

METODI E TECNICHE DELLA RICERCA IN PSICOLOGIA CLINICA E LABORATORIO

AA 2017/2018

PROF. V.P. SENESE

http://psiclab.altervista.org/MetTecPsicClinica2018/2017_2018.html

<https://goo.gl/YEDxH1>

Seconda Università di Napoli (SUN) – Facoltà di Psicologia – Dipartimento di Psicologia – METODI E TECNICHE DELLA RICERCA IN PSICOLOGIA CLINICA – Prof. V.P. Senese

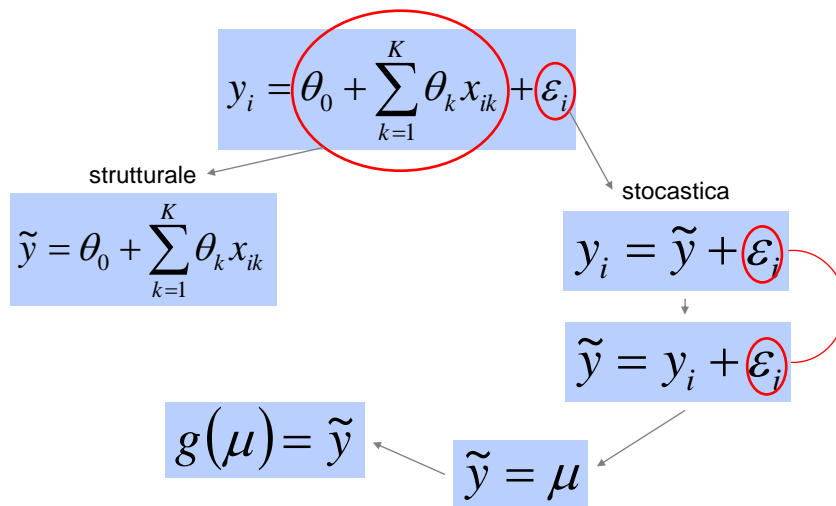
I MODELLI LINEARI GENERALIZZATI GLM

GLM

Secondo la teoria dei **Modelli Lineari Generalizzati (GLM)**, la **regressione lineare**, l'**analisi della varianza (ANOVA)**, la **regressione logistica** e i **modelli log-lineari** possono essere visti come casi speciali di una **classe più generale di modelli** che condividono: (a) alcune proprietà di base; (b) i metodi di stima dei parametri, e (c) le statistiche di *fit* (Miceli, 2001).

$$y_i = \theta_0 + \sum_{k=1}^K \theta_k x_{ik} + \varepsilon_i \quad \text{n. scalare}$$

$$y = X \cdot \theta + \varepsilon \quad \text{n. vettoriale}$$



Dove $g(\cdot)$ indica una generica funzione che viene detta "legame funzionale" (*link function*). Il modello lineare classico (**distribuzione gaussiana**) diventa così un caso particolare dei **GLM** dove il legame funzionale è quello dell'identità.

La **componente d'errore** può essere vista come la risultante delle variabili esplicative omesse (numerose), dell'errore casuale e dell'errore di misura casuale (Miceli, 2001).

GLM

Legame canonico	Legame funzionale	Distribuzione
• $\mu = \tilde{y}$	Identità	Normale (Gaussiana)
• $\log(\mu) = \tilde{y}$	Logaritmo	Poisson; Multinomiale; Prodotto multinomiale
• $\log\left(\frac{\mu}{1-\mu}\right) = \tilde{y}$	Logit	Binomiale; Multinomiale

LE ASSUNZIONI NEI GLM

ASSUNZIONI DEI GLM

- **misure**: tutte le **variabili indipendenti** sono misurate su scala ad **intervalli**, a **rapporti** o **dicotomica**;
- **modello**: la relazione tra **variabili indipendenti** e **dipendente** è **lineare**, **proporzionale** (invariante) e **additiva**;
- **specificazioni**: **tutti i predittori rilevanti** per la variabile dipendente sono stati inseriti nell'analisi, nessun **predittore irrilevante** è stato inserito (**parsimonia**);
- **valore atteso dell'errore**: gli errori sono esclusivamente di tipo casuale, relativi alla sola **variabile dipendente** e il valore atteso dell'errore ε (*epsilon*) è **0**;
- **omoschedasticità**: la varianza del termine d'errore ε è la stessa (o è costante) per tutti i valori delle variabili indipendenti;

ASSUNZIONI DEI GLM

- **no autocorrelazioni**: non ci devono essere correlazioni tra i termini dell'errore prodotti da ciascun predittore (matematicamente $\mathbf{E}(\varepsilon_i, \varepsilon_j) = 0$; oppure $\mathbf{COV}(\varepsilon_i, \varepsilon_j) = 0$) (**osservazioni indipendenti**);
- **no correlazioni tra errori e predittori**: i termini d'errore devono essere non correlati con le variabili indipendenti (matematicamente $\mathbf{E}(\varepsilon_j, X_i) = 0$);
- **assenza di perfetta multicollinearità**: nessuna delle variabili indipendenti deve essere una combinazione lineare perfetta delle altre variabili indipendenti (matematicamente, per ogni variabile i $R^2_i < 1$, dove R^2_i è la varianza della variabile indipendente X_i spiegata da tutti gli altri predittori (X_1, X_2, \dots, X_k)).

LA REGRESSIONE

LA REGRESSIONE

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 **covarianza** e **causazione**.

LA REGRESSIONE

COVARIAZIONE

(Covarianza, Correlazione o Associazione):
quando “semplicemente” osserviamo che due variabili presentano **variazioni concomitanti**.

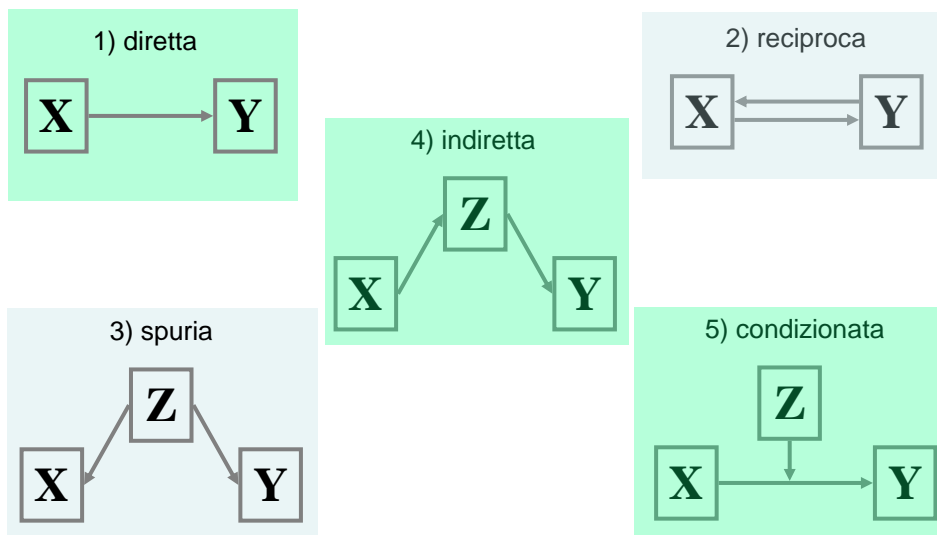
CAUSAZIONE:

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 covarianza è osservabile la causazione appartiene al dominio della teoria!!!

LA REGRESSIONE

I cinque fondamentali tipi di relazione causale fra due variabili:



LA REGRESSIONE LINEARE SEMPLICE

LA REGRESSIONE LINEARE

Quando la relazione si riferisce a **due variabili** di tipo **cardinale** (**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** aumenta conoscendo i valori assunti da una data variabile **x**.

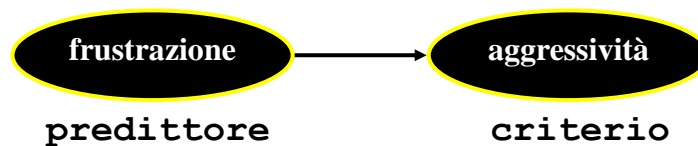
Se supponiamo che il punteggio y_i **dipende** dal punteggio x_i del soggetto, possiamo **prevedere** il valore in base alla seguente formula:

$$Y_i = \alpha + \beta X + \varepsilon$$

In pratica ipotizziamo che (mantenendo la **componente stocastica**) **se la teoria è vera**, allora la media di y è funzione di x .

