UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

Assunzioni (cap. 6)

Marcello Gallucci

marcello.gallucci@unimib.it

A
M
D

Modello Lineare Generale

- La regressione semplice e multipla e l'ANOVA sono sottocasi del modello lineare generale (GLM)

- La validità del GLM applicato ai propri dati dipende dalla soddisfazione di alcune assunzioni relative ai dati

- Se le assunzioni sono violate, i risultati saranno distorti

Assunzioni della Regressione Semplice

- Quando conduciamo una regressione o una ANOVA, facciamo implicitamente alcune assunzioni sui dati:

Scopo dell'operazione	Assunzione associata
Stimiamo gli effetti di relazione	La relazione è lineare
Stimiamo la varianza spiegata	La varianza di errore è uguale per tutti i valori predetti
Testiamo la significatività	Gli errori della regressione sono normalmente distribuiti

Assunzioni e Conseguenze

- La violazione di queste assunzioni (se non sono vere) porta a risultati non corretti

Assunzione	Se violata
La relazione è lineare	Non apprezziamo la relazione
La varianza di errore è uguale per tutti i valori predetti	La varianza spiegata sarà distorta
Gli errori della regressione sono normalmente distribuiti	Il valore-p sarà diverso dal vero rischio di commettere un errore nel rifiutare H0

Assunzione 1: Linearità

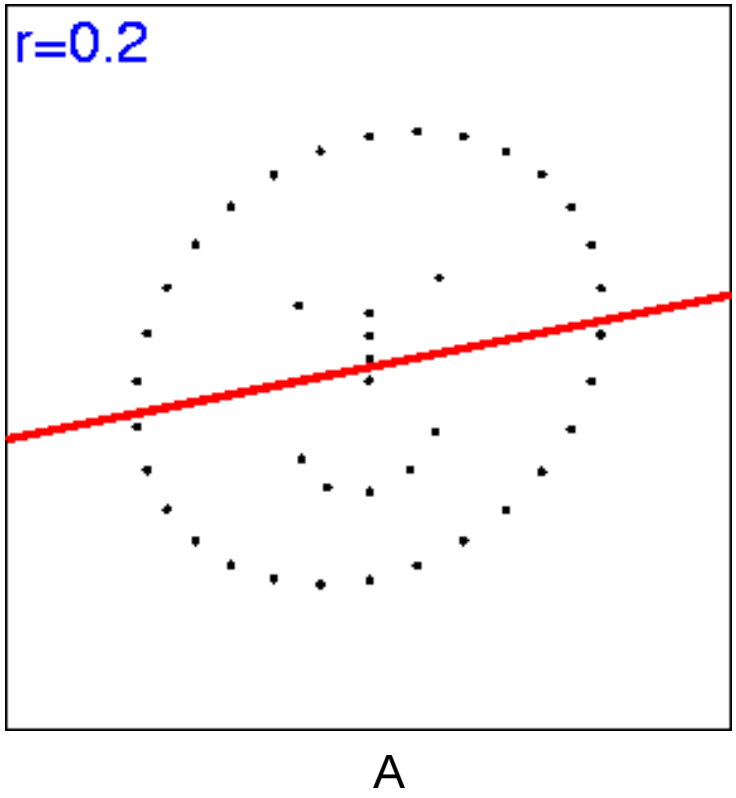
- Come visto precedentemente, la relazione che riusciamo a catturare con la regressione è una relazione lineare

Relazioni non lineari

- Le relazioni non lineari non sono catturate dalla correlazione/regressione

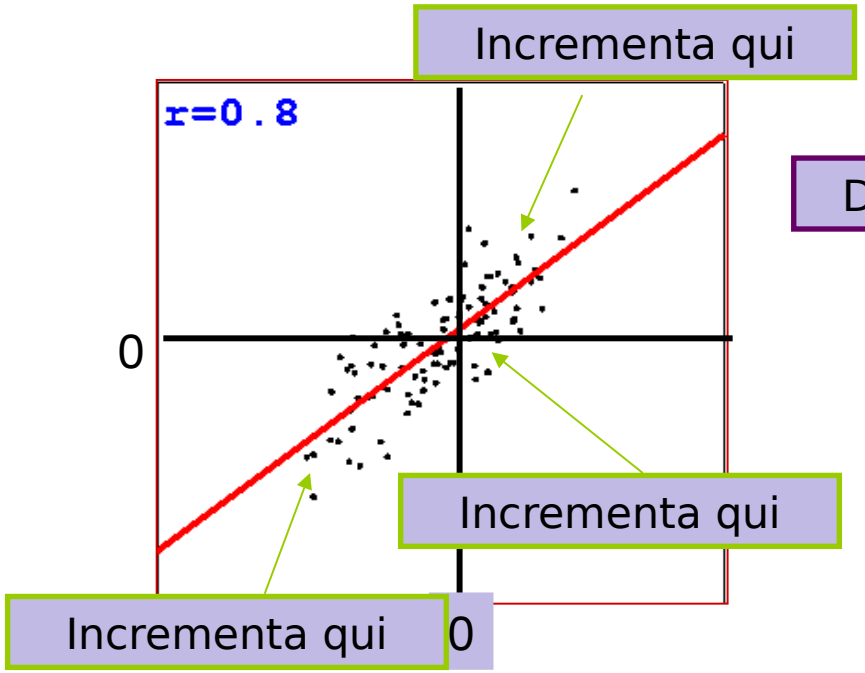
Le variabili A e B sono associate in maniera perfetta, eppure la loro correlazione è solo 0.2

La correlazione/regressione è in grado di quantificare solo le relazioni lineari

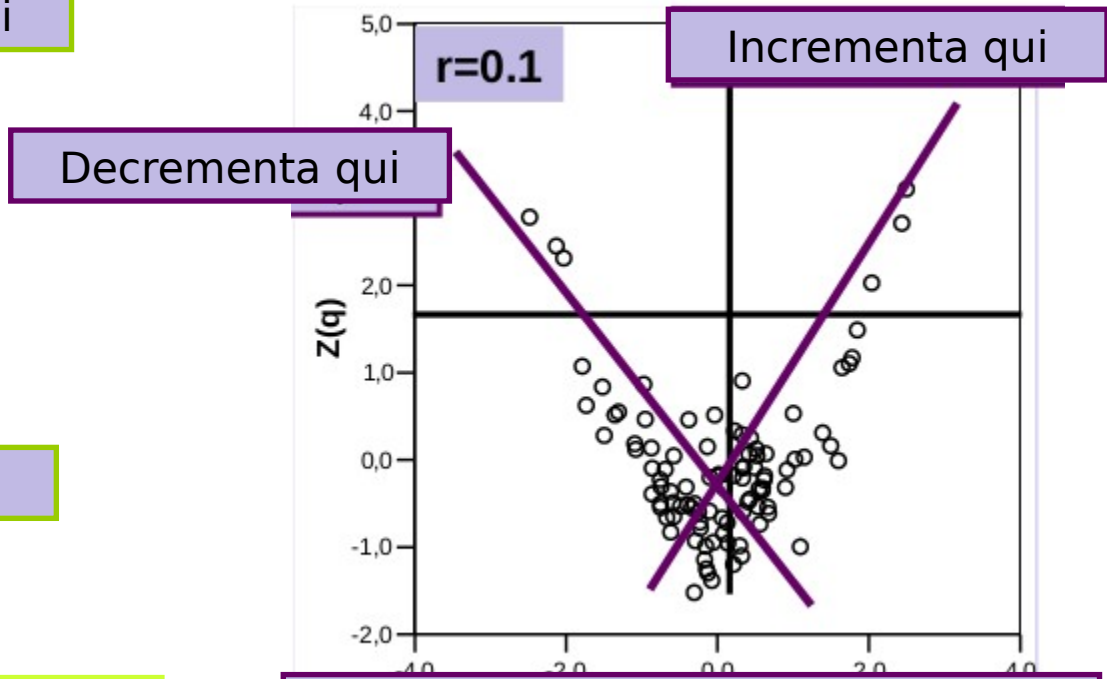


Relazioni non lineari

- La parte non lineare della relazione si perde in quanto le concordanza tra scostamenti (covarianza) è diversa ai diversi valori delle variabili



In media incrementa di .8 dev.stand. per ogni dev.stad. dell'altra



In media incrementa di solo .1 dev.stand. Per ogni dev.stad. dell'altra

Residui del modello

- Le assunzioni di Omoschedaticità e di normalità riguardano I residui (errori)

$$\hat{y}_i = a + b_{yx} x_i$$

predetti

$$y_i - \hat{y}_i = y_i - (a + b_{yx} x_i)$$

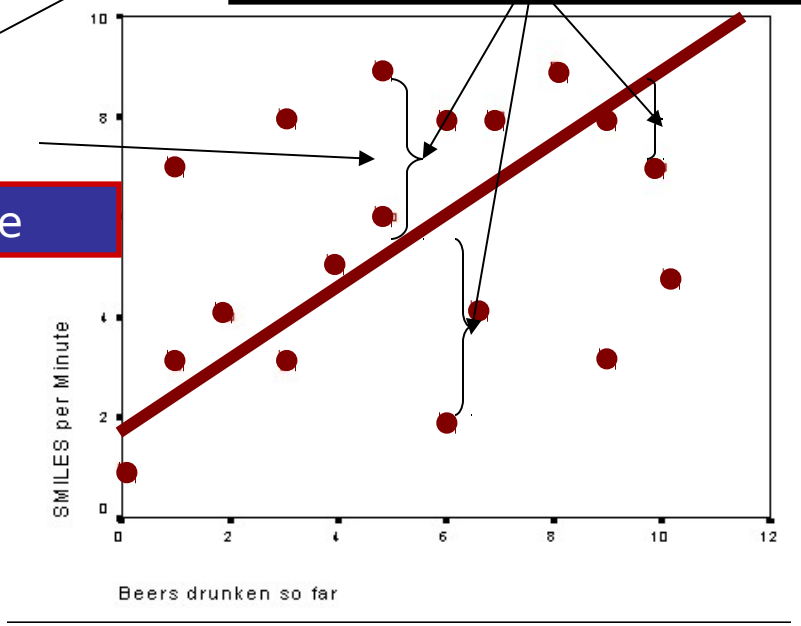
errore

Dunque i valori osservati di Y possono essere espressi come somma dei valori predetti e l'errore

