Modello logistico
(Modello di regressione logistica)

Prof. Giuseppe Verlato

Prof. Elisabetta Zanolin

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

PRINCIPALI MODELLI LINEARI UTILIZZATI IN MEDICINA

regressione lineare semplice — X ed Y sono var. quantitative

$$y = \beta_0 + \beta_1 x + \epsilon$$

regressione lineare multipla — X ed Y sono var. quantitative

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k + \epsilon$$

analisi della varianza (ANOVA) — Y quantitativa, X qualitative

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k + \epsilon$$

analisi della covarianza — Y quantitativa, X qualit. e quantit.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k + \epsilon$$

modello logistico — Y qualit., X quantit. e qualit.

$$\ln\left(\frac{p}{1-p}\right)\nu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k + \epsilon$$

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

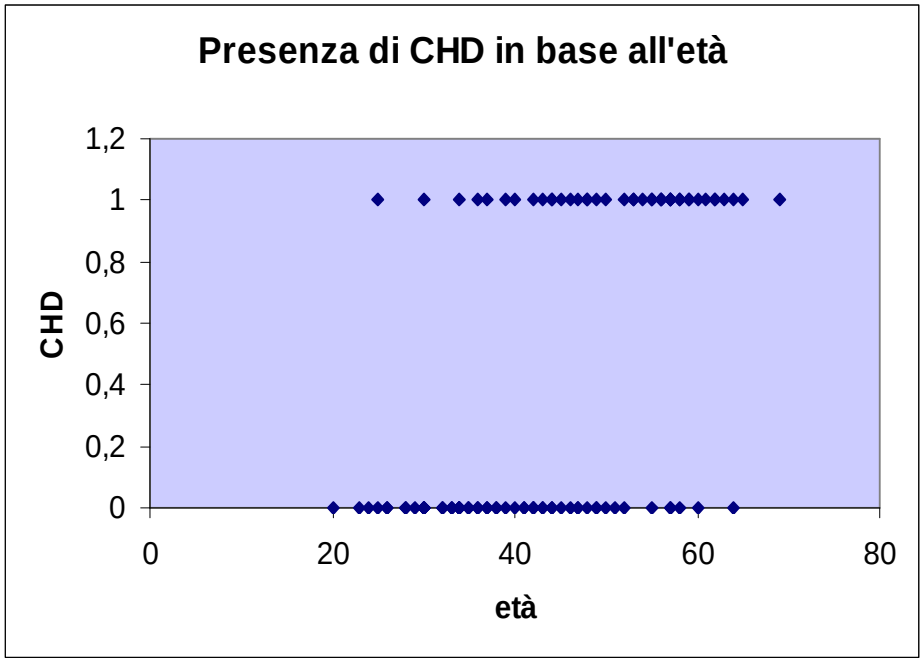
19 / 1510 persone-anno
0 / 90 persone-anno
11 / 630 persone-anno
6 / 103 persone-anno

inci-denze

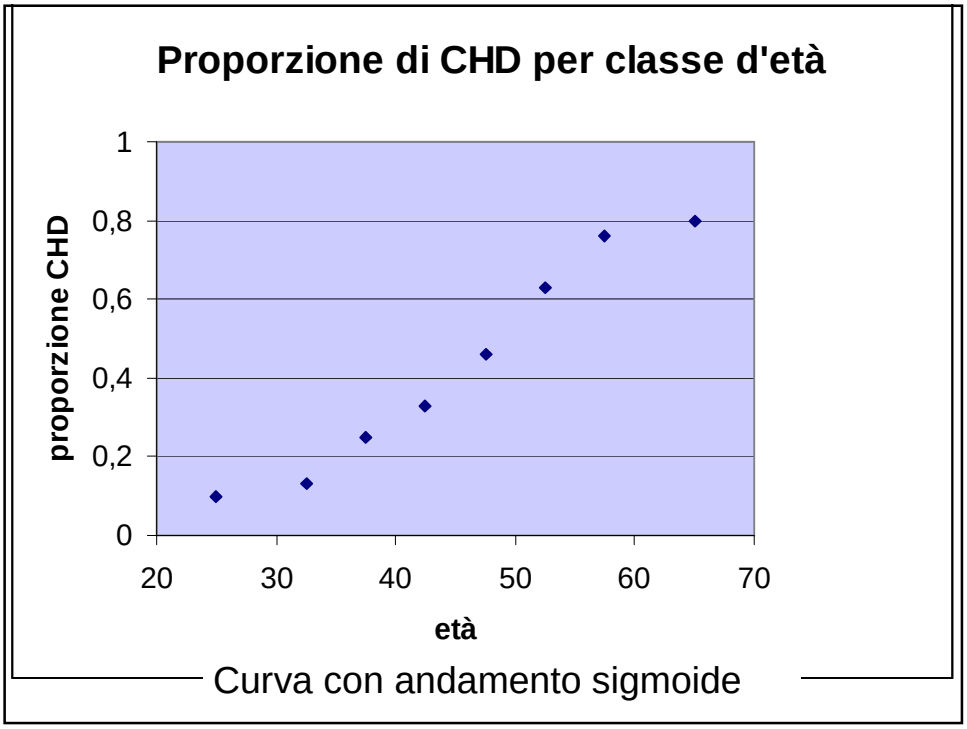
Dati relativi a 100 soggetti sulla presenza delle malattie ischemiche (CHD)