LA REGRESSIONE

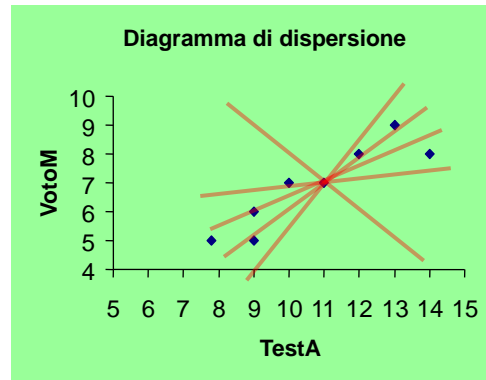
La **regressione lineare** si dice **semplice** quando abbiamo **una sola VD** (o **criterio**) e **una sola VI** (o **predittore**). L'ipotesi che viene formulata riguarda l'influenza della **VI** sulla **VD**.



$$\hat{Y} = \underbrace{\alpha}_{\text{costante}} + \underbrace{\beta}_{\text{coefficiente}} \underbrace{x}_{\text{predittore}} + \underbrace{\varepsilon}_{\text{errore}}$$

LA REGRESSIONE

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**.

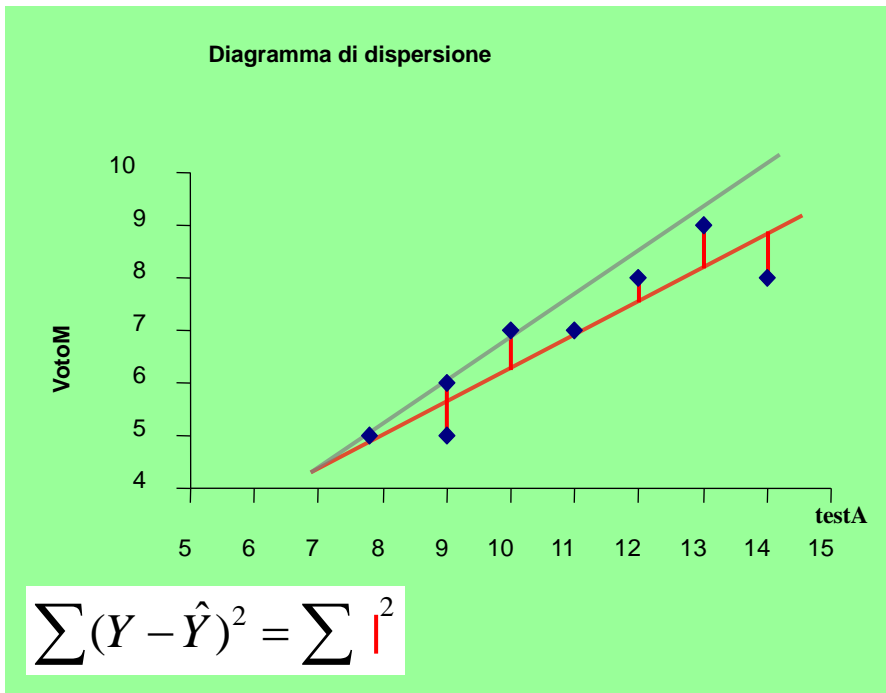
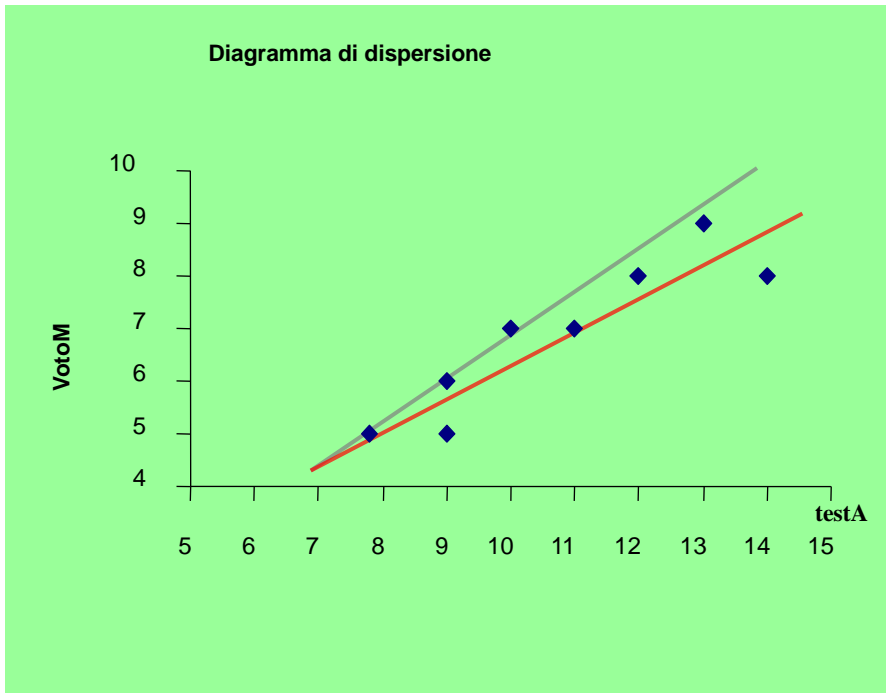


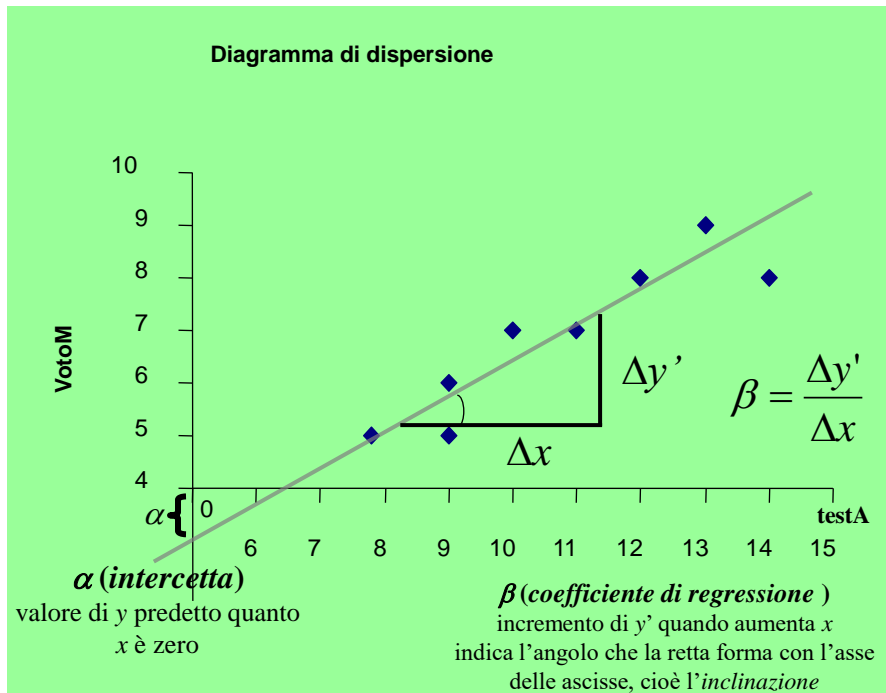
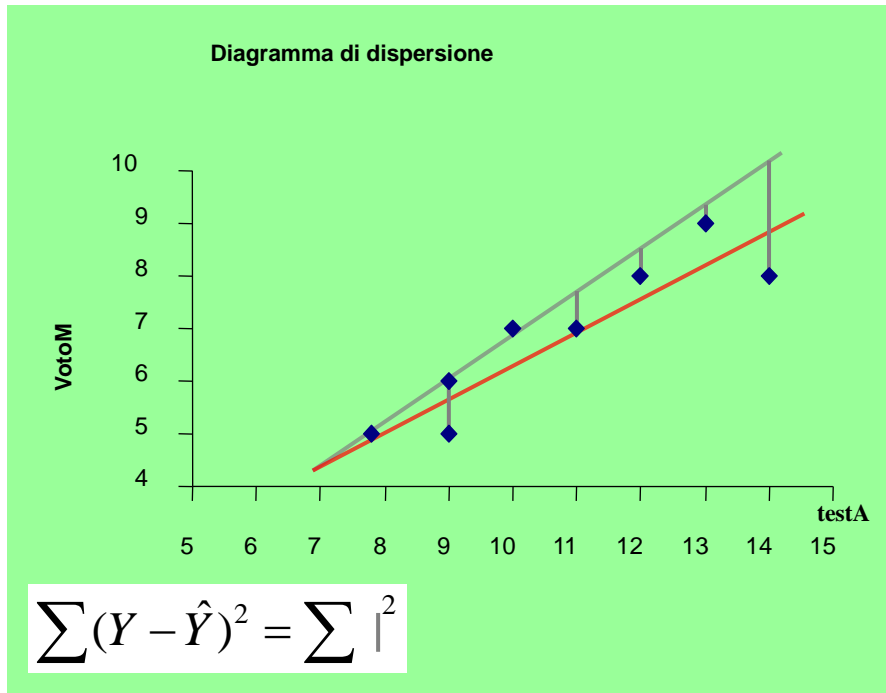
LA REGRESSIONE

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).

