

REGRESSIONE LOGISTICA

Introduzione

Con la regressione lineare esaminiamo modelli del tipo:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n,$$

La variabile risposta Y **P** continua e il nostro scopo **P** individuare una serie di variabili esplicative che ci aiutino a predirne il valore medio spiegando la variabilità osservata dei risultati.

In molte situazioni, per**b**, siamo interessati ad una variabile di risposta Y dicotomica¹. Il risultato di Y pub**b** assumere solo due possibili valori; in generale, il valore 1 rappresenta un 'successo' e il valore 0 un 'insuccesso'. La media della variabile casuale dicotomica, indicata con p , **P** la proporzione di volte in cui la variabile assume il valore 1.

In questo caso, vorremmo stimare la probabilità p e determinare i fattori o le variabili esplicative che ne influenzano il valore. A tal fine, utilizziamo una tecnica nota come regressione logistica.

Necessità del modello logistico

Pub**b** sorgere a questo punto spontanea una domanda: perché quando la variabile dipendente **P** di tipo dicotomico non possiamo utilizzare il modello di regressione lineare ?

Nel modello lineare la media dei valori di y_i per ogni valore della X (attesa condizionata di Y) risulta determinata da:

$$E(Y|x_i) = \beta_0 + \beta_1 x_i$$

In questo caso $E(Y|x)$ varia tra meno infinito e β_1 infinito. E' evidente che, nel caso di una variabile dicotomica, il modello non **P** adeguato². Si dovrà procedere a una riparametrizzazione della variabile dipendente in modo da consentire che il suo valore atteso

¹ In realtà il modello logistico, con alcuni aggiustamenti, si dimostra adeguato per studiare fenomeni in cui la variabile discreta di risposta (ordinale e non) sia classificata su più di due livelli.

² Nel caso di una variabile dicotomica la media condizionata deve essere compresa tra zero e uno

$$0 \leq E(Y|x) \leq 1$$

possa assumere un valore compreso tra π_j infinito e meno infinito.

$$\mathbf{B}(x) = E(Y|x)$$

$$\mathbf{B}(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

$$g(x) = \ln\left(\frac{\mathbf{B}(x)}{1 - \mathbf{B}(x)}\right)$$

$$g(x) = \beta_0 + \beta_1 x$$

La funzione logit , $g(x)$ è lineare nei suoi parametri, può essere continua, e può avere un range da più infinito a meno infinito.

Tuttavia nel modello di regressione lineare si assume anche che distribuzione condizionale della variabile dipendente abbia media $E(Y|x)$ e varianza costante. Nel caso di una variabile dicotomica, la varianza della distribuzione condizionale della variabile dipendente assume il valore $\mathbf{B}(x)[1 - \mathbf{B}(x)]$, ossia non è indipendente da x .

Nel modello di regressione si può dimostrare che, una volta assunta la normalità della distribuzione degli errori, i parametri stimati con il metodo dei minimi quadrati sono anche stime di massima verosimiglianza. Nel caso del modello logistico, venendo meno la normalità della distribuzione degli errori, non potremo usare la stima dei minimi quadrati ma dovremo stimare i parametri con il metodo più generale della massima verosimiglianza.

Stima di massima verosimiglianza

Si estragga un campione da una popolazione avente funzione di densità di probabilità $p(x; \mathbf{S})$ in cui \mathbf{S} è un qualunque vettore dei parametri della popolazione che occorre stimare. La densità di probabilità congiunta dell'intero campione³ si ottiene mediante moltiplicazione:

$$p(x_1, x_2, \dots, x_n; \mathbf{S}) = p(x_1; \mathbf{S}) p(x_2; \mathbf{S}) \dots p(x_n; \mathbf{S}) = \prod_{i=1}^n p(x_i; \mathbf{S})$$

Nel caso di una variabile dicotomica, il contributo alla funzione di verosimiglianza espresso da una coppia (y_i, x_i) vale:

$$. (x_i; \mathbf{S}) = \mathbf{B}(x_i)^{y_i} [1 - \mathbf{B}(x_i)]^{1-y_i}$$

$$l(\mathbf{S}) = \prod_{i=1}^n . (x_i; \mathbf{S})$$

Le stime di massima verosimiglianza dei parametri saranno quelle che massimizzano la funzione $l(\mathbf{S})$ ⁴

