Modello logistico
(Modello di regressione logistica)

Prof. Giuseppe Verlato

Sezione di Epidemiologia e Statistica Medica, Dipartimento di Sanità Pubblica e Medicina di Comunità, Università degli Studi di Verona

E per le variabili qualitative NOMINALI ?

2 VARIABILI (entrambe qualitative):
test del chi-quadrato, test esatto di Fischer

3 VARIABILI qualitative (2 var. + 1 var. di stratificazione): test di Mantel-Haenszel

MOLTE VARIABILI:
y dicotomica (malato/sano)→modello LOGISTICO

y politomica (fumatore, ex-fumatore, mai-fumatore)
→ modello MULTINOMIALE

MODELLO DI REGRESSIONE LOGISTICA

19 / (19+132)
0 / (0+9)
11 / (11+52)
6 / (6+97)

prevalenze

MODELLO LOG-LINEARE

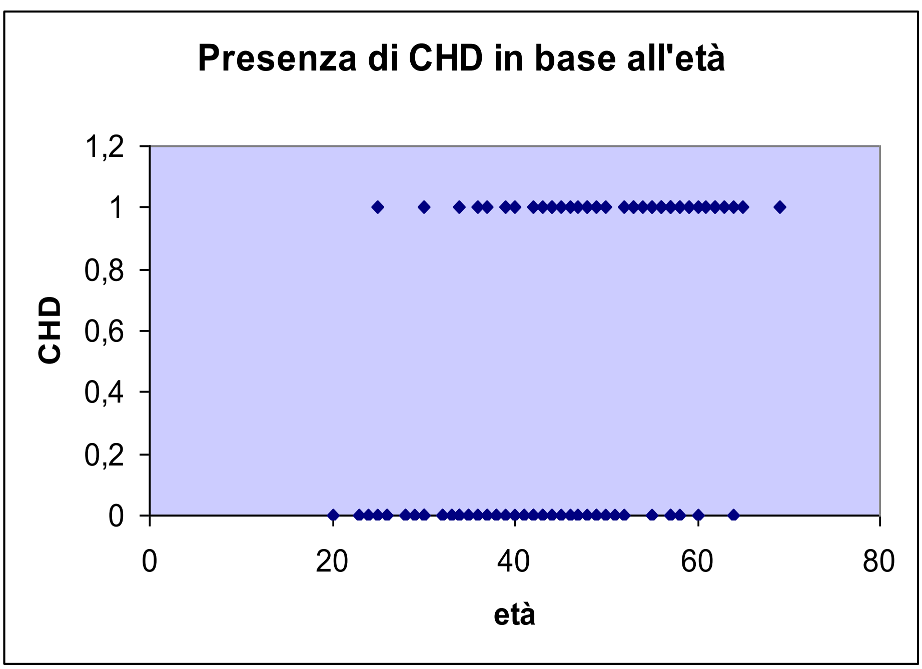
19	132
0	9
11	52
6	97

conteggi

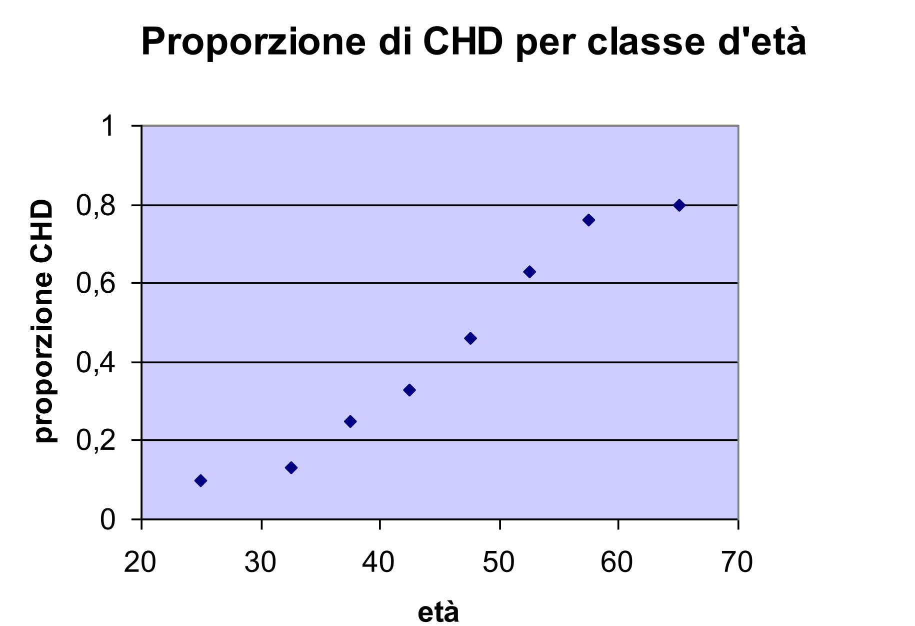
MODELLO DI POISSON

19 / 1510 persone-anno
0 / 90 persone-anno
11 / 630 persone-anno
6 / 103 persone-anno

inci-
denze



		CHD		
classe d'età	*N*	*assente*	*presente*	*proporzione*
20-29	10	9	1	0,1
30-34	15	13	2	0,13
35-39	12	9	3	0,25
40-44	15	10	5	0,33
45-49	13	7	6	0,46
50-54	8	3	5	0,63
55-59	17	4	13	0,76
60-69	10	2	8	0,8
totale	**100**	**57**	**43**	**0,43**



Curva con andamento sigmoide

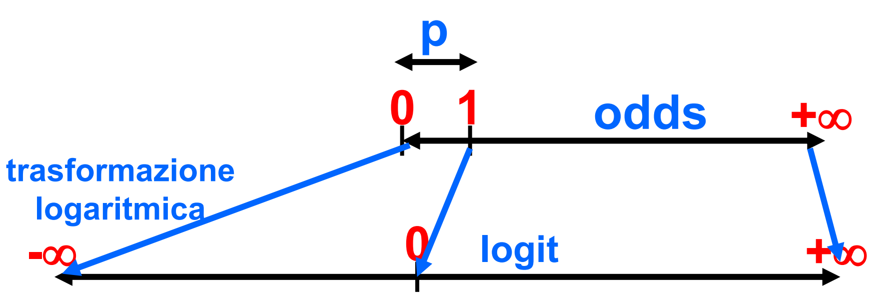
Nella regressione lineare multipla la Y varia tra -∞ e +∞

Nella regressione logistica

p(malattia) varia tra 0 e 1

odds(malattia) = p/(1-p) varia tra 0 e +∞

Logit = ln [p/(1-p)] varia tra -∞ e +∞



