TECNICHE DI ANALISI DEI DATI

AA 2017/2018

PROF. V.P. SENESE

Questi materiali sono disponibili per tutti gli studenti al seguente indirizzo:
https://goo.gl/hxL9zG

MODELLI LINEARI

LA REGRESSIONE

LA REGRESSIONE SEMPLICE (E MULTIPLA)

L'ANALISI DELLA VARIANZA

DISEGNI UNIVARIATI (DISEGNI FATTORIALI SEMPLICI E MISTI)

MODELLI LINEARI

Quando in una ricerca è possibile distinguere (in base alla teoria) tra **variabili indipendenti** e **variabili dipendenti** il ricercatore può essere interessato a verificare la presenza della **relazione *causale*** supposta (tra le variabili) nei dati raccolti (osservazioni campionarie).

Prima di iniziare un qualsiasi discorso sulle relazioni di causalità tra variabili dobbiamo ribadire la distinzione tra **covariazione** e **causazione**.

MODELLI LINEARI

COVARIAZIONE
(Covarianza, Correlazione o Associazione):
quando "semplicemente" osserviamo che due variabili presentano variazioni concomitanti.

CAUSAZIONE:
quando pensiamo che siano proprio le variazioni della variabile X a determinare le variazioni della variabile Y. Identifichiamo la DIREZIONALITÀ e l'esistenza del LEGAME DIRETTO tra le due variabili.

Mentre la covariazione è osservabile la causazione appartiene al dominio della teoria!!!

REGRESSIONE LINEARE

Quando la relazione si riferisce a due variabili di tipo **quantitativo** (I o R) l'analisi che può essere impiegata è l'analisi della **regressione lineare**.

In questo caso l'obiettivo è quello di voler verificare se la capacità di prevedere i valori di una data variabile **Y**, **E(Y)**, aumenta conoscendo i valori assunti da una data variabile **X**.

REGRESSIONE LINEARE

PREVISIONE DEI PUNTEGGI

Sappiamo che quando non conosciamo il punteggio Y_i di un soggetto, la migliore previsione che possiamo fare è usare come valore di riferimento il punteggio medio in Y:

$$Y_i = \mu_Y + \varepsilon$$

$$Y_i = E(Y) + \varepsilon$$

Ipotesi:

Questo modello assume che tutti le osservazioni vengono dalla stessa popolazione e che le differenze osservate sono dovute solo all'errore.

REGRESSIONE LINEARE

Se supponiamo che il punteggio $\mathbf{Y_i}$ **dipende dal punteggio** $\mathbf{X_i}$ del soggetto, possiamo provare a prevedere il valore di **Y** in base alla seguente formula:

$$E(Y_i) = \mu_Y + \beta X_i + \varepsilon_i$$

In pratica, ipotizziamo che (mantenendo la componente stocastica) se la teoria è vera, allora il valore atteso di **Y** è funzione **lineare** di **X**.

REGRESSIONE LINEARE

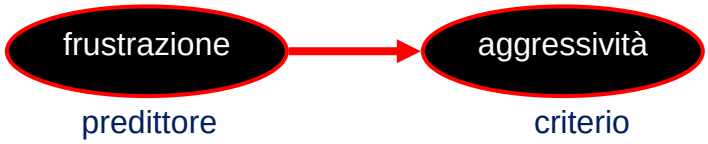
L'analisi di regressione (lineare) è una tecnica di analisi dei dati che esamina la relazione tra una (o più) variabili esplicative (VI o *predittori*) e una variabile criterio (o VD). Lo studio della relazione può avere un duplice scopo:

- **ESPLICATIVO**
  es. sottoporre a verifica un modello teorico

- **PREDITTIVO**
  es. individuare la combinazione lineare di variabili che consentono di stimare in modo ottimale la VD

REGRESSIONE LINEARE

La regressione lineare si dice **semplice** quando abbiamo una sola **VD** (o criterio) e una sola **VI** (o predittore). L'ipotesi che viene formulata riguarda l'influenza della VI sulla VD.



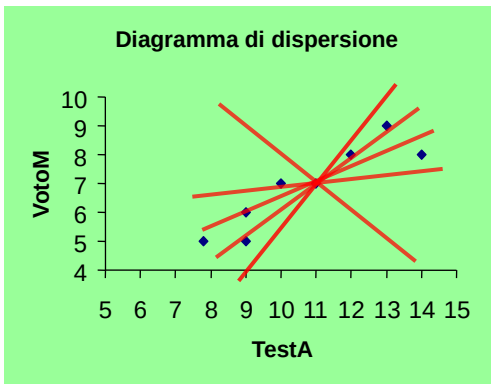
criterio coefficiente errore

$$\hat{Y} = \alpha + \beta x + \varepsilon$$

costante predittore

REGRESSIONE LINEARE

Da un punto di vista grafico viene individuata quella **retta** che, data la relazione tra le variabili, consente di **prevedere al meglio** i punteggi nella variabile dipendente a partire da quelli nella variabile indipendente.