$$y_i = (a + b_{yx} x_i) + (y_i - \hat{y}_i)$$

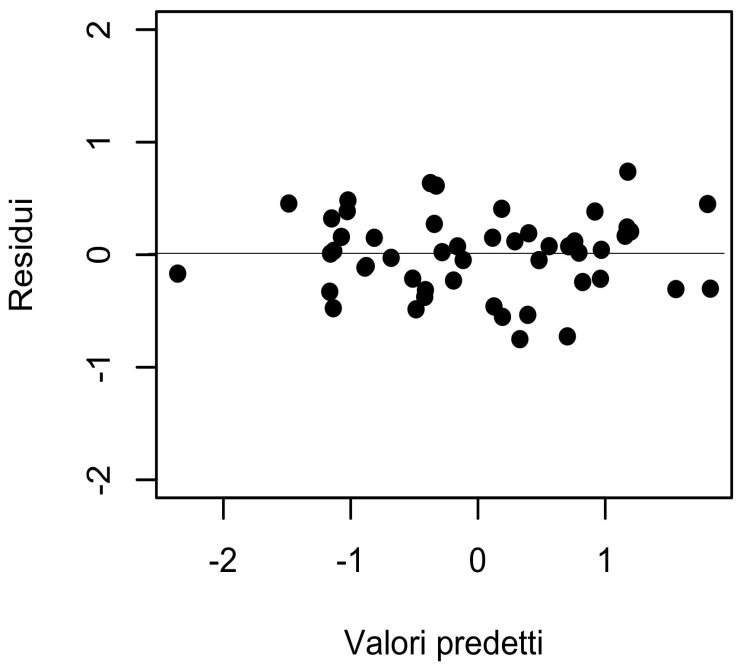
retta | errore

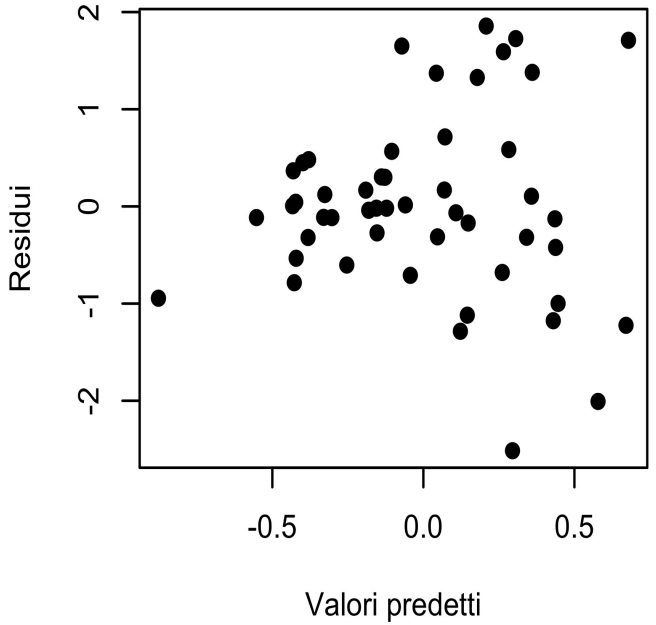
Discrepanza osservati-predetti



Assunzione 2: Omoschedasticità

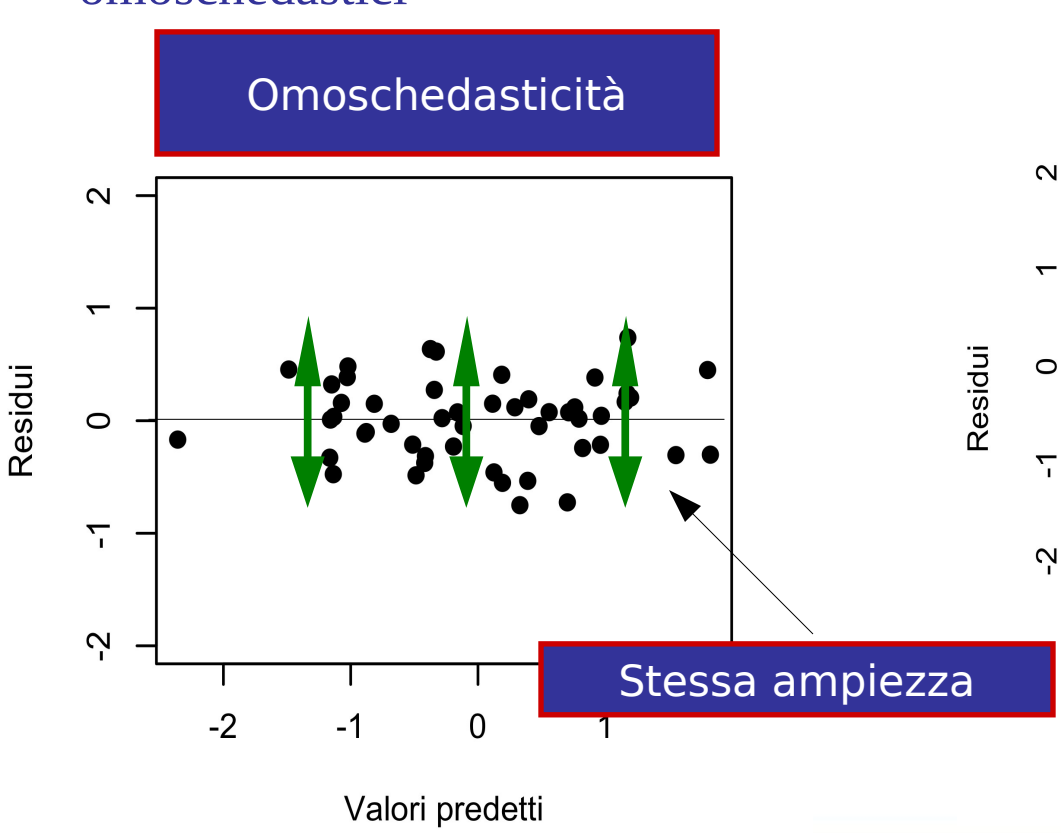
- Quando stimiamo la varianza spiegata assumiamo che la varianza di errore sia uguale per tutti i valori predetti, cioè gli errori siano omoschedastici

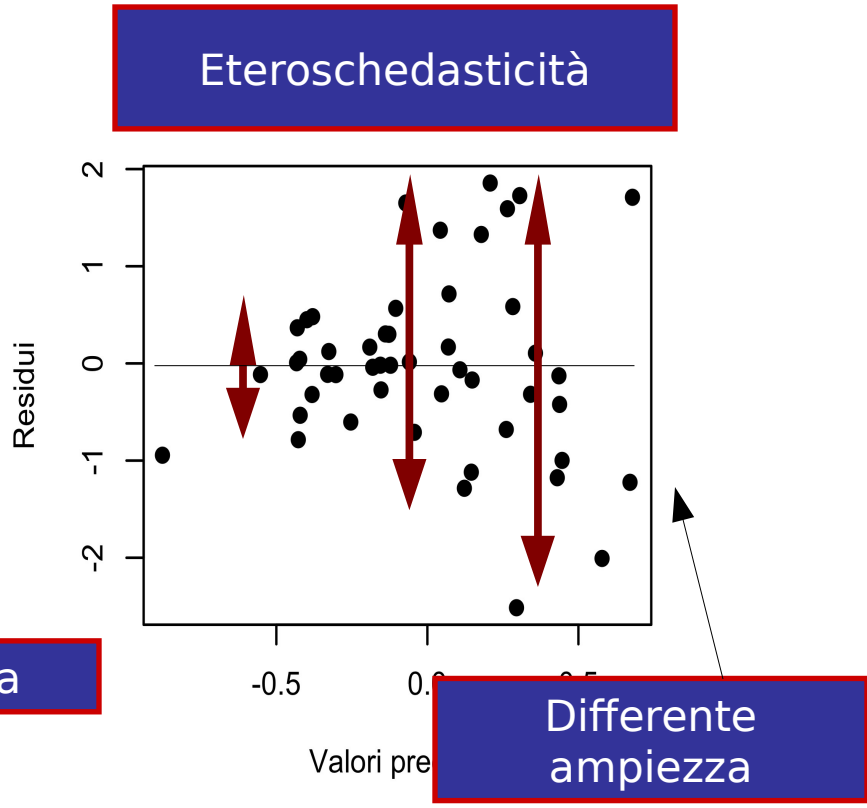




Assunzione 2: Omoschedasticità

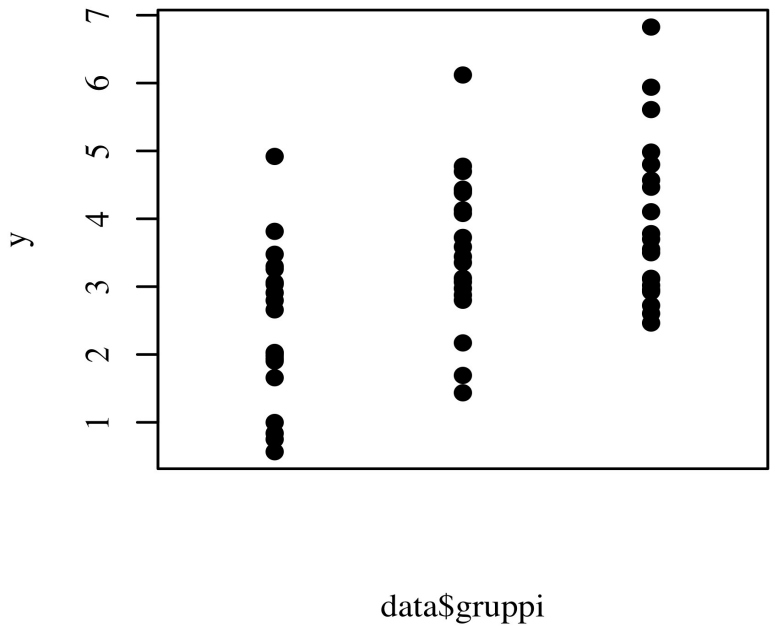
- Quando stimiamo la varianza spiegata assumiamo che la varianza di errore sia uguale per tutti i valori predetti, cioè gli errori siano omoschedastici

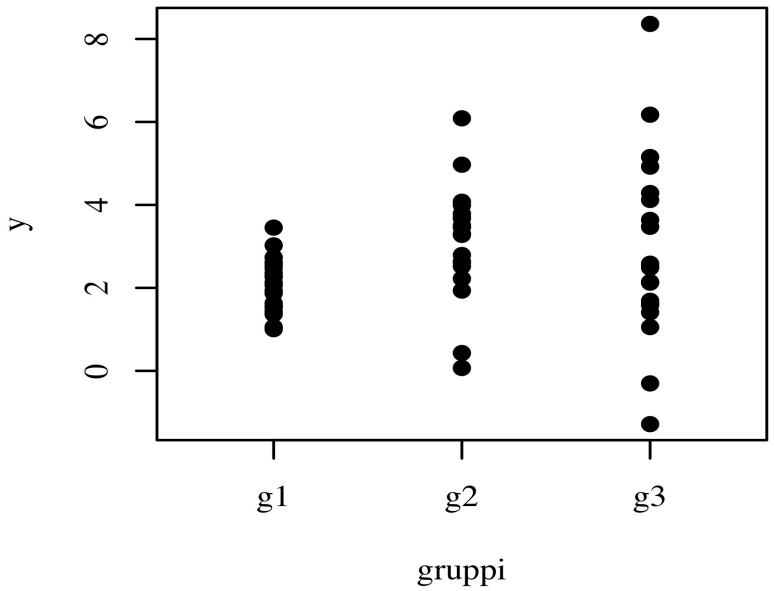




Assunzione 2: Omoschedasticità

- Quando stimiamo la varianza spiegata assumiamo che la varianza di errore sia uguale per tutti i valori predetti, cioè gli errori siano omoschedastici

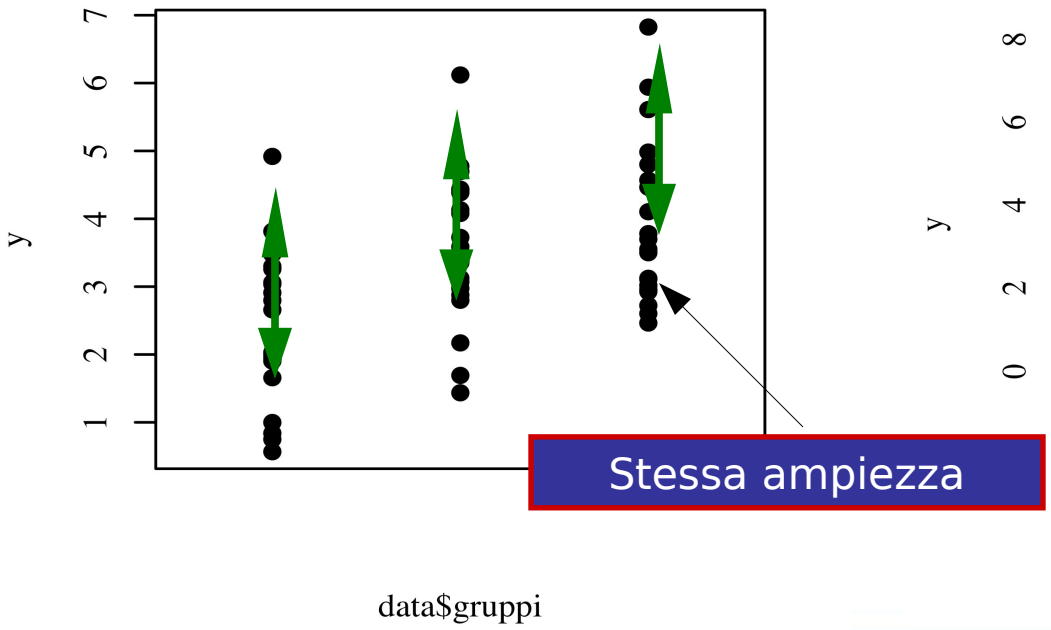




Assunzione 2: Omoschedasticità

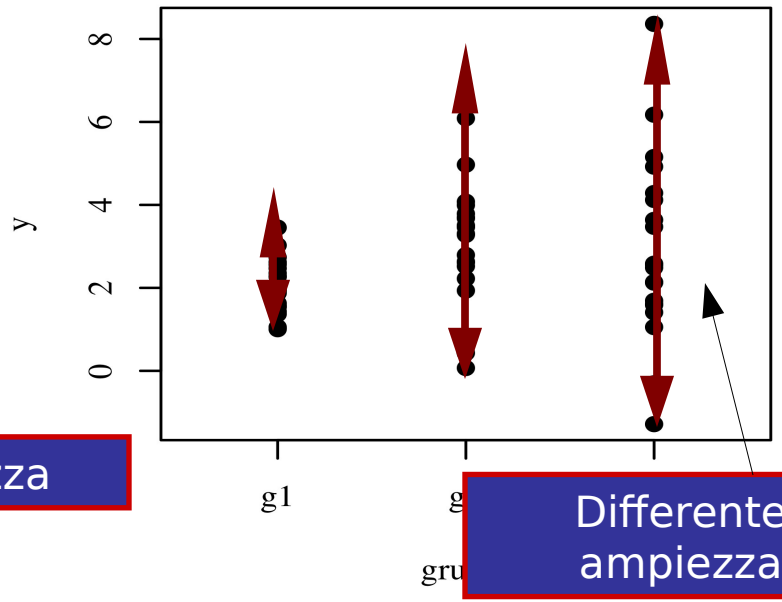
- Quando stimiamo la varianza spiegata assumiamo che la varianza di errore sia uguale per tutti i valori predetti, cioè gli errori siano omoschedastici

