METODI E TECNICHE DELLA RICERCA IN PSICOLOGIA CLINICA E LABORATORIO

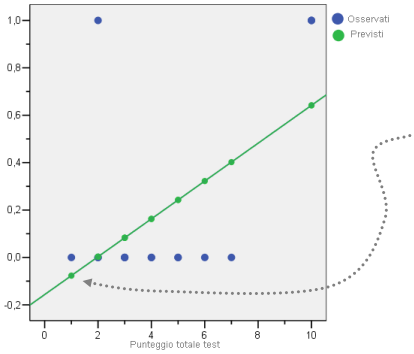
AA 2014/2015

PROF. V.P. SENESE

REGRESSIONE LOGISTICA

SEMPLICE E MULTIPLA

Il modello della regressione semplice può essere applicato anche quando la **variabile dipendente è dicotomica** (ad es. risposta giusta o sbagliata).

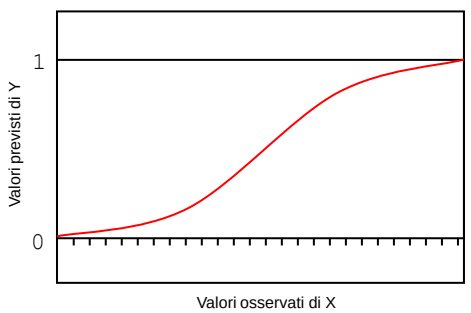


Tuttavia quando la variabile dipendente è dicotomica un **modello non lineare** sarebbe più appropriato.

Infatti, seguendo la linea di tendenza determinata dal modello lineare (**vedi grafico**) diviene evidente che, all'aumentare del punteggio totale, sono accettabili **valori previsti maggiori di** `1` o **minori di** `0`.

In generale, se la **variabile dipendente è dicotomica** e se è influenzata dalla variabile **`x`** allora si dovrebbe osservare che per valori molto alti di **`x`** (o molto bassi se la relazione è negativa) il valore in **`Y`** dovrebbe essere molto **vicino ad `1`** e non dovrebbe aumentare di molto.

In pratica la **curva** che rappresenta la relazione tra **`X`** e **`Y`** **dovrebbe essere** di tipo **logistico**.



La **non linearità** della relazione tra le variabili non consente di poter applicare il metodo **OLS** a meno che non si proceda ad opportune trasformazioni che rendano lineare la relazione. Si tratta di rendere lineare la relazione **nei termini dei parametri** (Berry & Feldman, 1985).

Una delle trasformazioni possibili è, ad esempio, la **trasformazione logaritmica** della variabile dipendente.

Dal momento che **la variabile dipendente** definisce l'appartenenza a un gruppo (o all'altro), i valori che vengono assegnati ai livelli sono arbitrari. Ciò che interessa, dunque, non è il valore atteso (predetto), **ma la probabilità** che un dato soggetto appartenga a meno a uno dei due gruppi.

Un modo per superare questo problema è quello di sostituire la probabilità (ad es. di **Y = 1**) con l'*odds*: ***odds*(Y = 1)**.

L'***odds*** è un modo di esprimere la **probabilità** mediante un **rapporto**. Si calcola facendo il **rapporto tra le frequenze osservate** in un **livello** con le frequenze osservate nell'**altro**.

Il valore dell'***odds*** esprime il **rapporto tra due categorie**.

Ad esempio, se ci sono **30 uomini** e **12 donne** (***n*** = 42) possiamo dire che la probabilità di essere uomini è .714, oppure che gli uomini sono il 71%, mentre l'***odds*** tra uomini e donne è 2.5, oppure che per ogni donna ci sono 2.5 uomini.

$$P(M) = \frac{30}{42} = .714$$

$$odds(M) = \frac{30}{12} = 2.5$$

Per esprimere se **la relazione tra due categorie** varia in funzione di un'altra variabile (valutare cioè l'**associazione** tra due variabili) è possibile utilizzare un altro indice chiamato ***odds ratio*** o rapporto tra gli ***odds***. Tale indice si ottiene facendo un rapporto tra gli ***odds*** ottenuti (rispetto ad una variabile) per ciascun livello della seconda variabile.