Si ottengono, in tal modo, delle equazioni che, a differenza del caso della regressione, non saranno lineari nei parametri e necessiteranno perciò di una soluzione numerica di tipo iterativo. Il metodo della massima verosimiglianza fornisce le stime dei parametri della popolazione che con maggiore probabilità \mathbf{P} in grado di determinare i valori campionari osservati, esso dà in un certo senso il 'valore' della popolazione che 'meglio si adatta' al campione osservato. Inoltre, sotto condizioni generali, la stima di massima verosimiglianza presenta le seguenti proprietà asintotiche⁵:

³ Si fa riferimento a un campione casuale (con reintroduzione o da popolazione infinita)

⁴ In realtà \mathbf{B} non si massimizza la funzione $l(\mathbf{S})$, bensì \mathbf{X} il suo logaritmo

⁵ Una proprietà asintotica \mathbf{P} tale quando \mathbf{P} verificata per grandi campioni. Nel caso di piccoli campioni perciò, non \mathbf{P} detto che la stima di massima verosimiglianza sia la miglior stima possibile dei parametri dell'universo

- 1) efficienza, poiché ha varianza più piccola di ogni altro stimatore;
- 2) consistenza, cioè, non distorsione asintotica, con varianza tendente a zero;
- 3) distribuzione normale

Valutazione della SIGNIFICATIVITA' dei coefficienti

La valutazione della significatività dei coefficienti si conduce sfruttando le funzioni di massima verosimiglianza calcolate in corrispondenza del modello completo e di quello ridotto⁶.

A titolo di esempio, si supponga di voler valutare se da un modello completo con 5 (p) covariate si possano eliminare, perché poco predittive, 2 (g) covariate. L'ipotesi nulla diventa in questo caso:

$$H_0: \beta = (0, 0, \beta_{g+1}, \dots, \beta_p)$$

Si procede nel seguente modo:

- 1) si calcola la funzione di massima verosimiglianza in corrispondenza del modello completo $l(\beta_0)$;*
- 2) si calcola la funzione di massima verosimiglianza in corrispondenza del modello ridotto $l(\beta_1)$;*
- 3) si calcola la funzione G così definita:*

$$G = -2 \ln \frac{l(\beta_1)}{l(\beta_0)}$$

E' ragionevole pensare che se il rapporto dei massimi delle funzioni di verosimiglianza con i due modelli tende all'unità, e quindi la differenza fra i rispettivi logaritmi tende a zero, il contributo delle g variabili sotto analisi sia praticamente trascurabile.

In effetti la funzione G si distribuisce, sotto H_0 , asintoticamente come una χ^2 con 2 (g) gradi di libertà.

⁶ I due modelli devono essere 'nested' ossia annidati. Il modello ridotto deve, in altri termini, contenere un subset di covariate ottenuto dal modello completo. Il modello ridotto non potrà mai prevedere l'inserimento di una covariata che non sia anche inserita nel modello completo

Il valore di G consente, confrontato con il valore corrispondente di P_2 , di rifiutare o non rifiutare l'ipotesi nulla H_0 .

Particolare attenzione occorre osservare quando il modello preveda l'inserimento di una covariata categorizzata su più di due valori. In questa situazione **P** possibile che le procedure automatiche selezionino come significativo un solo livello categorico escludendo i rimanenti. Accettando questa soluzione si commetterebbe un grave errore perché si avallerebbe un modello non ipotizzabile in partenza.

Le variabili categoriche devono sempre essere inserite o escluse in toto dal modello.