Omoschedasticità



Stessa ampiezza

Eteroschedasticità

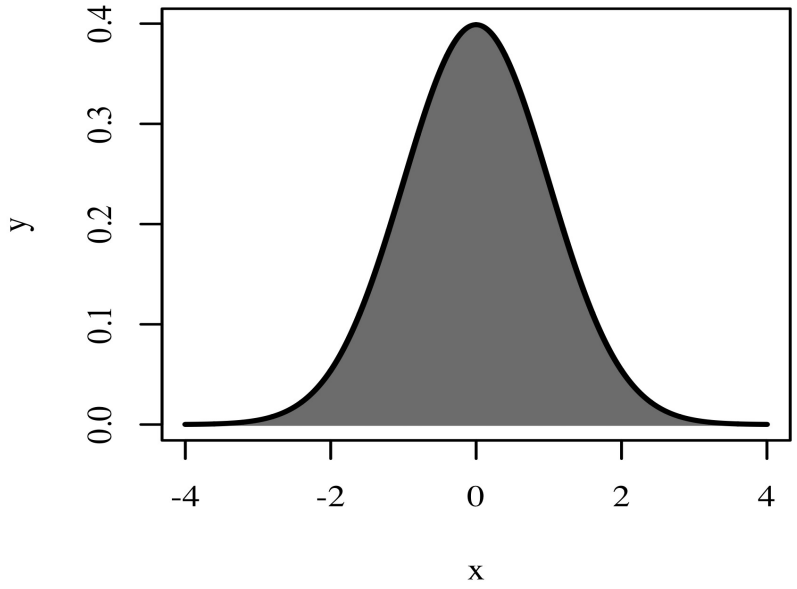


Differente ampiezza

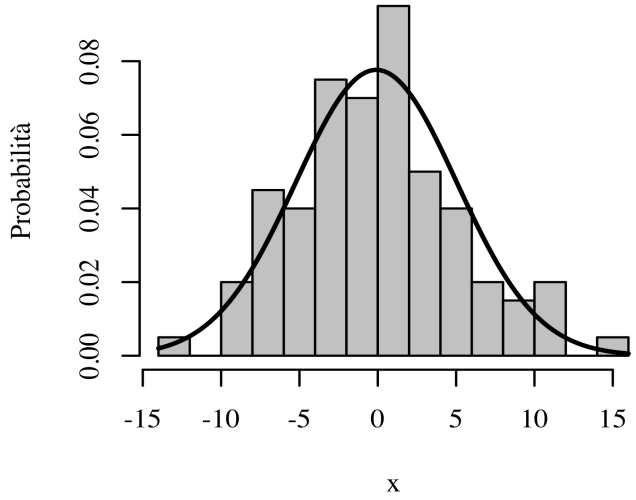
Assunzione 3: Normalità dei residui

- Si assume che i **residui** siano distribuiti normalmente. Cioè se facciamo un istogramma dei residui per tutti i soggetti, otteniamo uan distribuzione fatta a campana

Distribuzione normale



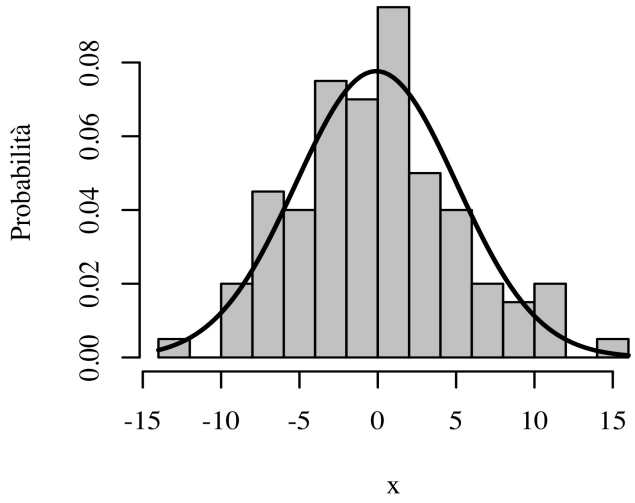
Residui normali



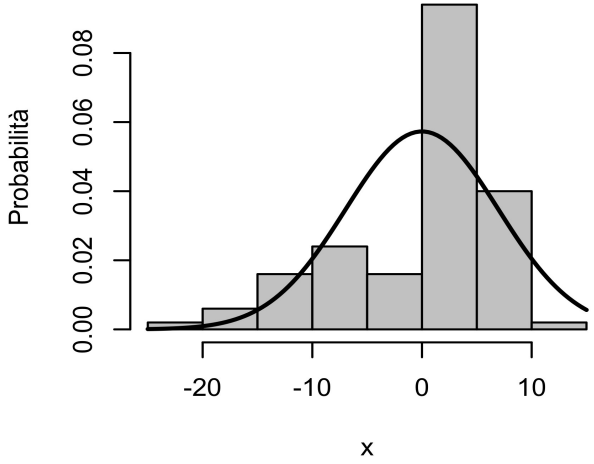
Assunzione 3: Normalità dei residui

- Si assume che i **residui** siano distribuiti normalmente. Cioè se facciamo un istogramma dei residui per tutti i soggetti, otteniamo uan distribuzione fatta a campana

Residui normali



Residui non normali

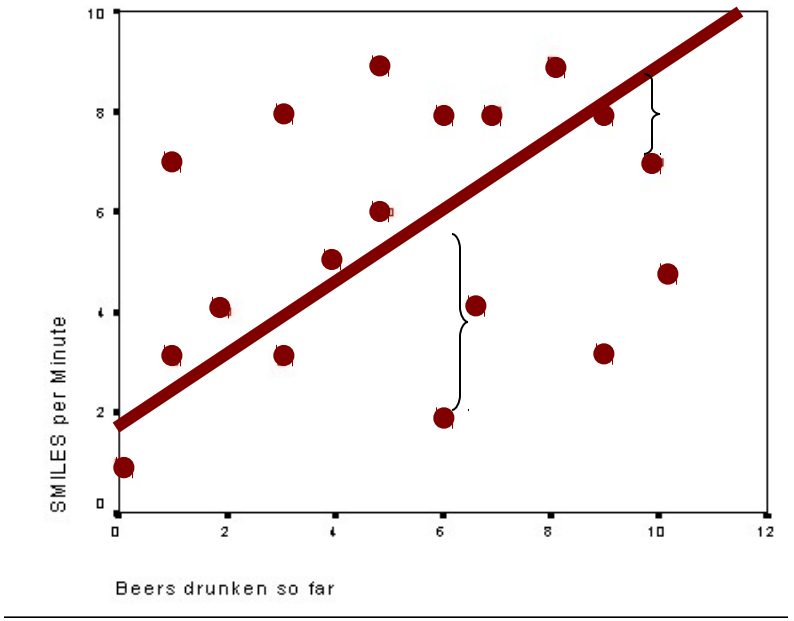


Test delle assunzioni

Analisi dei residui

- Per determinare se e quanto le assunzioni sono rispettare, è possible analizzare i residui della regressione/ANOVA

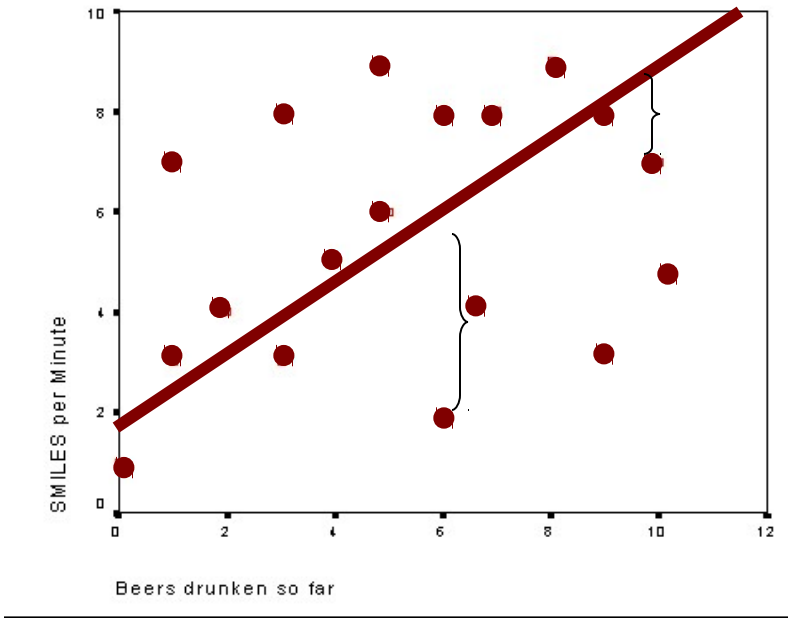
$$y_i - \hat{y}_i = y_i - (a + b_{yx} x_i)$$



Calcolare i residui

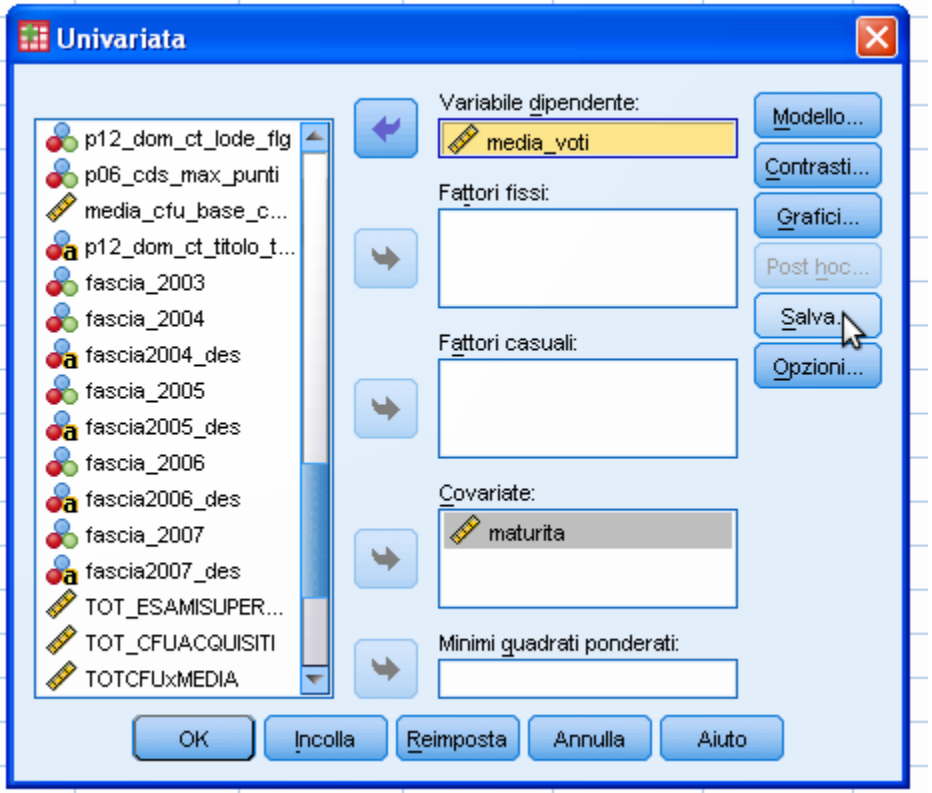
- Il calcolo dei residui (di norma fatto dal software automaticamente) consta nella mera sottrazione, per ogni soggetto, del punteggio predetto da quello osservato

$$y_i - \hat{y}_i = y_i - (a + b_{yx} x_i)$$



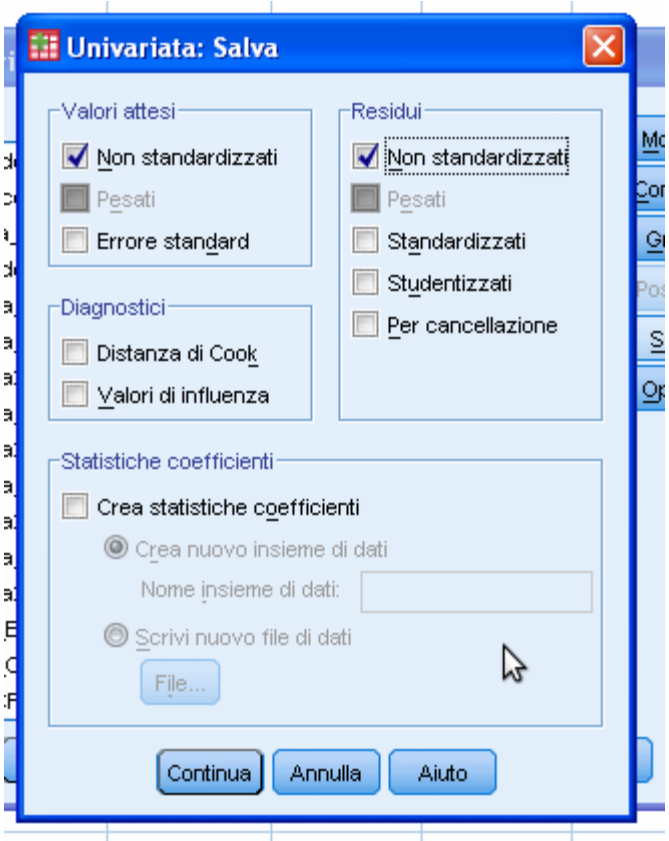
Calcolare i residui

- Nell'interfaccia SPSS, accediamo all'opzione “Salva”