ID	AGRP	AGE	CHD	ID	AGRP	AGE	CHD	ID	AGRP	AGE	CHD	ID	AGRP	AGE	CHD
1	1	20	0	26	3	35	0	51	4	44	1	76	7	55	1
2	1	23	0	27	3	35	0	52	4	44	1	77	7	56	1
3	1	24	0	28	3	36	0	53	5	45	0	78	7	56	1
4	1	25	0	29	3	36	1	54	5	45	1	79	7	56	1
5	1	25	1	30	3	36	0	55	5	46	0	80	7	57	0
6	1	26	0	31	3	37	0	56	5	46	1	81	7	57	0
7	1	26	0	32	3	37	1	57	5	47	0	82	7	57	1
8	1	28	0	33	3	37	0	58	5	47	1	83	7	57	1
9	1	28	0	34	3	38	0	59	5	47	0	84	7	57	1
10	1	23	0	35	3	38	0	60	5	48	1	85	7	57	1
11	2	30	0	36	3	39	0	61	5	48	1	86	7	58	0
12	2	30	0	37	3	39	1	62	5	48	1	87	7	58	1
13	2	30	0	38	4	40	0	63	5	49	0	88	7	58	1
14	2	30	0	39	4	40	1	64	5	49	0	89	7	59	1
15	2	30	0	40	4	41	0	65	5	49	1	90	7	59	1
16	2	30	1	41	4	41	0	66	6	49	0	91	8	60	0
17	2	32	0	42	4	42	0	67	6	50	1	92	8	60	1
18	2	32	0	43	4	42	0	68	6	50	0	93	8	61	1
19	2	33	0	44	4	42	0	69	6	51	0	94	8	62	1
20	2	33	0	45	4	42	1	70	6	52	1	95	8	62	1
21	2	34	0	46	4	43	0	71	6	52	1	96	8	63	1
22	2	34	0	47	4	43	0	72	6	53	1	97	8	64	0
23	2	34	1	48	4	43	1	73	6	53	1	98	8	64	1
24	2	34	0	49	4	44	0	74	7	54	0	99	8	65	1
25	2	34	0	50	4	44	0	75	7	55	1	100	8	69	1

CHD=0 malattia assente CHD=1 malattia presente



CHD				
classe d'età	*N*	*assente*	*presente*	*media(proporz.)*
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

Riassumendo…

- La media condizionale (E(Y|x)) deve essere compresa tra 0 e 1. Si utilizza quindi il modello di regressione logistica $\pi(x)$ che soddisfa questo requisito.
- La distribuzione bernoulliana descrive la distribuzione degli errori e quindi sarà la distribuzione su cui l'analisi statistica è incentrata.
- Nella regressione logistica, si utilizzeranno gli stessi principi seguiti nella regressione lineare

Odds

Il cavallo Varenne ha 20 probabilità su 100 di vincere una gara.

Il cavallo Varenne → 20 probabilità su 100 di vincere

Il cavallo Varenne → 80 probabilità su 100 di perdere

Odds di vittoria = 20 / 80 = 1 / 4 = 0,25

Pertanto Varenne viene dato 4 a 1 (1 a 4)

Chi scommette 1000 € su Varenne in ognuna di 100 gare,
vince 20 volte 4000 €, in tutto 80000 €,
perde 80 volte 1000 €, in tutto 80000 €,
per cui le perdite pareggiano le vincite.