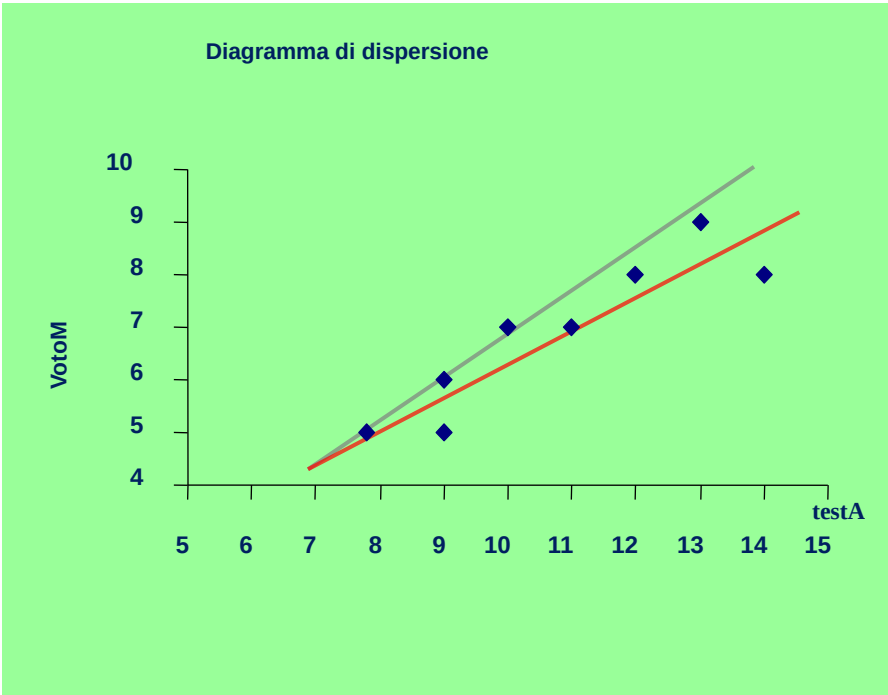


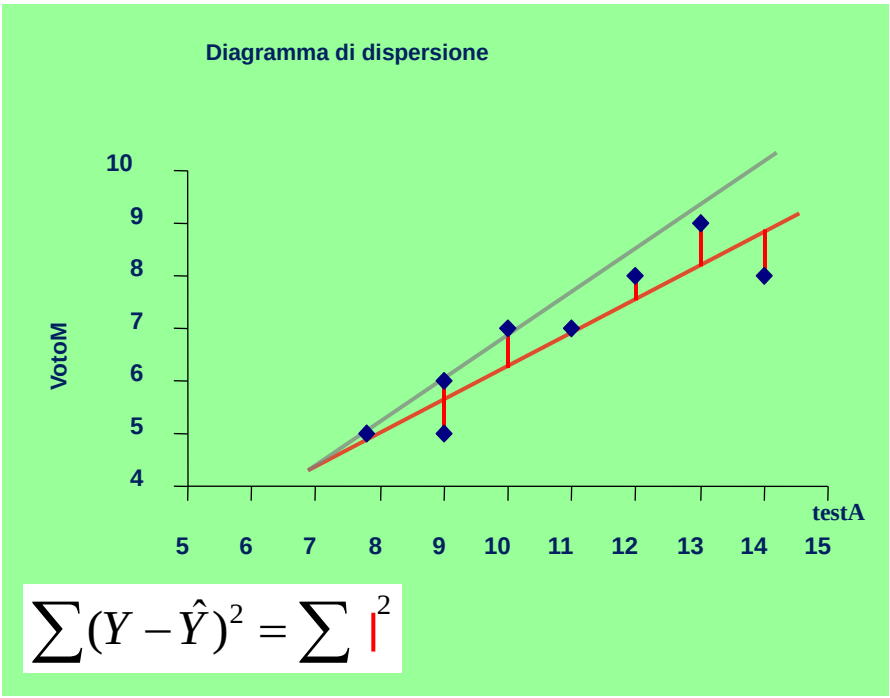
REGRESSIONE LINEARE

Dato un diagramma di dispersione tra due variabili, la retta di regressione è "**la migliore delle rette**" nel senso che è quella retta che passa più vicina a tutti i punti (minimizza tutte le distanze tra i punti e la retta).