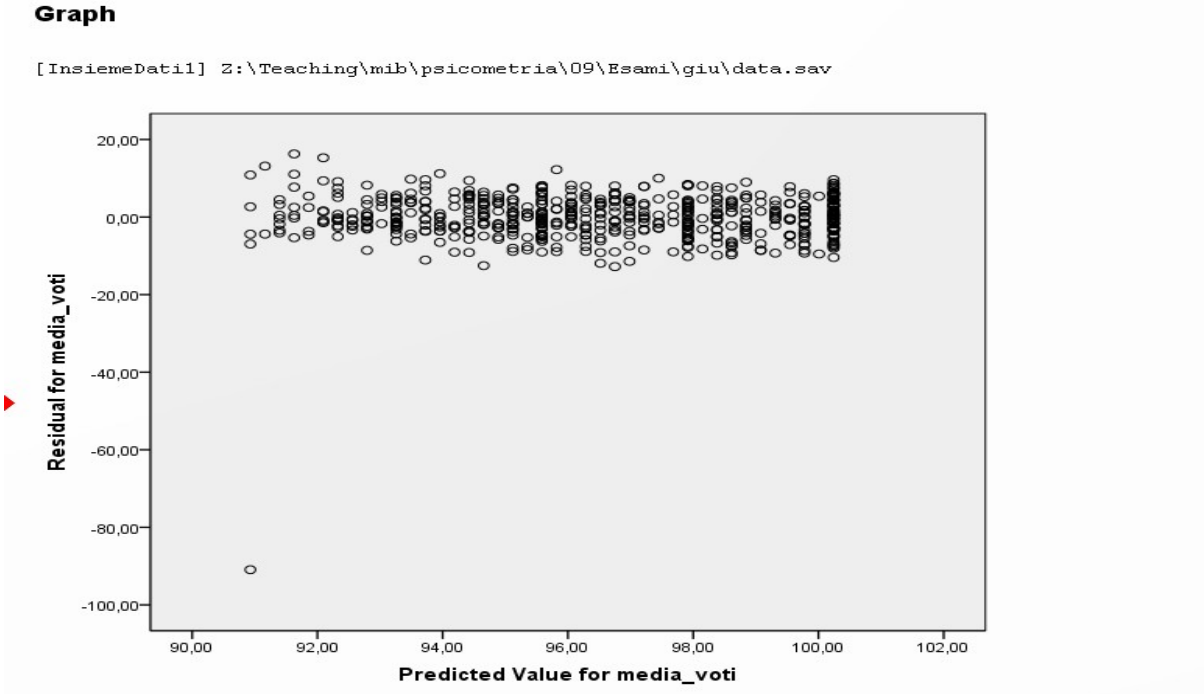
Calcolare i residui

- Chiediamo di salvare i residui ed i valori predetti
- Cosi' facendo verranno create due variabili
- PRE= valori predetti
- RES= valori residui

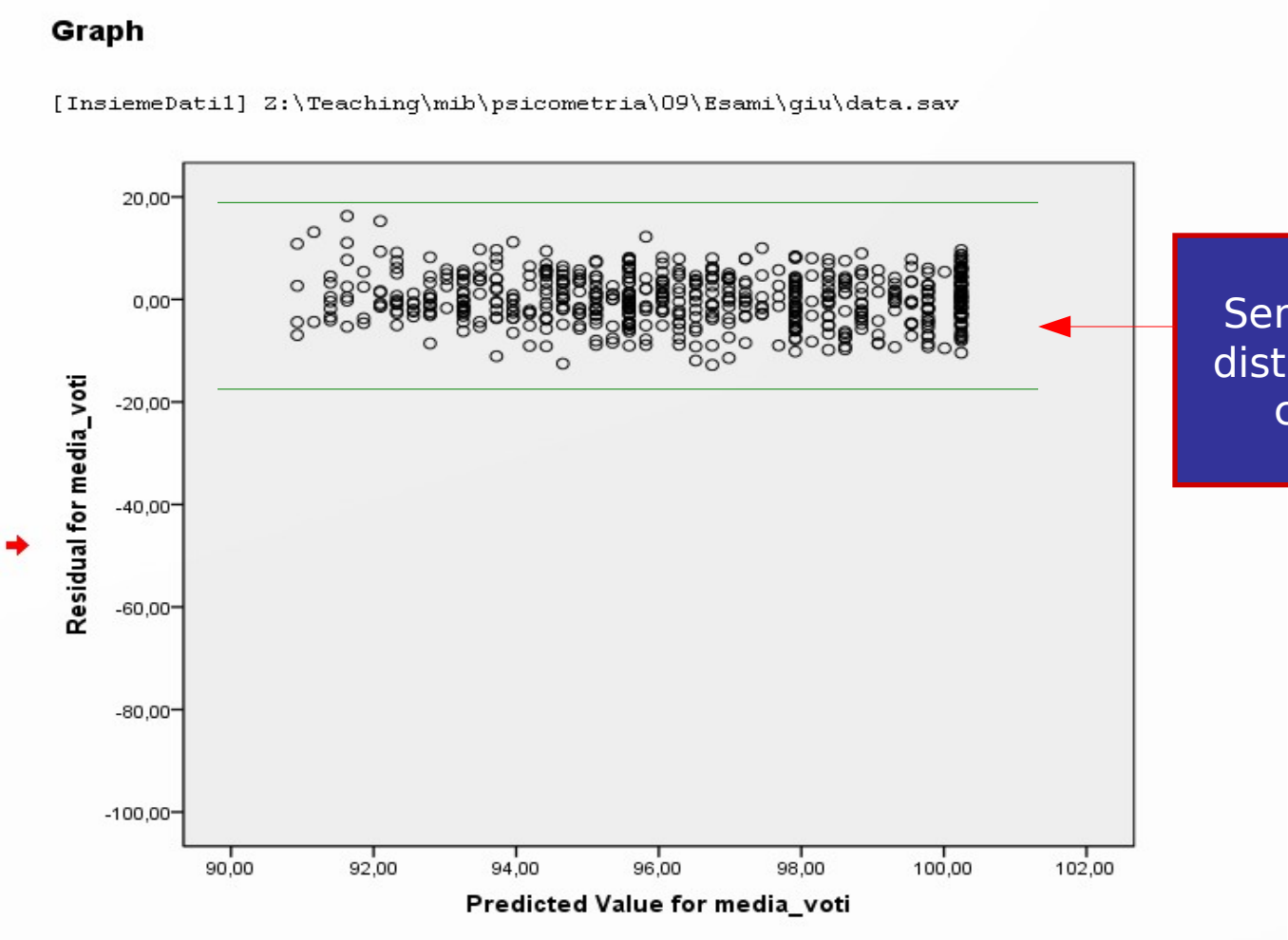


Controllo assunzioni

- Linearità e omoschedasticità: Se la relazione tra le variabili è lineare e l'assunzione di omoschedasticità è rispettata, lo scatterplot dei “predetti” vs “residui” deve avere una forma a “casuale rispetto all'asse delle X”