Ad esempio, se vogliamo valutare la relazione tra tipo di lavoro e sesso possiamo utilizzare una tabella di contingenza a doppia entrata e rappresentare la distribuzione di frequenze congiunte. Allora la domanda che possiamo porci è: il **rapporto** (***odds***) tra **uomini e donne** è uguale nei differenti **lavori**?

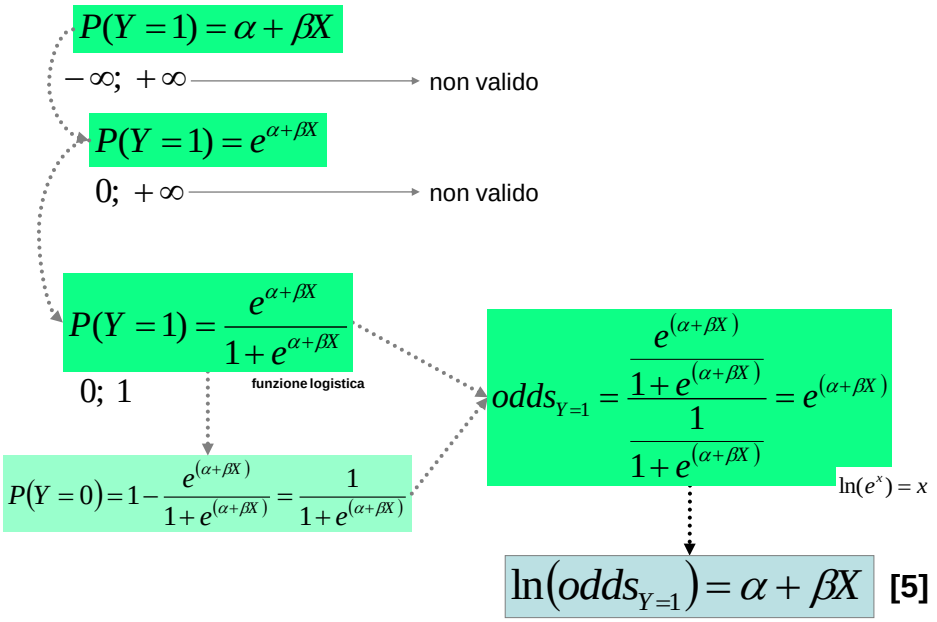
	Uomini	Donne	
Ingegneri	18	2	20
Insegnanti	12	10	22
	30	12	42

$$OR = \frac{18/2}{12/10} = \frac{18}{2} \cdot \frac{10}{12} = 7.5$$

Valori diversi da 1 indicano un'associazione tra le variabili. In questo caso si può dire che la proporzione degli uomini è 7.5 volte maggiore tra gli ingegneri rispetto agli insegnanti.

Relazione tra frequenza, *odds* e logaritmo dell'*odds*

	***Punteggio* (*x*)**	0	1	2	3	4	5	6
successo Y = 1	*f*	**2**	**3**	**5**	**2**	**5**	**6**	**7**
	%	.2	.3	.5	.2	.5	.6	.7
fallimento Y = 0	*f*	**8**	**7**	**5**	**8**	**5**	**4**	**3**
	%	.8	.7	.5	.8	.5	.4	.3
odds (*s*/*f*)		.25	.43	1	.25	1	1.5	2.3
log(odds (*s*/*f*))		-1.39	-.85	0	-1.39	0	.41	.85



Applicando queste trasformazioni, l'equazione della retta diviene:

$$P(Y=1)=\frac{e^{(\alpha+\beta_1+\beta_2+\ldots+\varepsilon)}}{1+e^{(\alpha+\beta_1+\beta_2+\ldots+\varepsilon)}} \qquad \textbf{[6]}$$

È importante sottolineare che la **probabilità**, l'***odds*** e il ***logit*** sono tre differenti modi di esprimere **esattamente le stesse cose**. La trasformazione in ***logit*** serve solo a garantire la correttezza matematica dell'analisi.

