



## Quinta giornata

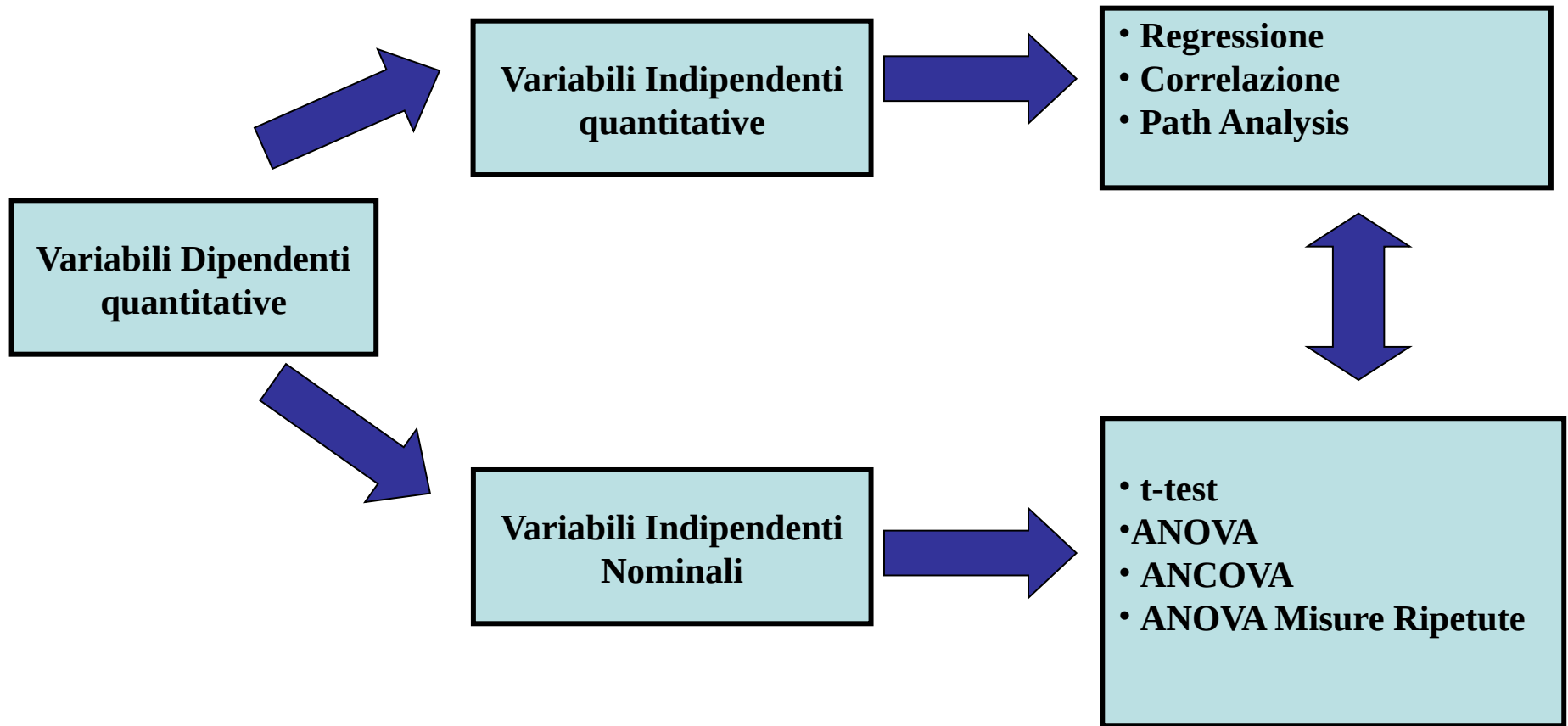
# Modelli generalizzati

Marcello Gallucci  
Univerisità Milano-Bicocca

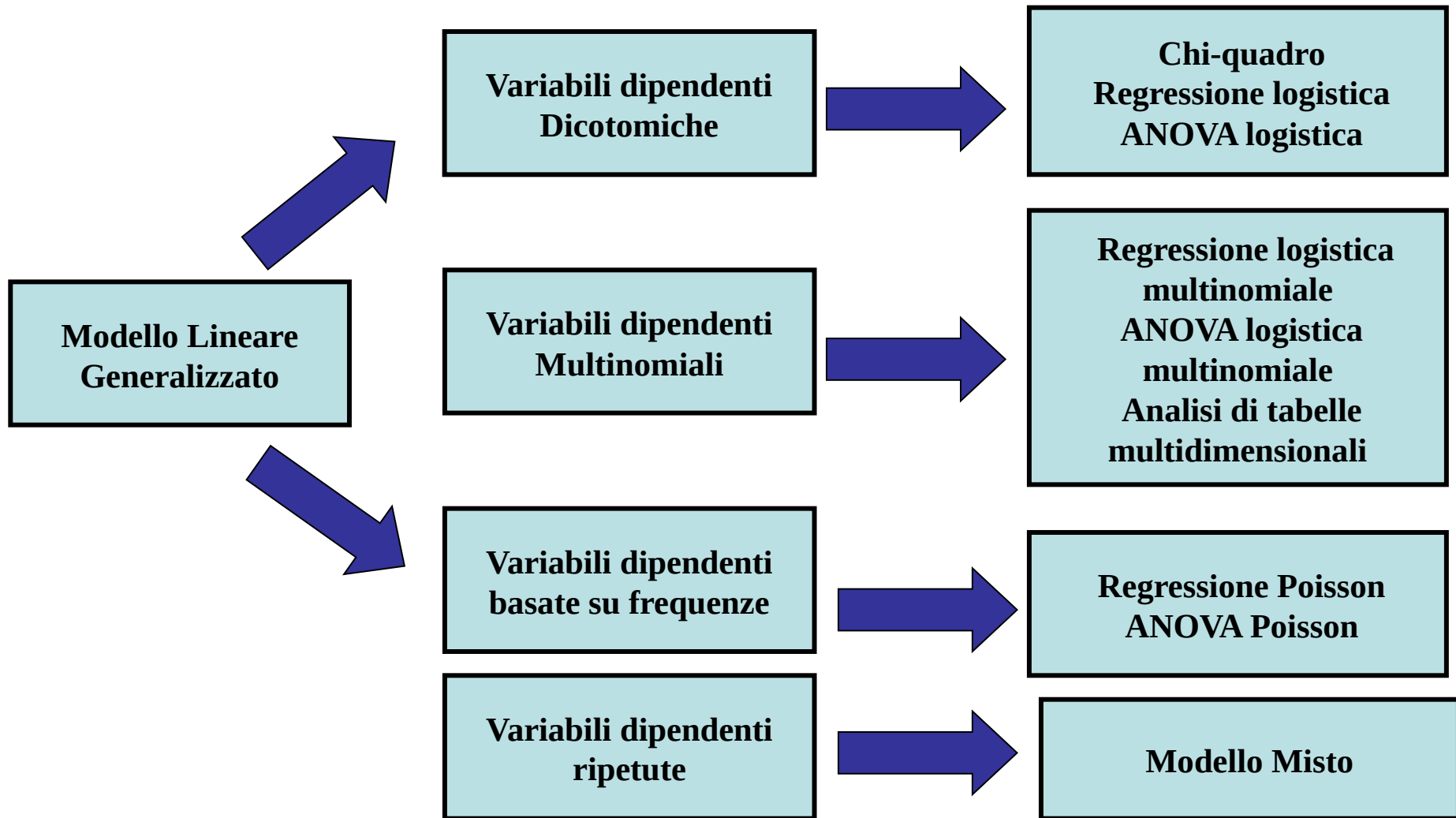
# Preludio

- La maggior parte delle comuni tecniche statistiche volte ad individuare le relazioni fra variabili, quali Correlazioni, Regressione, ANOVA, ANCOVA, sono riconducibili al **Modello Lineare Generale** (GLM)
- Il GLM ci permette di studiare gli effetti di variabili indipendenti di vario tipo su **variabili quantitative** (variabili dipendenti continue)
- La ricerca empirica è disseminata di variabili **dipendenti qualitative** (scelte dicotomiche, scelte multiple, frequenze di eventi, classificazioni, etc)
- I **Modelli Lineari Generalizzati** ci consentono di studiare gli effetti di variabili indipendenti su variabili dipendenti qualitative

# Modello Lineare Generale



# Modello Lineare Generalizzato

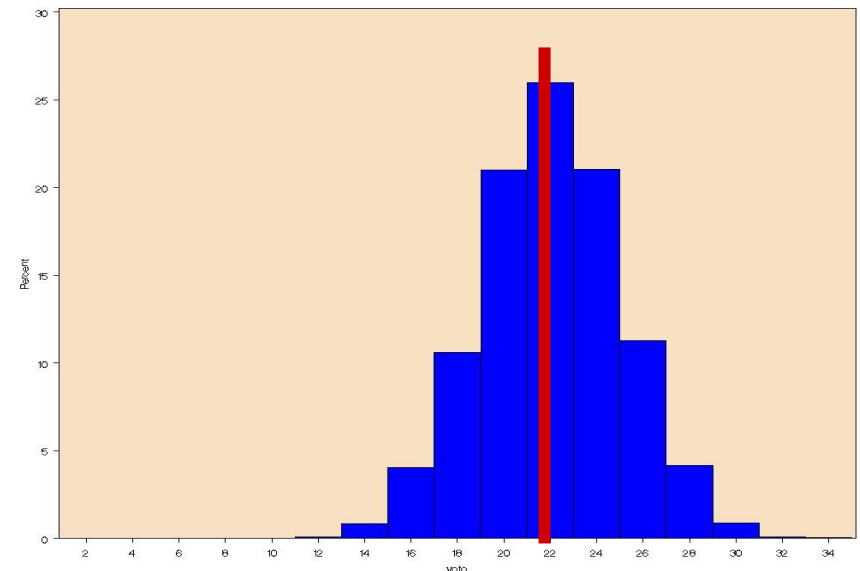


## Modello Lineare Generale

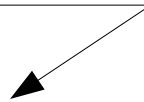
Il valore stimato della popolazione si definisce FISSO (fixed parameter)



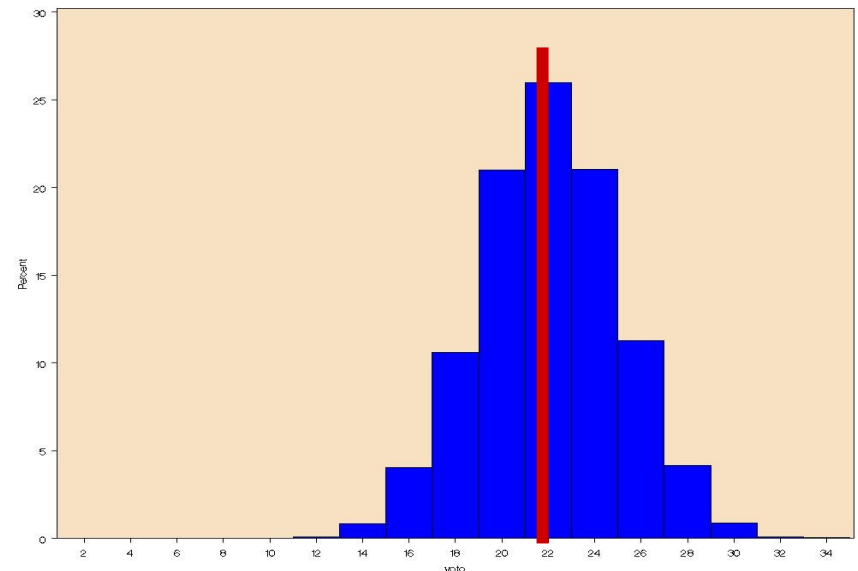
$$y_i = a + e_i$$
$$\text{corr}(e_i, e_j) = 0$$



Le variazioni casuali sono indipendenti l'una dall'altra



## Modello Lineare Generale



$$y_i = a + e_i$$
$$e_i \sim N(0, \sigma^2)$$

**I residui del modello sono distribuiti normalmente**

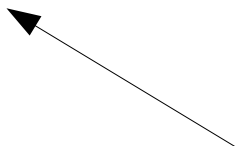
# Generalizzazioni

Useremo il **Modello Lineare Generalizzato** quando le assunzioni di normalità dei residui non può essere rispettata (variabili dipendenti categoriche)

## Generalized Linear Model

$$f(y_i) = a + b_1 \cdot x_{1i} + b_2 \cdot x_{2i} + \dots + b_k \cdot x_{ki} + e_i$$

**Variabile  
Dipendente**



An arrow points from the text 'Variabile Dipendente' to the  $y_i$  term in the equation above.

**distribuzione**



An arrow points from the text 'distribuzione' to the  $e_i$  term in the equation above.