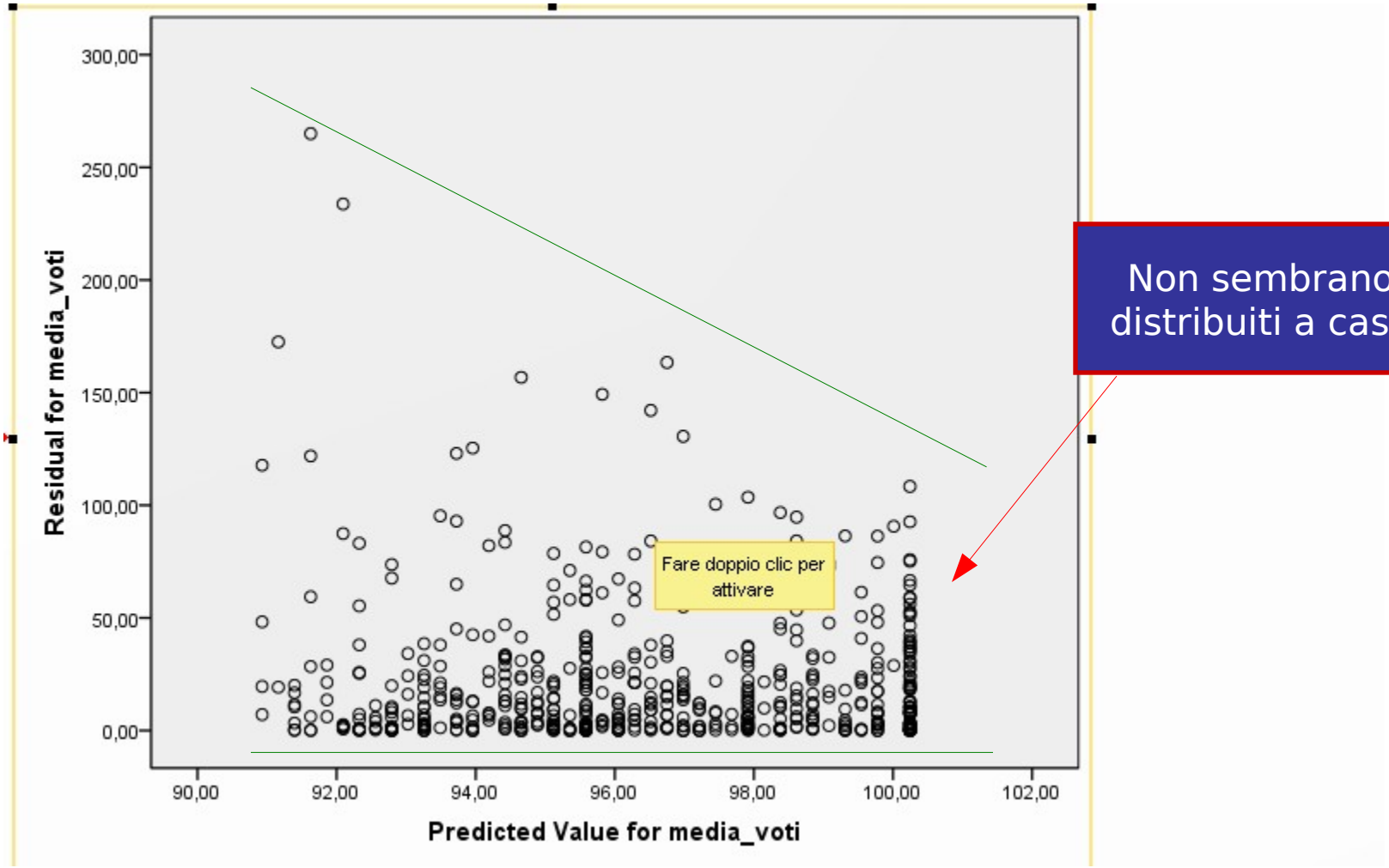


Omoschedastico



Sembrano distribuiti a caso

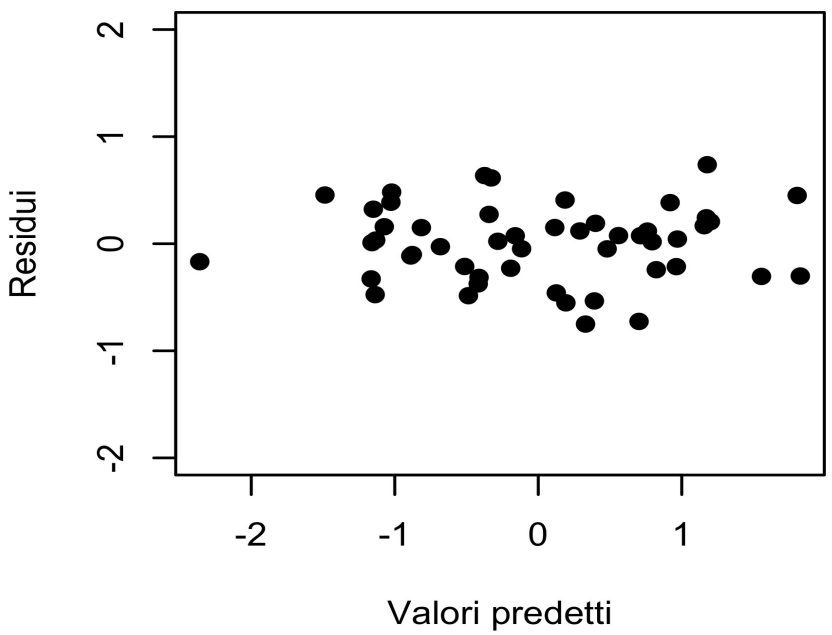
Possibili violazioni



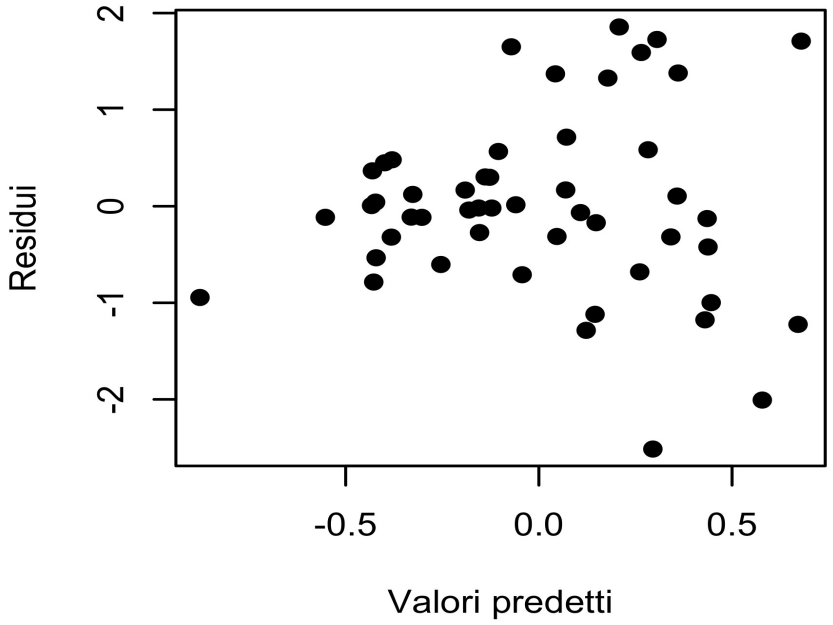
Non sembrano distribuiti a caso

Esempi VI Continua

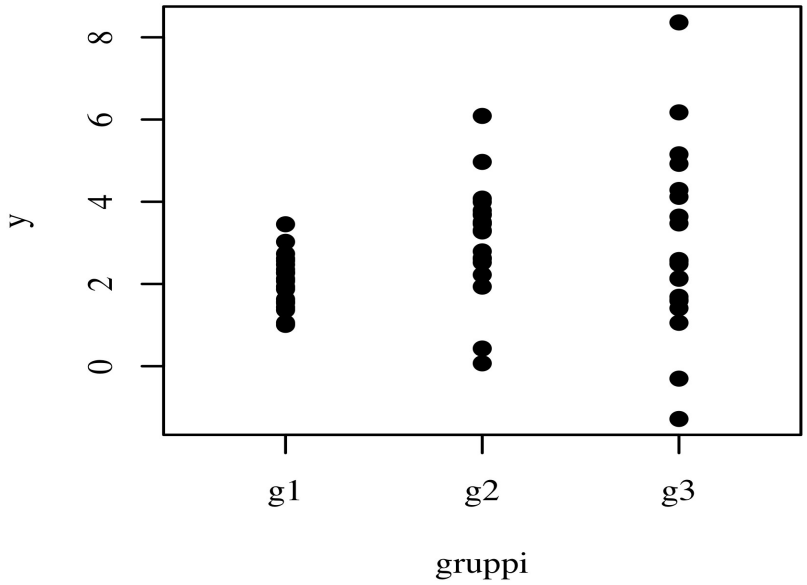
Sembrano distribuiti a caso



Non Sembrano distribuiti a caso



Esempi VI categorica

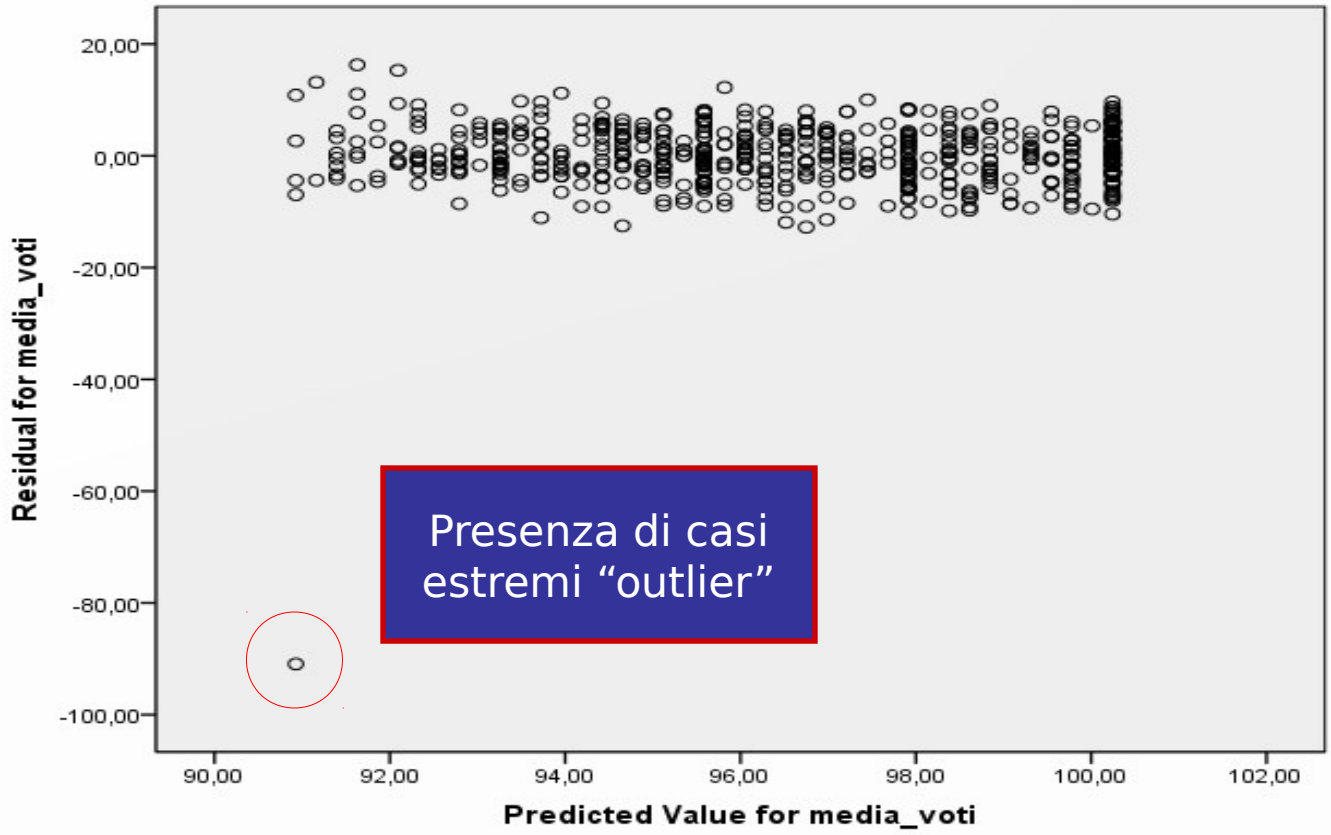


Non Sembrano distribuiti a caso

Possibili violazioni

Graph

[InsiemeDati1] Z:\Teaching\mib\psicometria\09\Esami\giu\data.sav



Outlier

- Outlier o “influential points” sono residui di molto discordanti con la distribuzione nel campione. Essi corrispondono a soggetti con valori estremi o nella variabile dipendente o nella indipendente.

- Gli outlier si eliminano dall'analisi

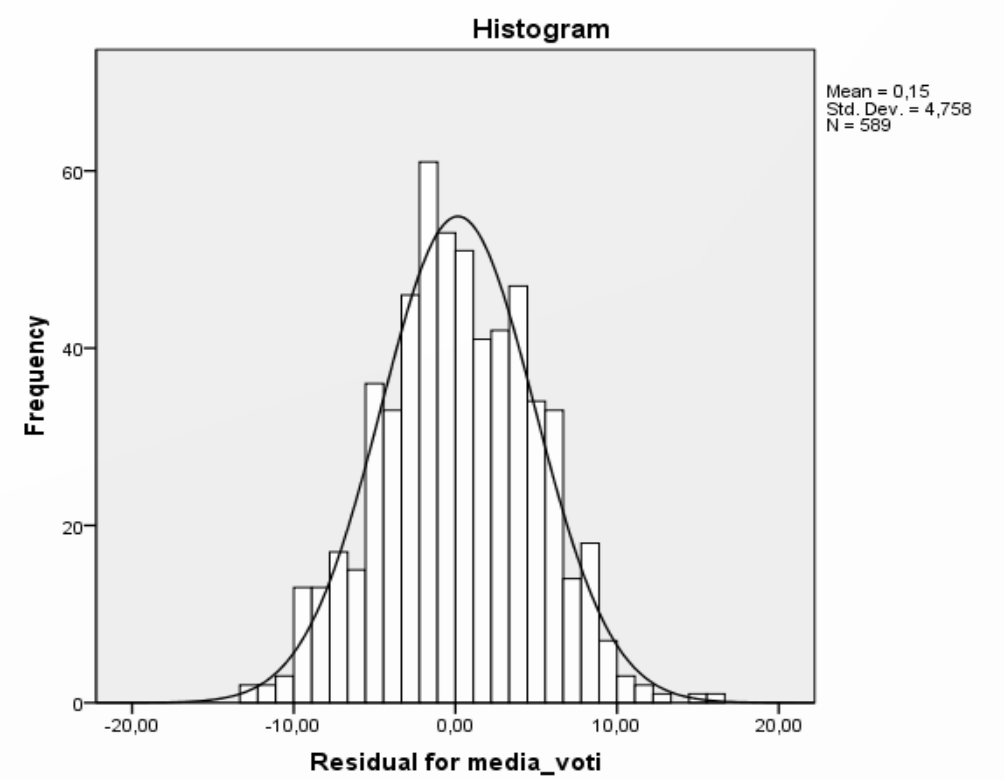
Outlier

- Outlier o “influential points” sono residui di molto discordanti con la distribuzione nel campione. Essi corrispondono a soggetti con valori estremi o nella variabile dipendente o nella indipendente.

- Gli outlier si eliminano dall'analisi

Normalità dei residui

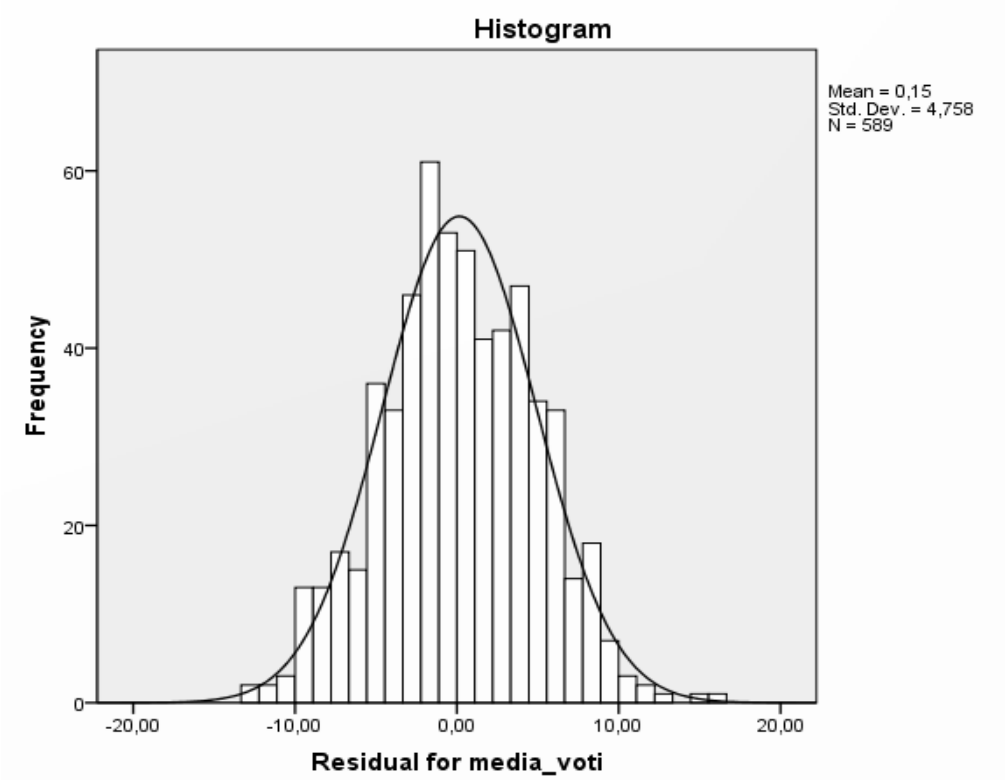
- Per verificare la normalità dei residui (cioè che si distribuiscano secondo una distribuzione Gaussiana normale), osserveremo l'istogramma