I MODELLI LINEARI GENERALIZZATI si differenziano per la distribuzione dell'errore (error function) e per la funzione legame (link function)

REGRESSIONE LINEARE MULTIPLA

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{13} x_1 x_3 + \varepsilon$$

La funzione legame (link-function) è l'IDENTITA'

L'errore segue la distribuzione NORMALE

MODELLO DI REGRESSIONE LOGISTICA

$$Ln\,[y/(1-y)] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{13} x_1 x_3 + \varepsilon$$

La funzione legame (link-function) è il LOGIT [LOG(ODDS)]

L'errore segue la distribuzione BERNOULLIANA

MODELLO LOG-LINEARE

$$Ln(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{13} x_1 x_3 + \varepsilon$$

La funzione legame (link-function) è il LOGARITMO

L'errore segue la distribuzione di POISSON

MODELLO DI REGRESSIONE LOGISTICA

Predittore lineare

$$\ln[\pi/(1-\pi)] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{13} x_1 x_3$$

Logit

Var.qualitative e/o quantitative

Termine d'interazione

$$\pi/(1-\pi) = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{13} x_1 x_3)$$

odds

$$\pi = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{13} x_1 x_3)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{13} x_1 x_3)}$$

prevalenza

Metodi di ottimizzazione

per trovare il modello che meglio si adatta ai dati

Se la funzione-legame è l'identità:
Regressione lineare semplice
Regressione lineare multipla
Analisi della varianza
Analisi della covarianza

