codevianza=covarianza*N

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

devianza di x

infine possiamo calcolare il coefficiente ***a*** :

$$a = \bar{y} - b\bar{x}$$

quindi esiste un'unica retta che rende minima la somma dei quadrati degli scarti. Inoltre questa retta implica che $\sum e_i = 0$ il che implica a sua volta che $\sum y_i = \sum \hat{y}_i$, cioè la media campionaria delle y_i osservate coincide per definizione con la media delle $\hat{y}_i$ calcolate mediante la retta dei minimi quadrati.

La tabella 1 riporta le misure del volume di una quantità di un gas a differenti temperature.

Tabella 1

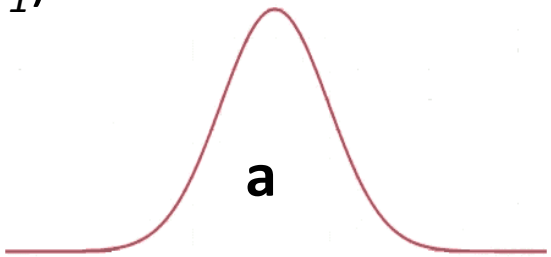
Temperatura	Volume
10	10,4
20	11,1
30	11,2
40	11,9
50	11,8
60	12,3

Calcolare il coefficiente di correlazione e verificare se esiste una dipendenza lineare del volume dalla temperatura.
(Usare la funzione CORRELAZIONE)
Determinare l'equazione della retta di regressione y=Ax+B e disegnarne il grafico

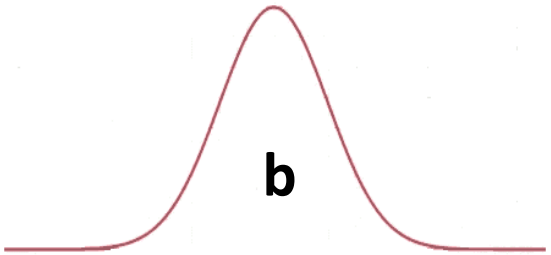
Il calcolo dei coefficienti della retta di regressione può essere fatto con le funzioni PENDENZA (coefficiente A) e INTERCETTA (coefficiente B)
Sintassi
PENDENZA(**y_nota**;**x_nota**)
INTERCETTA(**y_nota**;**x_nota**)
x_nota insieme dei valori della variabile indipendente X
y_nota insieme dei valori della variabile dipendente Y

Verifica del modello stimato

Quando si utilizzano dati campionari (e non dati della popolazione) calcoliamo **a** e **b** (stime dei parametri veri β_0 e β_1) che sono caratterizzati da distribuzioni statistiche di probabilità. In altre parole, al variare del campione, variano anche **a** e **b**. Ci domandiamo se **a** e **b** sono statisticamente uguali o diversi dai corrispondenti valori «veri» della popolazione (β_0 e β_1)



β_0 = **a** ?



β_1 = **b** ?

Ipotesi nulla H_0:

$\beta_0 = 0$
$\beta_1 = 0$

cioè ipotizzo che non vi sia relazione tra x e y. I parametri **a** e **b** sono stime di $\boldsymbol{\beta_0}$ e $\boldsymbol{\beta_1}$ ed hanno distribuzione campionaria.

Verifica del modello stimato

Introduciamo quindi l'errore standard per **a** e **b** con S_a e S_b

$$S_a = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2 \sum x^2}{N\,(N-2)\sum(x_i - \bar{x})^2}}$$

$$S_b = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{(N-2)\sum(x_i - \bar{x})^2}}$$

Test $H_0: \beta_0 = 0$

$$t = \frac{a - \beta_0}{S_a} = \frac{a}{S_a}$$

$H_0: \beta_1 = 0$

$$t = \frac{b - \beta_1}{S_b} = \frac{b}{S_b}$$

Test su β_1 è molto più importante di quello su β_0

Procedura: fisso il livello di significatività

Se decido 5%, avrò α=0.05, quindi $t_{critico}=t_{GL,\alpha}$ ossia $t_{GL,0.05}$