ODDS RATIO (OR) – 1

(rapporto crociato, stima indiretta del Rischio Relativo)

ESEMPIO:

Un fumatore ha 40 probabilità su 100 di essere iperteso a 60 anni.

Un non-fumatore ha 20 probabilità su 100 di essere iperteso a 60 anni.

1) Probabilità (p) → p (ipertensione / fumatore) = 40 / 100 = 0,4 = 40%

→ p(ipertensione / non-fumatore) = 20/100 = 0,2 = 20%

2) Odds (ω) = $\frac{p}{1-p}$ → odds di ipertensione nei fumatori = 40 / 60 = 0,67 = 67%

→ odds di ipertensione nei non-fumatori =20/80 =0,25 =25%

3) Odds Ratio (O.R.) = $\frac{\frac{p_1}{1-p_1}}{\frac{p_0}{1-p_0}}$

odds ratio di ipertensione nei fumatori rispetto ai non-fumatori = 0,67 / 0,25 = 2,67

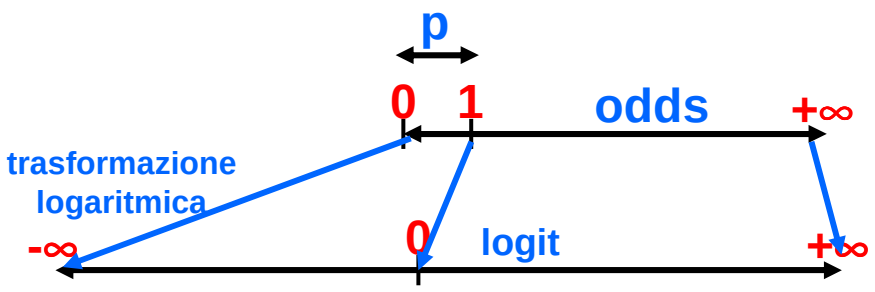
Nella regressione lineare multipla la Y varia tra -∞ e +∞

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 +∞



MODELLO DI REGRESSIONE LOGISTICA

$$\underbrace{Ln\left[\frac{\pi}{(1-\pi)}\right]}_{\text{Logit}} = \overbrace{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{13} x_1 x_3}^{\text{Predittore lineare}}$$

Var.qualitative e/o quantitative

Termine d'interazione

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

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

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

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

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

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

Notazione utilizzata

Paia di osservazioni:

(x_i, y_i) $i=1, 2, 3, 4 \ldots. N$

y_i=outcome *0*=assenza *1*=presenza

x_i=valore var. indipendente per il soggetto *iesimo*

Es. (età,CHD)

Soggetto 1 *: (20,0)*

Soggetto 23*: (34,1)*

•Il metodo della ***massima verosimiglianza*** ci fornisce i valori dei parametri ignoti che massimizzano la probabilità di ottenere i dati osservati.

•Per applicarlo dobbiamo costruire la **funzione di verosimiglianza,** che ci dà la probabilità di avere i dati osservati in funzione dei parametri ignoti.

•Scegliamo i parametri (**stime di massima verosimiglianza**) che massimizzano la funzione di verosimiglianza.

Come stimare i parametri ignoti nella regressione logistica tramite la funzione di verosimiglianza?

$\pi(x)$ ci dà la probabilità condizionale che y=1 per un dato valore di x

es. $\pi(x)$ ci dà la probabilità che un soggetto abbia CHD (y=1) all'età di 49 anni (x)

$1-\pi(x)$ ci dà la probabilità condizionale che y=0 per un dato valore di x

es. $1-\pi(x)$ ci dà la probabilità che un soggetto non abbia CHD (y=0) all'età di 49 anni (x)

per (x_i, y_i) dove $y_i=1$ il contributo alla funzione di verosimiglianza è $\pi(x_i)$

per (x_i, y_i) dove $y_i=0$ il contributo alla funzione di verosimiglianza è $1-\pi(x_i)$

Il contributo di una coppia alla funzione di verosimiglianza è:

$$\zeta(x_i) = \pi(x_i)^{y_i}\left[1-\pi(x_i)\right]^{1-y_i}$$

Assumendo che le osservazioni siano indipendenti, la funzione di verosimiglianza viene ottenuta come prodotto dei termini $\zeta(x_i)$:

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n}\zeta(x_i)$$

Per il principio della massima verosimiglianza, utilizziamo le stime di β_0 e β_1 che massimizzano $l(\boldsymbol{\beta})$, ma matematicamente il logaritmo di $l(\boldsymbol{\beta})$ è più facile da trattare: **log-likelihood**

$$L(\boldsymbol{\beta}) = \ln\left[l(\boldsymbol{\beta})\right] = \sum_{i=1}^{n}\left\{y_i \ln\left[\pi(x_i)\right] + (1-y_i)\ln\left[1-\pi(x_i)\right]\right\}$$

Per trovare i valori di β_0 e β_1 che massimizzano L($\boldsymbol{\beta}$), deriviamo L($\boldsymbol{\beta}$) rispetto a β_0 e β_1 e poniamo le espressioni risultanti=0. Per β_0 si ha:

$$\sum_{i=1}^{n}\left[y_i - \pi(x_i)\right] = 0$$

Per β_1 si ha:

$$\sum_{i=1}^{n} x_i\left[y_i - \pi(x_i)\right] = 0$$

Queste equazioni, non essendo lineari nei parametri, sono risolvibili tramite processi iterativi disponibili nei software statistici.

Si ottengono così le stime dei parametri $\hat{\beta}_0$ e $\hat{\beta}_1$, dove ^ indica la stima.

Esempio con i dati su **età e CHD**. Utilizzando un software statistico, otteniamo i risultati in tabella:

variabile	Coefficiente stimato	Errore standard	Coeff./ES
AGE	0.111	0.024	4.61
Constant	-5.310	1.134	-4.68

Log-likelihood=-53.677

Quindi …

… i valori predetti vengono dati dall'equazione:

$$\hat{\pi}(x) = \frac{e^{-5.31+0.111*age}}{1+e^{-5.31+0.111*age}}$$

e il logit stimato è:

$$\hat{g}(x) = -5.31+0.111*age$$

Log-likelihood=-53.6777 è ottenuto dall'equazione (che è stata massimizzata per ottenere le stima dei parametri) :

$$L(\boldsymbol{\beta}) = \ln[l(\boldsymbol{\beta})] = \sum_{i=1}^{n}\{y_i \ln[\pi(x_i)] + (1-y_i)\ln[1-\pi(x_i)]\}$$

L’interpretazione dei coefficienti (β) del modello di regressione logistica

Nella regressione lineare, i β ci dicono di quanto varia y al variare di x di un’unità.

$\beta_1 = y(x+1) – y(x)$

Analogamente anche per la regressione logistica:

$\beta_1 = g(x+1) – g(x)$

Il problema è dare un significato alla differenza tra questi 2 logit

Per scoprire il significato di questa differenza tra i due logit, consideriamo il caso in cui abbiamo una variabile indipendente x dicotomica, codificata come x=0 (non esposto) e x=1 (esposto)

$\beta_1 = g(x+1) - g(x) = g(1) - g(0) =$

$$= \ln\left[\frac{\pi(1)}{1-\pi(1)}\right] - \ln\left[\frac{\pi(0)}{1-\pi(0)}\right] =$$

$$= \ln\left\{\frac{\frac{\pi(1)}{1-\pi(1)}}{\frac{\pi(0)}{1-\pi(0)}}\right\} = \ln(OR)$$