# Modelli lineari generalizzati

- Il modello lineare si adatta ad una vasta gamma di tipologia di variabili dipendenti mediante due scelte: **link function** e **distribuzione**

$$f(Y) = a + b_{x.w} x_i + b_{w.x} w_i + e_i$$

Tipo di variabile	Link function	Distribuzione
Continue	identità	Normale
Dicotomiche	Logit dell'odd	Binomiale
categoriche	Logit del relative risk	Multinomiale
Ordinali	Logit cumulato	Multinomiale
Frequenze	LN delle frequenze	Poisson

## **Tecniche volte a studiare e quantificare gli effetti di una o più variabili indipendenti *continue o nominali* su una variabile dipendente *nominale (qualitativa)***

Tecniche statistiche:

- ◆ La regressione logistica: Variabile dipendente dicotomica
- ◆ La regressione ordinale: Variabile dipendente ordinabile
- ◆ La regressione multinomiale: Variabile dipendente politomica
- ◆ La regressione di Poisson: Variabile dipendente basata su frequenze
- ◆ Tutto ciò, anche a misure ripetute

# Violazioni assunzioni

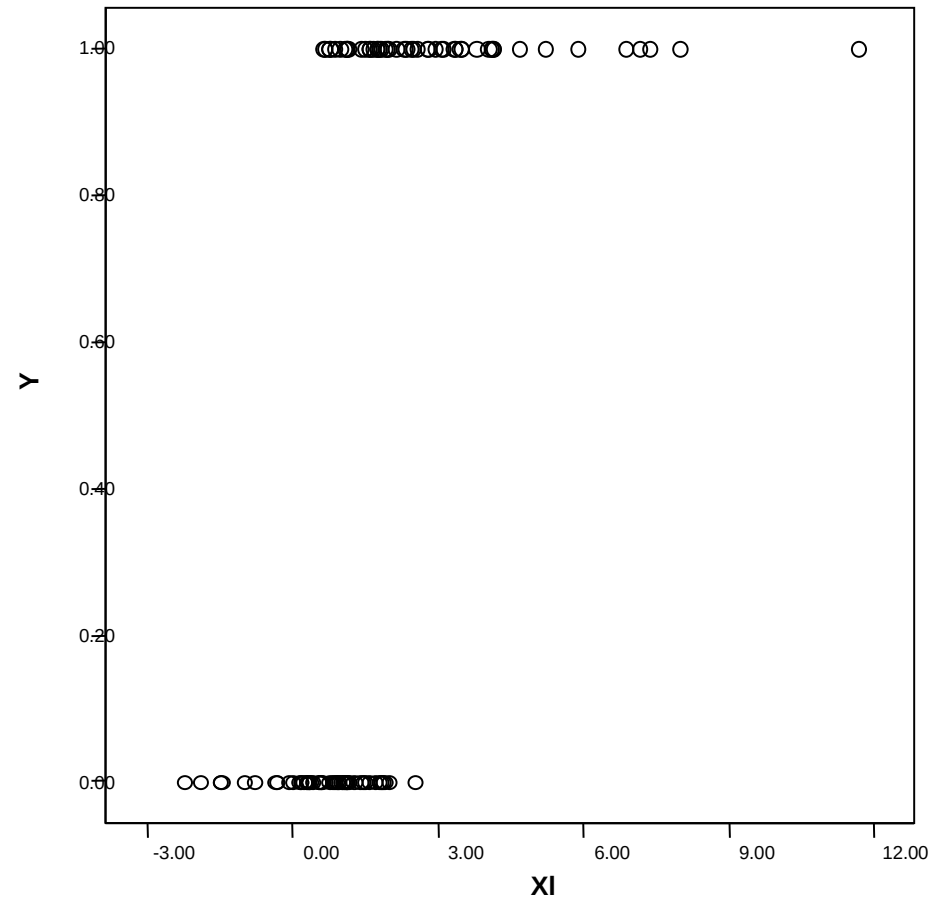
- ◆ Quando le assunzioni non sono soddisfatte, i risultati sono da considerarsi dubbi
- ◆ Se la violazione delle assunzioni è grave, la regressione/ANOVA non può essere applicata
- ◆ Nella pratica, con variabili dipendenti continue normalmente distribuite, queste assunzioni sono abbastanza semplici da soddisfare
- ◆ Ma cosa succede se volessimo usare una **variabile dipendente dicotomica**?

## Il modello logistico

# Il modello logistico

- ◆ Il modello logistico (regressione logistica) si propone di studiare e quantificare le relazioni tra una o più variabili indipendenti quantitative (es. età, salario, atteggiamenti, personalità) e una variabile **dipendente dicotomica** (es. stato civile, voto al referendum, appartenenza ad un gruppo, etc.)

# VD dicotomica



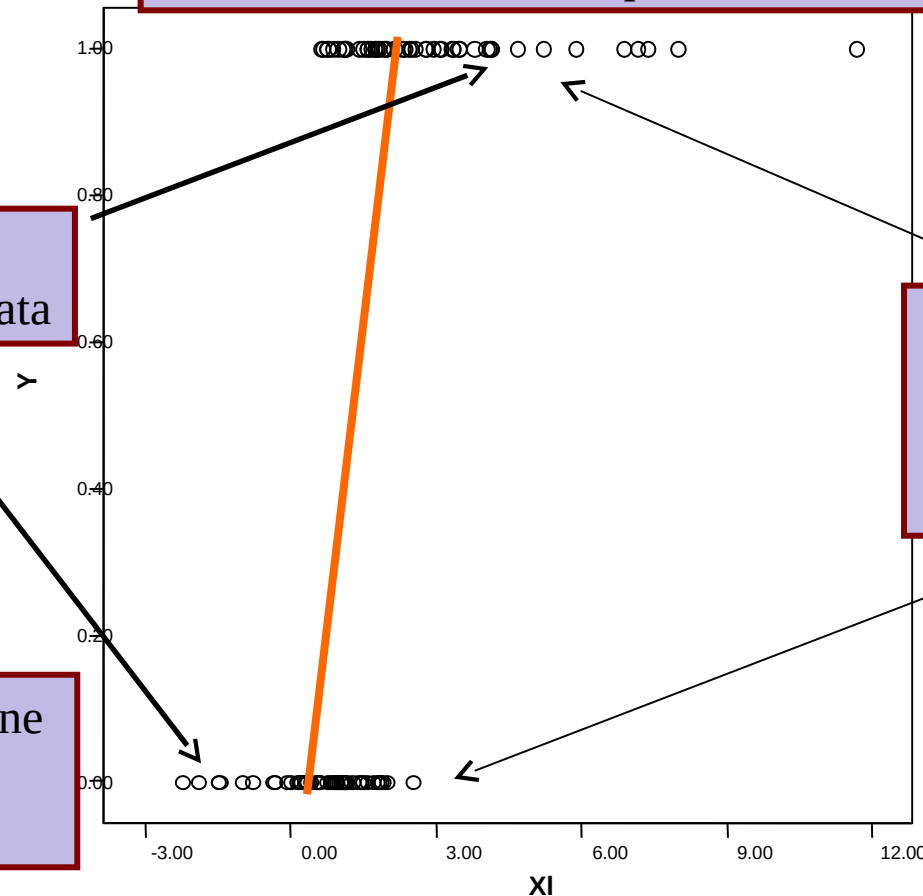
# VD dicotomica

La relazione non potrà mai essere lineare

Gran parte della  
varianza non è spiegata

La retta di regressione  
avrà moltissimo  
margine di errore

I punteggi saranno  
necessariamente  
raggruppati in due rette  
piatte



# Assunzione 1: Linearità

- Come visto precedentemente la relazione che riusciamo a catturare con la regressione è una relazione lineare
- Se la regressione è condotta con una variabile dipendente dicotomica, l'assunzione di linearità non può essere soddisfatta, creando problemi sia nella bontà della predizione, che nella sua interpretazione
- Dunque l'assunzione di linearità è sicuramente violata



# VD Categorical

- Quando abbiamo una **variabile dipendente dicotomica**, ogni soggetto ha o 1 o 0 come valore della variabile dipendente

VD=sexo (Maschi=0, Femmine=1), VD=acquisto (Si=1, No=0), voto al referendum (Si=1, No=0)

- La media della variabile dipendente è la probabilità di ottenere il valore 1

$$\bar{Y} = \frac{n_1}{n_{tot}}$$

- Ciò che prediciamo è la probabilità  $p$  di appartenere al gruppo con valore 1 (e  $1-p$  sarà la probabilità di appartenere al gruppo 0).

# Soluzione

- Necessitiamo dunque di un tipo di regressione che:
- Risolva il problema della omoschedasticità, linearità e normalità degli errori
- Ammetta valori non assurdi
- Ci esprima le relazioni sulla base di probabilità o qualcosa di comparabile
- **Cioè dobbiamo usare un modello che trasformi la variabile dipendente tale da linearizzare la relazione, rendere la variabile dipendente continua, e farla variare su tutto l'asse (valori positivi e negativi)**

# Soluzione: parte 1

- Intanto decidiamo di **non cercare di predire la probabilità**, ma il rapporto tra probabilità

- Tale rapporto è detto **odd (rapporto di probabilita')**

$$P_i = a + b_{yx} x_i \quad \Rightarrow \quad \frac{P_i}{1 - P_i} = a + b_{yx} x_i$$

# Odd

- L'odd è il rapporto tra la probabilità di un evento (appartenere ad un gruppo) rispetto alla probabilità del non evento (appartenere all'altro gruppo)

$$Odd = \frac{P_i}{1 - P_i}$$

- **Esempi: se la probabilità di avere una figlia femmina è .50**

$$Odd = \frac{.5}{1 - .5} = 1$$

- **se la probabilità di votare Si ad un referendum è .70**

$$Odd = \frac{.7}{1 - .7} = 2.33$$

# Odd

- L'odd indica quanto più probabile è un evento rispetto al suo complemento

$$\frac{P_i}{1-P_i}$$

- **Esempi: Una figlia femmina è tanto probabile quanto un maschio**

$$Odd = \frac{.5}{1-.5} = 1$$

- **Il voto Si è 2.33 volte più probabile del No**

$$Odd = \frac{.7}{1-.7} = 2.33$$

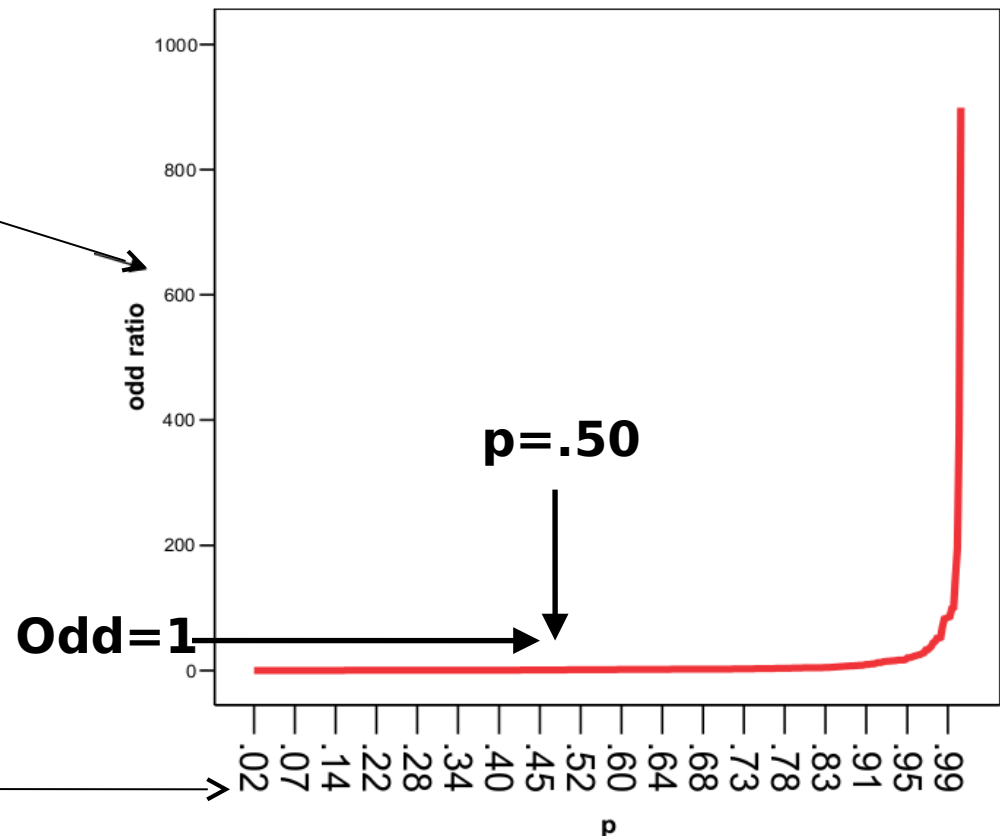
# Odd: Interpretazione

- L'odd consente di esprimere la probabilità mediante valori che variano da 0 ad infinito

Odd da 0 ad infinito

$$Odd = \frac{P_i}{1 - P_i}$$

Probabilità da 0 a 1



# Odd: Interpretazione

- L'odd varia da 0 ad infinito

Se gli eventi sono equiprobabili

$$p = .5 \rightarrow or = \frac{.5}{1 - .5} = 1$$

É maggiore di 1 se l'evento è più probabile del contrario

$$p = .7 \rightarrow or = \frac{.7}{1 - .7} = 2.33 > 1$$

É minore di 1 se l'evento è meno probabile del contrario

$$p = .2 \rightarrow or = \frac{.2}{1 - .2} = .25 < 1$$

# Problema con odd

- Se usassimo gli odd come variabile dipendente, potremmo ottenere predizioni impossibili, come predizioni di valori negativi

$$\frac{P_i}{1 - P_i} = a + b_{yx} x_i$$

Se  $a=1$ ,  $b=3$  e  $x=-2$

$$\frac{P_i}{1 - P_i} = 1 + 3 * (-2) = -5$$



## Soluzione: parte 2

- Decidiamo di non cercare di predire l'odd, ma il logaritmo dell'odd

$$\frac{P_i}{1 - P_i} = a + b_{yx} x_i \longrightarrow \ln\left(\frac{P_i}{1 - P_i}\right) = a + b_{yx} x_i$$

La trasformazione con il logaritmo si chiama **logit**

$$\log it = \ln\left(\frac{P_i}{1 - P_i}\right)$$

La regressione che cerca di predire il **logit** si chiama regressione **logistica**

# Vantaggi del Logaritmo

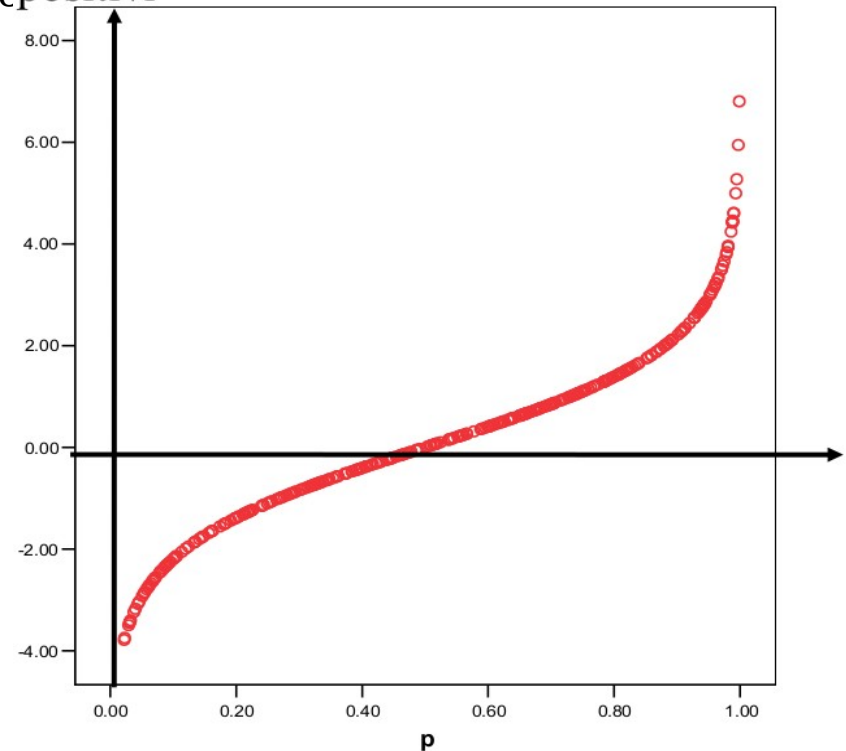
- La logistica usa il logaritmo in quanto:
  - Trasforma una variabile positiva (odd ratio) in negativi e positivi
  - É positivo se l'argomento è maggiore di 1  
(es.,  $\text{Ln}(5) = 1.6$        $2.718^{1.6} = 5$ )
  - É negativo se l'argomento è minore di 1  
(es.,  $\text{Ln}(0.2) = -1.6$        $2.718^{-1.6} = 0.2$ )
  - É zero se l'argomento è uguale ad 1  
( $\text{Ln}(1) = 0$        $2.718^0 = 1$ )