Nella stima dei parametri della regressione logistica il metodo **OLS** non può essere applicato (non sono verificati gli assunti), si utilizza l'**algoritmo di massima verosimiglianza** (*maximum likelihood* - ML) che stima i parametri in modo da **massimizzare la funzione** (*log-likelihood function*) che indica quanto è probabile ottenere il valore atteso di **Y** dati i valori delle variabili indipendenti.

La **soluzione ottimale** viene raggiunta partendo da dei **valori di prova** per i **parametri** (**arbitrari**) i quali successivamente vengono **modificati** per vedere se la funzione può essere migliorata. Il processo viene ripetuto (*iteration*) fino a quando la capacità di miglioramento della funzione è infinitesimale (*converge*).

Quando le assunzioni dell'OLS sono verificate, le stime dei parametri ottenute mediante il metodo OLS e il metodo ML sono identiche (Eliason, 1993). In questo senso il metodo OLS può essere considerato un caso particolare della ML; quando i parametri sono stimabili direttamente, senza iterazioni.

Immaginiamo di voler verificare la relazione tra una certa abilità (PLL_T) e la risposta ad un dato item che misura quella stessa abilità (PLLd3):

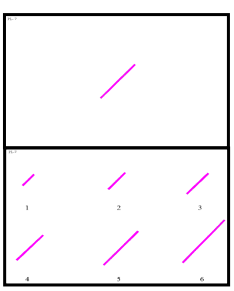
Percezione lunghezza linee

PLL_T

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00	1	3.3	3.3	3.3
	2.00	3	10.0	10.0	13.3
	3.00	3	10.0	10.0	23.3
	4.00	3	10.0	10.0	33.3
	5.00	4	13.3	13.3	46.7
	6.00	2	6.7	6.7	53.3
	7.00	2	6.7	6.7	60.0
	10.00	2	6.7	6.7	66.7
	11.00	2	6.7	6.7	73.3
	12.00	8	26.7	26.7	100.0
	Total	30	100.0	100.0	

PLLd3

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.00	17	56.7	56.7	56.7
	1.00	13	43.3	43.3	100.0
	Total	30	100.0	100.0	



REGRESSIONE SEMPLICE

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	-.316	.108		-2.937	.007
	PLL_T	.106	.013	.832	7.949	.000

a. Dependent Variable: PLLd3

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.832[a]	.693	.682	.28423

a. Predictors: (Constant), PLL_T
b. Dependent Variable: PLLd3

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	5.105	1	5.105	63.186	.000[a]
	Residual	2.262	28	.081		
	Total	7.367	29			

a. Predictors: (Constant), PLL_T
b. Dependent Variable: PLLd3

REGRESSIONE LOGISTICA

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	PLL_T	.786	.250	9.905	1	.002	2.194
	Constant	-6.128	1.974	9.635	1	.002	.002

a. Variable(s) entered on step 1: PLL_T.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	14.584[a]	.586	.786

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

Classification Table[a]

Observed		Predicted PLLd3 .00	Predicted PLLd3 1.00	Percentage Correct
Step 1	PLLd3 .00	17	0	100.0
	1.00	1	12	92.3
	Overall Percentage			96.7

a. The cut value is .500

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	26.470	1	.000
	Block	26.470	1	.000
	Model	26.470	1	.000

REGRESSIONE SEMPLICE

$$PLLd3 = -.316 + .106 \cdot (PLL\_T) + e$$

Es. PLL = 0

$$PLLd3 = -.316 + .106 \cdot (0) = -.316$$?