Assecondando questo principio, secondo la teoria classica, la **retta di regressione** si sceglie in base al **metodo dei minimi quadrati**. Si definisce "migliore" la retta che rende minima la **somma dei quadrati degli errori**, cioè:

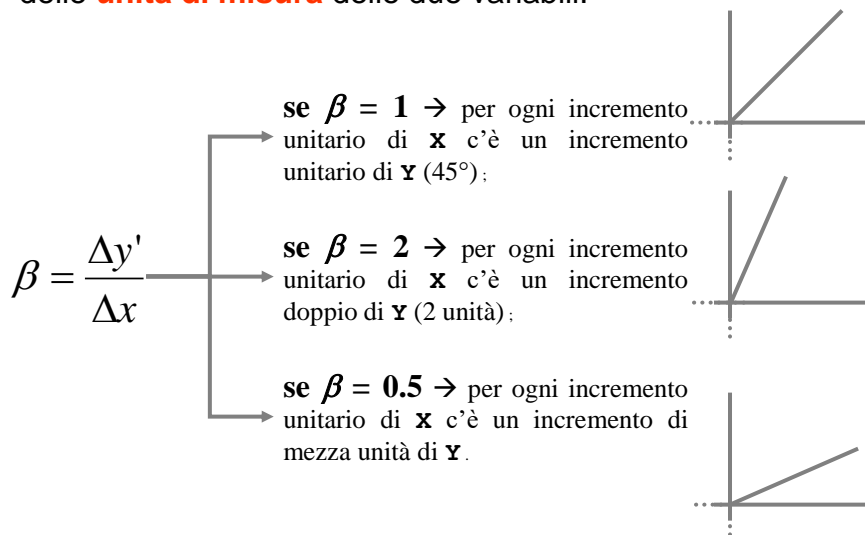
$$\sum (Y - \hat{Y})^2 = \text{più piccolo possibile}$$





COEFFICIENTE DI REGRESSIONE

Esprime la relazione tra x e y nei termini delle **unità di misura** delle due variabili.

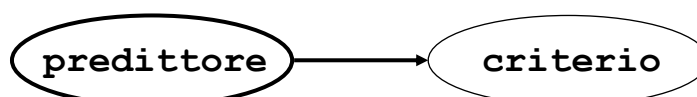


COEFFICIENTE DI REGRESSIONE STANDARDIZZATO

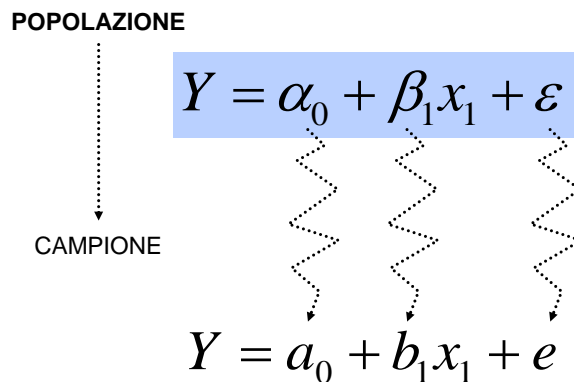
Il *coefficiente di regressione standardizzato* (β) esprime la relazione tra la variabile dipendente (y) e la variabile indipendente (x) in **unità di misura standard** (punti z).

COEFFICIENTE DI DETERMINAZIONE

Il *coefficiente di determinazione* (r^2) indica la percentuale di **varianza** (%) della variabile criterio (y) "spiegata" da quella predittore (x).



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



Il coefficiente di regressione è simboleggiato come:

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

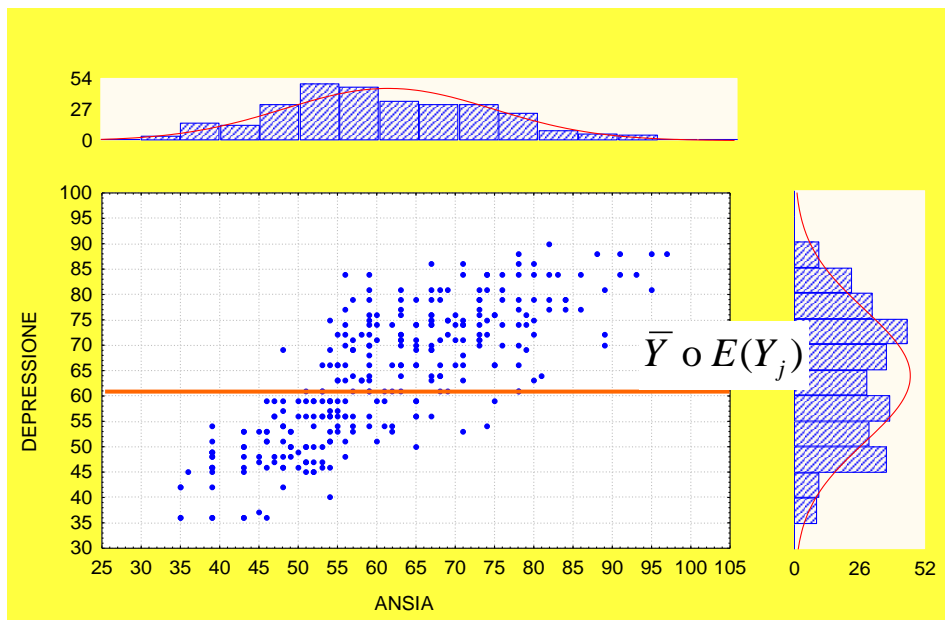
b quando ci si riferisce al coefficiente calcolato nel campione;

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

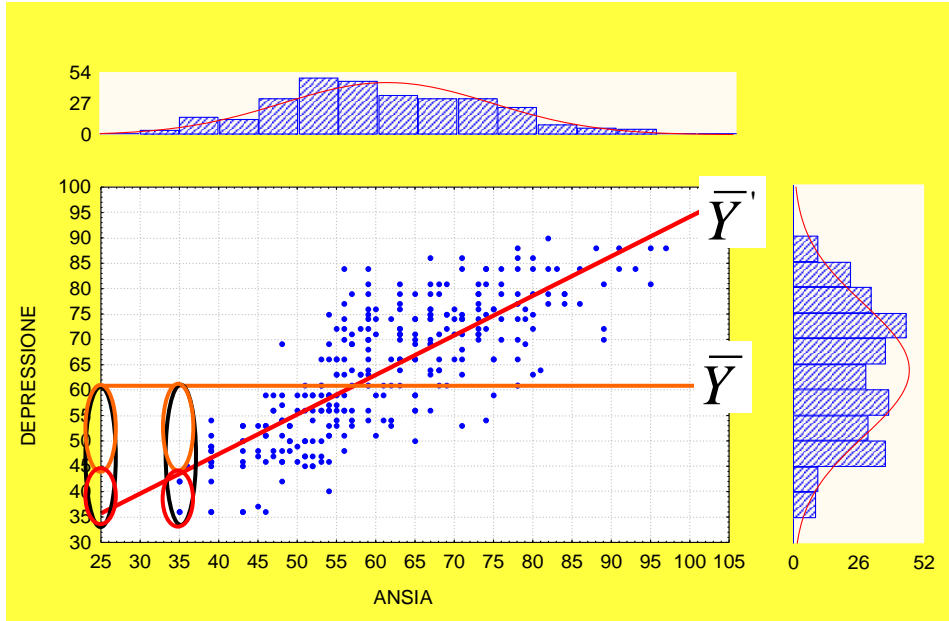
..PRIMA DI ENTRARE NEL VIVO...