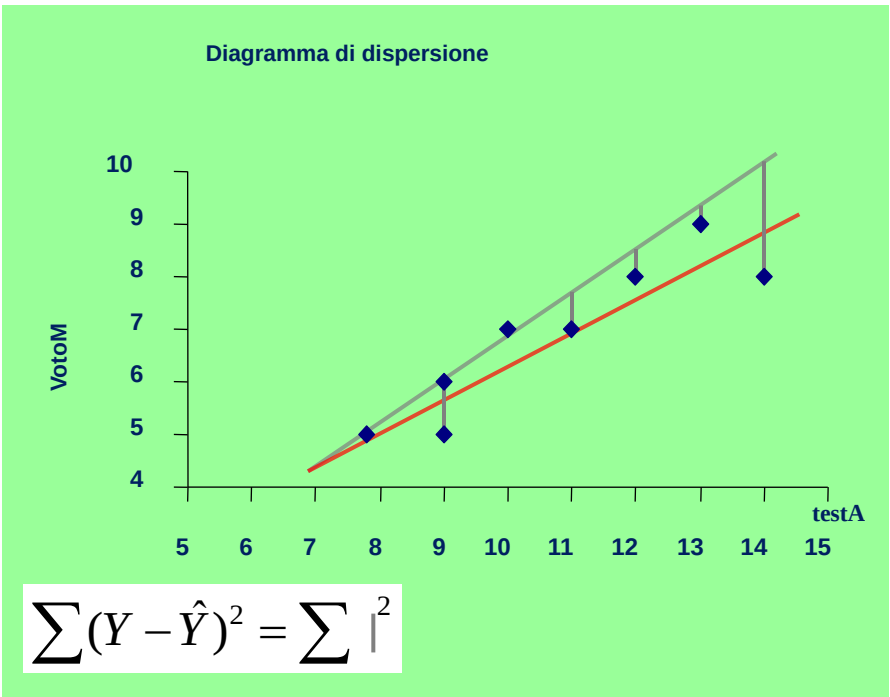
Si sceglie in base al **metodo dei minimi quadrati.** Si definisce "migliore" la retta che rende minima la **somma dei quadrati degli errori**, cioè:

$$\sum (Y_i - \hat{Y}_i)^2 = più\ piccolo\ possibile$$

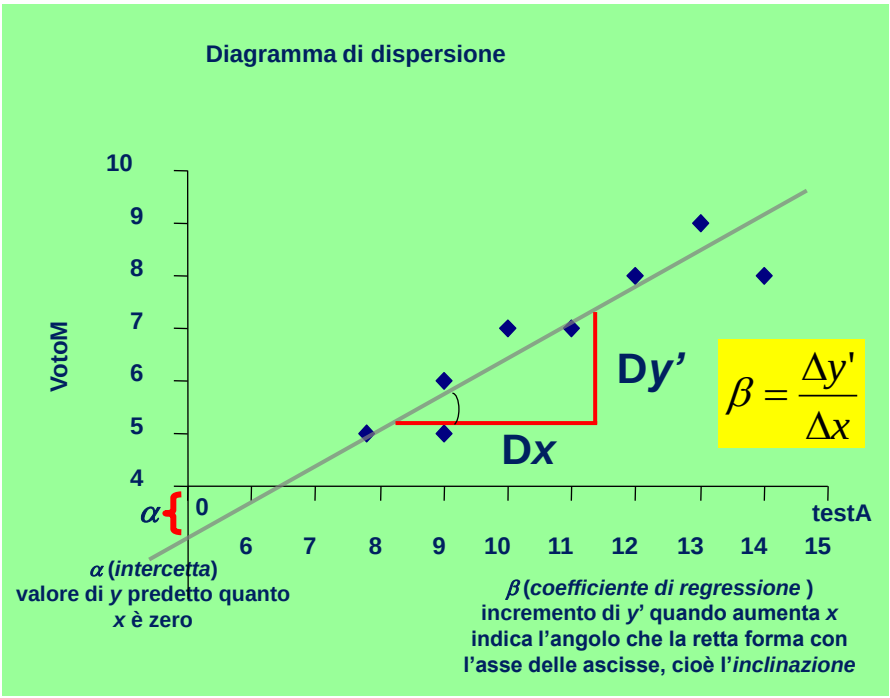




$$\sum (Y - \hat{Y})^2 = \sum |^2$$



$$\sum (Y - \hat{Y})^2 = \sum |^2$$

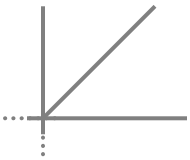


COEFFICIENTE DI REGRESSIONE

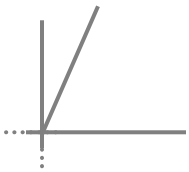
Esprime la relazione tra **X** e **Y** nei termini delle **unità di misura** delle due variabili. **Non è standardizzato** (± ∞) e si interpreta solo il segno.