REGRESSIONE LOGISTICA

$$\log it(PLLd3) = -6.128 + .786 \cdot (PLL\_T) + e$$

Es. PLL = 0

$$\log it(PLLd3) = -6.128 + .786 \cdot (0) = -6.128$$

Nell'interpretazione del modello della regressione logistica ci si avvale di **statistiche del tutto simili** alle statistiche **F** e **R^2** della regressione lineare.

Similmente alla somma dei quadrati, nella regressione logistica si utilizza il *log likelihood* come criterio per la scelta dei parametri del modello. In particolare, per ragioni matematiche, si utilizza il valore del *log likelihood* moltiplicato per –2, e abbreviato come **–2LL**. **Valori grandi e positivi indicano una bassa capacità di previsione del modello**.

Nel **modello con la sola intercetta** il valore della statistica **–2LL** rappresenta quello che nella regressione lineare corrisponde alla **devianza** (o somma dei quadrati totale, SST) e può essere indicata come **D_0**.

Se:

$n_{Y=1}$ numero di casi per i quali Y = 1

$n_{Y=0}$ numero di casi per i quali Y = 0

N numero totale di casi

$P(Y=1) = n_{Y=1} / N$ probabilità che Y = 1

allora:

$$D_0 = -2\{n_{Y=1} \ln[P(Y=1)] + n_{Y=0} \ln[P(Y=0)]\}$$

[7]

Y	*Freq.*
Y = 1	17
Y = 0	13
Tot.	N = 30

$$D_0 = -2\left\{17 \cdot \ln\frac{17}{30} + 13 \cdot \ln\frac{13}{30}\right\} =$$

$$D_0 = 41.054$$

Nel **modello che contiene sia l'intercetta sia la/le variabile/i indipendente/i**, il valore della statistica **–2LL** rappresenta la parte di **variabilità** dei dati che **non viene spiegata dal modello** (devianza d'errore) e viene indicata come D_M. Lo scarto tra D_0 e D_M rappresenta la parte di variabilità spiegata dalla/e variabile/i indipendente/i o **VARIABILITÀ SPIEGATA DAL MODELLO**; e viene indicata come G_M.

$$D_0 - D_M = G_M \qquad \textbf{[8]}$$

G_M viene anche chiamato *Chi-quadrato del modello* e indica la quantità di **riduzione dell'errore** dovuta al modello; ma solo se i modelli sono nidificati (*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.

$$M_A = a + b$$
$$M_B = a + b + c$$

La differenza tra i due **-2LL** (G_M), se calcolata su modelli ***nested***, può essere interpretata come statistica del χ^2 e utilizzata per la **verifica dell'ipotesi nulla del modello**:

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

se il G_M risulta statisticamente significativo (cioè quando il valore ha una $p < .05$) l'ipotesi H_0 può essere rifiutata; vale a dire che la previsione (**Y = 1**) può essere migliorata se consideriamo i predittori.

Per la verifica dell'ipotesi i **gradi di libertà** sono definiti dal **numero di predittori** (*gdl* = ***k***).

Se manteniamo la similitudine tra la statistica **–2LL** e la **devianza** della regressione, per ottenere una statistica simile all'**R^2** si può utilizzare il rapporto di verosimiglianza (*likelihood ratio*):

$$R_L^2 = \frac{G_M}{D_0} = \frac{G_M}{G_M + D_M} \qquad \textbf{[9]}$$

indice di McFadden (1974)

Analogamente a quanto avviene nella regressione, **R^2_L** può essere considerato come la **porzione di riduzione dell'errore** (**–2LL**) **dovuta al modello**. Detto in altri termini, indica **quanto considerare i predittori riduce la variazione nei dati** (stimata a partire dal modello nullo).

In aggiunta alle statistiche relative alla valutazione dell'**adeguatezza del modello** (*goodness of fit*), un ulteriore aspetto che viene preso in considerazione è la **capacità predittiva** del modello.