LA REGRESSIONE BIVARIATA DA UN PUNTO DI VISTA CONCETTUALE

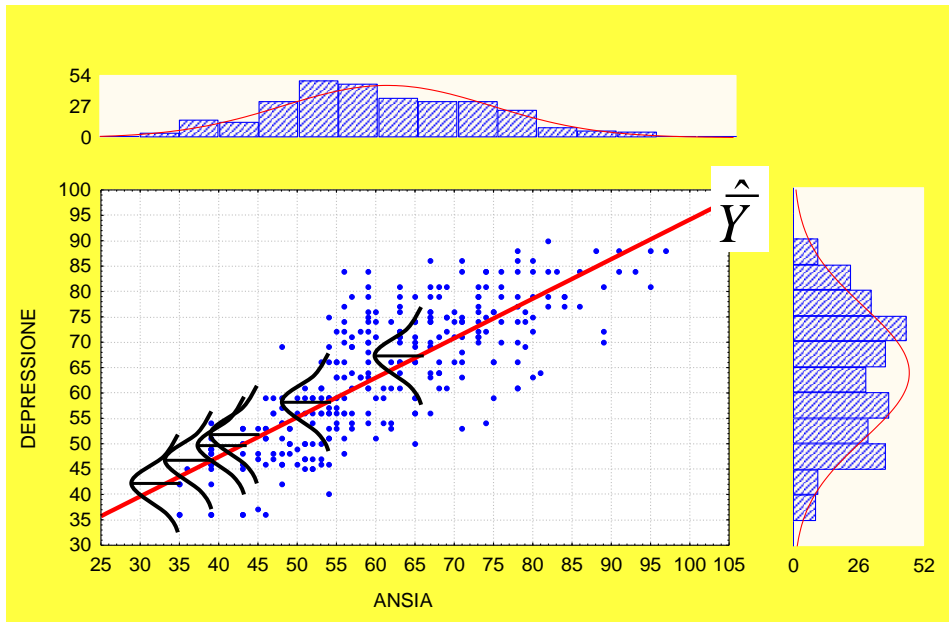
$$\bar{Y}_{depressione} = 61; ds = 13$$



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



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



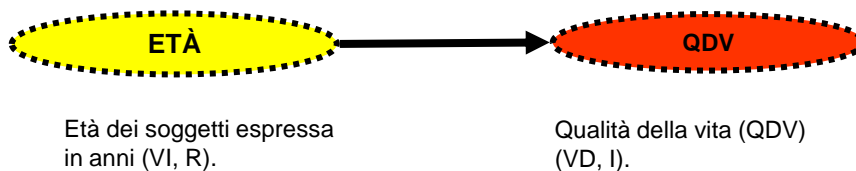
..ENTRIAMO NEL VIVO...

ESERCITAZIONE

REGRESSIONE SEMPLICE

Uno **psicologo** è interessato a verificare se la **qualità della vita** dipende dall'**età**. A tal scopo somministra ad un campione di **8 soggetti** una misura di **qualità della vita (QDV)** e rileva per ciascun soggetto l'**età**.

MODELLO TEORICO



Variables Entered/Removed^b

Model	Variables Entered	Variables Removed	Method
1	age ^a		Enter

a. All requested variables entered.
b. Dependent Variable: qdv

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.351 ^a	.123	-.023	2.961

a. Predictors: (Constant), age

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	7.407	1	7.407	.845	.393 ^a
	Residual	52.593	6	8.765		
	Total	60.000	7			

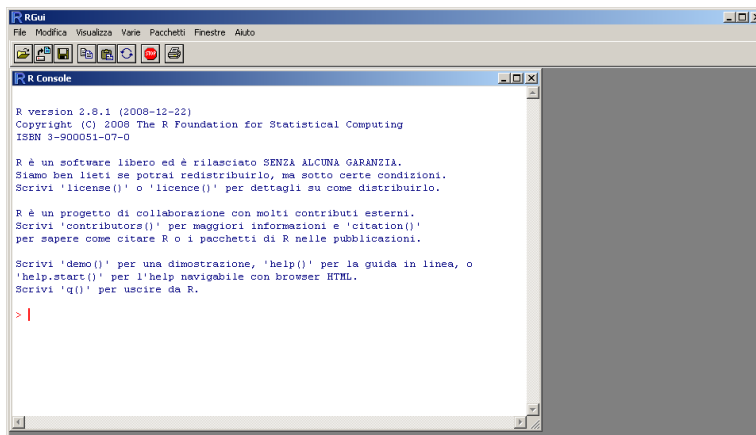
a. Predictors: (Constant), age
b. Dependent Variable: qdv

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	35.130	27.618		1.272	.250
	age	.741	.806	.351	.919	.393

a. Dependent Variable: qdv

Annotations:
 r^2 points to R Square (.123)
errore standard della stima points to Std. Error of the Estimate (2.961)
F points to F (.845)
 $\beta_{\text{standardizzato}}$ points to Standardized Coefficients (Beta) (.351)
correlazione points to R (.351)
b a points to Unstandardized Coefficients (B) (.741)



Notazione scientifica:

http://it.wikipedia.org/wiki/Notazione_scientifica

$$1.258e - 6 = 0.000001258$$

$$\left[\begin{array}{l} (x \cdot e - 3) = 0.00x \\ (y \cdot e - 5) = 0.0000y \end{array} \right.$$

```

> glm0<-glm(qdv ~ 1, family=gaussian) #Calcolo il modello nullo (H0)
>
> glm1<-glm(qdv ~ age, family=gaussian) #Calcolo il modello 1 (H1)
>
> anova(glm0, glm1, test="F") #Faccio il confronto tra i modelli (H0 vs H1)
Analysis of Deviance Table

Model 1: qdv ~ 1
Model 2: qdv ~ age
  Resid. Df Resid. Dev Df Deviance    F Pr(>F)
1         7    60.000
2         6    52.593  1     7.407 0.8451 0.3934
> summary(glm1)

Call:
glm(formula = qdv ~ age, family = gaussian)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.7963 -1.6389 -0.6852  1.5741  4.6852

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  35.1296    27.6180   1.272   0.250
age           0.7407     0.8058   0.919   0.393

(Dispersion parameter for gaussian family taken to be 8.765432)

    Null deviance: 60.000  on 7  degrees of freedom
Residual deviance: 52.593  on 6  degrees of freedom
AIC: 43.768

Number of Fisher Scoring iterations: 2

```

```

> cor.test(age, qdv)
Pearson's product-moment correlation
data: age and qdv
t = 0.9193, df = 6, p-value = 0.3934
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.4695737  0.8464570
sample estimates:
cor
0.3513642

```

..ENTRIAMO NEL VIVO...

INTRODUZIONE ALLA REGRESSIONE MULTIPLA

NELLA FORMA **GENERALE** DEL MODELLO DI REGRESSIONE LA VARIABILE DIPENDENTE Y VIENE CONSIDERATA COME FUNZIONE DI k VARIABILI INDIPENDENTI (X_{1j} ; X_{2j} ; X_{3j} ; ...; X_{kj}).

IL MODELLO DELLA REGRESSIONE LINEARE ASSUME CHE DATO UN SET DI VARIABILI INDIPENDENTI IL **VALORE MEDIO** (VALORE ATTESO) DELLA VARIABILE DIPENDENTE SI MODIFICA SECONDO LA SEGUENTE FORMULA:

$$E(Y_j) = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + \beta_3 X_{3j} + \dots + \beta_k X_{kj} + \varepsilon_j$$

$$E(Y_j) = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + \beta_3 X_{3j} + \dots + \beta_k X_{kj} + \varepsilon_j$$

$$\alpha, \beta_1, \beta_2, \beta_3, \beta_k, \varepsilon$$

LE **LETTERE GRECHE** RAPPRESENTANO I PARAMETRI CHE **ESPRIMONO LA RELAZIONE** TRA LE k VI E LA VD NELLA POPOLAZIONE

$$\beta_1, \beta_2, \beta_3, \beta_k$$

RAPPRESENTANO IL **COEFFICIENTE DI REGRESSIONE PARZIALE** TRA CIASCUNA DELLE k VI E LA VD **MANTENENDO COSTANTI** (**CONTROLLANDO**) TUTTE LE ALTRE VARIABILI.

$$E(Y_j) = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + \beta_3 X_{3j} + \dots + \beta_k X_{kj} + \varepsilon_j$$

α

È L'INTERCETTA E RAPPRESENTA IL VALORE ATTESO DI Y QUANDO **TUTTE LE VI SONO UGUALI A ZERO**.