$$\beta = \frac{\Delta y'}{\Delta x}$$

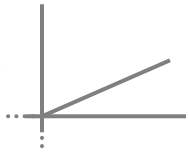
se β = 1 → per ogni incremento unitario di **X** c'è un incremento unitario di **Y** (45°);



se β = 2 → per ogni incremento unitario di **X** c'è un incremento doppio di **Y** (2 unità);



se β = 0.5 → per ogni incremento unitario di **X** c'è un incremento di mezza unità di **Y**.



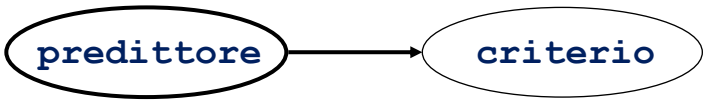
COEFFICIENTE DI REGRESSIONE STANDARDIZZATO

Il **coefficiente di regressione standardizzato** (β | $\pm$ 1) esprime la relazione tra la variabile dipendente (**Y**) e la variabile indipendente (**X**) in **unità di misura standard** (**punti** z).

N.B. **Solo nella regressione semplice** corrisponde al **coefficiente di correlazione**.

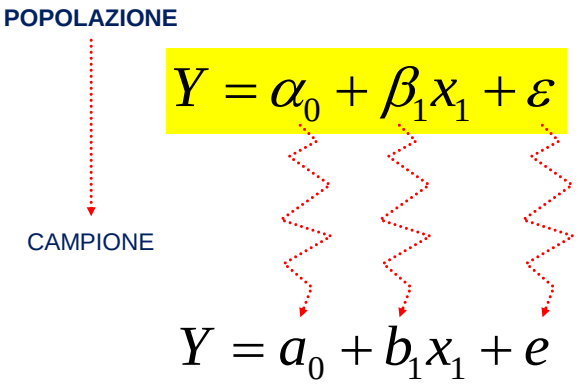
COEFFICIENTE DI DETERMINAZIONE

Il **coefficiente di determinazione** (r^2) indica la **proporzione di varianza** (%) della variabile criterio (**Y**) "spiegata" da quella del predittore (**X**). Il valore è compreso tra 0 e 1.



REGRESSIONE LINEARE

I coefficienti di regressione α e β della **popolazione** vengono **stimati** a partire dai **coefficienti di regressione campionari** a e b:



COEFFICIENTE DI REGRESSIONE

Il coefficiente di regressione è simboleggiato come:

β (**beta**) quando ci si riferisce al coefficiente **non standardizzato** della popolazione;

b quando ci si riferisce al coefficiente **non standardizzato** calcolato nel campione;

β (**beta**) quando ci si riferisce al **coefficiente standardizzato** (**punti** z) calcolato nel campione.

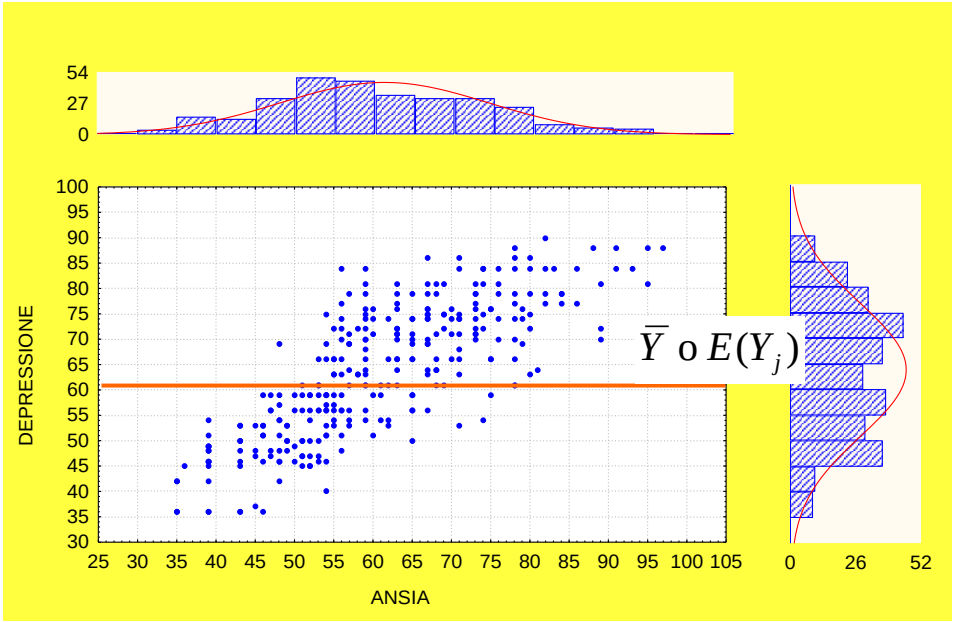
PARAMETRI

Nella **regressione semplice** le formule per il calcolo dei parametri sono le seguenti:

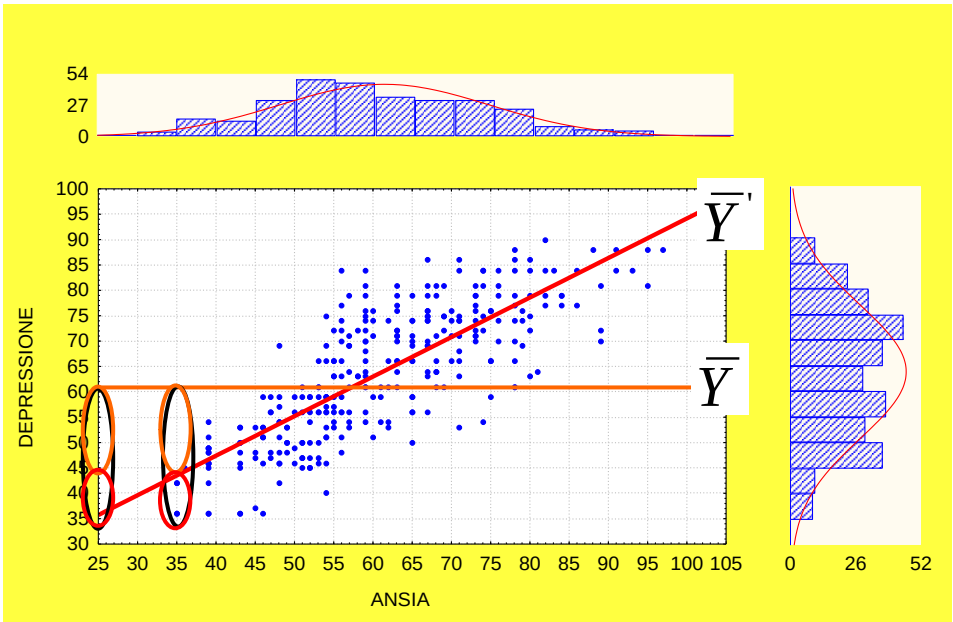
$$b = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

$$a = \overline{Y} - b\overline{X}$$

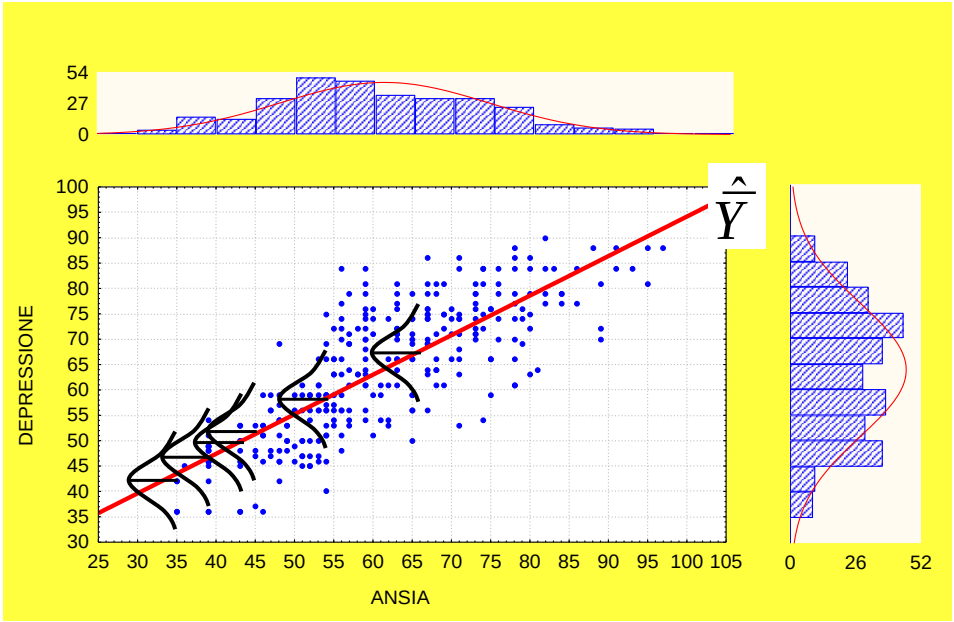
$\bar{Y}_{depressione} = 61; ds = 13$



$\bar{Y}_{depressione} = 61; ds = 13$



$\bar{Y}_{depressione} = 61; ds = 13$



SIGNIFICATIVITÀ DELLA PREVISIONE

Scomposizione **Devianza totale**, nelle componenti di **errore** e di "**effetto**":

$$SQ_{tot} = SQ_{reg} + SQ_{err}$$

La somma dei quadrati ***totale*** (***SQ_{tot}***) è data da una componente di ***errore*** (***SQ_{err}***) e da una componente ***spiegata dalla regressione*** (***SQ_{reg}***)