Test di Normalità

- E' possibile testare l'ipotesi nulla che la distribuzione dei residui sia normale: test di Kolmogorov-Smirnov

Il test di Kolmogorv-Smirnov testa la differenza tra la distribuzione dei residui e una normale gaussiana



Test di Normalità

- E' possibile testare l'ipotesi nulla che la distribuzione dei residui sia normale: test di Kolmogorov-Smirnov

Il test di Kolmogorv-Smirnov testa la differenza tra la distribuzione dei residui e una normale gaussiana

Se il test NON è significativo, l'assunzione di normalità è rispettata

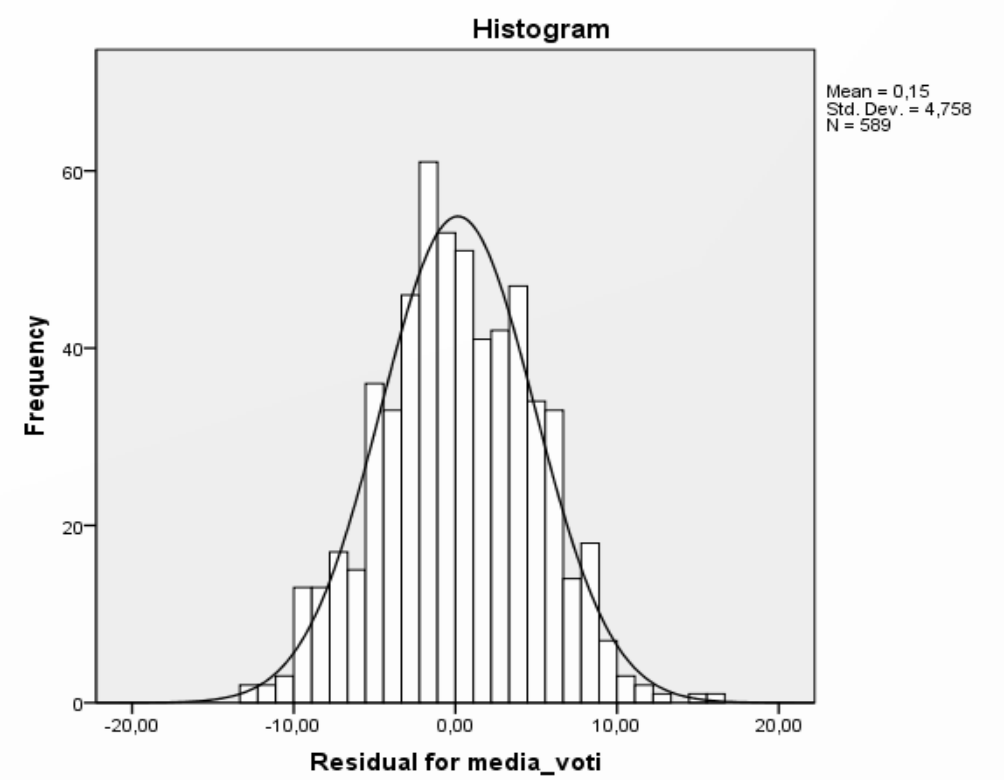
Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Residual for media_voti is normal with mean 0.154 and standard deviation 4.758.	One-Sample Kolmogorov-Smirnov Test	.819	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Normalità dei residui

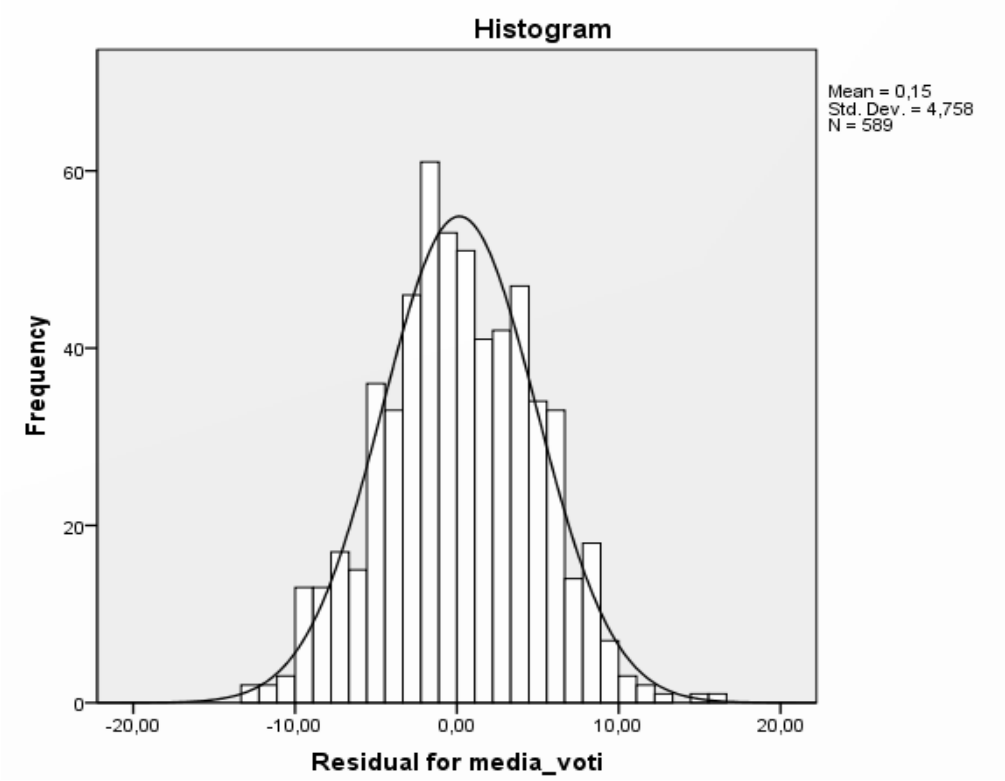
- Per verificare la normalità dei residui (cioè che si distribuiscano secondo una distribuzione Gaussiana normale), osserveremo l'istogramma



Test di Normalità

- E' possibile testare l'ipotesi nulla che la distribuzione dei residui sia normale: test di Kolmogorov-Smirnov

Il test di Kolmogorv-Smirnov testa la differenza tra la distribuzione dei residui e una normale gaussiana



Verifica assunzioni

- Per poter affermare che i risultati della nostra regressione/ANOVA sono validi, è necessario che i dati rispettino le assunzioni

- É possibile verificare le assunzioni analizzando i residui della regressione/ANOVA

- Il diagramma di dispersione che lega i valori predetti ai residui deve avere un andamento piatto, simetrico e regolare (banda costante senza outliers)

- La distribusione dei residui deve essere normale (test di Kolmogorov-Smirnov)

- Nella prossima lezione affronteremo dei possibili rimedi alla violazione delle assunzioni

Riepilogo

Assunzione	Effetto su:	Verifica	Rimedi
Indipendenza dei residui	Varianza di errore, R^2, test inferenziali	Controllo del disegno di ricerca	ANOVA a misure ripetute
Omoschedasticità	Varianza di errore, R^2, test inferenziali	Scatterplot dei valori residui e predetti,	**Trasformazione delle variabili, Test non-parametrici**
Outlier	Tutte le stime del modello	Istogramma dei residui, Scatterplot delle variabili, Scatterplot residui-predetti	Eliminare gli outlier
Normalità dei residui	Test inferenziali	Istogramma di frequenza, Test K-S	**Trasformazione, Test non-parametrici**, MLGZ
Linearità dei coefficienti	Coefficienti	Scatterplot tra variabili	Trasformazioni, Interpretazione, modello lineare generalizzato

MLGZ=modello lineare generalizzato

Soluzioni alla violazione

- Quando una delle assunzioni è violata, si possono analizzare i dati seguento tre approcci:

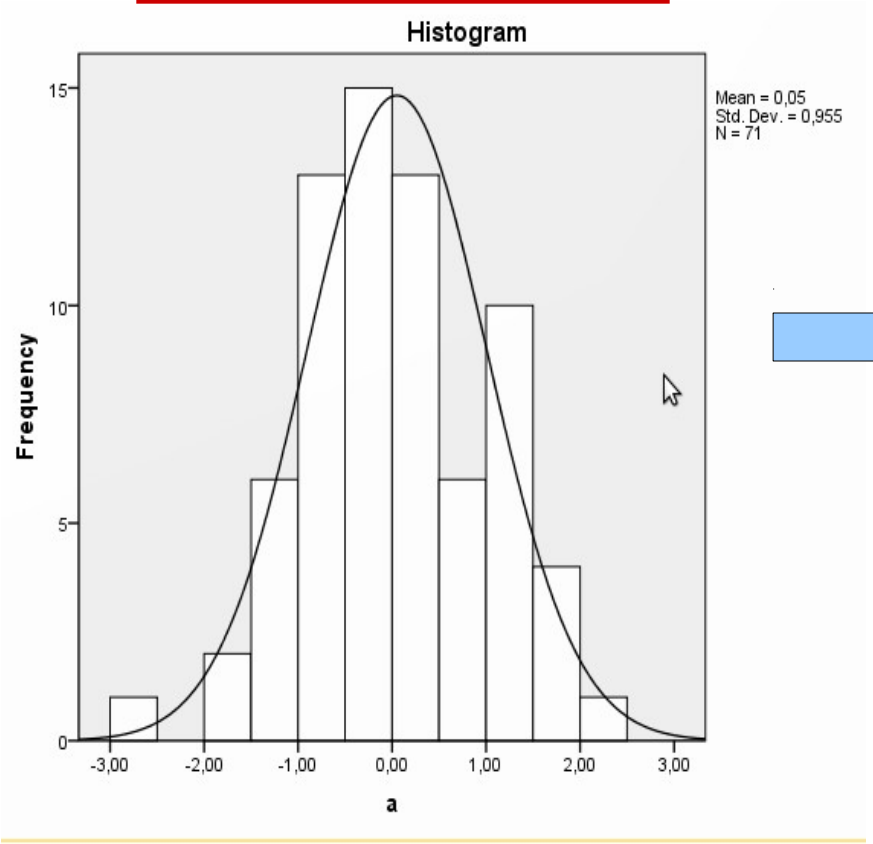
  - Cambiare le variabili: Trasformazione delle variabili
  - Cambiare test: test non parametrici
  - Cambiare modello: Modello lineare generalizzato (vedi lezioni succ.)

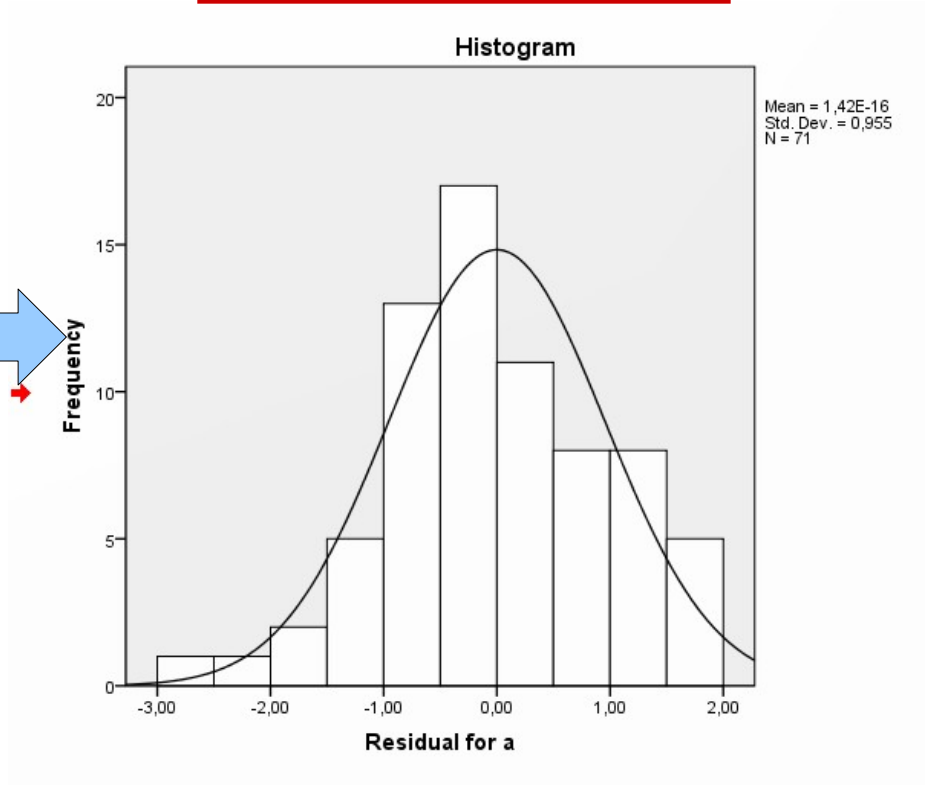
Trasformazione delle varibili dipendenti