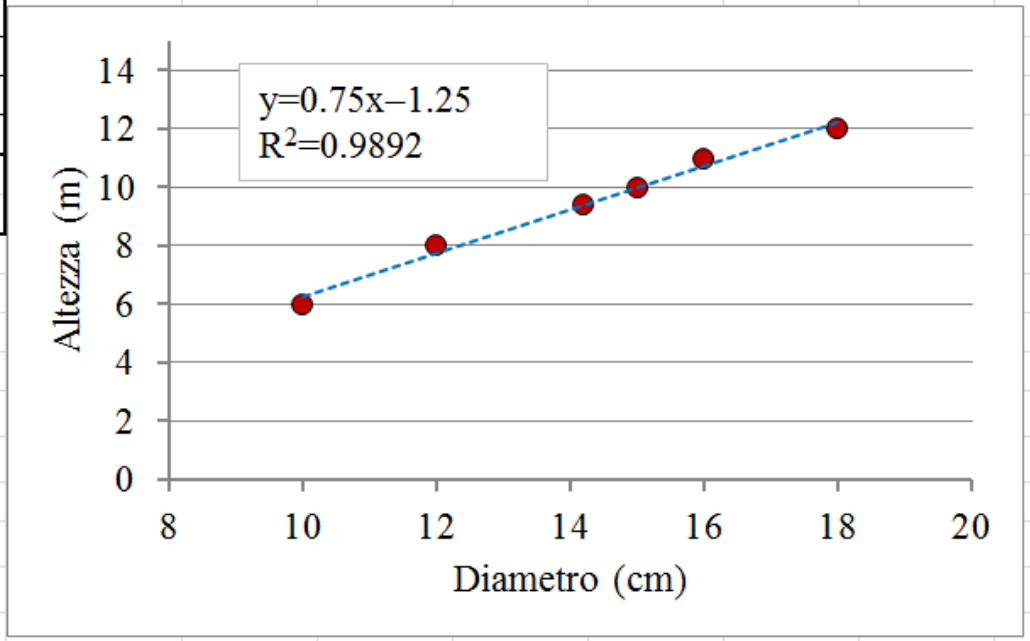
Caso 1: se ottengo $|t|<t_{GL,0.05}$, sto nella zona di accettazione quindi accetto l'ipotesi nulla H_0, pertanto non c'è proporzionalità diretta tra x e y (regressione NON valida)

Caso 2: invece… se ottengo $|t|> t_{GL,0.05}$, sto nella zona di rifiuto quindi non accetto l'ipotesi nulla H_0, pertanto esiste proporzionalità tra x e y (regressione valida)

Esempio: ho un campione di N=5 alberi con diametri (x) ed altezze (y) misurati. Calcolo anche i parametri dell'equazione di regressione che sono **a**=-1.25 e **b**=0.75. In questo caso eseguo il test solo su β_1

H_0: β_1 =0

	A	B	C	D	E
1	Diametro	Altezza			
2	x	y	yreg	(y-yreg)^2	(x-xmedio)^2
3	cm	m	m	m2	cm2
4	12	8	7.75	0.06	4.8
5	10	6	6.25	0.06	17.6
6	18	12	12.25	0.06	14.4
7	15	10	10.00	0.00	0.6
8	16	11	10.75	0.06	3.2
9	media			somma	
10	14.2	9.4		0.25	40.80
11	N	5			
12	GL	3			
13	a	-1.25	intercetta		
14	b	0.75	pendenza		
15	S_b	0.045	err. standard		
16	$t=b/S_b$	16.60			
17	α	0.05	livello signific.		
18	t_{crit}	3.18			
19	$t > t_{crit}$				
20	Rifiuto H_0 quindi la relazione tra x e y è altamente significativa				
21					



pertanto….al 95% β_1 =b±$t_{GL,\alpha}S_b$=0.75 ±3.18*0.045

Qualità della regressione

Decomposizione della devianza totale:

Devianza totale=devianza della regressione + devianza residua

Dev(y)=Dev(ŷ)+Dev(e)

Indice (o coefficiente) di determinazione (R^2) descrive la bontà di accostamento della retta di regressione ai dati campionari:

$$R^2 = \frac{Dev(\hat{y})}{Dev(y)} = 1 - \frac{Dev(e)}{Dev(y)}$$

Ad esempio R^2=0.86 indica che la relazione lineare tra x e y mediante la nostra retta di regressione «spiega» l'86% della variabilità di y, mentre esiste un residuo di variabilità «non spiegata» pari al 14% che è attribuibile alla variabile casuale *e*

Qualità della regressione

Coefficiente di determinazione

$$R^2 = \frac{\hat{\sigma}_y^2}{\sigma_y^2} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

devianza spiegata

devianza totale

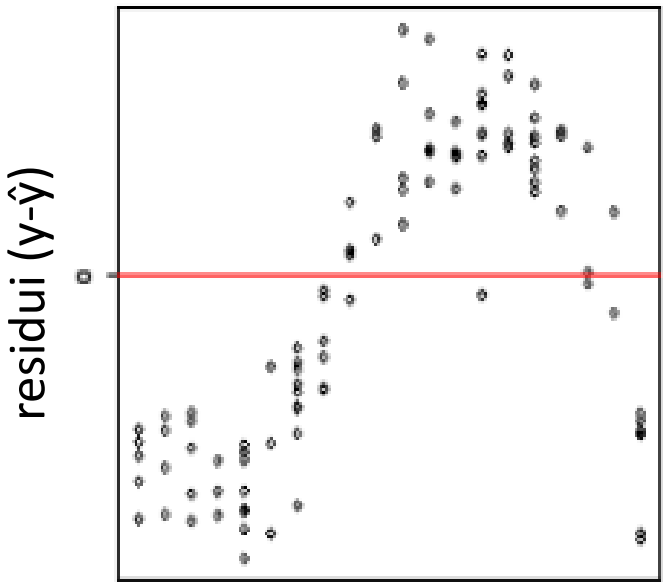
- misura la dispersione delle osservazioni attorno alla retta di regressione
- rappresenta la porzione della variazione in Y spiegata dalla regressione su X
- consente di valutare l'utilità dell'equazione di regressione ai fini della previsione sui valori della Y

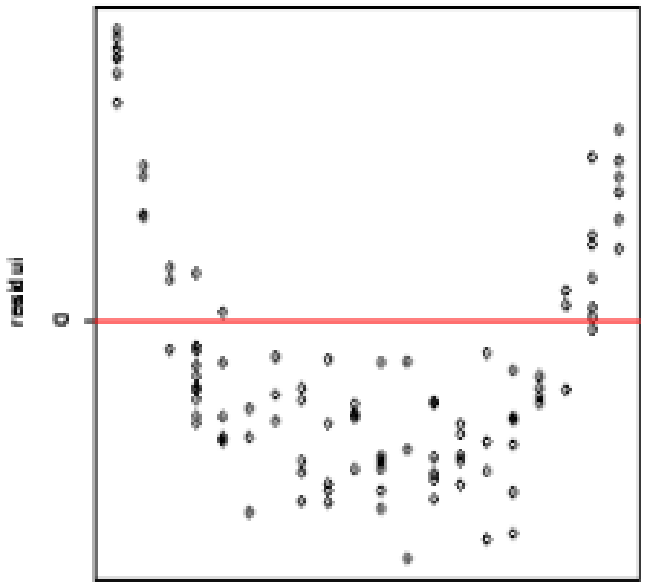
Coefficiente di determinazione

- Se c'è una perfetta relazione lineare tra X e Y
  - tutte le osservazioni cadono sulla retta di regressione
  - $\hat{\sigma}_y^2 = \sigma_y^2$, cioè nessun errore viene commesso nella predizione di Y a partire da X
  - $R^2=1$