$\beta_1 = \ln(OR)$

$e^{\beta 1} = OR$

Odds tra i non esposti

OR

Odds tra gli esposti

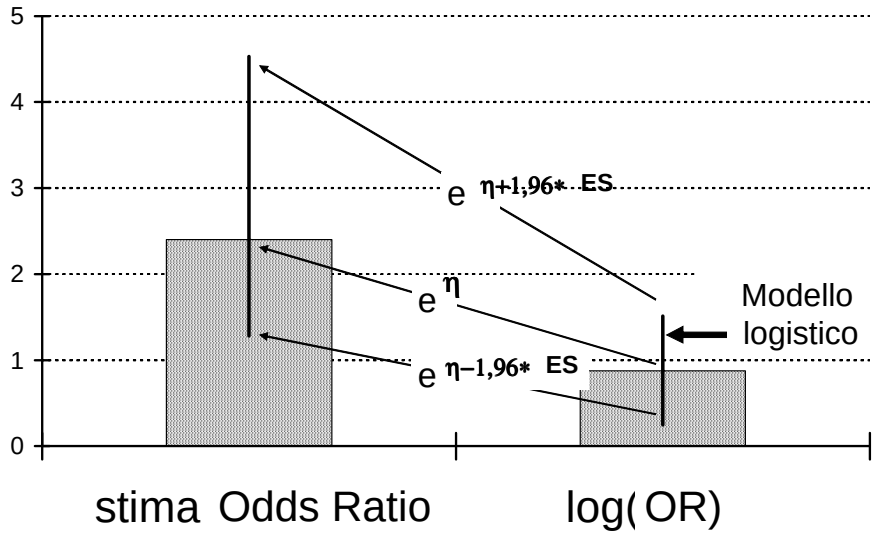
Quando abbiamo una sola variabile indipendente, possiamo verificare che il parametro β stimato dalla regressione logistica corrisponde al ln(OR) calcolato dalla tabellina corrispondente.

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

CONSEGUENZE della TRASFORMAZIONE LOGARITMICA:

L'intervallo di confidenza diventa asimmetrico



Costruendo ora la tabellina 2X2:

	AGE(x)		
CHD (y)	≥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

L'interpretazione dei coefficiente nel caso di una variabile indipendente classificata in più di 2 categorie è analoga.

Es. CHD e razza (bianca=1; nera=2; ispanica=3; altra=4)

	Coefficiente Stimato	Errore Standard	Coeff./ES	OR
RAZZA(2)	2.079	0.633	3.29	8.0
RAZZA(3)	1.792	0.646	2.78	6.0
RAZZA(4)	1.386	0.671	2.07	4.0
Costante	-1.386	0.500	-2.77	

In questo caso, l'OR è calcolato per tutte le razze rispetto alla razza bianca

Livello della variabile RAZZA	Variabili dummy (fittizie)		
	NERA	ISPANICA	ALTRO
RAZZA(1) Bianca	0	0	0
RAZZA(2) Nera	**1**	0	0
RAZZA(3) Ispanica	0	**1**	0
RAZZA(4) Altra	0	0	**1**

Anche quando x è continua l'interpretazione è analoga, ma β_1 dà il cambiamento del log-odds all'aumentare di 1 della variabile indipendente.

A volte può essere molto utile calcolare il cambiamento invece che per ogni unità di x, per ogni 10 UNITA' (ad es. invece di considerare solo 1 anno d'età, considerare 10 anni d'età) oppure per un incremento di UNA DEVIAZIONE STANDARD.

Una volta stimati i parametri, ci chiediamo: sono significativi?

Il modello che include la variabile in studio ci dà informazioni in più sull'outcome rispetto al modello che non la include?

Es. Considerare l'età come possibile fattore di rischio per l'insorgenza di CHD ha senso?

Test di significatività per i parametri

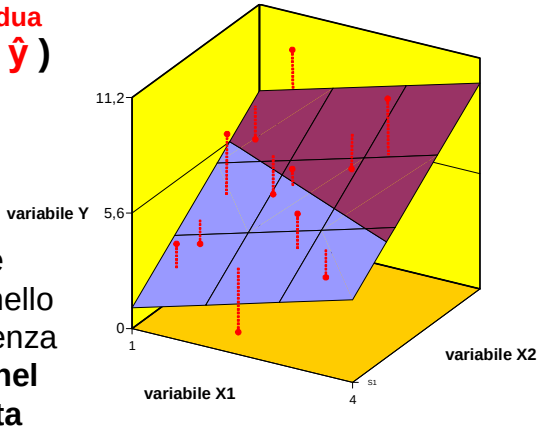
Regressione lineare multipla

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Variabilità totale — Variabilità residua

$$(y - \bar{y}) = (\hat{y} - \bar{y}) + (y - \hat{y})$$

Variabilità spiegata dalla regressione



La scomposizione delle devianza viene effettuata nello stesso modo: l'unica differenza è che **y atteso ($\hat{y}$) giace nel piano e non su una retta**

SCOMPOSIZIONE DELLA DEVIANZA nella Regressione lineare multipla - 2

Variabilità totale — Variabilità residua

$$(y - \bar{y}) = (\hat{y} - \bar{y}) + (y - \hat{y})$$

Variabilità spiegata dalla regressione

Si può dimostrare che:

Devianza totale, SST — Devianza residua, SSE

$$\Sigma(y - \bar{y})^2 = \Sigma(\hat{y} - \bar{y})^2 + \Sigma(y - \hat{y})^2$$

Devianza spiegata dalla regressione, SSR

Anche nella **regressione logistica**, si confrontano i valori **y osservati** con quelli **previsti** dal modello con e senza la variabile di regressione.