SIGNIFICATIVITÀ DELLA PREVISIONE

$$SQ_{tot} = SQ_{reg} + SQ_{err}$$

DEVIANZA SPIEGATA
SQ_{reg}

$$\sum \left(Y_i - \bar{Y}\right)^2 = \sum \left(\hat{Y}_i - \bar{Y}\right)^2 + \sum \left(Y_i - \hat{Y}_i\right)^2$$

SQ_{tot}
DEVIANZA TOTALE

SQ_{err}
DEVIANZA NON SPIEGATA o RESIDUA

SIGNIFICATIVITÀ DELLA PREVISIONE

Per verificare la significatività della previsione, si confrontano le due varianze. La previsione è significativa se la **varianza spiegata** dalla regressione è **maggiore** di quella **residua**.

Le **varianze** si calcolano **dividendo le devianze** per i **gradi di libertà** opportuni.

$$GDL_{tot} = GDL_{reg} + GDL_{err}$$

$$N - 1 = (k) + (N - k - 1)$$

N = numero di osservazioni
k = numero di predittori

SIGNIFICATIVITÀ DELLA PREVISIONE

Per **confrontare la due varianze** e verificare se quella spiegata dalla regressione è maggiore di quella residua, si calcola la statistica **F**. La **varianza spiegata** dalla regressione va al numeratore, quella **residua** al denominatore.

$$\mathrm{F} = \frac{\text{var spiegata}}{\text{var errore}}$$

$$F = \frac{Var_{reg}}{Var_{res}} = \frac{\frac{Dev_{reg}}{k}}{\frac{Dev_{res}}{N-k-1}}$$

H_0 : la varianza spiegata è uguale a quella residua (casuale)

$\mathrm{H}_0 : F = 1 \Leftrightarrow \mathrm{H}_1 : F > 1$

$$\mathrm{gdl}_{\mathrm{F}} = \frac{k}{n-k-1}$$

SIGNIFICATIVITÀ DELLA PREVISIONE

La verifica dell'ipotesi nulla (**H_0**) fatta utilizzando la statistica **F** riguarda il **modello complessivo**; si assume che tutte le ***k*** variabili indipendenti **non influenzino** in modo significativo la variabile dipendente:

$$\mathrm{H}_0 \Rightarrow \beta_1 = \beta_2 = \beta_3 = \ldots = \beta_k = 0$$

$$\mathrm{H}_1 \Rightarrow \beta_1 \; o \; \beta_2 \; o \; \beta_3 \; \mathrm{o} \ldots \mathrm{o} \; \beta_k \neq 0$$

SIGNIFICATIVITÀ DELLA PREVISIONE

Se la **F** è significativa (**H_1**) allora l'analisi prosegue per verificare quale **predittore** ha determinato l'effetto. Viene quindi definita una specifica ipotesi nulla (**H_0**) per ciascun predittore.

$$H_0 \Rightarrow \beta_i = 0$$

$$H_1 \Rightarrow \beta_i \neq 0$$

Solo nella regressione semplice questo test è ridondante dal momento che c'è un solo predittore.

Il test statistico appropriato per la verifica è il valore ***t*** (un campione):

$$t = \frac{b_i - \beta_{i_{H_0}}}{s_{b_i}} \Rightarrow \frac{b_i}{s_{b_i}}$$

$$gdl_t = n - k - 1$$

BONTÀ DI ADATTAMENTO

La statistica maggiormente impiegata per la valutazione della **bontà di adattamento** del modello (***goodness-of-fit***) è l'**R^2** (***effect size***) che viene stimato con la seguente formula:

$$R^2 = \frac{\text{dev spiegata}}{\text{dev totale}}$$

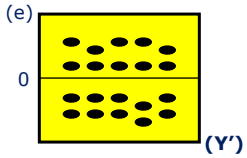
$$R^2 = \frac{\sum_{j=1}^{n} (\hat{Y}_j - \bar{Y})^2}{\sum_{j=1}^{n} (Y_j - \bar{Y})^2}$$

Il **coefficiente di determinazione** (***R^2***) indica la **proporzione di varianza** (**%**) della variabile criterio (**Y**) "spiegata" da quella del predittore (**X**). Il valore è compreso tra 0 e 1.