# Perché il logaritmo?

- Il logaritmo di una variabile che varia da 0 ad infinito (come gli odd), varia per tutti i valori possibili, da negativi a positivi . -

Il logaritmo dell'Odd  
permette di esprimere la  
probabilità mediante valori  
sia positivi che negativi

$$\log it = \ln\left(\frac{p}{1-p}\right)$$

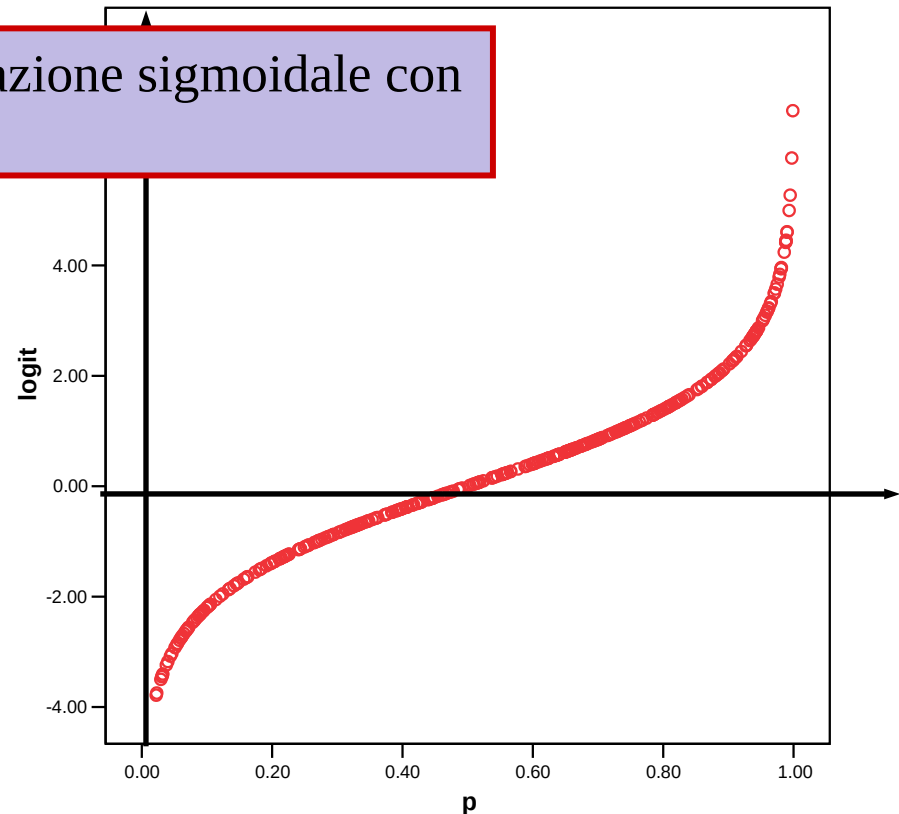


# Perché il logaritmo?

- Il logaritmo di una variabile che varia da 0 ad infinito (come gli odd ratio), varia per tutti i valori possibili, da negativi a positivi

Il logaritmo dell'OR sta in relazione sigmoideale con la probabilità

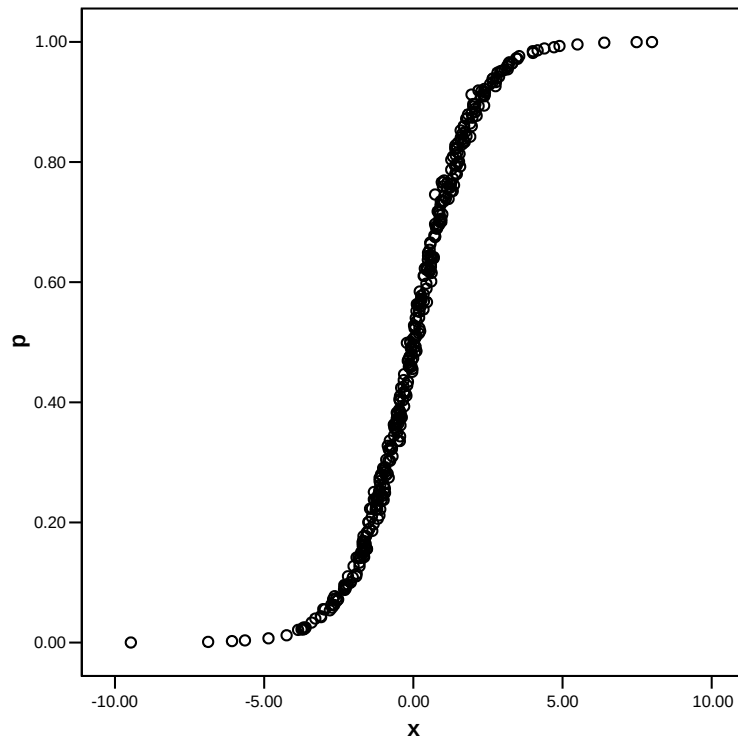
$$\log it = \ln\left(\frac{p}{1-p}\right)$$



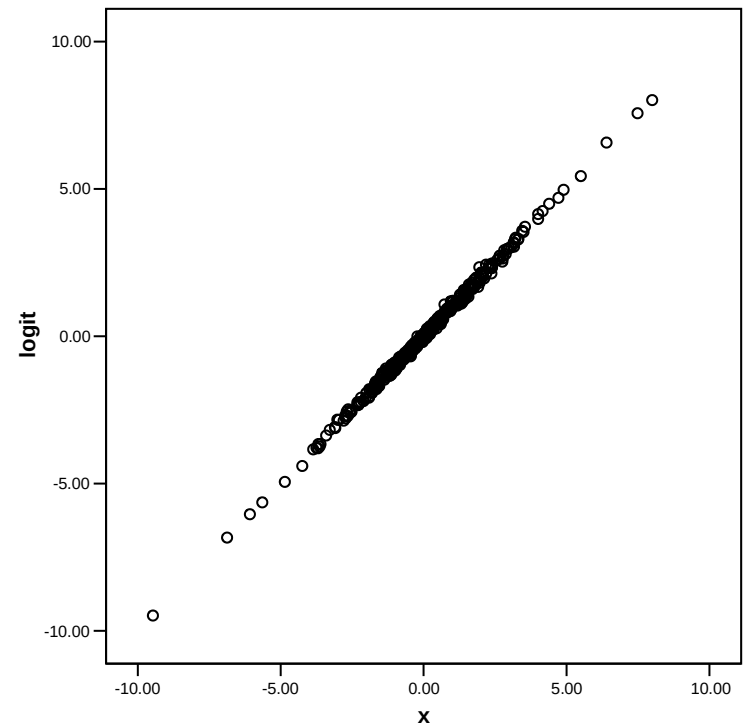
# Linearizzazione della relazione

- Grazie al fatto che il **logit** sta in rapporto sigmoidale con la probabilità, il logit sarà in rapporto lineare con le variabili dipendenti

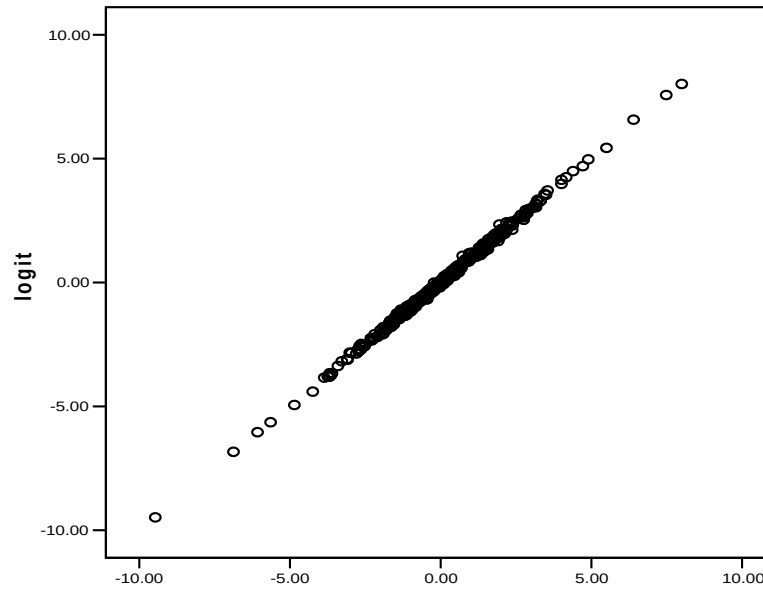
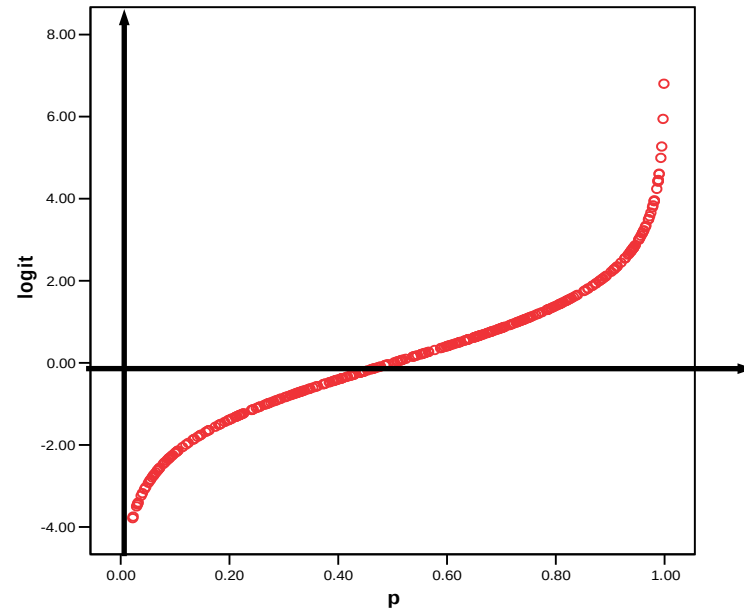
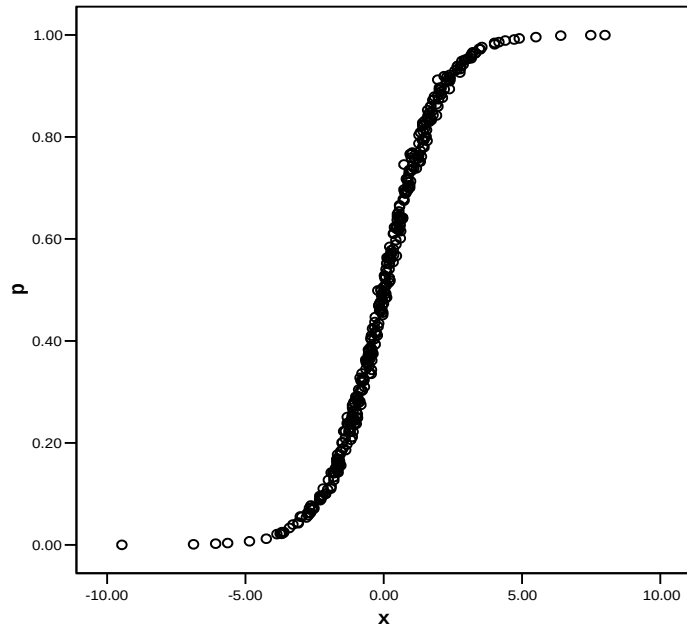
Se  $X$  predice  $P$  grazie ad una sigmoidale



$X$  predurrà LOGIT grazie ad una retta

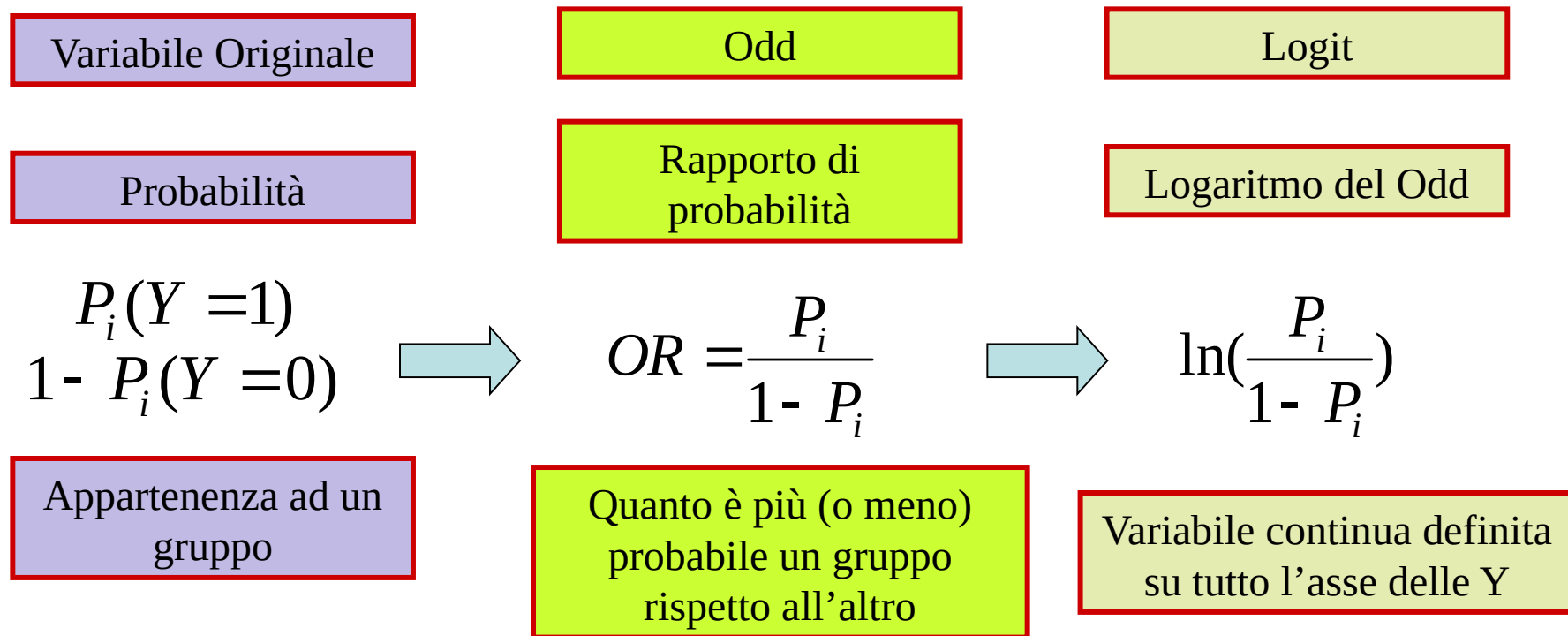


# Linearizzazione della relazione II



# Trasformazione Logit

- Per ovviare a ciò, la regressione logistica non predice la variabile dipendente così come è, ma la trasforma