Omoschedasticità
$\varepsilon \sim N(0;\sigma^2)$

Metodo dei minimi quadrati (least-square method)

Regressione logistica

Non-Omoschedasticità
$\varepsilon \sim B(0;\pi(1-\pi))$

Metodo della massima verosimiglianza (maximum likelihood)
Metodo iterativo

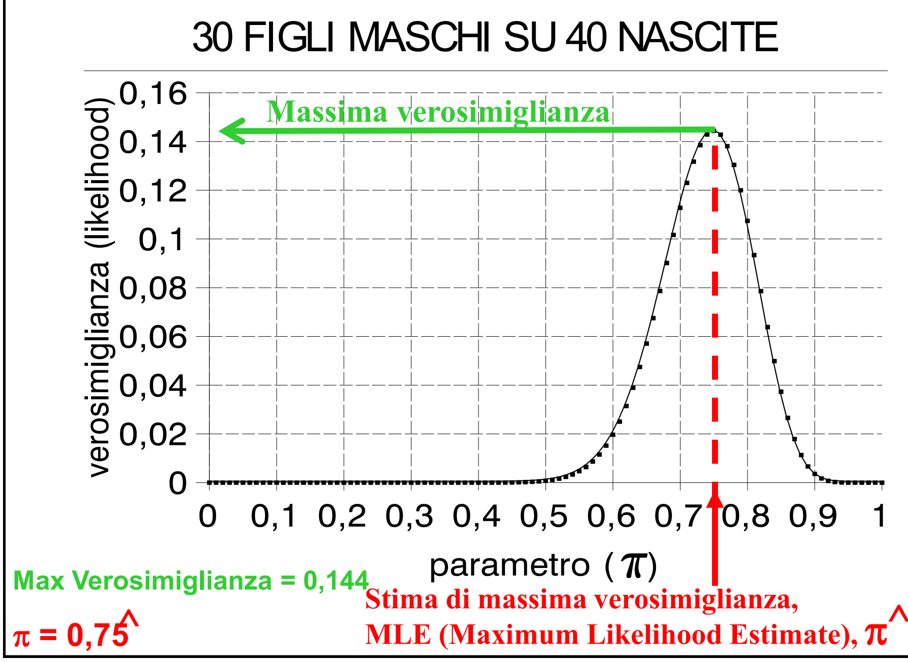


Table 4. Risk factors for e-cigarette ever use.

Risk factors	OR of e-cigarette ever use (95%IC)	p-value
Smoking Habits		
Never Smokers	1	
Ex-Smokers	**5.97 (3.46 – 10.33)**	**<0.001**
Occasional Smokers	**7.75 (5.07 – 11.85)**	**<0.001**
Regular Smoker	**24.6 (16.62 – 36.50)**	**<0.001**
Sex		
Male	1	
Female	**0.45 (0.32 – 0.63)**	**<0.001**
Family history of smoking		
None	1	
Relative who smoked e-cig	0.68 (0.37 – 1.25)	0.219
Relative who smoked tobacco	1.19 (0.92 – 1.53)	0.178
Housemates currently smoking		
None	1	
Housemates smoking e-cig	**2.54 (1.15 – 5.59)**	**0.021**
Housemates smoking tobacco	1.16 (0.94 – 1.43)	0.163
Centre		
Verona	1	
Vicenza	**1.12 (1.01 - 1.24)**	**0.027**
Legnago	**0.89 (0.85 – 0.94)**	**<0.001**
Trento	**0.82 (0.73 – 0.91)**	**<0.001**
Bolzano	**0.66 (0.63 – 0.69)**	**<0.001**
University class		
1st year	1	
2st year	1.01 (0.73 – 1.41)	0.939
3st year	0.79 (0.56 – 1.11)	0.173

OR (Odds Ratio), 95% IC, *P*-values were computed by logistic regression model. Significant results were highlighted in bold.