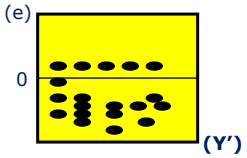
ASSUNZIONI

Oltre all'assunzione di **linearità** la regressione multipla si basa sulle seguenti assunzioni:

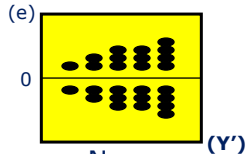
- Nessun errore di **specificazione** (no variabili irrilevanti, si variabili rilevanti);
- Nessun errore di **misurazione** della/e variabile/i indipendente/i;
- La **media** degli errori di predizione (e) attorno ad ogni valore (Y') predetto deve essere uguale a **0**;
- Gli errori di predizione (e) attorno ad ogni valore (Y') predetto debbono essere **indipendenti** e distribuiti **normalmente**;
- La **varianza** degli errori di predizione (e) attorno ad ogni valore (Y') predetto deve essere **uguale** (omoschedastica).



OK

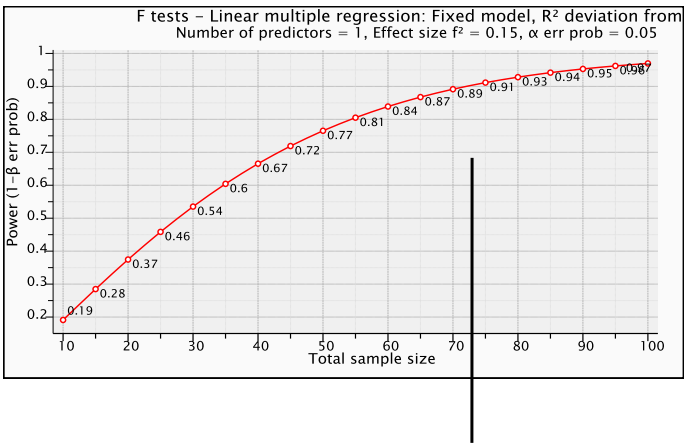


Non Normalità

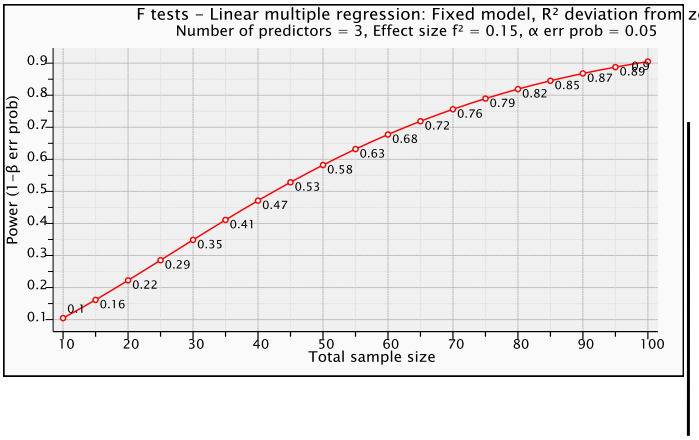


Non Omoschedasticità

POWER (*k* = 1)



POWER (*k* = 3)



ESEMPIO #1

Il voto medio in matematica predice il voto al test di statistica?

Regressione semplice con una variabile indipendente (MBQ; VI-I) e una variabile dipendente (sqz; VD-I).