Distribuzione dei residui

Variabile dipendente

Residui

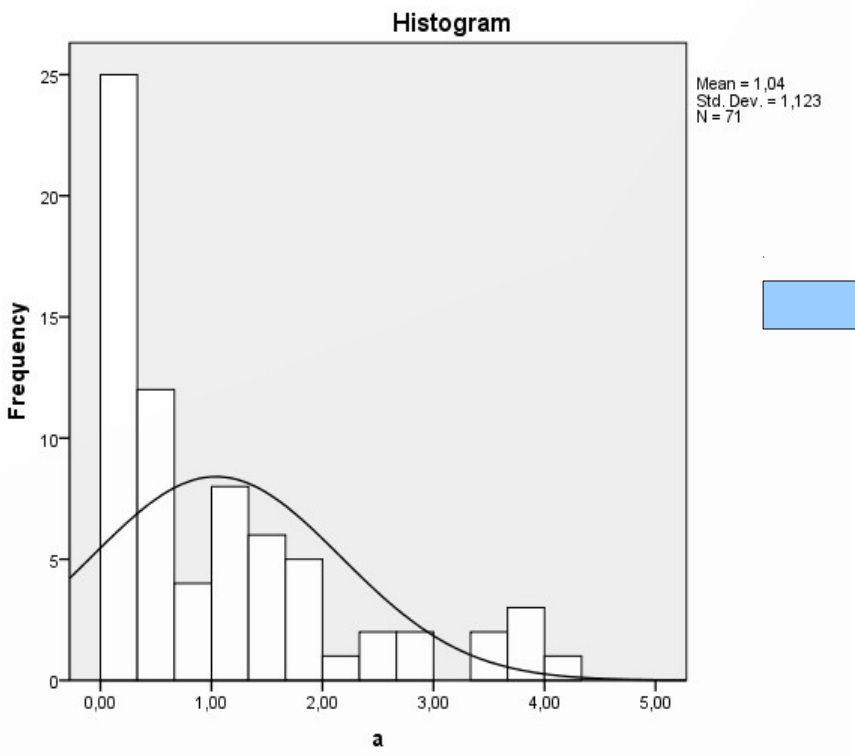


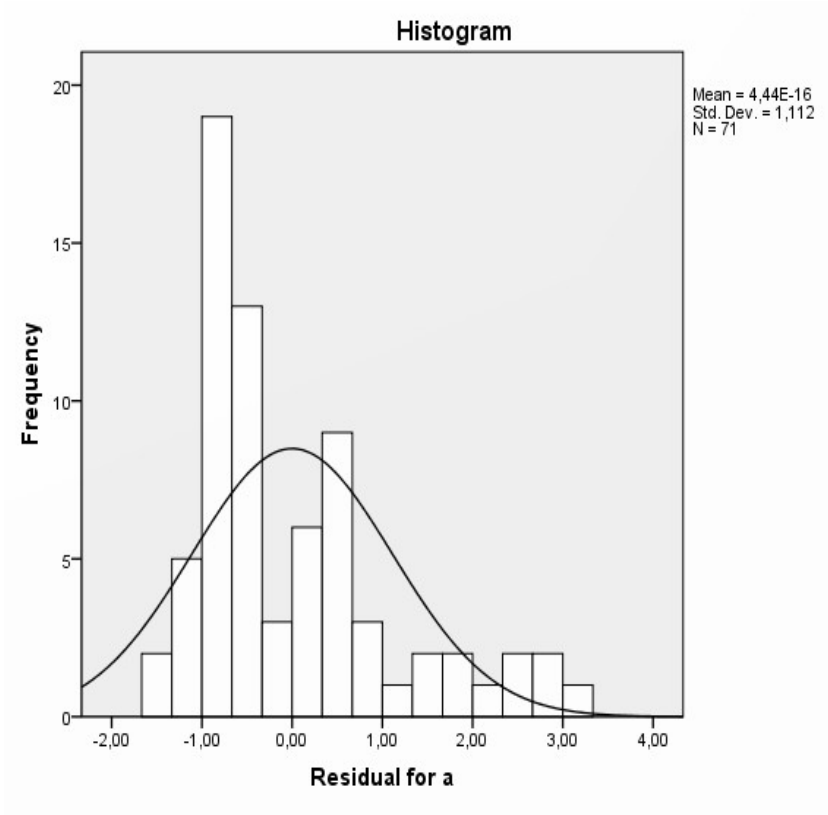


Distribuzione dei residui

Variabile dipendente

Residui





Trasformazione variabili

- Quando la variabile dipendente non è distribuita normalmente, si può operare una trasformazione della variabile al fine di modificarne la forma della distribuzione

- Esistono vari tipi di trasformazioni, suddivisibili in due classi

  1) Trasformazioni volte a normalizzare la variabile

  2) Trasformazioni in ranghi (ranks)

Normalizzazione

- Le trasformazioni volte a normalizzare la distribuzione hanno come scopo quello di rendere la nuova variabile dipendente “più normale” dell'originale: Ogni formula può funzionare, purchè non cambi l'ordine dei punteggi.

$$Y'_i = f(Y_i)$$

Esempi

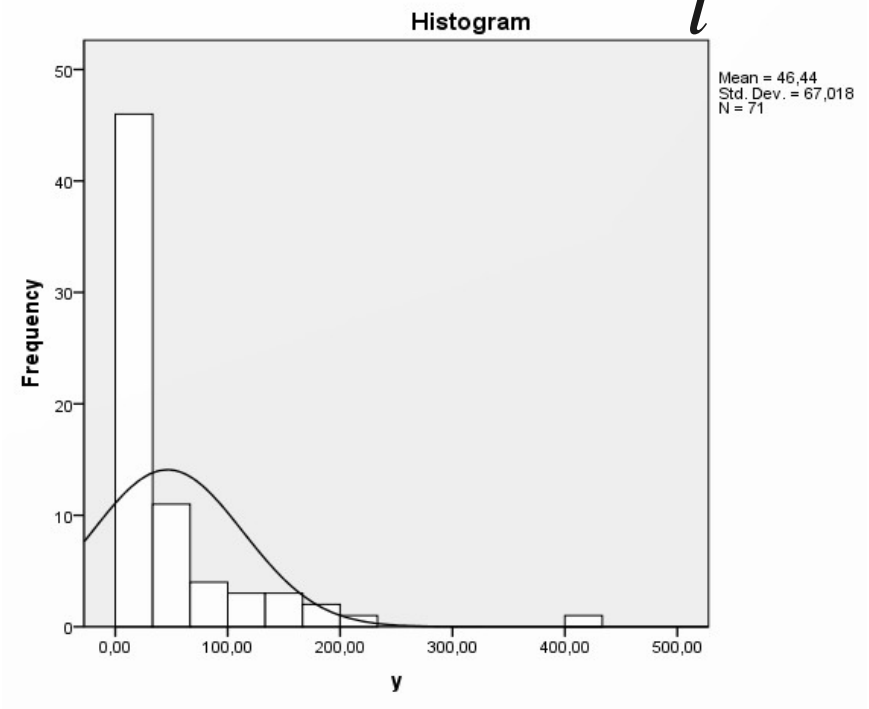
$$Y'_i = Y_i^2 \qquad Y'_i = \ln Y_i \qquad Y'_i = 1/Y_i$$

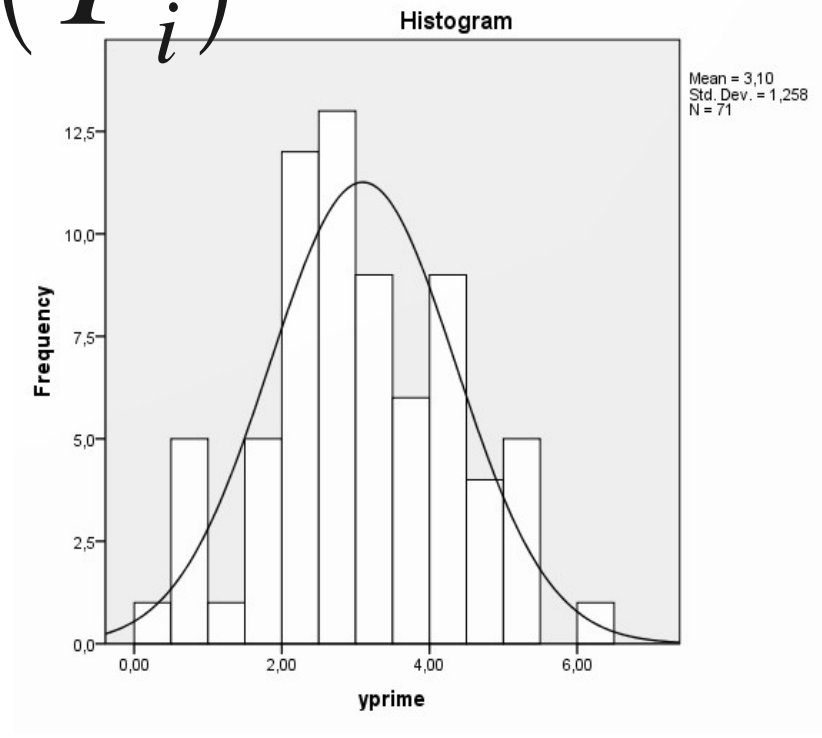
Normalizzazione

- Se la trasformazione funziona, la nuova variabile sarà una normale

(testate, ad esempio, con il Kolmogov-Smirnof)

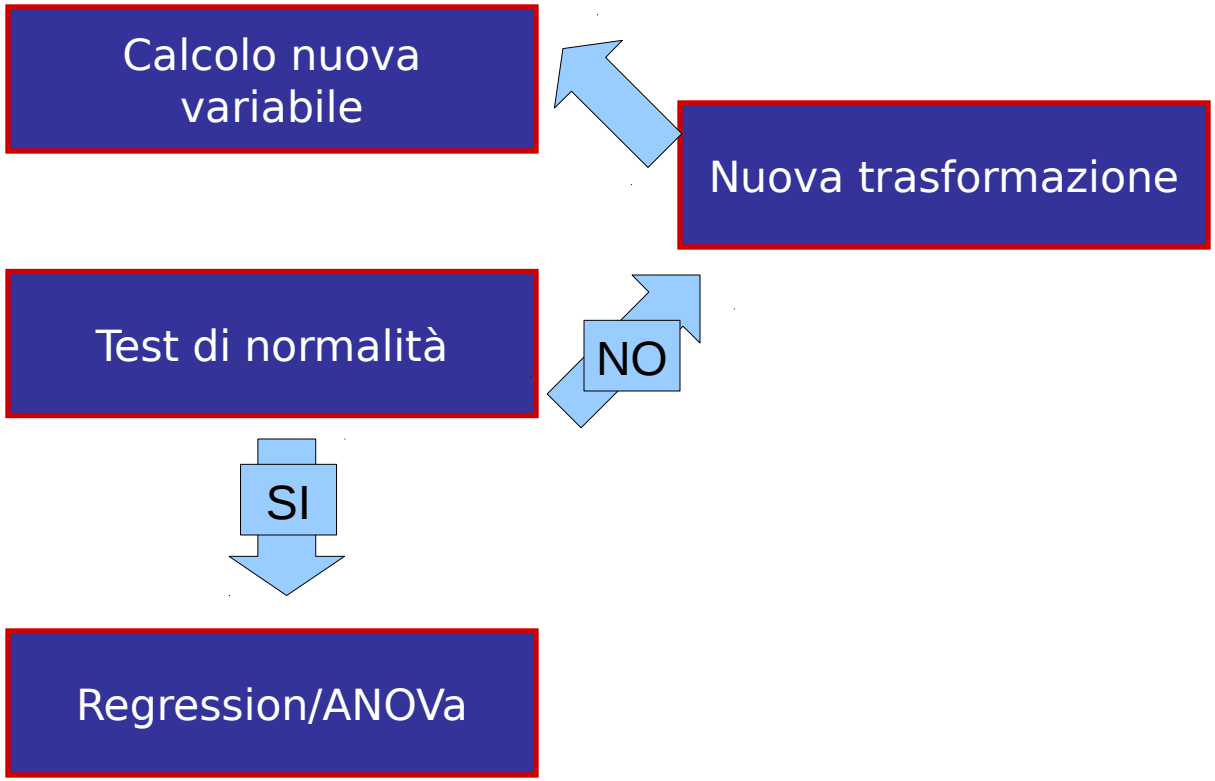
$$Y'_i = \ln(Y_i)$$





Scelta della Trasformazione Normalizzante

- Non esiste una regola precisa per scegliere la trasformazione: La trasformazione che normalizza la variabile è quella che funziona