ε

PER OGNI VALORE Y PREDETTO (MANTENENDO COSTANTI I VALORI NELLE VI) IL MODELLO PREVEDE UNA COMPONENTE D'**ERRORE** (COMPONENTE STOCASTICA) QUESTO TERMINE D'ERRORE RAPPRESENTA: 1) L'EFFETTO SULLA VD Y NON ESPLICITAMENTE INCLUSO NEL MODELLO; 2) UN RESIDUO CASUALE NELLA VARIABILE DIPENDENTE.

$$E(Y_j) = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + \beta_3 X_{3j} + \dots + \beta_k X_{kj} + \varepsilon_j$$

SEBBENE SIA IMPLICITO NELLA FORMULAZIONE DEL MODELLO È **IMPORTANTE SOTTOLINEARE** CHE LA RELAZIONE TRA $E(Y_j)$ E CIASCUN X_{kj} È CONCEPTA COME **LINEARE** E CHE GLI EFFETTI DELLE k VI SONO **ADDITIVI**.

PER UNA CORRETTA APPLICAZIONE DEL MODELLO DELLA REGRESSIONE, QUINDI, PER UNA CORRETTA STIMA DEI PARAMETRI DELLA POPOLAZIONE E PER LA VERIFICA DELLE IPOTESI È **NECESSARIO** CHE ALCUNE **ASSUNZIONI** SIANO VERIFICATE.

NELL'APPLICAZIONE DELLA REGRESSIONE MULTIPLA **NON** CI TROVIAMO NELLA CONDIZIONE DI CONOSCERE I **PARAMETRI DELLA POPOLAZIONE** DIRETTAMENTE, MA SI **STIMANO** A PARTIRE DA UN NUMERO FINITO DI OSSERVAZIONI: LE **OSSERVAZIONI CAMPIONARIE**

PER DISTINGUERE LA **REGRESSIONE CAMPIONARIA** DA QUELLA DELLA **POPOLAZIONE** IL MODELLO DI REGRESSIONE VIENE SCRITTO IN QUESTO MODO:

$$E(Y_j) = a + b_1 X_{1j} + b_2 X_{2j} + b_3 X_{3j} + \dots + b_k X_{kj} + e_j$$

DOVE LE **LETTERE LATINE** INDICANO I PARAMETRI DEL MODELLO **STIMATI A PARTIRE DAL CAMPIONE** (n)

PER LA **STIMA DEI PARAMETRI** a E b_i ($i = 1, 2, \dots, k$) IL METODO PIÙ FREQUENTEMENTE IMPIEGATO È IL CRITERIO DEI MINIMI QUADRATI (**ORDINARY LEAST SQUARE – OLS**).

LO SCOPO È QUELLO DI STIMARE I PARAMETRI a E b_i IN MODO TALE CHE SI **RIDUCA AL MINIMO** LA DISTANZA AL QUADRATO TRA VALORE **PREDETTO** (\hat{Y}_j) E VALORE **OSSERVATO** (Y_j)

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

NELLA **REGRESSIONE BIVARIATA** LE FORMULE SONO LE SEGUENTI:

$$b_i = \frac{\sum_{j=1}^n (X_{ij} - \bar{X}_i)(Y_j - \bar{Y})}{\sum_{j=1}^n (X_{ij} - \bar{X}_i)^2} \quad a = \bar{Y} - b_i \bar{X}$$

NELLA **REGRESSIONE MULTIPLA** LE FORMULE PER IL CALCOLO DEI PARAMETRI RICHIEDONO L'ALGEBRA MATRICIALE.

DAL MOMENTO CHE SI TRATTA DI **STIME CAMPIONARE** DEI PARAMETRI È NECESSARIO CONOSCERE L'EFFETTO DELL'**ERRORE** SULLA STIMA. PER FARE CIÒ È NECESSARIO CALCOLARE L'**ERRORE STANDARD** (s_{b_i}) DEL COEFFICIENTE STIMATO:

$$s_{b_i} = \sqrt{\frac{\sum_{j=1}^n (Y_j - \hat{Y}_j)^2}{\sum_{j=1}^n (X_{ij} - \bar{X}_i)^2 (1 - R_i^2) (n - k - 1)}}$$

DOVE: n È L'AMPIEZZA CAMPIONARIA; k È IL NUMERO DI VI DEL MODELLO; R_i^2 È LA CORRELAZIONE MULTIPLA AL QUADRATO DELLA VI_i SU TUTTE LE ALTRE VI .

$$s_{b_i} = \sqrt{\frac{\sum_{j=1}^n (Y_j - \hat{Y}_j)^2}{\sum_{j=1}^n (X_{ij} - \bar{X}_i)^2 (1 - R_i^2) (n - k - 1)}}$$

DELLA FORMULA È UTILE NOTARE CHE L'ERRORE DI STIMA DI b_i (s_{b_i}) **SI RIDUCE** SE:

- AL NUMERATORE: È MINORE L'ERRORE DI STIMA DI Y_j
- AL DENOMINATORE:
 - È MAGGIORE LA VARIANZA DI X_i
 - È MINORE LA CORRELAZIONE DI X_i CON LE ALTRE V_i
 - È MAGGIORE IL NUMERO DELLE OSSERVAZIONI n
(SE IL NUMERO DI PREDITTORI AUMENTA E SI APPROSSIMA ALL'AMPIEZZA CAMPIONARIA, s AUMENTA NOTEVOLMENTE)

UN ALTRO ASPETTO UTILE ALLA VALUTAZIONE DEL MODELLO DI REGRESSIONE È LA VALUTAZIONE DELLA **BONTÀ DI ADATTAMENTO** DEL MODELLO (*goodness-of-fit*). LA STATISTICA MAGGIORMENTE IMPIEGATA È L' R^2 , CHE VIENE STIMATA CON LE SEGUENTI FORMULE:

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

o

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

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

$$R^2 = 1 - \frac{\text{dev errore}}{\text{dev totale}}$$

L' R^2 VARIA **SEMPRE** TRA 0 E 1. PUÒ ESSERE INTERPRETATO COME LA **PERCENTUALE DI VARIANZA** (%) DELLA **VD SPIEGATA** DALLE **VI** CONSIDERATE NEL MODELLO. OPPURE COME LA **% DI RIDUZIONE DELL'ERRORE** NELLA PREVISIONE DELLA **VI**.

NELL'UTILIZZO DELL' R^2 DUE ASPETTI DEVONO ESSERE SOTTOLINEATI:

- **È DIPENDENTE DAL CAMPIONE**, DUE MODELLI APPLICATI SU DUE CAMPIONI POSSONO AVERE DEI PARAMETRI b IDENTICI MA R^2 DIFFERENTI; QUESTO È DETERMINATO DALLA DIVERSA VARIANZA DI Y;
- **È INFLUENZATO DAL NUMERO DI PREDITTORI**, A PARITÀ DI CAMPIONE PER CONFRONTARE DUE MODELLI È NECESSARIO CALCOLARE UN VALORE CORRETTO (**ADJUSTED R^2**) (WONNACOTT, WONNACOTT, 1979).

$$\bar{R}^2 = \left(R^2 - \frac{k}{n-1} \right) \left(\frac{n-1}{n-k-1} \right)$$

UN'ALTRA STATISTICA COMUNEMENTE IMPIEGATA PER LA VALUTAZIONE DELLA **BONTÀ DI ADATTAMENTO** DEL MODELLO (*goodness-of-fit*) È L'**ERRORE STANDARD DELLA STIMA**, CHE VIENE STIMATO CON LA SEGUENTE FORMULA:

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

$$s_e = \sqrt{\frac{\text{devianza_residua}}{\text{gdl}(dev_res)}}$$

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** dalla regressione SQ_{reg}

DEVIANZA **TOTALE** SQ_{tot}

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

SQ_{err}
DEVIANZA **NON SPIEGATA**
o **RESIDUA** (somma di e)

NELLA RICERCA PSICOLOGICA NON SIAMO INTERESSATI ESCLUSIVAMENTE ALLA **STIMA DEI PARAMETRI** DELLA POPOLAZIONE, MA SIAMO INTERESSATI A VOLER **VERIFICARE SE I PARAMETRI CAMPIONARI SONO VICINI A QUELLI DELLA POPOLAZIONE**, VALE A DIRE ALLA **VERIFICA DELLE IPOTESI**.