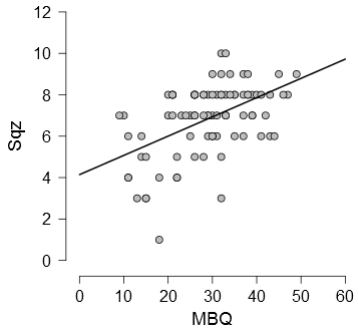
$$H_0 \Rightarrow \beta_i = 0$$

$$H_1 \Rightarrow \beta_i \neq 0$$

$$\alpha = .05$$

ESEMPIO #1

Analisi grafica della relazione



Correlazione

Pearson Correlations

		MBQ	Sqz
MBQ	Pearson's r	—	0.508
	p-value	—	< .001
Sqz	Pearson's r		—
	p-value		—

ESEMPIO #1

Model Summary

Model	R	R^2	Adjusted R^2	RMSE
1	0.508	0.258	0.249	1.504

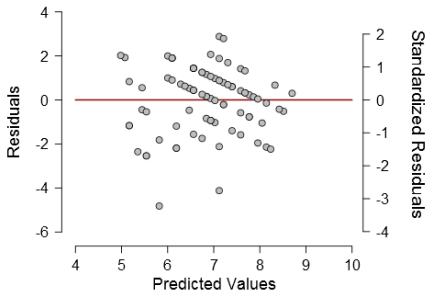
ANOVA

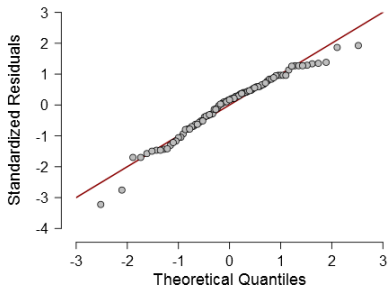
Model		Sum of Squares	df	Mean Square	F	p
1	Regression	65.26	1	65.260	28.85	< .001
	Residual	187.75	83	2.262		
	Total	253.01	84			

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p	2.5%	97.5%
1	intercept	4.144	0.529		7.834	< .001	3.092	5.196
	MBQ	0.093	0.017	0.508	5.371	< .001	0.059	0.127

ESEMPIO #1





ESEMPIO #1

Questo risultato ci porta a respingere l'ipotesi nulla e a supportare l'ipotesi alternativa.

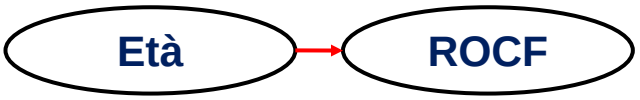
$$H_0 \Rightarrow \beta_i = 0$$

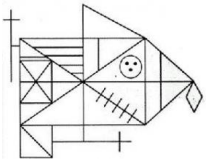
$$H_1 \Rightarrow \beta_i \neq 0$$

Il voto medio in matematica (MBQ) influenza significativamente il voto al test di statistica (sqz), ***F*(1,83) = 28.85, *p* < .001, R^2 = .258**. In particolare, i dati evidenziano una relazione positiva tra le due variabili, ***b* = 0.093, β = .508, 95%CI_b [0.059;0.127]**, ovvero coloro che hanno un voto in matematica più alto hanno un voto maggiore al test di statistica.

ESEMPIO #2

L'età influenza la capacità di copiare la figura di Rey?





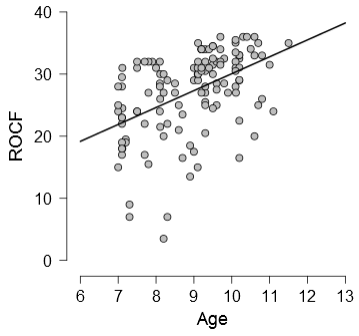
Regressione semplice con una variabile indipendente (Età; VI-R) e una variabile dipendente (ROCF; VD-R).

$$H_0 \Rightarrow \beta_i = 0$$

$$H_1 \Rightarrow \beta_i \neq 0$$

$$\alpha = .05$$

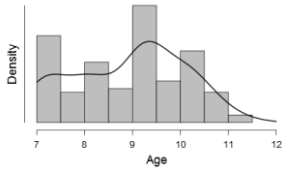
ESEMPIO #2



Pearson Correlations

		Age	ROCF
Age	Pearson's r	—	0.470
	p-value	—	< .001
ROCF	Pearson's r		—
	p-value		—

Dati reali
***N* = 127**



ESEMPIO #2

Model Summary

Model	R	R^2	Adjusted R^2	RMSE
1	0.470	0.221	0.215	5.941

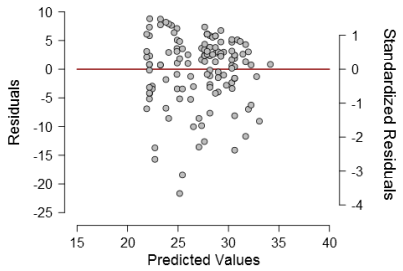
ANOVA

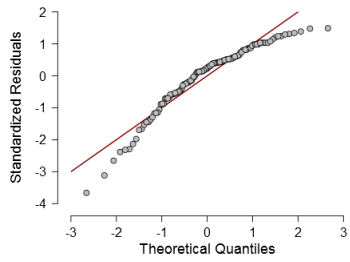
Model		Sum of Squares	df	Mean Square	F	p
1	Regression	1251	1	1250.92	35.44	< .001
	Residual	4412	125	35.29		
	Total	5663	126			

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p	2.5%	97.5%
1	intercept	2.833	4.120		0.687	0.493	−5.322	10.987
	Age	2.723	0.457	0.470	5.953	< .001	1.818	3.629

ESEMPIO #2





ESEMPIO #2