# Regressione logistica

La regressione logistica è una regressione in cui la variabile dipendente è dicotomica, e dunque si predice mediante una regressione lineare il logaritmo del rapporto tra la probabilità di essere in un gruppo piuttosto che l'altro

$$\ln\left(\frac{P_i}{1 - P_i}\right) = a + b_{yx} x_i$$



# Modello logistico

- Il modello lineare si adatta ad una vasta gamma di tipologia di variabili dipendenti mediante due scelte: **link function** e **distribuzione**

$$f(Y) = a + b_{x.w} x_i + b_{w.x} w_i + e_i$$

Tipo di variabile

Link function

Distribuzione

Dicotomiche

Logit dell'odd

Binomiale

# Regressione Logistica

- Dato che la variabile è stata trasformata, la regressione ora è possibile
- Rispetto alla regressione che già conosciamo, cambierà:
  - **Come interpretare i coefficienti**
  - Il test di significatività:
    - R/jamovi: Chi-quadro e z-test
    - SPSS F-test e Wald test
  - **Come interpretare l' $R^2$**

# Logistica in pratica

● Ci proponiamo di studiare in un campione di 100 studenti soggetti, la relazione tra la carriera dello studente e le probabilità di essere promosso all'esame di statistica

● La variabile dipendente è:

- 0 bocciato
- 1 promosso

● La variabile indipendente è:

- la media agli esami precedenti

## Descriptives

Descriptives

	promosso	votip
N	100	100
Missing	0	0
Mean		22.7
Median		22.8
Minimum		18.0
Maximum		26.9





## Frequencies

Frequencies of promosso

Levels	Counts	% of Total	Cumulative %
0	32	32.0 %	32.0 %
1	68	68.0 %	100.0 %

# Logistica in pratica

## ● GAMLj

Clipboard		Variables		Rows	
	 votip	 ostudio	 promosso	 votip18	
1	22.112	1.477	1	4.112	
2	21.059	1.312	0	3.059	
3	21.819	1.518	0	3.819	
4	25.042	1.631	0	7.042	
5	22.506	1.574	0	4.506	
6	23.377	1.574	1	5.377	
7	24.665	1.629	0	6.665	
8	21.802	1.322	1	3.802	
9	25.592	1.723	1	7.592	
0	23.386	1.650	1	5.386	
1	22.006	1.348	0	4.006	
2	23.322	1.645	1	5.322	
3	18.000	0.942	0	0.000	
4	22.082	1.477	1	4.082	
5	21.042	1.422	1	3.042	

# Scelta del modello

## ● Linear Models: Generalized linear model

Generalized Linear Models

Continuous dependent variable    Categorical dependent variable

☐ Linear    ☒ Logistic

☐ Poisson    ☐ Probit

☐ Poisson (overdispersion)    ☐ Multinomial

☐ Negative Binomial

Dependent Variable

→

Factors

→

Covariates

→

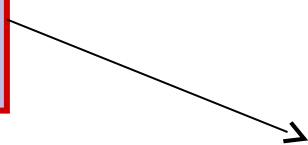
Effect Size    Confidence Intervals

☒ Odd Ratios (expB)    ☒ Confidence interval    Interval  %

Scegliamo prima quale  
tipo di modello stimare

## ●Recap del modello e R-quadro

McFadden's R squared:  
Proporzione di errore  
ridotto dal modello



### Generalized Linear Models

#### Model Info

Info	Value	Comment
Model Type	Logistic	Model for binary y
Link function	logit	log odd of promosso=1
Distribution	Binomial	Dichotomous event distribution of y
R-squared	0.158	Proportion of reduction of error
AIC	109.512	Less is better
Deviance	105.512	Less is better
Residual DF	98	
Converged	yes	A solution was found

- Come in GLM abbiamo I test sugli effetti generali (Chi-quadro) e la stima dei coefficienti

Analysis of Deviance: Omnibus Tests

	X <sup>2</sup>	df	p
votip	19.9	1	< .001

Model Coefficients (Parameter Estimates)

		Estimate	SE	95% Confidence Interval		exp(B)	z	p
	Contrast			Lower	Upper			
(Intercept)	Intercept	0.923	0.249	0.453	1.437	2.52	3.70	< .001
votip	votip	0.629	0.165	0.331	0.983	1.88	3.82	< .001



Il coefficiente è espresso nella scala logit!!

# Interpretazione

- Essendo una regressione, possiamo interpretare i coefficiente come al solito

Valore atteso nel logaritmo di Odd, quando la VI è zero

Model Coefficients (Parameter Estimates)

		Estimate	SE	95% Confidence Interval		exp(B)	z	p
Contrast				Lower	Upper			
(Intercept)	Intercept	0.923	0.249	0.453	1.437	2.52	3.70	< .001
votip	votip	0.629	0.165	0.331	0.983	1.88	3.82	< .001

Cambiamento atteso nel logaritmo di Odd, per uno spostamento nella di una unità nella VI

Significatività (valore-p): se minore di 0.05, rifiutiamo l'ipotesi nulla di  $B=0$



# Interpretazione: Problema

- Il problema sta nel fatto che tutte le informazioni (come in ogni regressione) sono espresse nell'unità di misura della VD
- Nel caso del logaritmo di Odd, questa unità è non intuitiva e poco informativa

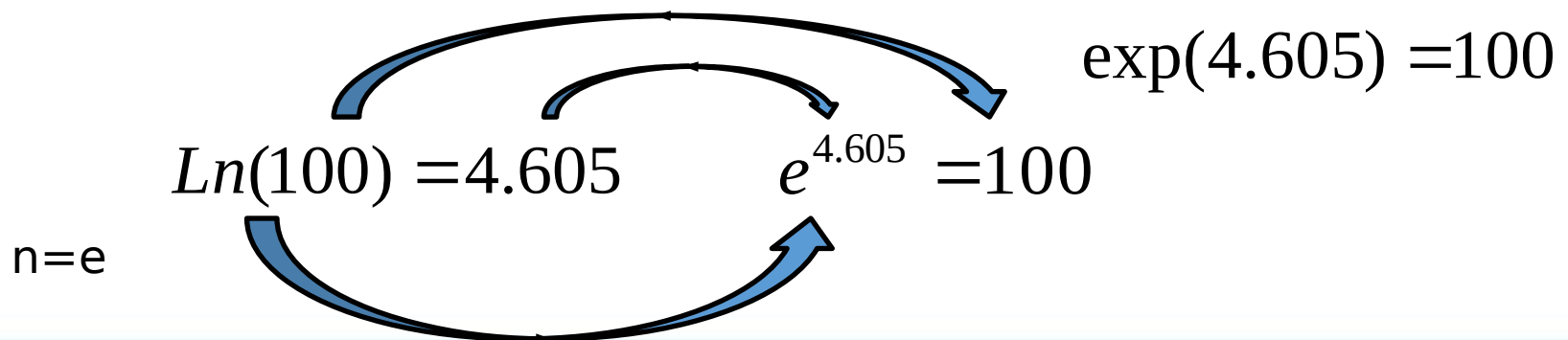
Model Coefficients (Parameter Estimates)

		Estimate	SE	95% Confidence Interval		exp(B)	z	p
	Contrast			Lower	Upper			
(Intercept)	Intercept	0.923	0.249	0.453	1.437	2.52	3.70	< .001
votip	votip	0.629	0.165	0.331	0.983	1.88	3.82	< .001

Per ogni voto medio in più, ci aspettiamo un aumento del logaritmo di Odd di .629! Ma è tanto o poco?

# Svantaggi del Logaritmo

- Il problema del logaritmo è che la sua unità di misura non è intuitivamente interpretabile
  - Una differenza di .629 nella scala logaritmica è tanto o poco in termini di probabilità?
- Per ovviare a ciò, le quantità espresse su scala logaritmiche possono essere riportate all'unità originale mediante la funzione esponenziale



# Unità più comprensibili

- Dato che nella logistica le informazioni sono ottenute sulla base di una VD logaritmica, la funzione esponenziale le riporta all'unità precedente (*funzione inversa*)
- L'unità precedente è l'odd ratio

Logit

Odd ratio

$\underset{\substack{\nearrow \text{base} \\ \uparrow \text{argomento}}}{\text{Ln}\left(\frac{P_i}{1 - P_i}\right)}$

$\longleftrightarrow$

$\underset{\substack{\nearrow \text{funzione esponenziale} \\ \uparrow \text{argomento}}}{\text{exp}\left(\text{Ln}\left(\frac{P_i}{1 - P_i}\right)\right)} = \frac{P_i}{1 - P_i}$

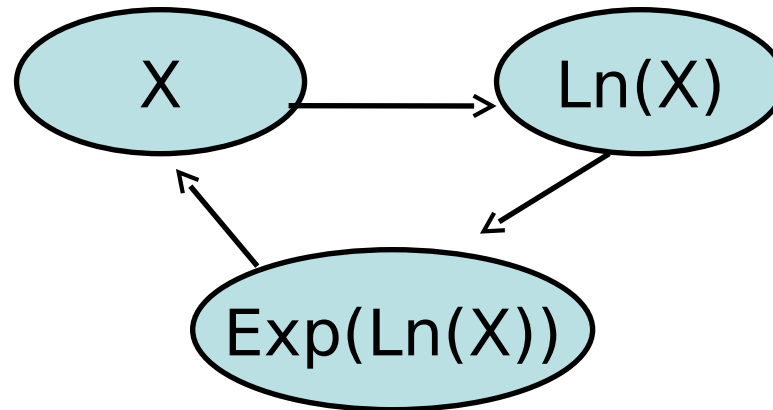
logaritmo

In generale

$\text{exp}(\text{Ln}(q)) = q$

# L'esponenziale

- La funzione esponenziale di un logaritmo ci dà l'argomento originale del logaritmo



# Relazioni tra unità di misure

- Al fine interpretativo è importante ricordare che:
- La somma tra due logaritmi, equivale al prodotto tra gli argomenti

Logit

Odd ratio

$$q = \text{Ln}(a) + \text{Ln}(b)$$

$$\text{Ln}(2) + \text{Ln}(3) = 1.79$$

$$B = \text{Ln}(\text{Odd}_1) + \text{Ln}(\text{Odd}_2)$$



$$\exp(q) = a * b$$

$$\exp(1.79) = 3 \times 2 = 6$$

$$\exp(B) = \text{Odd}_1 * \text{Odd}_2$$

# B espresso come OR

- Per facilitare l'interpretazione, il legame tra VD e VI si esprime mediante l'esponenziale di B

Model Coefficients (Parameter Estimates)

		Estimate	SE	95% Confidence Interval		exp(B)	z	p
	Contrast			Lower	Upper			
(Intercept)	Intercept	0.923	0.249	0.453	1.437	2.52	3.70	< .001
votip	votip	0.629	0.165	0.331	0.983	1.88	3.82	< .001

Exp(B) trasforma il B espresso in scala logaritmica in un B espresso in termini di odd ratio

# B espresso come OR

- Per facilitare l'interpretazione, il legame tra VD e VI si esprime mediante l'esponenziale di B

Model Coefficients (Parameter Estimates)

		Estimate	SE	95% Confidence Interval		exp(B)	z	p
	Contrast			Lower	Upper			
(Intercept)	Intercept	0.923	0.249	0.453	1.437	2.52	3.70	< .001
votip	votip	0.629	0.165	0.331	0.983	1.88	3.82	< .001

Per ogni unità in più di *votip*, il rapporto di probabilità tra votare promosso e bocciato (odd) aumenta di 1.88 volte

# Relazioni tra unità di misure

- Ma la somma di due logaritmi equivale alla esponenziale dei prodotti degli argomenti

Logit

Odd ratio

$$q = \text{Ln}(a) + \text{Ln}(b)$$



$$\exp(q) = a * b$$



# Interpretazione di $\exp(B)$

- Aumentando X di 1, aumento in termini di logaritmo di .629
- L'esponenziale di .629 è 1.88

Model Coefficients (Parameter Estimates)

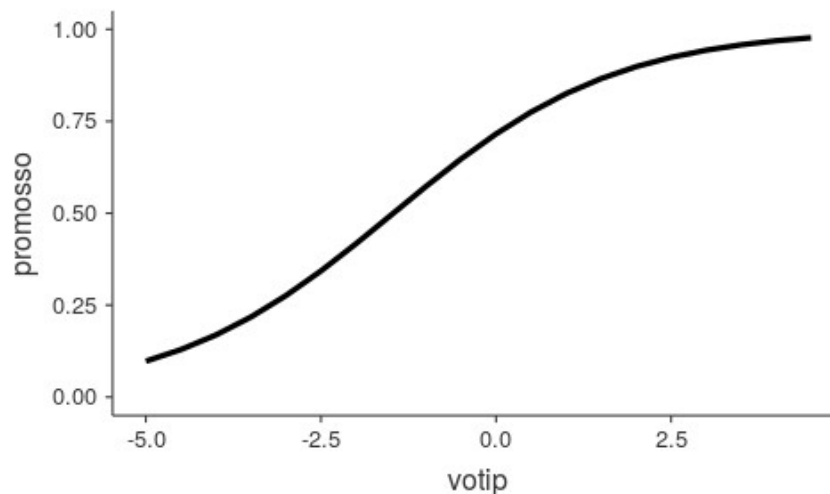
		Estimate	SE	95% Confidence Interval		$\exp(B)$	z	p
	Contrast			Lower	Upper			
(Intercept)	Intercept	0.923	0.249	0.453	1.437	2.52	3.70	< .001
votip	votip	0.629	0.165	0.331	0.983	1.88	3.82	< .001