Trasformazione normalizzante

- Se si trova la trasformazione che normalizza la variabile dipendente, i risultati della regressione/anova saranno più attendibili

- Si deve però fare attenzione che le unità di misura sono cambiate, dunque si interpreteranno preferibilmente i coefficienti standardizzati

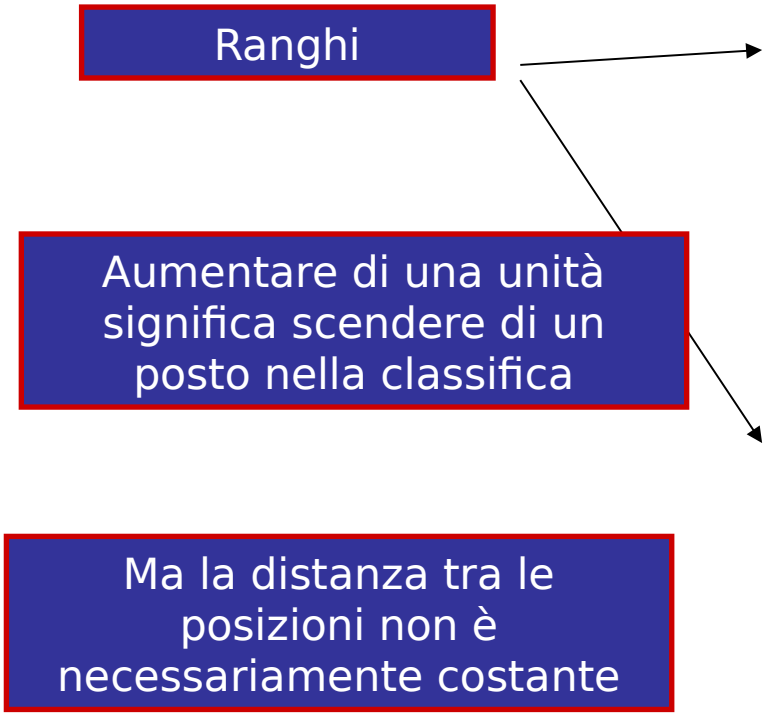
Trasformazione in ranghi

- Un'altra classe di trasformazioni prevede di calcolare i ranghi delle variabili continue inserite nelle analisi

- La trasformazione in ranghi modifica i test del GLM (regressione/correlazione/anova) in test **non parametrici**

Ranghi

- I ranghi rappresentano la posizione in una classifica ordinata secondo i punteggi della VD



1.Harvard University

2.Stanford University

3.Massachusetts Institute of Technology (MIT)

4.University of California, Berkeley

5.University of Cambridge

6.California Institute of Technology

7.Princeton University

8.Columbia University

9.University of Chicago

10.University of Oxford

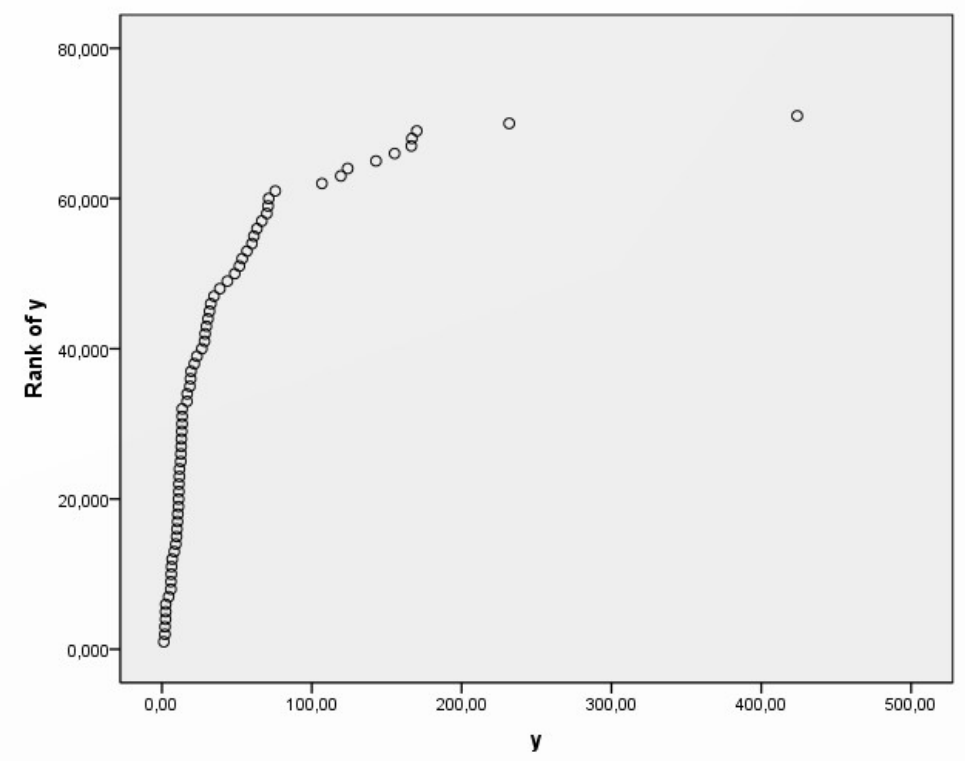
Trasformazione in ranghi

VD	y	Ry	Ranghi
	1,23	1,000	
	1,95	2,000	
	2,01	3,000	
	2,45	4,000	
	2,49	5,000	
	2,59	6,000	
	4,47	7,000	
	5,97	8,000	
	5,99	9,000	
	6,19	10,000	
	6,36	11,000	
	6,88	12,000	
	8,18	13,000	
	9,21	14,000	
	9,76	15,000	
	9,98	16,000	
	10,36	17,000	
	10,58	18,000	

Trasformazione in ranghi

Ranghi

I ranghi rispettano l'ordine dei punteggi, non la loro intensità relativa

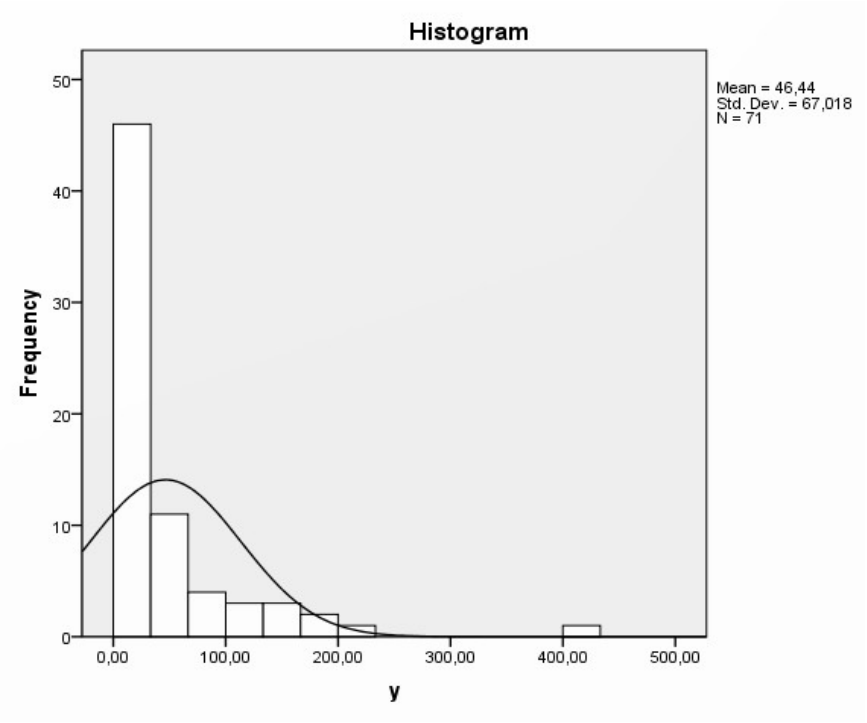


VD

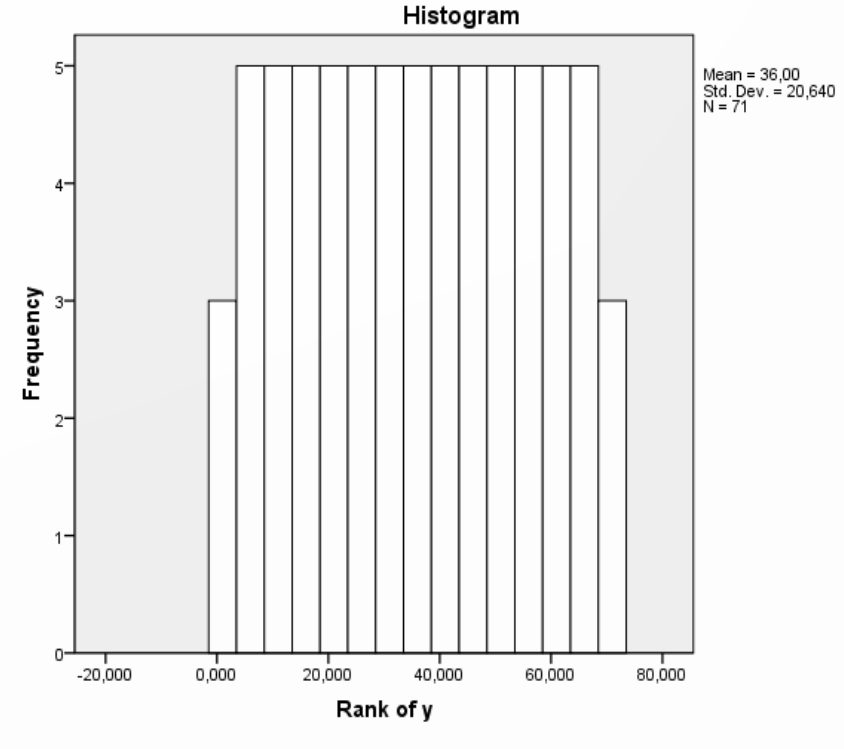
Trasformazione in ranghi

Ed uniformano la distribuzione dei punteggi

VD



Ranghi



Test non parametrici

- I test non parametrici (che studiamo in questo corso) equivalgono alle tecniche statistiche studiate fino ad ora operate dopo aver trasformato le variabili continue nei loro rispettivi ranghi.

Scopo	**Tecnica MLG**	**Test non-parametrico**
Relazione tra due variabili	Correlazione Regressione semplice standardizzata	Correlazione di Spearman
Relazione tra una dipendente e una o più indipendenti continue	Regressione multipla	Regressione non parametrica
Confronto fra due gruppi	t-test	Mann-Whitney
Confronto fra due o più gruppi	ANOVA	Kruskal-Wallis
Confronto fra gruppi a misure ripetute	ANOVA misure ripetute	Friedman test

Correlazione di Spearman

- Consta nel calcolare la correlazione (quella che conosciamo) sui ranghi (R) delle variabili

$$\rho_i = \frac{COV(R_Y, R_X)}{STD(R_Y) * STD(R_X)}$$

Indica il grado di monotonicità della relazione tra due variabili

Correlations

			a	b
Spearman's rho	a	Correlation Coefficient	1,000	,144
		Sig. (2-tailed)	.	,230
		N	71	71
	b	Correlation Coefficient	,144	1,000
		Sig. (2-tailed)	,230	.
		N	71	71

Regressione non-parametrica

- Consta nel calcolare regressione sui ranghi delle variabili

Indica il grado di monotonicità della relazione tra due variabili espressa come cambiamento del rango in Y per un rango in più di X

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	.873	1.124		.777	.457
	Rank of Birre	.855	.166	.864	5.158	.001

a. Dependent Variable: Rank of Sorrisi

Mann-Whitney

- Equivale a calcolare il t-test sui ranghi delle variabili

$$U = \frac{ttest\,(N-1)}{\sqrt{ttest^2+1}}$$

Indica il grado differenza delle distribuzioni di Y tra due gruppi

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of y is the same across categories of Drop.	Independent-Samples Mann-Whitney U Test	.248	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Kruskal-Wallis

- Equivale a calcolare il ANOVA-one way (la F-test) sui ranghi delle variabili

Kruskal-Wallis Test

Indipendente

dipendente

Ranks

	Drop	N	Mean Rank
y	yes	49	37,90
	no	22	31,77
	Total	71	

Test Statistics[a,b]

	y
Chi-square	1,337
df	1
Asymp. Sig.	,248

a. Kruskal Wallis Test

b. Grouping Variable: Drop

$$KW = \frac{Ftest\,(N-1)(K-1)}{N-K-Ftest\,(N-1)}$$

Viene però valutato con il Chi-quadro, invece che con la distribuizione F

Indica il grado differenza delle distribuzioni di Y tra vari gruppi

Fine della Lezione X