APPLICAZIONE: esempio n. 1

I dati si riferiscono a 100 soggetti con presenza eventuale di Coronary Heart Disease (Hosmer & Lemeshow Table 1.1 pag.3)

COD codice identificativo
AGE eta' del soggetto
CHD 1 Coronary Heart Disease presente 0 assente

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

. logit CHD AGE

Iteration 0: Log Likelihood =-68.331491
Iteration 1: Log Likelihood =-54.170558
Iteration 2: Log Likelihood =-53.681645
Iteration 3: Log Likelihood =-53.676547
Iteration 4: Log Likelihood =-53.676546

Logit Estimates Number of obs = 100
chi2(1) = 29.31
Prob > chi2 = 0.0000
Pseudo R2 = 0.2145
Log Likelihood = -53.676546

CHD	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
AGE	.1109211	.0240598	4.610	0.000	.0637647	.1580776
_cons	-5.309453	1.133655	-4.683	0.000	-7.531376	-3.087531

. logistic CHD AGE

Logit Estimates Number of obs = 100
chi2(1) = 29.31
Prob > chi2 = 0.0000
Pseudo R2 = 0.2145
Log Likelihood = -53.676546

CHD	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
AGE	1.117307	.0268822	4.610	0.000	1.065842	1.171257

. logit CHD

Iteration 0: Log Likelihood =-68.331491

Logit Estimates Number of obs = 100
chi2(0) = 0.00
Prob > chi2 = .
Pseudo R2 = 0.0000
Log Likelihood = -68.331491

CHD	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
_cons	-.2818512	.2019893	-1.395	0.163	-.6777429	.1140406

Cerchiamo, sulla base di quanto **P** stato detto in precedenza di rispondere al seguente quesito: l'et**B** (AGE) d**B** un contributo significativo nella previsione della presenza di una malattia coronarica (CHD)?

Mettiamo percib a confronto due modelli:

modello 0 : logistic CHD AGE Log Likelihood = - 53.676546

modello 1 : logistic CHD Log Likelihood = - 68.331491 (senza covariate)

La funzione G vale:

$$G = -2[68.332 - (-53.677)] = 29.31$$

Poiché i due modelli (nested) differiscono per una sola covariata il valore di G dovrà essere confrontato con il valore relativo di P^2 con un grado di libertà.

$$G = 29.31 \geq 3.84 = P^2_{(0.05, 1df)}$$

Con una fiducia del 95% respingiamo l'ipotesi nulla, ovvero l'ipotesi che l'età (AGE) non sia un fattore predittivo di malattia coronarica (CHD)

Il test può essere effettuato in modo automatico sfruttando il comando lrtest (likelihood-ratio test)

```
. quietly logistic CHD AGE
. lrtest, saving(0)
. quietly logistic CHD
. lrtest
```

```
Logistic: likelihood-ratio test                chi2(1)      =      29.31
                                                Prob > chi2 =      0.0000
```

Il significato dei coefficienti della regressione logistica

I coefficienti della regressione lineare logistica consentono di risalire, nel caso di una variabile dicotomica, al valore dell'ODDS RATIO (OR)⁷ relativo secondo la seguente relazione:

$$OR_i = e^{b_i}$$

E poiché i coefficienti stimati (stimati con il metodo di massima verosimiglianza) si distribuiscono in modo normale risulta agevole anche la determinazione dell'intervallo di confidenza dell'OR

$$OR_{iL} = e^{[b_i - 1.96SE(b_i)]} \quad OR_{iU} = e^{[b_i + 1.96SE(b_i)]}$$

Nel caso di una variabile continua, il relativo coefficiente della regressione logistica esprime la variazione, in termini di log odds, corrispondente alla modificazione unitaria della variabile indipendente.

Ovvero, ad una crescita di m unit \mathbf{B} della variabile indipendente, il log odds ratio \mathbf{P} uguale a m volte il coefficiente della regressione logistica.

Quando il modello include p_j di una variabile indipendente, ciascun coefficiente consente di stimare il log odds ratio 'adjusted' per tutte le rimanenti variabili.

⁷ In una tabella 2x2 l'OR corrisponde al valore del prodotto crociato e, **nel caso di eventi rari**, approssima il valore del Rischio Relativo