Il paragone viene effettuato tramite la log-likelihood.

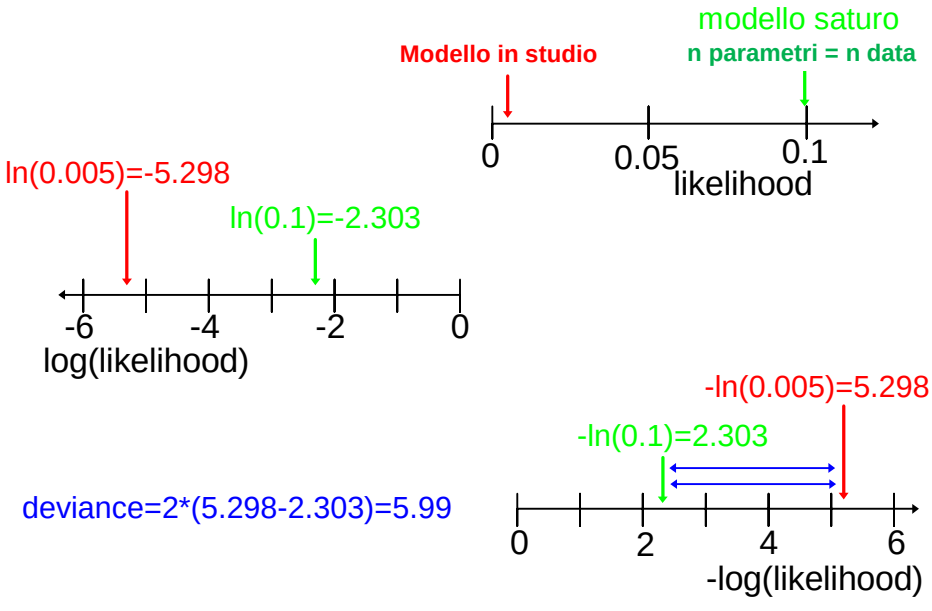
Pensiamo agli osservati come dati previsti da un **modello saturo** (= modello che contiene tanti parametri quanti i dati).

$$D = -2\ln\left[\frac{(verosimiglianza.modello.corrente)}{(verosimiglianza.modello.saturo)}\right] =$$

$$= 2\left[\ln(veros.mod.saturo) - \ln(veros.mod.corrente)\right]$$

D viene chiamata ***devianza*** e ha lo stesso ruolo di $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ la devianza residua nella regressione lineare

CONCETTO DI DEVIANCE



$$D = -2\ln\left[\frac{(verosimiglianza.modello.corrente)}{(verosimiglianza.modello.saturo)}\right]$$

Rapporto di verosimiglianza

Likelihood ratio test

D è distribuita come χ^2 con g.l.=n° osservazioni-n° parametri

Per testare la significatività di una variabile indipendente (x):

G = *D*(per il modello **senza** la variabile) – *D*(per il modello **con** la variabile)

Sostituendo *D* diventa:

$$G = -2\ln\left[\frac{verosimiglianza.senza.variabile}{verosimiglianza.con.variable}\right] =$$

$$= 2[\ln(verosim.con.\text{var}) - \ln(verosim.senza.\text{var})]$$

Sotto l'ipotesi che β_1=0, la statistica G segue χ^2 con 1 g.l.