- Dunque in termini di Odd c'è un aumento di 1.88
- Cioè per ogni voto in più in media l'odd di essere promosso aumenta di 1.88 volte

# Interpretazione del grafico

- Possiamo sempre chiedere il grafico dei valori predetti dal modello. In GAMLj il grafico è espresso in probabilità (del gruppo 1)

Effects Plots

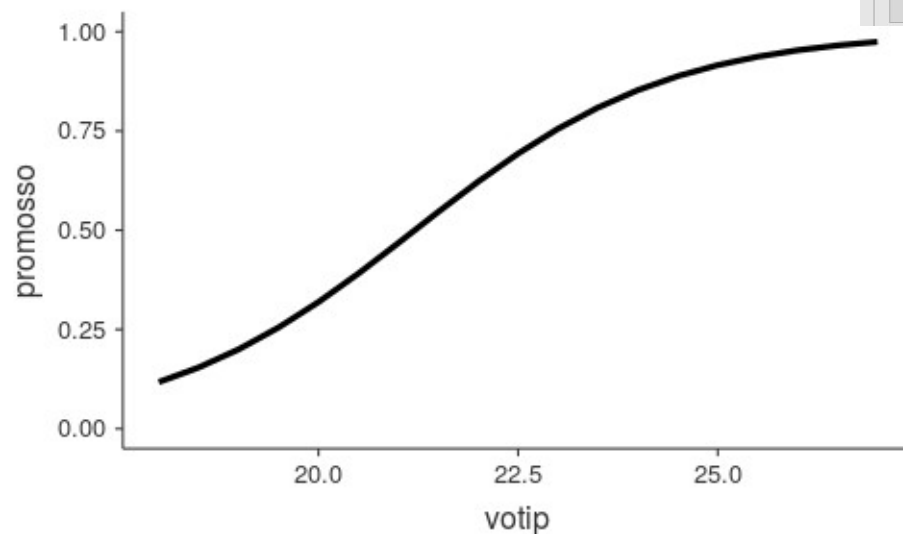


Ricordiamoci che  
GAMLj centra la X

# Interpretazione del grafico

- Se settiamo il “covariates scaling” to “none”, otteniamo un grafico più comprensibile

Effects Plots



The image shows a screenshot of the SPSS software interface. The 'Covariates Scaling' dropdown menu is open, showing the following options: none, none, centered, and standardized. The 'none' option is selected. To the right of the dropdown menu, there is a small plot showing a line with a positive slope. Below the dropdown menu, there are buttons for 'Post Hoc Tests' and 'Plots'. On the far right, there is a 'Descriptives' panel showing the following table:

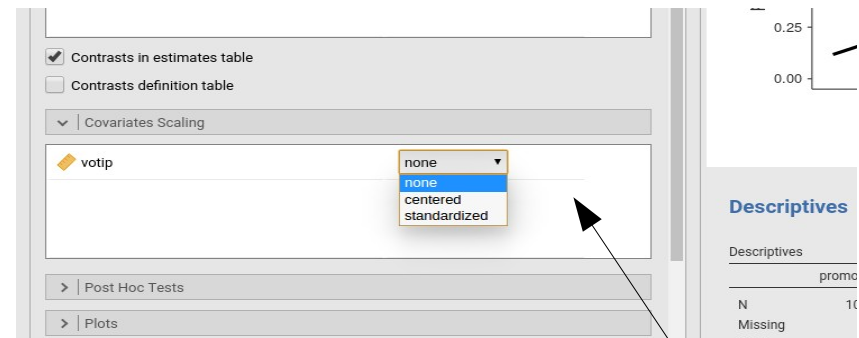
Descriptives	
	promc
N	11
Missing	

# Effetti standardizzati

- Notiamo che nella regressione logistica non è previsto il coefficiente di regressione standardizzato
- Ma noi sappiamo che per standardizzare i coefficienti basta standardizzare le variabili
- Il logit non lo possiamo standardizzare, ma la variabile indipendente si

# Effetti standardizzati

## ● GAMLj:



Model Coefficients (Parameter Estimates)

		95% Confidence Interval				exp(B)	z	p
	Contrast	Estimate	SE	Lower	Upper			
(Intercept)	Intercept	0.923	0.249	0.453	1.44	2.52	3.70	< .001
votip	votip	1.112	0.291	0.586	1.74	3.04	3.82	< .001

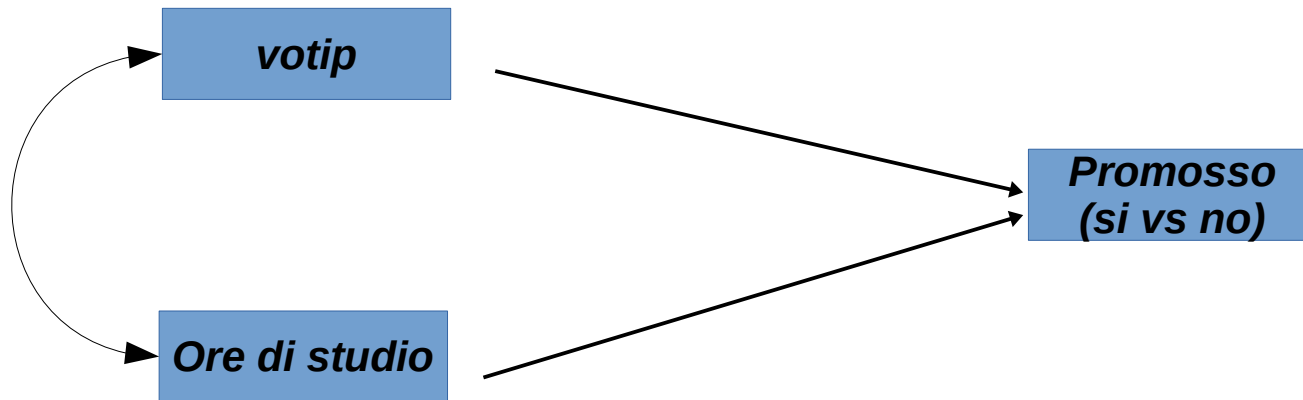
Per ogni **deviazione standard** in più di *votip*, il rapporto di probabilità tra votare promosso e bocciato (odd) aumenta di 3.04 volte

# Morale

- Tutto ciò che sappiamo sulla regressione lineare (interazione, effetti parziali, mediazione, path analysis) rimane concettualmente equivalente per la logistica
- Cambia cosa si predice ed il calcolo dei coefficienti

# Regressione logistica multipla

- Aggiungiamo al modello di prima la variabile predittrice “ore di studio a settimana” otteniamo un modello logistico multiplo



# Logistica multipla

● Aggiungiamo al modello di prima la variabile predittrice “ore di studio a settimana” otteniamo un modello logistico multiplo

Generalized Linear Models

**Continuous dependent variable**

☐ Linear

**Categorical dependent variable**

☒ Logistic

☐ Probit

☐ Multinomial

**Frequencies**

☐ Poisson

☐ Poisson (overdispersion)

☐ Negative Binomial

**Dependent Variable**

→

**Factors**

→

**Covariates**

→

**Effect Size**

☒ Odd Ratios (expB)

**Confidence Intervals**

☒ Confidence interval Interval  %



# Recap del modello

- R-squared

## Generalized Linear Models

### Model Info

Info	Value	Comment
Model Type	Logistic	Model for binary y
Link function	logit	log odd of promosso=1
Distribution	Binomial	Dichotomous event distribution of y
R-squared	0.173	Proportion of reduction of error
AIC	109.631	Less is better
Deviance	103.631	Less is better
Residual DF	97	
Converged	yes	A solution was found

### Analysis of Deviance: Omnibus Tests

	X <sup>2</sup>	df	p
votip	5.31	1	0.021
ostudio	1.88	1	0.170

# Coefficienti

- I coefficienti sono espressi nella scala logaritmica [**B**] e nella scala dei rapporti di probabilit  o odd ratio [**exp(B)**]

Analysis of Deviance: Omnibus Tests

	X <sup>2</sup>	df	p
votip	5.31	1	0.021
ostudio	1.88	1	0.170

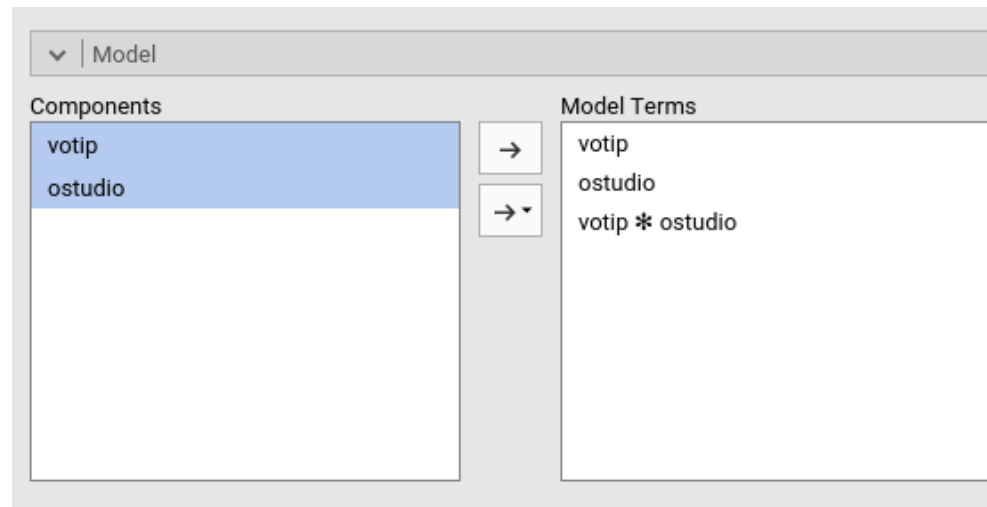
Model Coefficients (Parameter Estimates)

	Contrast	Estimate	SE	95% Confidence Interval		exp(B)	z	p
				Lower	Upper			
(Intercept)	Intercept	0.939	0.253	0.4637	1.463	2.56	3.71	< .001
votip	votip	0.452	0.204	0.0661	0.879	1.57	2.21	0.027
ostudio	ostudio	1.767	1.302	-0.7553	4.402	5.86	1.36	0.175

Interpretati come prima, ma aggiungendo “al netto di ...”

# Moderazione

- Ovviamente possiamo inserire anche una interazione, come in qualunque altro modello lineare



# Moderazione

- Ovviamente possiamo inserire anche una interazione, come in qualunque altro modello lineare

Analysis of Deviance: Omnibus Tests

	X <sup>2</sup>	df	p
votip	5.078	1	0.024
ostudio	1.299	1	0.254
votip * ostudio	0.725	1	0.395

Model Coefficients (Parameter Estimates)

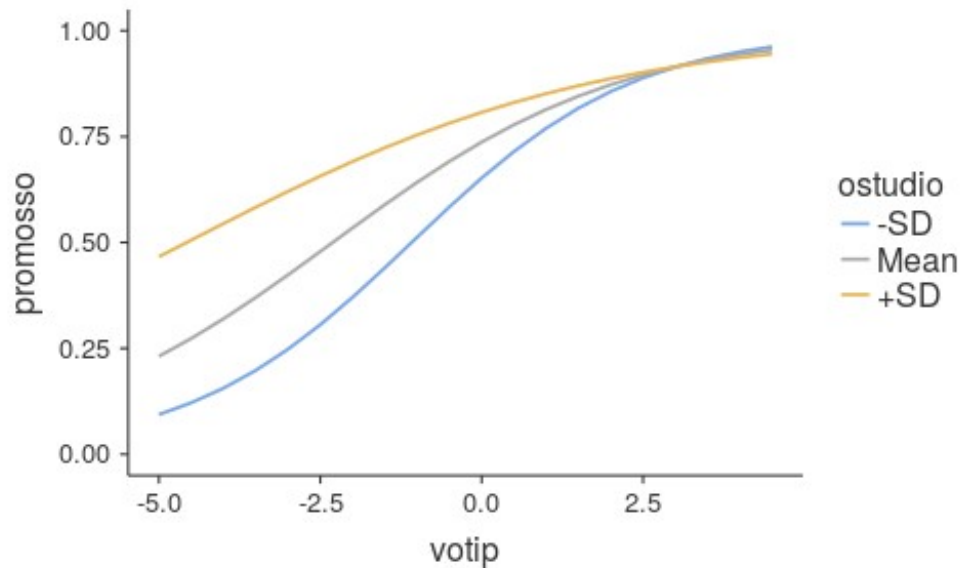
	Contrast	Estimate	SE	95% Confidence Interval		exp(B)	z	p
				Lower	Upper			
(Intercept)	Intercept	1.031	0.276	0.5092	1.599	2.803	3.731	< .001
votip	votip	0.446	0.205	0.0571	0.875	1.562	2.172	0.030
ostudio	ostudio	1.504	1.329	-1.0811	4.190	4.499	1.132	0.258
votip * ostudio	votip * ostudio	-0.497	0.589	-1.6709	0.632	0.608	-0.843	0.399

Effetti principali e interazione

# Moderazione

- Volendo possiamo plottare le simple slopes (esprese in probabilità)

Effects Plots



In questo caso non è significativa, ma in generale nel grafico vediamo come l'andamento delle probabilità associate a X cambiano per diversi valori del moderatore

# Morale

- La **Regressione Logistica** e' una regressione con una VD binaria
- Si focalizza sulla probabilita' di appartenenza al gruppo
- I coefficienti sono espressi in scala logaritmica (B) come Odd Ratio  $\exp(B)$
- Il  $\exp(B)$  e' la quantita' per la quale OR viene moltiplicato quando muoviamo la VI di 1 unita'
- La bonta' dell'equazione complessiva e' espressa con  $R^2$
- logica di fondo e' come per la Regressione Lineare Multipla
- E' comprensibile come un caso particolare di Modello Lineare Generalizzato

## Il modello multinomiale

# Il modello multinomiale

- ◆ Il modello multinomiale (regressione multinomiale) si propone di studiare e quantificare le relazioni tra una o più variabili indipendenti quantitative (es. età, salario, atteggiamenti, personalità) e una variabile **dipendente categorica** (con più di due gruppi)
- ◆ Sostanzialmente generalizza la logistica predicendo delle dummy che deconpongono la variabile dipendente categorica



- ◆ Un campione di studenti (USA) può scegliere tra tre curricula scolastici: Accademic, general, vocational.
- ◆ Il ricercatore ha a disposizione un test di performance scolastica per ogni soggetto, capacità di scrittura (*write*), e lo stato socio economico (*ses*) del soggetto
- ◆ Vogliamo capire come e se queste variabili predicono la scelta del curriculum scolastico

# Esempio

- ◆ Un campione di studenti (USA) può scegliere tra tre curricula scolastici: Accademic, general, vocational.