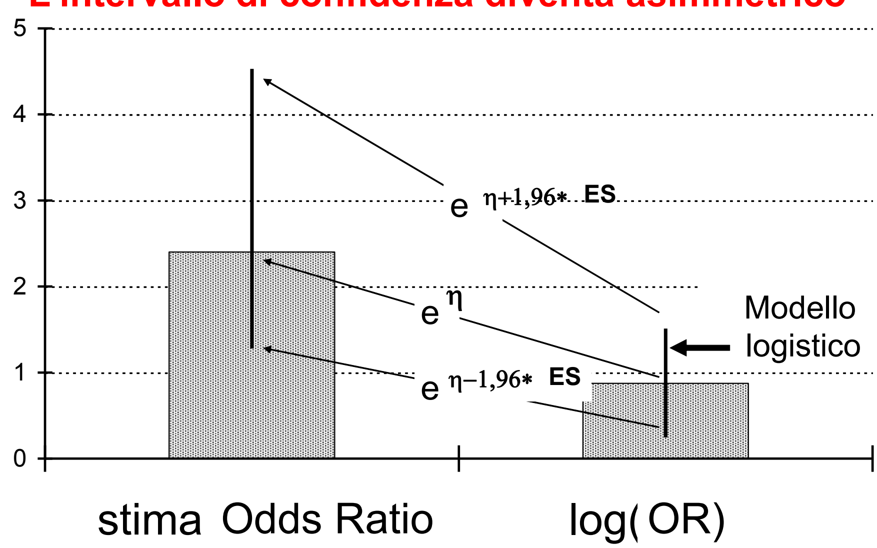
Canzan et al, BMC Public Health 2019

Logistic regression   Number of obs = 1,354
LR chi2(14) = 476.34
Prob > chi2 = 0.0000
Log likelihood = -597.07448   Pseudo R2 = 0.2852

provatoecg	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_IsmokHab_1	1.787651	.2388358	7.48	0.000	1.319541	2.25576
_IsmokHab_2	2.047776	.2235388	9.16	0.000	1.609648	2.485904
_IsmokHab_3	3.203935	.1918879	16.70	0.000	2.827841	3.580028
_Isesso_2	-.8049849	.1684709	-4.78	0.000	-1.135182	-.4747881
_Ifamiliari_1	-.3819719	.2876365	-1.33	0.184	-.9457291	.1817854
_Ifamiliari_2	.1720737	.1825327	0.94	0.346	-.1856837	.5298312
_Iconvivent_1	.9307554	.3162551	2.94	0.003	.3109068	1.550604
_Iconvivent_2	.1494138	.1920308	0.78	0.437	-.2269596	.5257873
_Isede_2	.1159116	.2095446	0.55	0.580	-.2947882	.5266115
_Isede_3	-.1109284	.2274942	-0.49	0.626	-.5568088	.334952
_Isede_4	-.2020076	.2462749	-0.82	0.412	-.6846976	.2806824
_Isede_5	-.4146958	.2101358	-1.97	0.048	-.8265544	-.0028371
_Iannodicor_2	.0129026	.1907115	0.07	0.946	-.360885	.3866903
_Iannodicor_3	-.2393992	.1827322	-1.31	0.190	-.5975477	.1187494
_cons	-1.872191	.2466523	-7.59	0.000	-2.355621	-1.388762

CONSEGUENZE della TRASFORMAZIONE LOGARITMICA:

L’intervallo di confidenza diventa asimmetrico



Quando abbiamo una sola variabile indipendente, possiamo verificare che il parametro β stimato dalla regressione logistica corrisponde al ln(OR) calcolato dalla tabellina corrispondente.

Es. Età e CHD: dividiamo l'età in 2 categorie <55(x=0) e $\geq$55(x=1)

	Coefficiente Stimato	Errore Standard	Coeff./ES	OR
AGE	2.094	0.529	3.96	8.1
Constant	-0.841	0.255	-3.30	

Costruendo ora la tabellina 2X2:

	AGE(x)		
CHD (y)	$\geq$55(1)	<55	
Presente (1)	21	22	43
Assente (0)	6	51	57
Totale	27	73	100

OR=21*51/(6*22)=8.11

che quindi corrisponde a quanto trovato con la regressione logistica