CIÒ AVVIENE MEDIANTE IL **TEST DELLA SIGNIFICATIVITÀ STATISTICA** CHE VALUTA LO SCOSTAMENTO DEL PARAMETRO OSSERVATO DAL VALORE ATTESO SECONDO L'**IPOTESI NULLA** (H_0).

Per verificare se la previsione è significativa la **varianza spiegata** dalla regressione deve essere **maggiore** di quella residua.

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

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

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

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 $\Rightarrow F_{critico(k, N-k-1)}$.

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

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

PER IL **MODELLO COMPLESSIVO**, CON k VI, L'IPOTESI NULLA (H_0) è LA SEGUENTE:

$$H_0 \Rightarrow \beta_1 = \beta_2 = \beta_3 = \dots = \beta_k = 0$$

$$H_1 \Rightarrow \beta_1 \text{ o } \beta_2 \text{ o } \beta_3 \text{ o } \dots \text{ o } \beta_k \neq 0$$

Un modo **alternativo** per definire il **test statistico** della verifica è mediante il valore dell' R^2 :

$$F = \frac{\frac{R^2}{k}}{\frac{(1-R^2)}{(n-k-1)}}$$

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

$$gdl_F = \frac{k}{n-k-1}$$

PER **CIASCUN PREDITTORE** VIENE POI DEFINITA UNA SPECIFICA IPOTESI NULLA (H_0).

$$H_0 \Rightarrow \beta_i = 0$$

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

IL TEST STATISTICO APPROPRIATO È IL VALORE t :

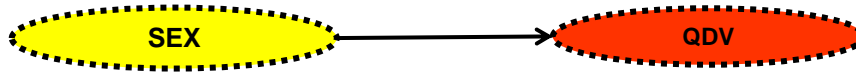
$$t = \frac{b_i - \beta_{iH_0}}{s_{b_i}} \Rightarrow \frac{b_i}{s_{b_i}}$$

$$\text{gdl}_t = n - k - 1$$

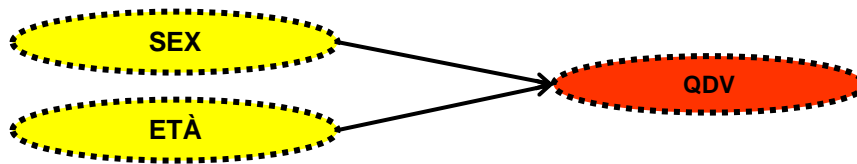
ESERCITAZIONE INTRODUZIONE ALLA REGRESSIONE MULTIPLA

ESEMPIO

MODELLO TEORICO 1



MODELLO TEORICO 2



VARIABILI DUMMY

Il modello della regressione lineare può essere esteso facilmente per inserire predittori misurati su scala dicotomica, inclusi set di variabili dicotomizzate o **variabili dummy** (si veda Lewis-Beck, 1980; Berry e Feldman, 1985; Hardy, 1993).

Es.

SESSO: 1=M; 2=F;

LIVELLO SOCIO-ECONOMICO: 1=Basso; 2=Medio; 3=Alto.

COD	SESSO	LSE
1	1	1
2	2	2
...	...	
100	1	3

COD	M	MEDIO	ALTO
1	1	0	0
2	0	1	0
...	...		
100	1	0	1

NdE -1

Qualità della vita (QDV)

COD	SEX	QDV
1	0	61
2	1	61
3	0	59
4	1	57
5	1	63
6	0	57
7	1	60
8	0	58

$$QDV = \alpha + \beta(SEX)$$

$$QDV = \alpha + \beta(0) = \alpha \quad \text{Media femmine}$$

$$QDV = \alpha + \beta(1) = \alpha + \beta \quad \text{Media maschi}$$

Group Statistics

	sex	N	Mean	Std. Deviation	Std. Error Mean
qdv	0	4	58.00	.816	.408
	1	4	63.00	1.633	.816

OUTPUT REGRESSIONE

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.913 ^a	.833	.806	1.291

a. Predictors: (Constant), sex

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	58.000	.645		89.853	.000
	sex	5.000	.913	.913	5.477	.002

a. Dependent Variable: qdv

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	50.000	1	50.000	30.000	.002 ^a
	Residual	10.000	6	1.667		
	Total	60.000	7			

a. Predictors: (Constant), sex

b. Dependent Variable: qdv

$$QDV_F = \alpha = 58$$

$$QDV_M = \alpha + \beta = 58 + 5 = 63$$

MODELLO TEORICO 1

```
> aggregate(prova[,c(3,4)],list(prova$sex),mean)#Medie differenziate per fattore
  Group.1  age qdv
1      0 33.75  58
2      1 34.75  63
```

```
> anova(glm0, glm2, test="F") #Confronto tra i modelli
Analysis of Deviance Table

Model 1: qdv ~ 1
Model 2: qdv ~ sex
  Resid. Df Resid. Dev Df Deviance  F    Pr(>F)
1         7         60          1      30 0.001547 **
2         6         10          1      50 0.001547 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> summary(glm2)

Call:
glm(formula = qdv ~ sex, family = gaussian)

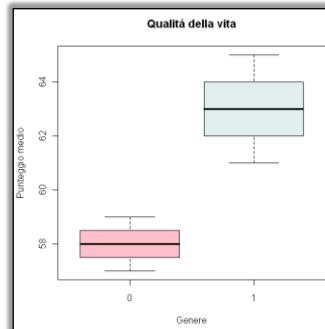
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.000e+00 -2.500e-01  3.553e-15  2.500e-01  2.000e+00

Coefficients:
(Intercept)  58.0000    0.6455  89.853  1.28e-10 ***
sex           5.0000    0.9129  5.477  0.00155 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 1.666667)

Null deviance: 60 on 7 degrees of freedom
Residual deviance: 10 on 6 degrees of freedom
AIC: 30.488

Number of Fisher Scoring iterations: 2
```



MODELLO TEORICO 2

```
> glm3<-glm(qdv ~ age + sex, family=gaussian) #Calcola il modello 3 (H1)
```

```
> anova(glm0, glm3, test="F") #Confronto tra i modelli
Analysis of Deviance Table

Model 1: qdv ~ 1
Model 2: qdv ~ age + sex
  Resid. Df Resid. Dev Df Deviance    F    Pr(>F)
1         7         60      2      50 12.5 0.01134 *
2         5         10      2       5  1.0 0.00000
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> summary(glm3) #Riassume il modello 3

Call:
glm(formula = qdv ~ age + sex, family = gaussian)

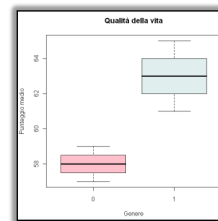
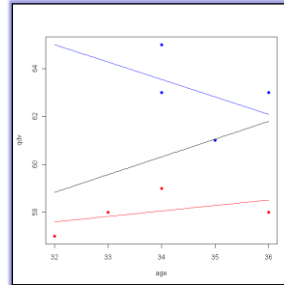
Deviance Residuals:
    1     2     3     4     5     6 
1.421e-14 -2.000e+00  1.000e+00  2.000e+00  7.105e-15 -1.000e+00
    7     8 
0.000e+00  0.000e+00

Coefficients:
(Intercept)  5.800e+01  1.409e+01  4.116  0.00921 **
age          -2.674e-15  4.170e-01 -6.41e-15  1.000000
sex           5.000e+00  2.083e+00  4.611  0.00576 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 2)

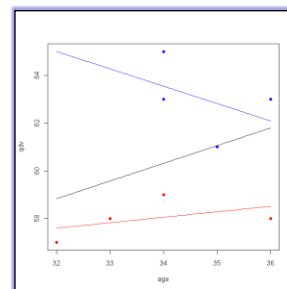
Null deviance: 60 on 7 degrees of freedom
Residual deviance: 10 on 5 degrees of freedom
AIC: 32.488

Number of Fisher Scoring iterations: 2
```