## Contingency Tables

Contingency Tables

prog	ses			Total
	high	low	middle	
academic	42	19	44	105
general	9	16	20	45
vocation	7	12	31	50
Total	58	47	95	200

## Descriptives

Descriptives

	write
N	200
Missing	0
Mean	52.8
Median	54.0
Minimum	31
Maximum	67

- ◆ Write è un test di performance con punteggio continuo

# Multinomiale

- La **regressione multinomiale** equivale a G-1 regressioni che cercano di predire la variabile dipendente decomposta in G-1 dummy

Nuove variabili dipendenti

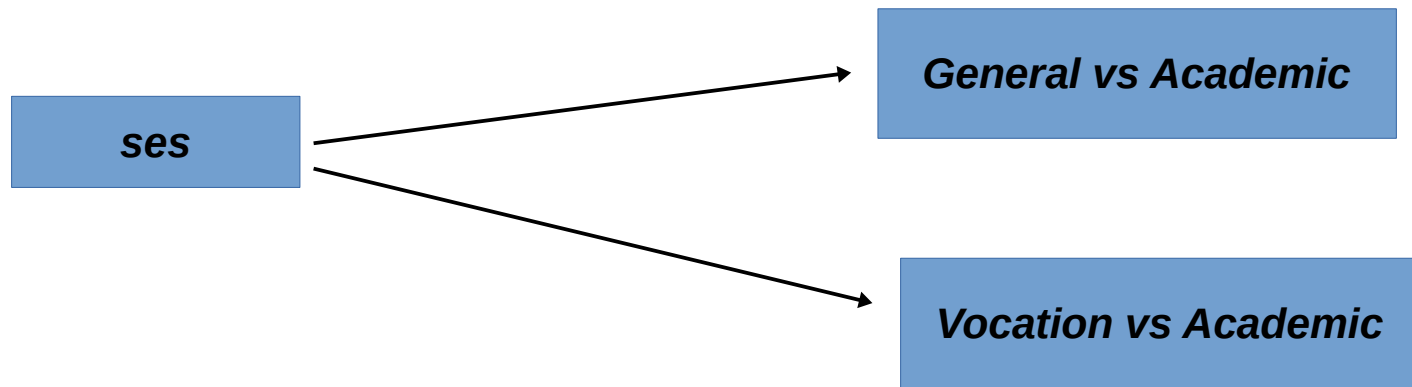
Var Indip	Categoria	Dip1	Dip2
Programma	academic	0	0
	general	1	0
	vocation	0	1

## Frequencies

Frequencies of prog	
Levels	Counts
academic	105
general	45
vocation	50

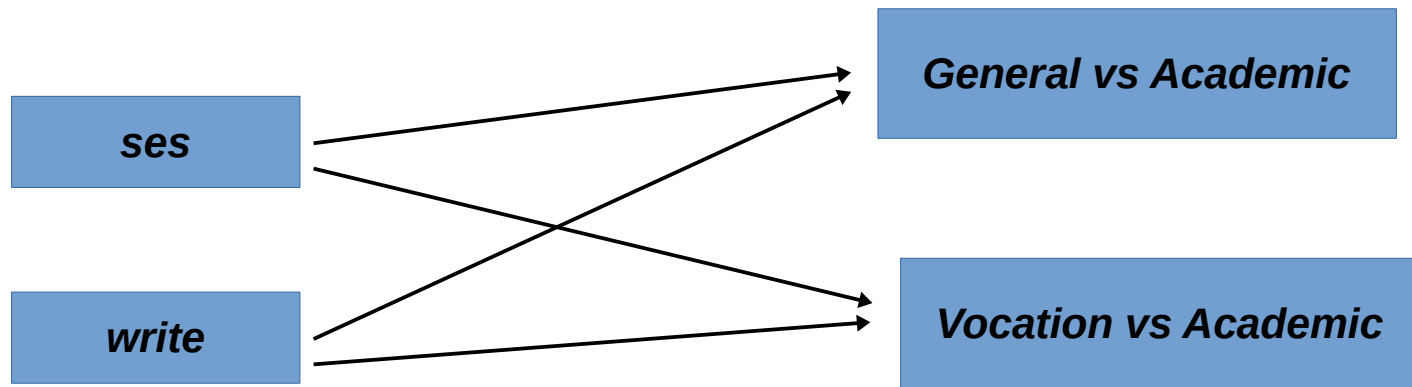
# Logica della multinomiale

- Utilizzando una serie di (pseudo) logistiche, prediciamo le dummies che decompongono la categoria



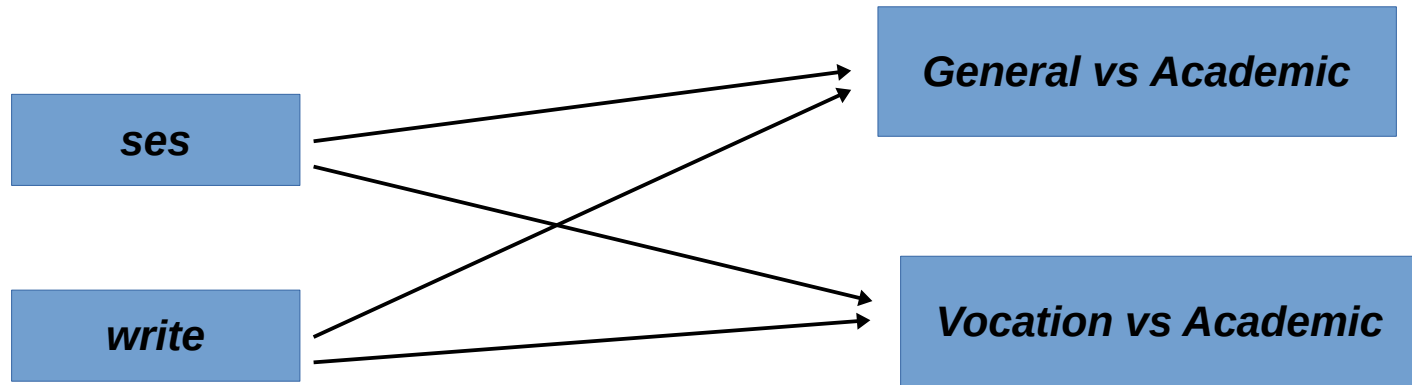
# Logica della multinomiale

- Utilizzando una serie di (pseudo) logistiche, prediciamo le dummies che decompongono la categoria



# Logica della multinomiale

- Interpretaremo i risultati come abbiamo fatto per la regressione logistica, ricordando che la predizione è sulle dummies della dipendente



# Relative risk

- La **regressione multinomiale** la probabilità di essere in un gruppo rispetto al gruppo reference è detta relative risk (una sorta di odd)

		Relative risk general/academic	Relative risk vocation/academic
Var Indip	Categoria	Dip1	Dip2
Programma	academic	0	0
	general	1	0
	vocation	0	1

# Modello multinomiale

- Il modello lineare si adatta ad una vasta gamma di tipologia di variabili dipendenti mediante due scelte: **link function** e **distribuzione**

$$f(Y) = a + b_{x.w} x_i + b_{w.x} w_i + e_i$$

Tipo di variabile

Link function

Distribuzione

categoriche


Logit del relative risk

Multinomiale



# Logistica multinomiale

● Selezioniamo tipo di modello “multinomial” e procediamo a settare le variabili

Generalized Linear Models 

**Continuous dependent variable**      **Categorical dependent variable**

☐ Linear      ☐ Logistic

☐ Poisson      ☐ Probit

☐ Poisson (overdispersion)      ☒ Multinomial

☐ Negative Binomial

**Frequencies**

☐ Poisson

☐ Poisson (overdispersion)

☐ Negative Binomial

**Dependent Variable**

→

**Factors**

→

**Covariates**

→

**Effect Size**      **Confidence Intervals**

☒ Odd Ratios (expB)      ☒ Confidence interval      Interval  %

**Variables List:**

- A
- id
- female
- ses
- schtyp
- read
- math
- science
- socst
- honors
- awards

# Logistica multinomiale

- Recap e R-quadro

## Generalized Linear Models

### Model Info

Info	Value	Comment
Model Type	Multinomial	Model for categorical y
Link function	logit	Log of the odd of each category over y=0
Distribution	Multinomial	Multi-event distribution of y
R-squared	0.0911	Proportion of reduction of error
AIC	379.0217	Less is better
Deviance	371.0217	Less is better
Residual DF	4.0000	
Converged	yes	A solution was found

# Logistica multinomiale

## ● Tests e coefficienti

Test omnibus: Effetto principale di  
“write” sulla categorica (in generale)

Analysis of Deviance: Omnibus Tests

	X <sup>2</sup>	df	p
write	37.2	2	< .001

Model Coefficients (Parameter Estimates)

Response Groups		Contrast	Estimate	SE	95% Confidence Interval		exp(B)	z	p
					Lower	Upper			
general - academic	(Intercept)	Intercept	-0.7711	0.1849	-1.134	-0.4086	0.463	-4.17	< .001
	write	write	-0.0660	0.0210	-0.107	-0.0248	0.936	-3.14	0.002
vocation - academic	(Intercept)	Intercept	-0.8584	0.1995	-1.249	-0.4673	0.424	-4.30	< .001
	write	write	-0.1178	0.0216	-0.160	-0.0754	0.889	-5.45	< .001

# Logistica multinomiale

## ● Tests e coefficienti

Analysis of Deviance: Omnibus Tests

	X <sup>2</sup>	df	p
write	37.2	2	< .001

Model Coefficients (Parameter Estimates)

Response Groups		Contrast	Estimate	SE	95% Confidence Interval		exp(B)	z	p
					Lower	Upper			
general - academic	(Intercept)	Intercept	-0.7711	0.1849	-1.134	-0.4086	0.463	-4.17	< .001
	write	write	-0.0660	0.0210	-0.107	-0.0248	0.936	-3.14	0.002
vocation - academic	(Intercept)	Intercept	-0.8584	0.1995	-1.249	-0.4673	0.424	-4.30	< .001
	write	write	-0.1178	0.0216	-0.160	-0.0754	0.889	-5.45	< .001

Effetto (in logit) di “write” sulla scelta  
“general vs academic”

Effetto (in logit) di “write” sulla scelta  
“vocation vs academic”

# Logistica multinomiale

## ● Tests e coefficienti

Analysis of Deviance: Omnibus Tests

	$\chi^2$	df	p
write	37.2	2	< .001

Model Coefficients (Parameter Estimates)

Response Groups		Contrast	Estimate	SE	95% Confidence Interval		exp(B)	z	p
					Lower	Upper			
general - academic	(Intercept)	Intercept	-0.7711	0.1849	-1.134	-0.4086	0.463	-4.17	< .001
	write	write	-0.0660	0.0210	-0.107	-0.0248	0.936	-3.14	0.002
vocation - academic	(Intercept)	Intercept	-0.8584	0.1995	-1.249	-0.4673	0.424	-4.30	< .001
	write	write	-0.1178	0.0216	-0.160	-0.0754	0.889	-5.45	< .001

Effetto (in odd) di “write” sulla scelta  
“general vs academic”

Effetto (in odd) di “write” sulla scelta  
“vocation vs academic”

# Logistica multinomiale

## ● Esempio di interpretazione

Analysis of Deviance: Omnibus Tests

	X <sup>2</sup>	df	p
write	37.2	2	< .001

Model Coefficients (Parameter Estimates)

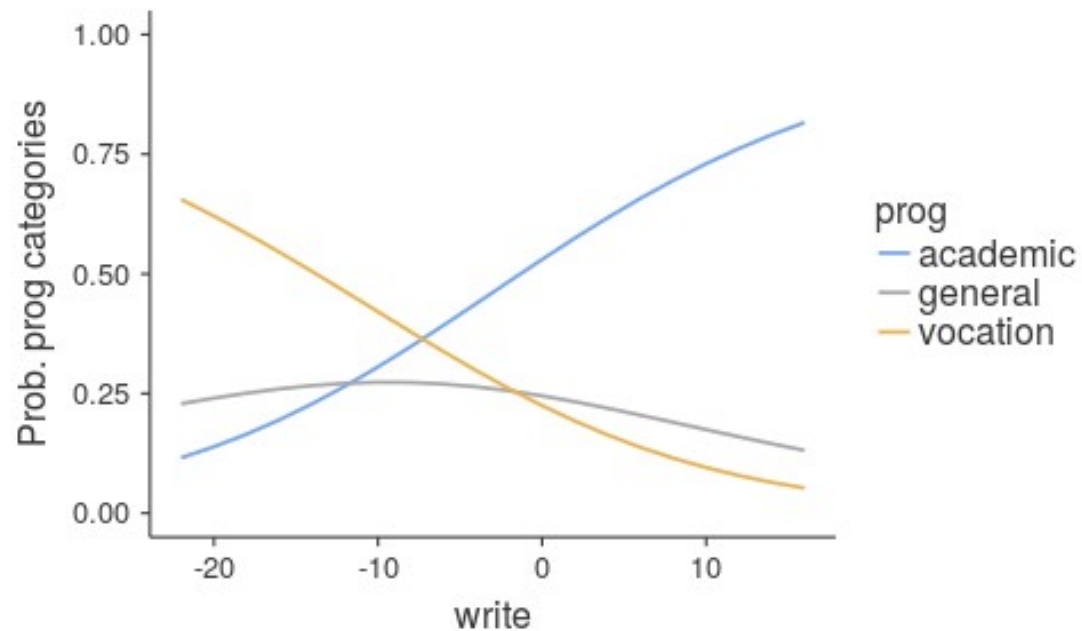
Response Groups		Contrast	Estimate	SE	95% Confidence Interval		exp(B)	z	p
					Lower	Upper			
general - academic	(Intercept)	Intercept	-0.7711	0.1849	-1.134	-0.4086	0.463	-4.17	< .001
	write	write	-0.0660	0.0210	-0.107	-0.0248	0.936	-3.14	0.002
vocation - academic	(Intercept)	Intercept	-0.8584	0.1995	-1.249	-0.4673	0.424	-4.30	< .001
	write	write	-0.1178	0.0216	-0.160	-0.0754	0.889	-5.45	< .001