APPLICAZIONE: esempio n. 2

I dati si riferiscono a 71 pazienti leucemici, classificati come responder e non responder alla terapia (Lee Table 11.1 pag 283)

cod codice identificativo del paziente
 resp 1 responder 0 non responder
 age eta' in anni
 agec 1 < 50 anni 0 >= 50 anni

cod	resp	age	agec	cod	resp	age	agec
1	1	20	1	36	1	56	0
2	1	25	1	37	1	19	1
3	1	26	1	38	0	27	1
4	1	26	1	39	0	33	1
5	1	27	1	40	0	34	1
6	1	28	1	41	0	37	1
7	1	28	1	42	0	43	1
8	1	31	1	43	0	45	1
9	1	33	1	44	0	45	1
10	1	33	1	45	0	47	1
11	1	36	1	46	0	48	1
12	1	40	1	47	0	51	0
13	1	40	1	48	0	52	0
14	1	45	1	49	0	53	0
15	1	45	1	50	0	57	0
16	1	50	0	51	0	59	0
17	1	50	0	52	0	59	0
18	1	53	0	53	0	60	0
19	1	56	0	54	0	60	0
20	1	62	0	55	0	61	0
21	1	71	0	56	0	61	0
22	1	74	0	57	0	61	0
23	1	75	0	58	0	63	0
24	1	77	0	59	0	65	0
25	1	18	1	60	0	71	0
26	1	19	1	61	0	73	0
27	1	22	1	62	0	73	0
28	1	26	1	63	0	74	0
29	1	27	1	64	0	80	0
30	1	28	1	65	0	21	1
31	1	28	1	66	0	28	1
32	1	28	1	67	0	36	1
33	1	34	1	68	0	55	0
34	1	37	1	69	0	59	0
35	1	47	1	70	0	62	0
				71	0	83	0

. tabulate resp agec

resp	agec		Total
	0	1	
0	22	12	34
1	10	27	37
Total	32	39	71

L'OR⁸ si pub calcolare direttamente dalla tabella come prodotto crociato:

$$OR = \frac{27}{12} \frac{22}{10} = 4.95$$

```
. logit resp agec
Iteration 0: Log Likelihood =-49.150051
Iteration 1: Log Likelihood =-43.958747
Iteration 2: Log Likelihood =-43.947194
Iteration 3: Log Likelihood =-43.947193
```

```
Logit Estimates                                     Number of obs =      71
                                                    chi2(1)           =   10.41
                                                    Prob > chi2       = 0.0013
                                                    Pseudo R2        = 0.1059

Log Likelihood = -43.947193
```

resp	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
agec	1.599388	.5155821 ⁹	3.102	0.002	.5888652	2.60991
_cons	-.7884574	.381385	-2.067	0.039	-1.535958	-.0409564

```
. logistic resp agec
Logit Estimates                                     Number of obs =      71
                                                    chi2(1)           =   10.41
                                                    Prob > chi2       = 0.0013
                                                    Pseudo R2        = 0.1059

Log Likelihood = -43.947193
```

resp	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
agec	4.95	2.552131	3.102	0.002	1.801942	13.59783

$$OR = e^b = e^{1.5994} = 4.95$$

$$OR_L = e^{0.5888652} = 1.801942 \quad OR_U = e^{2.60991} = 13.59783$$

⁸ In questo caso **P** facile constatare che OR sovrastima il valore di RR

$$RR = (27/39)(32/10) = 2.21$$

⁹ L'errore standard del ln(OR), in una tabella 2x2 **P** stimato da:

$$ES(\ln(OR)) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

Con i nostri dati si ha:

$$ES(\ln(OR)) = \sqrt{\frac{1}{22} + \frac{1}{12} + \frac{1}{10} + \frac{1}{27}} = 0.05155821$$

APPLICAZIONE: esempio n.3

Interpretazione dei coefficienti del modello logistico relativi a variabili discrete codificate su p_{ij} di due livelli

I dati si riferiscono a uno studio sui fattori di rischio associati a un basso peso alla nascita. (Hosmer & Lemeshow data Appendix 1)

```
. tabulate low race
```

birth weight<2500 g	race			Total
	white	black	other	
0	73	15	42	130
1	23	11	25	59
Total	96	26	67	189
OR	w/w 1	b/w 2.33	o/w 1.89	
OR	w/b 0.43	b/b 1	o/b 0.812	

In questo data set la razza **P** codificata come segue:

```
race 1 white
race 2 black
race 3 other
```

Il comando xi, unito all'indicazione i.race, consente la costruzione automatica delle variabili dummy.