> Nella maggior parte dei casi siamo interessati a conoscere se il modello è in grado di prevedere adeguatamente $P(Y_i = 1)$. Tuttavia in altri casi possiamo essere interessati a verificare se il modello è in grado di prevedere adeguatamente l'appartenenza dei casi ad un gruppo o ad un altro, quindi siamo molto **più interessati alla tabella delle classificazioni**.

L'**indice** per la valutazione della capacità predittiva del modello maggiormente impiegato si basa sulla valutazione della **riduzione dell'errore in percentuale** (*proportional change in error*):

$$\text{Efficienza predittiva} = \frac{(\text{errori senza il modello}) - (\text{errori con il modello})}{(\text{errori senza il modello})}$$

Al pari della regressione lineare, anche nella regressione logistica siamo interessati a valutare il **contributo specifico** di ogni **variabile indipendente** sulla **variabile dipendente**, testandone la sua **significatività**.

Per la valutazione del contributo di ciascuna variabile si considerano i **coefficienti di regressione**. A tal scopo possiamo considerare sia i **coefficienti non standardizzati** (se siamo interessati alle unità di misura) sia i **coefficienti di regressione standardizzati** (che esprimono la relazione tra le variabili nei termini delle **deviazioni standard**).

Il modo più utilizzato per valutare il contributo di ciascun predittore sulla variabile dipendente è mediante la statistica di **Wald** (W_k):

$$W_k^2 = \left(\frac{b_k}{s_{b_k}} \right)^2$$

Tale statistica segue la distribuzione **Chi-quadro.**

Gdl = 1

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	PLL_T	.786	.250	9.905	1	.002	2.194
	Constant	-6.128	1.974	9.635	1	.002	.002

a. Variable(s) entered on step 1: PLL_T.

$$W_k^2 = \left(\frac{b_k}{s_{b_k}} \right)^2$$

$$W_{PLL}^2 = \left(\frac{.786}{.250} \right)^2 = (3.144)^2 = 9.8847$$

Corrisponde ad un *punto z*

$\chi^2_{(1)} = 9.8847$; $p = .001666$

$$\log it(\text{PLLd3}) = -6.128 + .786 \cdot (\text{PLL\_T}) + e$$

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	PLL_T	.786	.250	9.905	1	.002	2.194
	Constant	-6.128	1.974	9.635	1	.002	.002

a. Variable(s) entered on step 1: PLL_T.

$$\log it(\text{PLLd3}) = -6.128 + .786 \cdot (\text{PLL\_T}) + e$$

$$P(Y=1) = \frac{e^{(\alpha+\beta_1+\beta_2+\ldots+\varepsilon)}}{1+e^{(\alpha+\beta_1+\beta_2+\ldots+\varepsilon)}}$$

Percezione lunghezza linee (0-12)

PLL_T

		Frequency	Percent
Valid	1.00	1	3.3
	2.00	3	10.0
	3.00	3	10.0
	4.00	3	10.0
	5.00	4	13.3
	6.00	2	6.7
	7.00	2	6.7
	10.00	2	6.7
	11.00	2	6.7
	12.00	8	26.7
	Total	30	100.0