All'aumentare dello score "write" di una unità, l'odd di scegliere general rispetto a academic aumenta di .936 volte, dunque diminuisce

# Grafico


- Il grafico degli effetti è espresso in probabilità di appartenere ad un gruppo (fare un scelta di curriculum) in funzione della X

Effects Plots














# Logistica multinomiale


- Per una (o più) indipendenti categoriche procediamo come in tutti i modello lineari

Generalized Linear Models 


☐ Negative Binomial

 A  
 id  
 female  
 schtyp  
 read  
 write  
 math  
 science  
 socst  
 honors  
 awards

→

Dependent Variable  
 prog

→

Factors  
 ses

←

Covariates

**Effect Size** **Confidence Intervals**

☒ Odd Ratios (expB) ☒ Confidence interval Interval  %



# Logistica multinomiale

- Interpretiamo i coefficienti ricordando che l'effetto associato ad una VI categorica è sempre dato dalla differenza fra gruppi definiti dalla VI

Analysis of Deviance: Omnibus Tests

	X <sup>2</sup>	df	p
ses	16.8	4	0.002

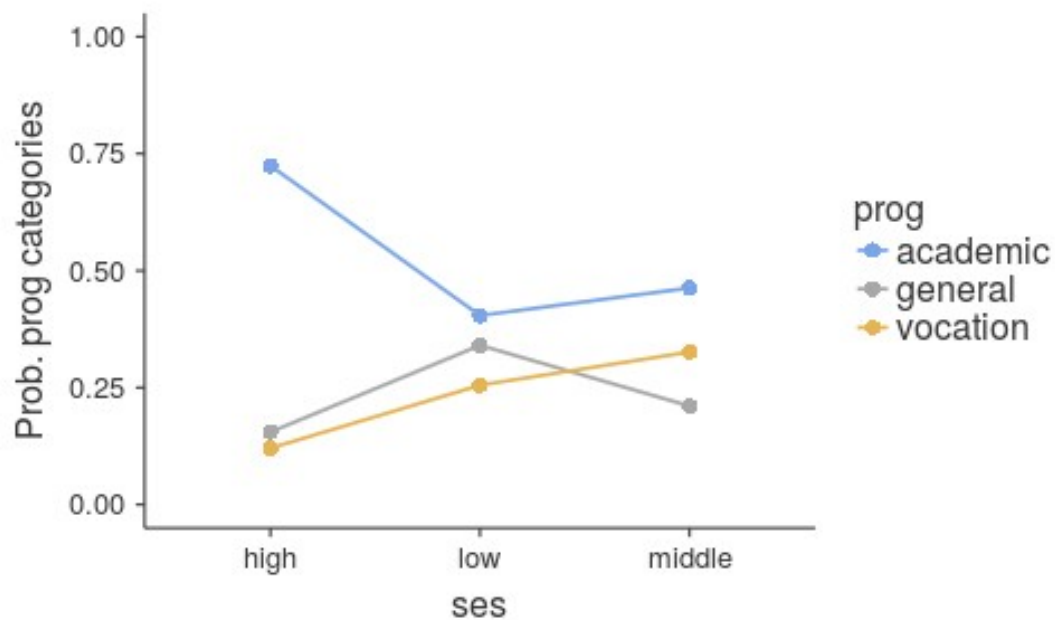
Model Coefficients (Parameter Estimates)

Response Groups		Contrast	Estimate	SE	95% Confidence Interval		exp(B)	z	p
					Lower	Upper			
general - academic	(Intercept)	Intercept	-0.8336	0.189	-1.2048	-0.462	0.434	-4.402	< .001
	ses1	low - ( high, low, middle )	0.6617	0.272	0.1277	1.196	1.938	2.429	0.015
	ses2	middle - ( high, low, middle )	0.0451	0.245	-0.4354	0.526	1.046	0.184	0.854
vocation - academic	(Intercept)	Intercept	-0.8672	0.199	-1.2579	-0.476	0.420	-4.350	< .001
	ses1	low - ( high, low, middle )	0.4076	0.292	-0.1640	0.979	1.503	1.398	0.162
	ses2	middle - ( high, low, middle )	0.5170	0.241	0.0447	0.989	1.677	2.145	0.032

# Logistica multinomiale

- Interpretiamo il grafico in termini di probabilità

Effects Plots



# Modello di Poisson

(variabili dipendenti a frequenze)

# Modelli lineari generalizzati

- Il modello lineare si adatta ad una vasta gamma di tipologia di variabili dipendenti mediante due scelte: **link function** e **distribuzione**

$$f(Y) = a + b_{x.w} x_i + b_{w.x} w_i + e_i$$

Tipo di variabile

Link function

Distribuzione

Frequenze

LN delle frequenze

Poisson

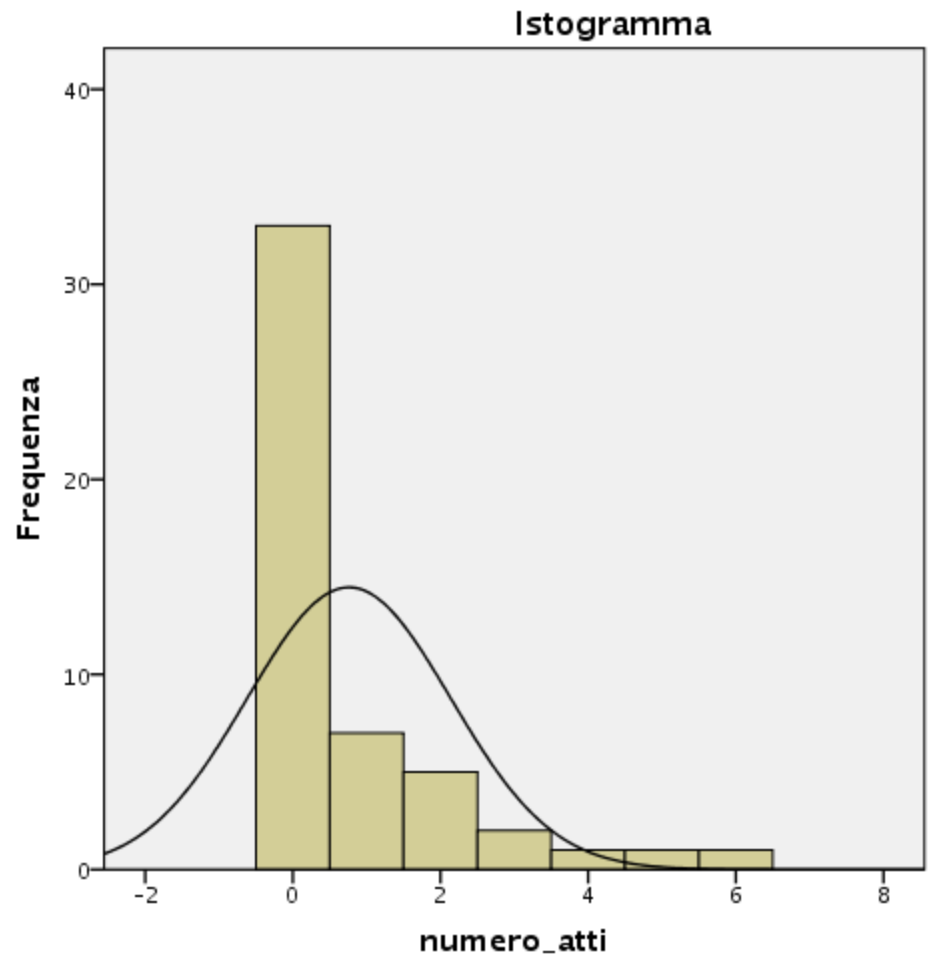
# Variabili dipendenti a frequenza

Un campione di bambini viene misurato mediante un test di aggressività basato sulla valutazione delle maestre dell'asilo o della scuola. Per capire la validità della misura vengono predisposte delle osservazioni sui bambini durante le ore di gioco libero a scuola o all'asilo. Le registrazioni della sessione di gioco vengono visionate da un esperto che conta il numero di atti aggressivi compiuti da ogni singolo bambino. La variabile che viene prodotta è dunque il numero (la frequenza) di atti aggressivi nelle sessioni per ogni bambino.

numero_atti					
		Frequenza	Percentuale	Percentuale valida	Percentuale cumulata
Validi	0	33	66.0	66.0	66.0
	1	7	14.0	14.0	80.0
	2	5	10.0	10.0	90.0
	3	2	4.0	4.0	94.0
	4	1	2.0	2.0	96.0
	5	1	2.0	2.0	98.0
	6	1	2.0	2.0	100.0
	Totale	50	100.0	100.0	

# Variabili dipendenti a frequenza

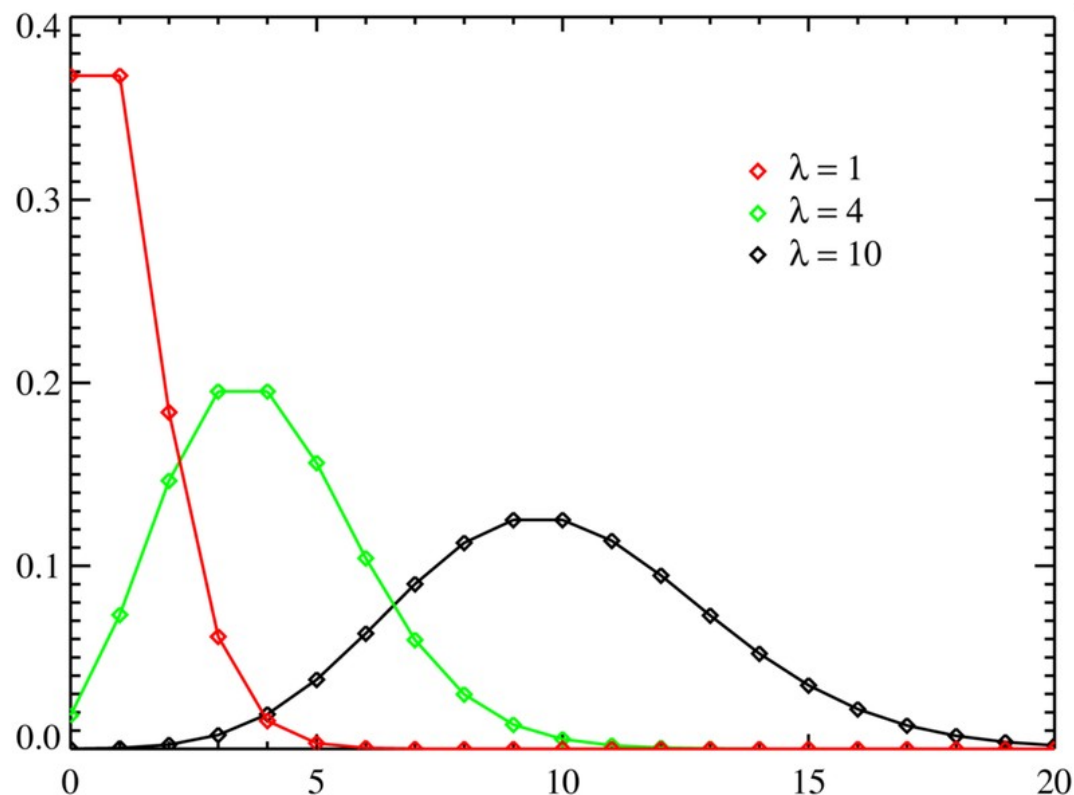
La variabile dipendente è  
chiaramente non normale!  
Segue invece un'altra distribuzione  
nota detta **distribuzione di Poisson**



# Distribuzione Variabili “counts”

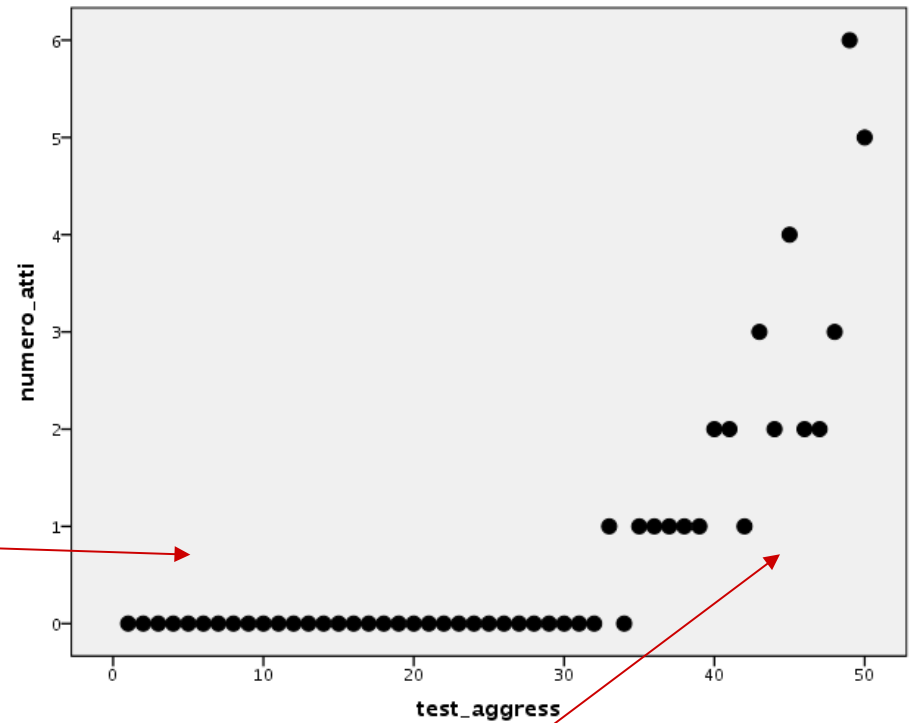
- La distribuzione di Poisson indica la probabilità di osservare un evento con differenti frequenze, data un certo tasso ( $\lambda$ ) di incidenza dell'evento