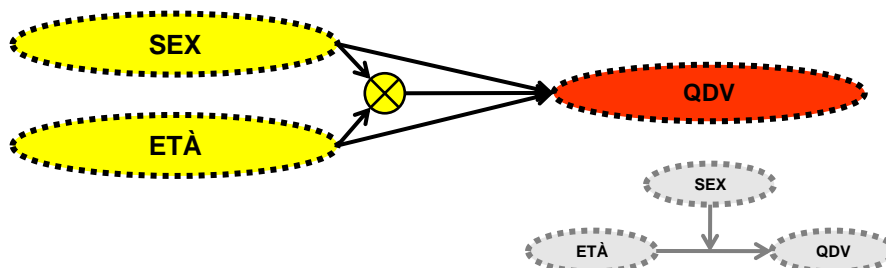


ESEMPIO

Nel modello 1 e nel modello 2 la relazione tra età e qualità della vita è stata **forzata** essere uguale per gli **uomini** e per le **donne**. Tuttavia è possibile (vedi Figura) che ci sia una differenza significativa. Un modo per verificare questa ipotesi è inserire l'**interazione tra le variabili**. Vale a dire una nuova variabile che è il prodotto tra le due (**sessoxetà**)



MODELLO TEORICO 3



MODELLO TEORICO 3

```
> glm4<- (glm(qdv ~ z.age*sex, family=gaussian)) #Calcolo il modello 4 (H1)
```

```
> anova(glm0, glm4, test="F") #Faccio il confronto tra i modelli
Analysis of Deviance Table

Model 1: qdv ~ 1
Model 2: qdv ~ z.age * sex
Resid. Df Resid. Dev Df Deviance   F Pr(>F)
1         7    60.000
2         4    8.088 3    51.912 8.5575 0.0325 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> summary(glm4) #Riassumo il modello trovato

Call:
glm(formula = qdv ~ z.age * sex, family = gaussian)

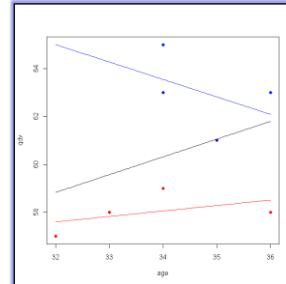
Deviance Residuals:
    1     2     3     4     5     6     7     8
-0.5143 -1.8182  0.9429  1.4545  0.9091 -0.6000 -0.5455  0.1714

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  58.1143    0.7505  77.431 1.67e-07 ***
z.age        0.3174    0.6676   0.475 0.65925
sex          5.2494    1.1192   4.690 0.00938 **
z.age:sex    -1.3274    1.3652  -0.972 0.38595
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 2.022078)

Null deviance: 60.0000 on 7 degrees of freedom
Residual deviance: 8.0883 on 4 degrees of freedom
AIC: 32.791

Number of Fisher Scoring iterations: 2
```



TECNICHE DI REGRESSIONE MULTIPLA

NELLA PRATICA LA REGRESSIONE MULTIPLA PUÒ ESSERE USATA UTILIZZANDO **DIVERSE STRATEGIE**.

TALI STRATEGIE DIFFERISCONO PREVALENTEMENTE NEL CRITERIO CHE DEFINISCE **L'ORDINE DI "ENTRATA" DELLE VI NELL'EQUAZIONE** DI REGRESSIONE.

L'ORDINE DI INTRODUZIONE, INFATTI, DETERMINA LA PARTE DI VARIANZA DELLA VD UTILIZZATA PER LA VERIFICA DELLE IPOTESI SU CIASCUNA VI.

LA **PRIMA** VARIABILE HA A DISPOSIZIONE TUTTA LA VARIANZA DELLA **VD** (**100%**), LA SECONDA AVRÀ A DISPOSIZIONE SOLO LA **VARIANZA RESIDUA**, E COSÌ PER TUTTE LE SUCCESSIVE VARIABILI.

TRE SONO LE STRATEGIE MAGGIORMENTE IMPIEGATE NELLA PRATICA:

- **LA REGRESSIONE STANDARD** (**ESPLICATIVA**). CONSENTE DI VERIFICARE L'ENTITÀ DELLA **RELAZIONE COMPLESSIVA** TRA **VI** E **VD**, E IL CONTRIBUTO SPECIFICO DI CIASCUNA **VI** CONTROLLATO PER TUTTE LE **VI** IN EQUAZIONE.
- **LA REGRESSIONE GERARCHICA** (**COMPARATIVA**). CONSENTE DI VALUTARE QUAL È IL CONTRIBUTO AGGIUNTIVO DELLA/E VARIABILE/I X_2 INSERITA/E DOPO X_1 .
- **LA REGRESSIONE STATISTICA** (**PREDITTIVA**). CONSENTE DI IDENTIFICARE LA **MIGLIORE COMBINAZIONE PREDITTIVA** TRA LE **VI** CONSIDERATE.

REGRESSIONE MULTIPLA STANDARD

Tutte le **VI** vengono inserite **contemporaneamente**. ognuna, infatti, è trattata **come se fosse l'ultima**.

ad ogni **VI** corrisponde solo quella parte di variabilità che condivide "**UNICAMENTE**" con la **VD**.

Viene quindi interpretato il modello complessivo e il contributo di ciascun predittore sulla **VD**. questa seconda interpretazione si avvale dell'utilizzo dei **coefficienti di regressione parziale**.

L'ampiezza dell'**R²** è determinata dalla **porzione unica** di ciascun predittore e dalla **porzione comune** a tutte le lariabili che aumenta all'aumentare della **collinearità** tra le **VI**.

REGRESSIONE MULTIPLA GERARCHICA

L'**ordine** di inserimento delle variabili viene **specificato dal ricercatore**. Ogni **VI** è valutata per **quanto aggiunge** nella spiegazione della **VD** rispetto a quanto è stato spiegato dalle variabili inserite precedentemente.

L'**ordine** viene stabilito dal ricercatore **in funzione delle considerazioni teoriche** o **logiche**.

Il cambiamento viene valutato mediante le variazioni osservate nei termini dell'**R²** la cui significatività e poi valutata mediante il valore **F**.

REGRESSIONE MULTIPLA GERARCHICA

Nella **regressione gerarchica**, i modelli sono confrontabili quando sono gerarchicamente organizzabili o **nidificati** o *nested*.

Un **modello A** (M_A) si dice *nested* in un **modello B** (M_B) se il **modello A** è composto da alcuni dei termini contenuti nel **modello B**, e non ve ne sono di diversi, mentre nel **modello B** vi sono anche **termini aggiuntivi**.

$$M_A = a + b$$

$$M_B = a + b + c$$

$$M_A - M_B = \Delta M_B$$

PER **PORRE A CONFRONTO DIFFERENTI MODELLI** È POSSIBILE UTILIZZARE LA STATISTICA **F** PER VALUTARE SE IL **CONTRIBUTO DIFFERENZIALE È SIGNIFICATIVO**.

$$E(Y_j) = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + \beta_3 X_{3j} + \beta_k X_{kj} + \beta_{k+1} X_{k+1j} + \beta_{k+2} X_{k+2j} + \beta_{k+r} X_{k+rj} + \varepsilon_j$$

IN QUESTO CASO SI È INTERESSATI A VERIFICARE L'EFFETTO CHE L'AGGIUNTA DEGLI r PREDITTORI HA NELLA FUNZIONALITÀ DEL MODELLO:

$$F = \frac{\frac{R^2 - R_m^2}{r}}{\frac{(1 - R^2)}{(n - k - r - 1)}}$$

$$H_0 \Rightarrow \beta_{k+1} = \beta_{k+2} = \beta_{k+r} = 0$$

DOVE R_m^2 CORRISPONDE AL COEFFICIENTE R^2 OTTENUTO **SENZA GLI r PREDITTORI**.

a	b	c	d	$Dev_{Totale} = a + b + c + d$
a	b	c	d	$R^2_{Totale} = \frac{a + b}{a + b + c + d}$
a	b	c	d	$R^2_{Semi-Parziale} = \frac{b}{a + b + c + d}$
a	b	c	d	$R^2_{Parziale} = \frac{b}{b + c + d}$

POSSIBILI R^2 E RELATIVE INTERPRETAZIONE

R^2 **totale** del modello → si ottiene facendo il rapporto tra $DEV_{spiegata}$ e DEV_{totale} della VD. Corrisponde alla capacità esplicativa **totale** di tutte le variabili nel modello (indistintamente).