Per default la dummy omessa **P** quella associata al valore minimo della variabile. Nel data set in esame verrà omessa la dummy associata a race=1 (white); **ci** significa che race=white sarà assunta come razza di riferimento.

```
. xi: logistic low i.race
i.race          Irace_1-3      (naturally coded; Irace_1 omitted)
```

```
Logit Estimates                                Number of obs =    189
                                                chi2(2)          =    5.01
                                                Prob > chi2      = 0.0817
Log Likelihood = -114.83082                    Pseudo R2       = 0.0214
```

low	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
Irace_2	2.327536	1.078613	1.823	0.068	.9385073 5.772385
Irace_3	1.889234	.6571342	1.829	0.067	.9554577 3.735597

```
. list race Irace_2 Irace_3 (controllo codifica dummy)
```

```
1.   race   Irace_2   Irace_3
    white         0         0
.....
97.  black         1         0
.....
123. other         0         1
.....
189. other         0         1
```

Possiamo comunque imporre una qualsiasi altra categoria come riferimento. Imponiamo ad esempio che il modello utilizzi come razza di riferimento race=2 (black)

```

. char race[omit]2

. xi: logistic low i.race
i.race          Itrace_1-3      (naturally coded; Itrace_2 omitted)

Logit Estimates                                Number of obs =    189
                                                chi2(2)          =     5.01
                                                Prob > chi2     = 0.0817
Log Likelihood = -114.83082                    Pseudo R2       = 0.0214

```

	low	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
Itrace_1		.4296389	.1991007	-1.823	0.068	.1732386	1.065522
Itrace_3		.8116883	.3819123	-0.443	0.657	.3227642	2.041236

```

. list race Itrace_1 Itrace_3 (controllo codifica dummy)

```

```

      race   Itrace_1   Itrace_3
1.    white           1           0
.....
97.   black           0           0
.....
123.  other           0           1
.....
189.  other           0           1

```

Per ritornare alla razza bianca come categoria di riferimento:

```

. char race[omit] 1

. xi: logistic low i.race
i.race          Itrace_1-3      (naturally coded; Itrace_1 omitted)

Logit Estimates                                Number of obs =    189
                                                chi2(2)          =     5.01
                                                Prob > chi2     = 0.0817
Log Likelihood = -114.83082                    Pseudo R2       = 0.0214

```

	low	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
Itrace_2		2.327536	1.078613	1.823	0.068	.9385073	5.772385
Itrace_3		1.889234	.6571342	1.829	0.067	.9554577	3.735597

APPLICAZIONE: esempio n. 4

Riprendiamo l'esempio n.1 a pag.6

Interpretazione di un coefficiente relativo ad una variabile continua

Quale interpretazione possiamo dare al coefficiente relativo all'etB?

CHD	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
AGE	.1109211	.0240598	4.610	0.000	.0637647	.1580776
_cons	-5.309453	1.133655	-4.683	0.000	-7.531376	-3.087531

$$\hat{\xi}(x) = -5.309 + 0.1109 AGE$$

L'OR relativo ad una differenza di etB di 10 anni vale:

$$OR_{10} = e^{(10 \cdot 0.1109)} = 3.03$$

La validitB di tale conclusione P tuttavia discutibile. Infatti la variazione di rischio da 30 a 40 anni sarB, in generale, sicuramente diversa dalla variazione di rischio che si incontra nel passaggio da 50 a 60 anni.

Quando percib si ritiene non accettabile il legame lineare tra logit e una covariata continua, P opportuno discretizzare quest'ultima facendo uso di adeguate variabili dummy.

Stima dell'OR in presenza di interazione

Consideriamo un modello in cui sia inserito un fattore di rischio F, una variabile continua X, e la loro interazione. Il logit sarà espresso come segue

$$g(f, x) = \beta_0 + \beta_1 f + \beta_2 x + \beta_3 fx$$

Il log odds per F=f1 versus F=f0, con X=x è:

$$\ln(OR) = g(f_1, x) - g(f_0, x) \approx b_1(f_1 - f_0) + b_3(f_1 - f_0)x$$

La varianza di ln(OR) vale:

$$\text{var}(\ln(OR)) = \text{var}(b_1)(f_1 - f_0)^2 + \text{var}(b_3)[x(f_1 - f_0)]^2 + 2\text{cov}(b_1, b_3)x(f_1 - f_0)$$