Più l'evento è raro, meno la distribuzione assomiglia ad una normale



# Variabili “counts”

- Sono raramente in relazione lineare con le VI



Tendono ad essere basse  
per molti valori della VI


Incrementano rapidamente  
per altri valori della VI



# Distribuzione Variabili “counts”

- Le variabili a frequenza tendono ad avere una distribuzione di Poisson
- La distribuzione di Poisson indica la probabilità di osservare un evento con differenti frequenze, data un certo tasso ( $\lambda$ ) di incidenza dell'evento

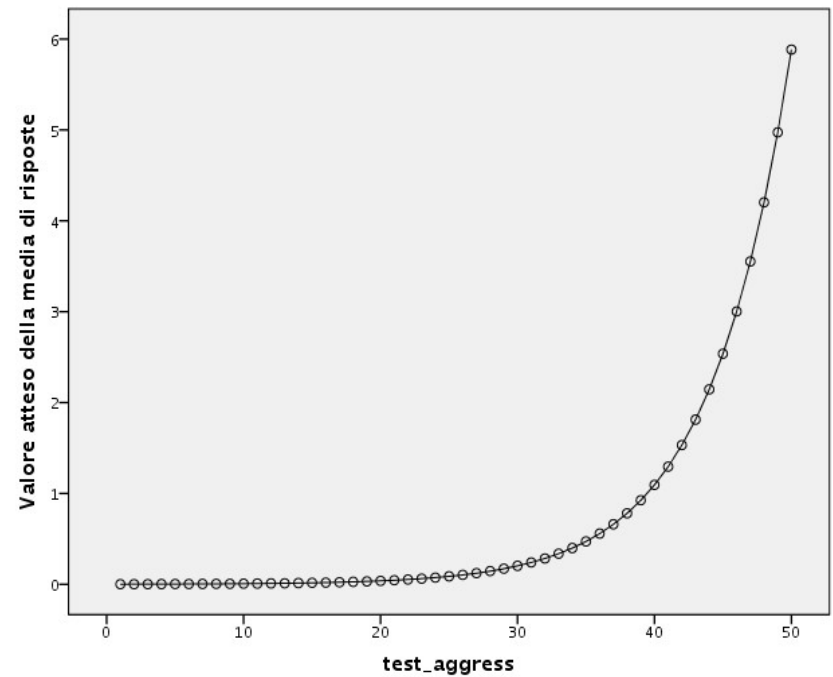
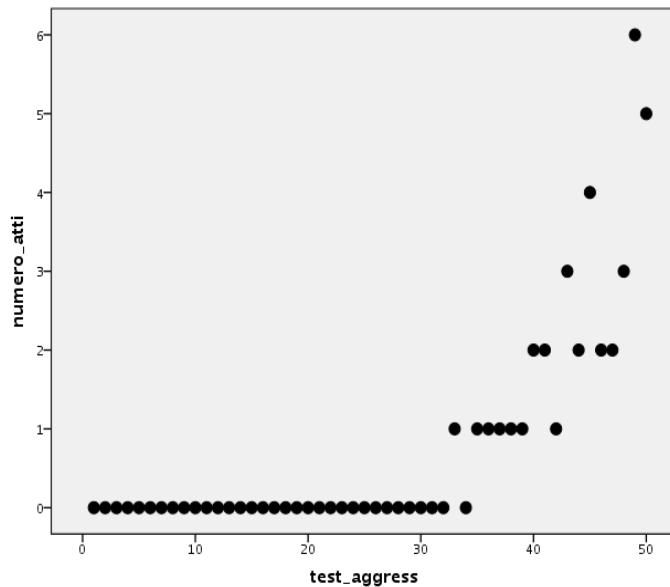
$$y = \text{Poisson}(\lambda)$$



Frequenza media attesa  
dell'evento

# Trasformazione dei “counts”


- Per catturare la relazione non lineare tra la X e la frequenza di Y usiamo la trasformazione logartimica delle frequenze, rende la forma della relazione come segue:




# Regressione di Poisson

- Il modello lineare diventa una regressione di Poisson che stima l'effetto delle VI sul logaritmo delle frequenze dell'evento della VD

La trasformazione è il  
*logaritmo* delle frequenze



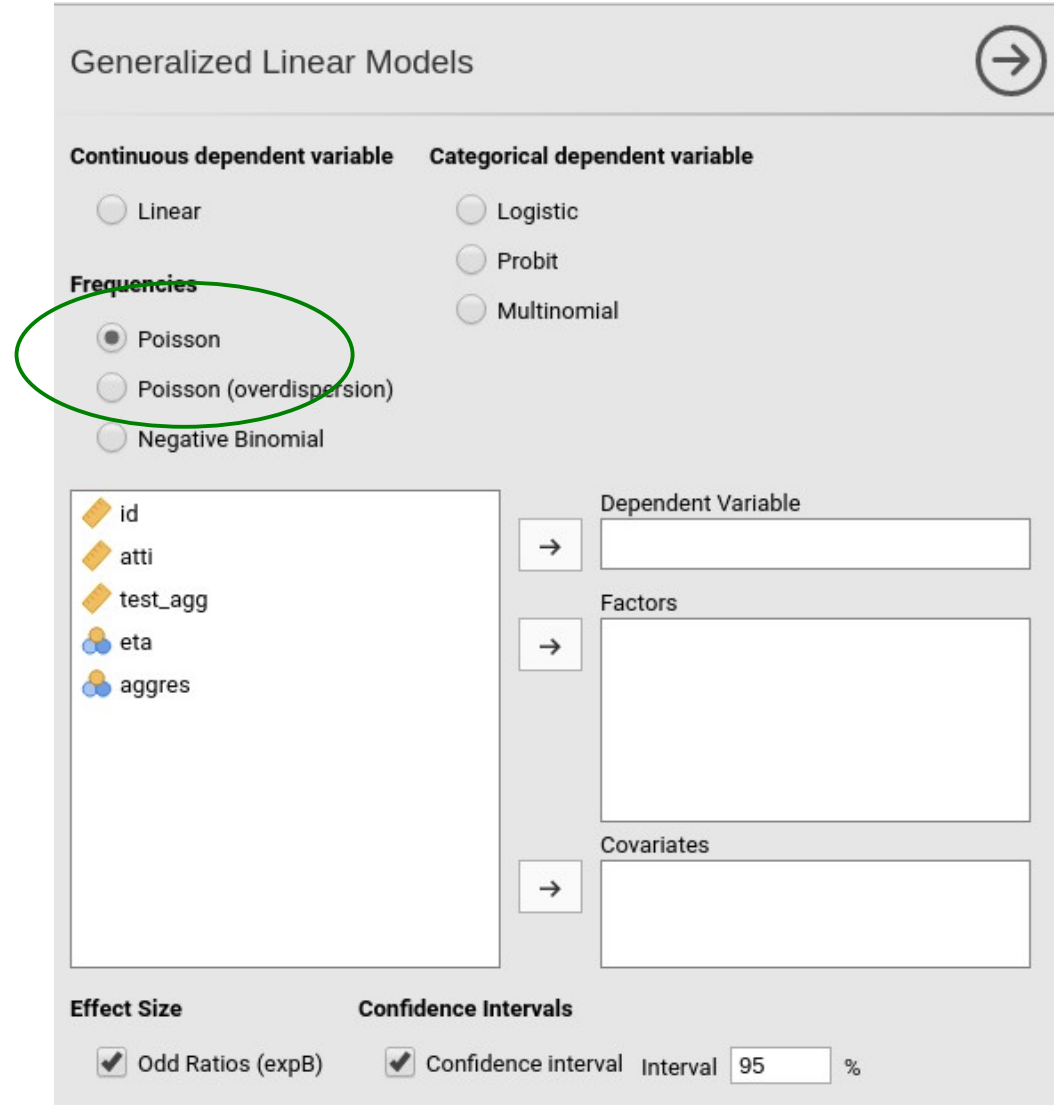
La distribuzione degli  
errori è Poisson



$$\ln(y) = a + b_x x_i + b_w w_i + e_i$$

# Regressione di Poisson

- Per ottenere il modello in SPSS useremo “Modelli Lineari Generalizzati” (come per la logistica) ma selezioneremo il modello opportuno



The image shows the 'Generalized Linear Models' dialog box in SPSS. The 'Continuous dependent variable' section has 'Linear' selected. The 'Categorical dependent variable' section has 'Logistic', 'Probit', and 'Multinomial' options. The 'Frequencies' section has 'Poisson' selected, which is circled in green. Below this, there are three empty boxes for 'Dependent Variable', 'Factors', and 'Covariates'. At the bottom, there are checkboxes for 'Effect Size' and 'Confidence Intervals'. The 'Effect Size' checkbox is checked, and the 'Confidence Intervals' checkbox is also checked, with a value of '95' entered in the 'Interval' field.

Generalized Linear Models

Continuous dependent variable

Linear

Categorical dependent variable

Logistic

Probit

Multinomial

Frequencies

Poisson

Poisson (overdispersion)

Negative Binomial

id

atti

test\_agg

eta

aggres

Dependent Variable

Factors

Covariates

Effect Size


Confidence Intervals

Odd Ratios (expB)

Confidence interval Interval 95 %

# Regressione di Poisson

- Per il resto faremo come per la logistica: settiamo la variabile dipendente e indipendente

Generalized Linear Models 

**Continuous dependent variable**      **Categorical dependent variable**

☐ Linear      ☐ Logistic

☐ Poisson      ☐ Probit

☐ Poisson (overdispersion)      ☐ Multinomial

☐ Negative Binomial

**Frequencies**

☒ Poisson

☐ Poisson (overdispersion)

☐ Negative Binomial

**Dependent Variable**

→

**Factors**

→

**Covariates**

→

**Effect Size**      **Confidence Intervals**

☒ Odd Ratios (expB)      ☒ Confidence interval      Interval  %

# Esempio in SPSS

Risultati: Controllo che il modello sia giusto

Model Info

Info	Value	Comment
Model Type	Poisson	Model for count data
Link function	log	Coefficients are in the log(y) scale
Distribution	Poisson	Rare events distribution of y
R-squared	0.613	Proportion of reduction of error
AIC	58.129	Less is better
Deviance	10.673	Less is better
Residual DF	48	
Value/DF	0.195	Close to 1 is better
Converged	yes	A solution was found

Risultati: Test sull'effetto dei predittori (corrispondente al test F dell'ANOVA/Regression)

Analysis of Deviance: Omnibus Tests

	X <sup>2</sup>	df	p
test_agg	85.9	1	< .001

Risultati:  
Coefficienti

Model Coefficients (Parameter Estimates)

	Contrast	Estimate	SE	95% Confidence Interval		exp(B)	z	p
				Lower	Upper			
(Intercept)	Intercept	-2.349	0.5492	-3.575	-1.408	0.0955	-4.28	< .001
test_agg	test_agg	0.168	0.0275	0.119	0.228	1.1832	6.11	< .001

# Esempio in SPSS

Model Coefficients (Parameter Estimates)

		Estimate	SE	95% Confidence Interval		exp(B)	z	p
	Contrast			Lower	Upper			
(Intercept)	Intercept	-2.349	0.5492	-3.575	-1.408	0.0955	-4.28	< .001
test_agg	test_agg	0.168	0.0275	0.119	0.228	1.1832	6.11	< .001

Coefficienti in scala  
logaritmica

Per ogni unità in più di x il logaritmo  
della frequenza aumenta di .115 unità

# Esempio in SPSS

- Notiamo che la dipendente è  $\ln(Y)$ , dunque  $\exp(B)$  – che toglie il logaritmo al coefficiente – è nell'unità di misura di  $Y$

Model Coefficients (Parameter Estimates)

		Estimate	SE	95% Confidence Interval		$\exp(B)$	z	p
	Contrast			Lower	Upper			
(Intercept)	Intercept	-2.349	0.5492	-3.575	-1.408	0.0955	-4.28	< .001
test_agg	test_agg	0.168	0.0275	0.119	0.228	1.1832	6.11	< .001

Coefficienti  $\exp(B)$

Per ogni unità in più di  $X$ , la frequenza degli atti aumenta di **1.183 volte**



# Interpretazione

- Se al logaritmo della frequenza rimuovo il logaritmo (mediante  $\exp(B)$ ), mi ritrova la scala di frequenza originale

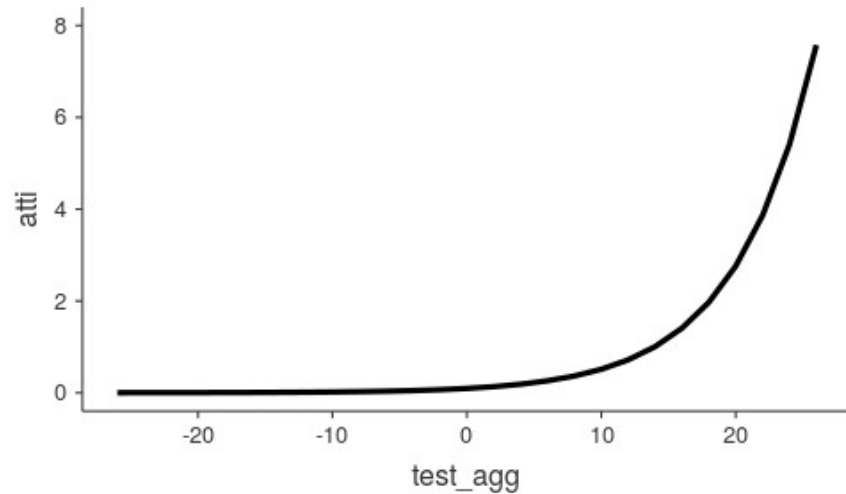
$$\exp(\ln(y)) = y$$

- Rimuovendo il logaritmo, ciò che si sommava ora si moltiplica

# Esempio

- Plot: espresso in frequenza (dipendente)

Effects Plots



Secondo il modello questa è come cambia la frequenza degli atti aggressivi in funzione del punteggio al test

# Esempio

- Considerando i valori osservati e quelli predetti, il fit è molto buono