(quando x è continua o dicotomica)

Altro test per testare la significatività delle variabili:

Wald test

$$W = \frac{\hat{\beta}_1}{\hat{ES}(\hat{\beta}_1)}$$

Sotto l'ipotesi nulla $\beta_1=0$, W segue una distribuzione normale

esempio

variabile	Coefficiente stimato	Errore standard	Coeff./ES
AGE	0.111	0.024	4.61
Constant	-5.310	1.134	-4.68

Log-likelihood=-53.677

Log-likelihood ratio test:

Log-likelihood senza la variabile AGE (modello nullo solo con la costante) = -68.322

G=2(-53.677 -(-68.322))=29.31 χ^2con 1 g.l. P<0.001

Wald test: z=4.61 P<0.001 AGE è significativa

Per il passaggio alla regressione logistica multipla, aggiungiamo le variabili di interesse al modello in studio:

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

$$\pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}}$$

Test di significatività per i parametri – regressione logistica multipla

Es. Studio sui possibili fattori di rischio per **basso peso alla nascita** (peso<2.5kg): y=1 bimbo a basso peso; y=0 bimbo peso normale

Dati su 189 donne di cui 59 hanno avuto bimbi con basso peso alla nascita.

Variabili indipendenti (x) considerate:

- Età
- Peso della madre all'ultimo ciclo
- Razza (0=bianca, 1=Nera, 2=altra razza)
- N. di visite dal medico nel I°trimestre

	Coefficiente stimato	Errore standard	Coeff./Se
Età	-0.024	0.034	-0.71
Peso-madre	-0.014	0.652E-02	-2.14
razza1	1.004	0.497	2.02
razza2	0.433	0.362	1.20
n.visite I°trim.	-0.049	0.167	-0.30
costante	1.295	1.069	1.21

Log-likelihood=-111.286

$$G = -2\ln\left[\frac{verosimiglianza.senza.variabile}{verosimiglianza.con.variable}\right] =$$

$$= -2\left[\ln(veros.senza.\mathrm{var}) - \ln(veros.con.\mathrm{var})\right]$$

Non abbiamo la log-likelihood del modello senza variabili: fittiamo un modello con solo il termine costante: Log-likelihood=-117.336

G= -2[(-117.336) – (-111.286)]=12.1

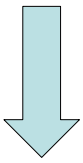
Sotto H_0 che tutti i parametri=0, G si distribusce come χ^2 con gradi di libertà pari al n. di parametri stimati (in questo caso 5).

$P[\chi^2(5)]<0.05$

Il rifiuto dell'ipotesi nulla in questo caso significa che **almeno uno dei parametri stimati è diverso da 0.**

Per vedere quali delle variabili potrebbero essere escluse dal modello, facciamo il **Wald test** su ciascun parametro (ultima colonna tabella)

La razza e il peso della madre sono le due variabili che risultano significative



Possiamo costruire un modello con un minor numero di variabili

	Coefficiente stimato	**Errore standard**	**Coeff./ES**
Peso-madre	-0.015	0.642E-02	-2.37
razza1	1.081	0.487	2.22
razza2	0.481	0.356	1.35
Costante	0.806	0.843	0.96

Log-likelihood= -111.630

Confrontiamolo con il modello precedente:

G= -2[(-111.630) - (-111.286)] = 0.688 con 2 g.l.

P[χ^2 (2)]>0.05 non signif.

Il presente modello è un buon modello

	Coefficiente stimato	Errore standard	Coeff./ES
Peso-madre	-0.015	0.642E-02	-2.37
razza1	1.081	0.487	2.22
razza2	0.481	0.356	1.35
Costante	0.806	0.843	0.96

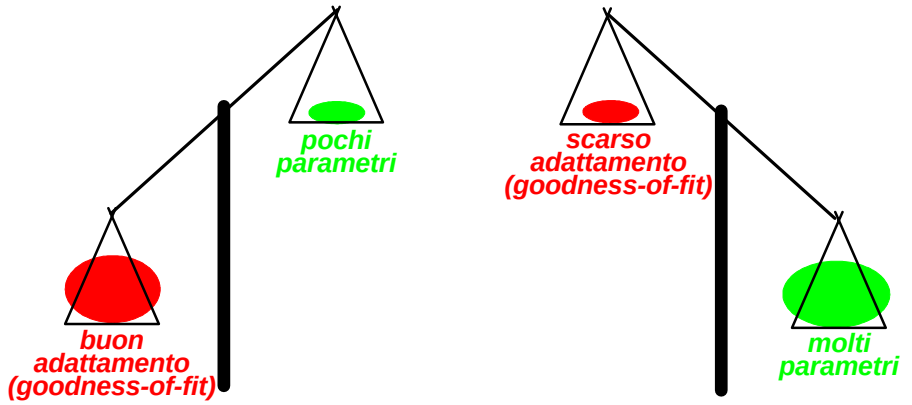
Problema: solo uno dei due coefficienti di **Razza** (razza1) risulta significativo con il Wald test; la variabile sarà complessivamente 'statisticamente significativa' per il modello?