R^2 è il quadrato del coefficiente di correlazione lineare $\rho(x,y)$

Bontà di adattamento





Andamenti come questi indicano che il modello lineare non è adatto a spiegare il legame tra le variabili

Regressione non lineare

- esponenziale $y = ae^{bx}$
- logaritmica $y = \alpha \ln x + \beta$
- polinomiale $y = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + .. + \alpha_m x^m$

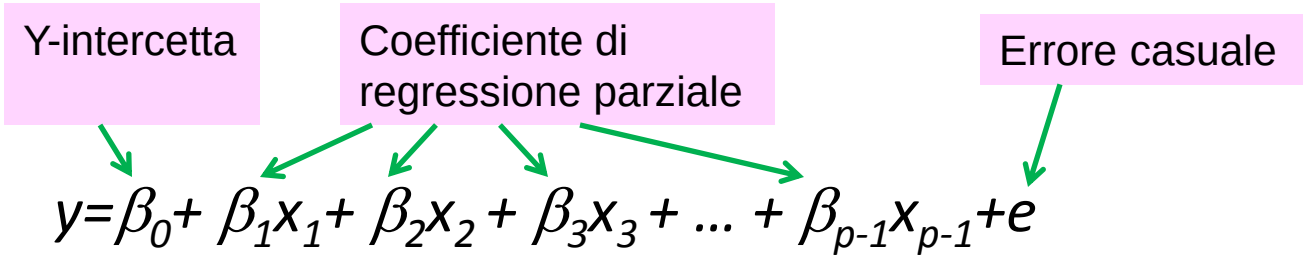
Esempi di linearizzazione

$y=ax^b$ ⟹ $\ln(y)=\ln(a)+b\ln(x)$

$y=a + \exp(b/x)$ ⟹ $\ln(y)=\ln(a)+b/x$

Regressione multipla

La variabile dipendente y è una funzione lineare delle p-1 variabili indipendenti (esplicative) x_1, x_2,…x_{p-1}:



Per la validità del modello di regressione multipla valgono le stesse ipotesi della regressione semplice. Inoltre si aggiunge l'ipotesi che le variabili indipendenti esplicative non siano correlate tra di loro

Esempio con p=3 (2 variabili indipendenti x_1 e x_2)

$$y=\beta_0+\beta_1x_1+\beta_2x_2+e$$

Se la correlazione tra x_1 e x_2 fosse alta, avremmo $x_1= a+bx_2$ e quindi le due variabili fornirebbero le stesse informazioni su y (una tra x_1 e x_2 risulterebbe ridondante)

Esempio su 20 famiglie. Voglio spiegare il consumo alimentare (C) utilizzando due variabili esplicative:
reddito (R) e dimensione della famiglia (DF)

N	C	R	DF
1	5.2	28	3
2	5.1	26	3
3	5.6	32	2
4	4.6	24	1
5	11.3	54	4
6	8.1	59	2
7	7.8	44	3
8	5.8	30	2
9	5.1	40	1
10	18	82	6
11	4.9	42	3
12	11.8	58	4
13	5.2	28	1
14	4.8	20	5
15	7.9	42	3
16	6.4	47	1
17	20	112	6
18	13.7	85	5
19	5.1	31	2
20	2.9	26	2

$C_i=\beta_0+\beta_1 R_i+\beta_2 DF_i+e_i$ *dove i=1,2,3…,20*

Come risolvo?