R^2 **semi-parziale** (ΔR^2) → si ottiene facendo il rapporto tra $DEV_{spiegata}$ da una singola VI e DEV_{totale} della VD. Corrisponde alla capacità esplicativa **unica** di una singola VI rispetto alla variabilità totale della VD. Ovvero la parte della varianza spiegata attribuibile unicamente dalla variabile considerata.

R^2 **parziale** → si ottiene facendo il rapporto tra $DEV_{spiegata}$ dalla singola variabile e $(DEV_{totale} - DEV_{spiegata_dalle_altre_VI})$. Corrisponde alla capacità esplicativa **unica** di una singola VI rispetto alla variabilità della VD **non spiegata dalle altre** VI. Ovvero la proporzione della varianza residua del modello precedente spiegata dalla VI considerata.

REGRESSIONE MULTIPLA STATISTICA

L'**ordine** di inserimento delle variabili viene **determinato algebricamente**. generalmente il termine di riferimento è il **coefficiente di correlazione parziale**.

Esistono **tre principali tecniche**: *forward* (in cui si aggiungono le VI significativamente associate alla VD); *backward* (in cui si eliminano le VI non associate significativamente alla VD); *stepwise* (in cui si aggiungono le vi associate significativamente alla VD, ma se ai passaggi successivi perdono la forza associativa vengono eliminate).

MEDIANTE L'**R²** SI VALUTA IL **MODELLO FINALE**, SI VALUTA L'**ORDINE DI INGRESSO** DELLE VARIABILI E IL **CONTRIBUTO DI CIASCUNA**.

LE ASSUNZIONI NELLA REGRESSIONE MULTIPLA

INDIPENDENTEMENTE DALLA TECNICA SCELTA, PER UNA **CORRETTA APPLICAZIONE** DEL MODELLO DELLA REGRESSIONE, QUINDI, PER UNA **CORRETTA STIMA DEI PARAMETRI** DELLA POPOLAZIONE E PER LA **VERIFICA DELLE IPOTESI** È **NECESSARIO** CHE LE ASSUNZIONI PREVISTE DAL MODELLO SIANO VERIFICATE.

IN CASO DI VIOLAZIONE, IL RISCHIO IN CUI SI PUÒ INCORRERE DIPENDE DAL TIPO DI VIOLAZIONE OSSERVATA.

..ASSUNZIONI...

- **TUTTE LE VARIABILI** DEVONO ESSERE MISURATE SU SCALA ALMENO AD **INTERVALLI** E **SENZA ERRORE**
- LA VARIABILE DIPENDENTE È **FUNZIONE LINEARE** DELLA COMPONENTE DETERMINISTICA ($X_{1j} + X_{2j} + X_{3j} + \dots + X_{kj}$)
- PER OGNI SET DELLE k VARIABILI INDIPENDENTI ($X_{1j}; X_{2j}; X_{3j}; \dots; X_{kj}$), **$E(\varepsilon_j) = 0$**
- PER OGNI SET DELLE k VARIABILI INDIPENDENTI, **$VAR E(\varepsilon_j) = \sigma^2$** (COSTANTE)

..ASSUNZIONI...

- PER OGNI COPPIA DELLE k VARIABILI INDIPENDENTI, $\text{COV}(\varepsilon_j, \varepsilon_h) = 0$ (GLI ERRORI NON DEVONO ESSERE COMUNI)
- PER OGNI VARIABILE INDIPENDENTE X_i , $\text{COV}(X_i, \varepsilon) = 0$
- NON CI DEVE ESSERE UNA **PERFETTA COLLINEARITÀ** TRA LE V_i NEL MODELLO
- PER OGNI SET DELLE k VARIABILI INDIPENDENTI ε_j DEVE ESSERE NORMALMENTE DISTRIBUITO

Se i primi **6 assunti** sono rispettati (in base al teorema di **Gauss-Markov**) è possibile affermare che le formule di stima derivate dal principio dei minimi quadrati (LS) sono efficienti e senza *bias*; e vengono dette **BLUE (BEST LINEAR UNBIASED ESTIMATOR)**. Il teorema, tuttavia, vale solo se gli assunti sono rispettati.

In genere, il **metodo più utile** per verificare l'adeguatezza del modello è l'**analisi dei residui** dei valori stimati dalla regressione per ogni valore osservato:

$$e = Y_j - \hat{Y}_j$$

L'ASSUNTO DELLA MULTICOLLINEARITÀ

UNA PRIMA DISTINZIONE DEVE ESSERE FATTA TRA LA PERFETTA MULTICOLLINEARITÀ E LE FORME MENO ESTREME DI MULTICOLLINEARITÀ.

LA **PERFETTA COLLINEARITÀ** ESISTE QUANDO UNA O PIÙ **VI** È PERFETTAMENTE CORRELATA ($r = 1$) AD UNA O PIÙ DELLE ALTRE **VI** NELL'EQUAZIONE.

$$X_1 = 2.3X_2 + 3 \quad \text{oppure} \quad X_2 = 2X_3$$

FORTUNATAMENTE NELLA PRATICA PSICOLOGICA NON CAPITANO QUASI MAI CASI DI QUESTO TIPO (DOVE LA STIMA DEI PARAMETRI RISULTA NON POSSIBILE).

MOLTO PIÙ SPESSO ABBIAMO A CHE FARE CON IL CASO IN CUI SI ASSISTE A **FORME MENO ESTREME DI COLLINEARITÀ**.

NEGLI **ESPERIMENTI**, AD ESEMPIO, QUESTO PROBLEMA VIENE PERFETTAMENTE RISOLTO DAL MOMENTO CHE LE VARIABILI SONO MANIPOLATE DALLO SPERIMENTATORE IN MODO DA RENDERLE INDIPENDENTI.

NELLA PRATICA È BENE CONSIDERARE LA **COLLINEARITÀ** COME UN **GRADIENTE**.

La presenza della **multicollinearità** non altera la validità dell'ols, ma **influisce sull'interpretazione della significatività** delle **stime** dei **coefficienti parziali**.

Infatti, quando due o più variabili indipendenti sono altamente correlate è **IMPOSSIBILE conoscere il contributo di ciascuna** delle due variabili sulla **variabile dipendente**.

Da un punto di vista statistico l'influenza della **collinearità** si osserva nella **stima** del coefficiente d'**errore** (**s**) che inevitabilmente **aumenta** e nei conseguenti **test di significatività** (**t**) dove si osserva una **riduzione** dei valori.

Gli effetti della multicollinearità sono **IRRILEVANTI** se il nostro modello si pone come obiettivo la predizione della vd (**MODELLO PREDITTIVO**); diventano molto **più SERI** se l'obiettivo della regressione è quello di definire la **rilevanza** dei singoli predittori (**MODELLO INTERPRETATIVO**).

Tranne nel caso della perfetta multicollinearità, nella pratica non esistono **test** che consentono di **definire** se questo problema esiste o meno.

FORTUNATAMENTE, PERÒ, ESISTONO DEI **SEGNALI** CHE POSSONO PORTARCI SOSPETTARNE LA PRESENZA.

ALCUNI **SEGNALI** POSSONO ESSERE RICONOSCIUTI QUANDO:

- **IL MODELLO** MOSTRA UN BUON FIT CON I DATI E TUTTAVIA SI OSSERVA CHE **TUTTI** I COEFFICIENTI PARZIALI SONO **NON SIGNIFICATIVI**;
- LE **STIME** DEI COEFFICIENTI PARZIALI **NON SONO STABILI** NEI DIVERSI CAMPIONI O NELLO STESSO CAMPIONE A SEGUITO DI LEGGERE VARIAZIONI DEL MODELLO.

SE I SI RILEVANO TALI SEGNALI È POSSIBILE IMPIEGARE ALCUNI **TEST** PER PROCEDERE AD UNA PIÙ DIRETTA VERIFICA.