Effettuiamo il **likelihood ratio test** che confronta il modello con **Peso-madre e Razza** con quello contenente solo **Peso-madre**:

G= -2[(-114.345) - (-111.630)] = 5.43 con 2 g.l.

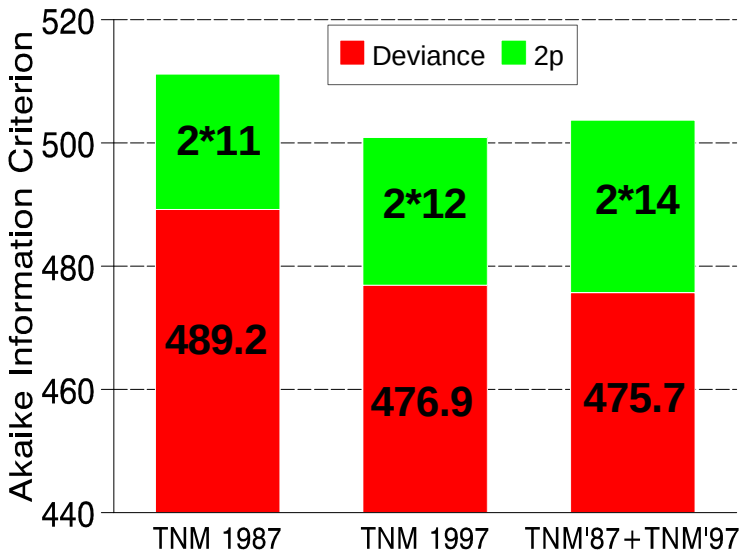
P[χ^2 (2)]=0.066 >0.05 non signif., teoricamente dovremmo escludere **Razza** dal modello, **MA** è biologicamente importante, perciò NON la escludiamo.

E per confrontare modelli con variabili diverse ?



modello ottimale ***modello inadeguato***

L'AIC (Akaike Information criterion) tiene conto sia della bontà dell'adattamento (deviance) che della parsimonia del modello (gradi di libertà)



Applicazioni del Modello di Regressione Logistica -1

1) **Ricerca di variabili esplicative**: quali sono i fattori di rischio di una determinata malattia?

I modelli più utilizzati nell'epidemiologia clinica sono:

a) il **modello logistico** se la variabile di risposta è dicotomica (malattia presente/assente).

b) il **modello di Cox** se dobbiamo tenere presente, oltre ad una variabile di risposta dicotomica (evento presente/assente, vivo/morto), anche l'intervallo di tempo intercorso prima del verificarsi dell'evento (tempo di sopravvivenza)

Modello	Tipo di studio	Variabile di risposta
Modello logistico	Trasversale (Analisi della prevalenza)	Variabile dicotomica
Modello di Cox	Longitudinale (analisi della sopravvivenza)	Variabile dicotomica + tempo di sopravvivenza

2) Stima della probabilità di appartenenza:

Se dobbiamo stimare la probabilità che un soggetto (o un'altra unità statistica) appartenga ad un gruppo (Y=0) oppure ad un altro (Y=1), possiamo cercare la combinazione lineare di variabili esplicative che crea la maggiore discriminazione fra le unità del primo e del secondo gruppo.

3) Previsione: Sviluppare un modello che non solo descriva in modo adeguato la variabile di risposta nel campione in studio, ma possa essere applicato anche ad altri dati.

Ad esempio, vogliamo predire se un soggetto ha la cirrosi epatica, senza ricorrere alla biopsia e all'esame istologico, ma semplicemente sulla base di esami ematochimici. Per scegliere il modello migliore, si utilizzano i dati di soggetti con diagnosi certa, che vengono assegnati casualmente ad un gruppo di apprendimento (training set) o di validazione (validation set). Il modello viene costruito sul primo gruppo e successivamente la sua capacità predittiva viene verificata sul secondo gruppo.

Testi consigliati

- Hosmer DW Jr, Lemeshow S. Applied logistic regression. John Wiley & Sons, New York 1990
- Clayton D, Hills M: Statistical models in epidemiology. Oxford Science Publication; Oxford 1993