minimo 0

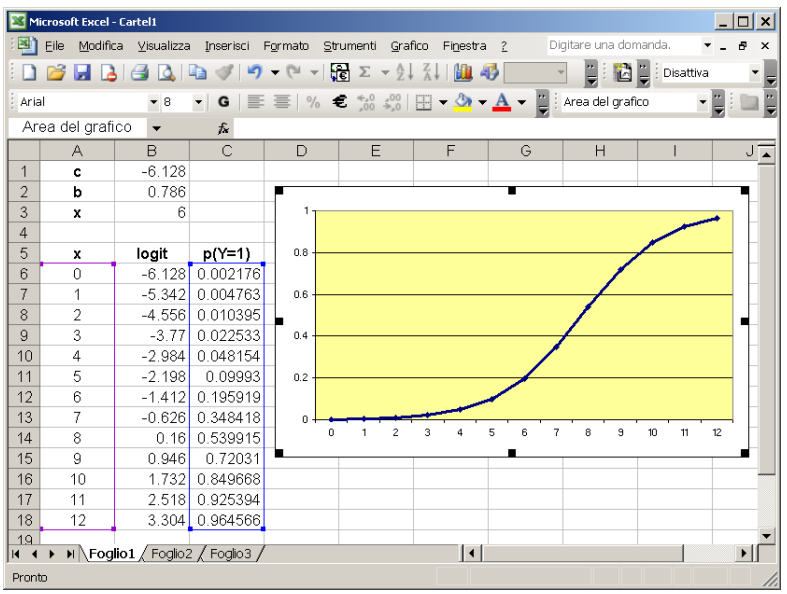
$$\log it(\text{PLLd3}) = -6.128 + .786 \cdot (0) = -6.128$$

$$P(Y=1) = \frac{e^{-6.128}}{1+e^{-6.128}} = .002 = 0.2\%$$

massimo 12

$$\log it(\text{PLLd3}) = -6.128 + .786 \cdot (12) = 3.304$$

$$P(Y=1) = \frac{e^{+3.304}}{1+e^{+3.304}} = .964 = 96.4\%$$



	A	B	C
1	c	-6.128	
2	b	0.786	
3	x	6	
4			
5	x	logit	p(Y=1)
6	0	-6.128	0.002176
7	1	-5.342	0.004763
8	2	-4.556	0.010395
9	3	-3.77	0.022533
10	4	-2.984	0.048154
11	5	-2.198	0.09993
12	6	-1.412	0.195919
13	7	-0.626	0.348418
14	8	0.16	0.539915
15	9	0.946	0.72031
16	10	1.732	0.849668
17	11	2.518	0.925394
18	12	3.304	0.964566

Per porre a confronto variabili che hanno delle unità di misura differenti è necessario calcolare i coefficienti di regressione standardizzati.

Un **coefficiente standardizzato** è un coefficiente che è stato calcolato su variabili che hanno come unità di misura la deviazione standard. Tali coefficienti indicano di quante deviazioni standard varia la varabile dipendente per ogni variazione unitaria (standard) della variabile indipendente.

Nel caso della regressione logistica i coefficienti standardizzati (b^*_{YX}) indicano di quante deviazioni standard si modifica il *logit* della Y_i per ogni variazione standard della variabile X_{ki}. La formula per il calcolo è la seguente:

b_{YX} = coeff. di regressione non stand.
s_X = dev. st. di X
$R_{\hat{Y}Model}$ = coeff. di regressione lineare
$s_{logit(\hat{Y})}$ = dev. st. di $logit(\hat{Y})$ stimato

$$b^*_{YX} = \frac{\left(b_{YX} \cdot s_X\right) \cdot R}{s_{\log it\left(\hat{Y}\right)}}$$

Un ulteriore parametro che può essere utilizzato per l'interpretazione della relazione tra le variabili è l'*odds ratio* che nell'output viene riportato come *exp*(B).

Tale valore esprime la variazione della variabile dipendente in funzione di variazioni della variabile indipendente.

Se il valore è **superiore ad 1** significa che all'aumentare della variabile indipendente **aumenta la probabilità** di **Y = 1**. Al contrario, se il valore è **inferiore ad 1** significa che ad aumentare della variabile indipendente **decresce la probabilità** **che Y = 1**.

È importante sottolineare sia che l'*odds ratio* ha la stessa interpretazione del coefficiente di regressione sia che per confrontare i differenti livelli di probabilità (Y = 1) nei diversi livelli delle variabili indipendenti è necessario calcolare la probabilità e non basta rifarsi ai valori dell'*odds*.

ESERCITAZIONE REGRESSIONE LOGISTICA