1) Strumenti / Analisi Dati... / Regressione
2) Dati / Risolutore

CONDIZIONE DEI MINIMI QUADRATI ORDINARI (OLS):

$$\sum_{i=1}^{n} e_i^2 = \min$$

Risultati della regressione multipla

$$C = -1.118 + 0.148\,R + 0.793(DF)$$

1	OUTPUT RIEPILOGO								
2									
3	*Statistica della regressione*								
4	R multiplo	0.966747321							
5	R al quadrato	0.934600383							
6	R al quadrato corrett	0.926906311							
7	Errore standard	1.261013533							
8	Osservazioni	20							
9									
10	ANALISI VARIANZA								
11		*gdl*	*SQ*	*MQ*	*F*	*Significatività F*			
12	Regressione	2	386.3128628	193.1564314	121.470181	8.55833E-11			
13	Residuo	17	27.03263721	1.59015513					
14	Totale	19	413.3455						
15									
16		*Coefficienti*	*Errore standard*	*Stat t*	*Valore di significatività*	*Inferiore 95%*	*Superiore 95%*	*feriore 95.0*	*eriore 95.0%*
17	**Intercetta**	**-1.118**	0.655	-1.708	0.106	-2.500	0.263	-2.500	0.263
18	**R**	**0.148**	0.016	9.049	0.000	0.114	0.183	0.114	0.183
19	**DF**	**0.793**	0.244	3.245	0.005	0.277	1.309	0.277	1.309
20									

L'equazione di regressione multipla

$$C = -1.118 + 0.148\,R + 0.793(DF)$$

Dove
- C è in Euro*1000
- R è in Euro*1000
- DF è in numero di componenti.

β_1 **= 0.148**: il consumo alimentare aumenta, in media, di **148** Euro all'anno all'aumentare di **1000 Euro del REDDITO**, al netto (fermo restando) degli effetti dovuti alle variazioni di **DF**

β_2 **= 0.793**: il consumo alimentare aumenta, in media, di **793** Euro all'anno all'aumentare di **1 di DF** , al netto (fermo restando) degli effetti dovuti alle variazioni del **REDDITO**

Quale variabile ha la maggior influenza sul consumo ?

Il confronto non è possibile in quanto y e x_1 e x_2 hanno unità di misura diversa. Per rendere possibile il confronto è necessario fare ricorso a dei coefficienti di regressione parziali che sono numeri puri e ottenuti partendo da una equazione di regressione multipla in termini di variabili standardizzate Z.

$$Z_y = \alpha_1 Z_1 + \alpha_2 Z_2 + u$$

$$\alpha_j = \frac{\sigma_{x_j}}{\sigma_Y} \times \beta_j$$

Il coefficiente di regressione è moltiplicato per il rapporto delle deviazioni standard della variabile indipendente X_j e della variabile dipendente Y

Rimozione delle ipotesi sul modello di regressione

1) Omissione di una variabile esplicativa nella regressione multipla: otteniamo stimatori distorti

2) Regressione non lineare
Se i modelli non sono linearizzabili, si ricorre all'ottimizzazione parametrica iterativa con il metodo dei minimi quadrati. La ricerca del minimo è fatta per tentativi (RISOLUTORE su Excel)

3) Eteroschedasticità nel modello di regressione, cioè la varianza varia al variare di x. Le stime risultano non distorte, consistenti ma poco efficienti (poco credibili). Quindi si trasformano le variabili in modo da ottenere un nuovo modello di regressione con residui omoschedastici. Procedura dei minimi quadrati ponderati indicati con WLS (*Weighted Least Squares*)

Rimozione delle ipotesi sul modello di regressione

4) Autocorrelazione degli errori. L'ipotesi è che gli errori siano incorrelati fra di loro e quindi indipendenti. Caso raro nelle serie storiche: $e_t=\rho e_{t-1}+v_t$ per t=1,2,…,N

↳ la correlazione temporale ρ in genere è incognito

dove v sono le variabili casuali incorrelate

5) multicollinearità delle variabili esplicative. Il metodo dei minimi quadrati non funziona! Bisogna eliminare le variabili collineari o riformulare il modello di regressione multipla

6) Non-Normalità dei residui. Si risolve solo con N campionario molto grande e i test di confidenza con limiti fiduciali si approssimano col Teorema del Limite Centrale