Ovviamente le espressioni si semplificano molto qualora F sia in fattore di rischio dicotomico (0/1)

$$\ln(OR) = b_1 + b_3x$$

$$\text{var}(\ln(OR)) = \text{var}(b_1) + \text{var}(b_3)x^2 + 2\text{cov}(b_1, b_3)x$$

Gli estremi dell'intervallo di confidenza di OR sono:

$$\exp(b_1 + b_3x) \pm z_{(1-\alpha)/2} SE[\ln(OR)]$$

$$SE[\ln(OR)] = \sqrt{\text{var}(\ln(OR))}$$

APPLICAZIONE: esempio n.5

Determinazione dell'OR in presenza di interazione

Usiamo ancora il data set lbw

```
. use lbw
Ricodifichiamo la variabile continua lwt tramite una variabile categorica lwd
. generate lwd=0
. replace lwd=1 if lwt < 110
```

MODELLO 0

```
. logit low
```

Iteration 0: Log Likelihood = -117.336

```
Logit Estimates                                     Number of obs =    189
                                                    chi2(0)          =     0.00
                                                    Prob > chi2      =     .
Log Likelihood = -117.336                          Pseudo R2       = 0.0000
```

low	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
_cons	-.789997	.156976	-5.033	0.000	-1.097664	-.4823297

MODELLO 1

```
. logit low lwd
```

Iteration 0: Log Likelihood = -117.336
Iteration 1: Log Likelihood = -113.161
Iteration 2: Log Likelihood = -113.12058
Iteration 3: Log Likelihood = -113.12058

```
Logit Estimates                                     Number of obs =    189
                                                    chi2(1)          =     8.43
                                                    Prob > chi2      = 0.0037
Log Likelihood = -113.12058                          Pseudo R2       = 0.0359
```

low	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
lwd	1.053762	.3615635	2.914	0.004	.3451102	1.762413
_cons	-1.053762	.1883882	-5.594	0.000	-1.422996	-.6845277

MODELLO 2

```
. logit low lwd age
```

Iteration 0: Log Likelihood = -117.336
Iteration 1: Log Likelihood = -112.19831
Iteration 2: Log Likelihood = -112.14339
Iteration 3: Log Likelihood = -112.14338

```
Logit Estimates                                     Number of obs =    189
                                                    chi2(2)          =    10.39
                                                    Prob > chi2      = 0.0056
Log Likelihood = -112.14338                          Pseudo R2       = 0.0443
```

low	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
lwd	1.010122	.3642627	2.773	0.006	.2961806	1.724064
age	-.044232	.0322248	-1.373	0.170	-.1073913	.0189274
_cons	-.026891	.7621481	-0.035	0.972	-1.520674	1.466892

MODELLO 3

```
. xi : logit low i.lwd age i.lwd*age
i.lwd          Ilwd_0-1      (naturally coded; Ilwd_0 omitted)
i.lwd*age      IlXage_#     (coded as above)
```

Note: Ilwd_1 dropped due to collinearity.
 Note: age dropped due to collinearity.
 Iteration 0: Log Likelihood = -117.336
 Iteration 1: Log Likelihood = -110.71804
 Iteration 2: Log Likelihood = -110.57024
 Iteration 3: Log Likelihood = -110.56997

```
Logit Estimates                                Number of obs =    189
                                                chi2(3)          =   13.53
                                                Prob > chi2     =  0.0036
Log Likelihood = -110.56997                    Pseudo R2       =  0.0577
```

low	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
Ilwd_1	-1.944089	1.724804	-1.127	0.260	-5.324643	1.436465
age	-.0795722	.0396343	-2.008	0.045	-.157254	-.0018904
IlXage_1	.1321967	.0756982	1.746	0.081	-.0161691	.2805626
_cons	.7744952	.9100949	0.851	0.395	-1.009258	2.558248

Tabella riassuntiva dei coefficienti relativi ai vari modelli

modello	costante	lwd	age	lwd*age
0	-0.79			
1	-1.054	-1.054		
2	-0.027	1.01	-0.044	
3	0.774	-1.944	-0.080	0.132

Evidenziamo ora la matrice di varianza e covarianza relativa al modello 3

```
. matrix vc =get(VCE)
. matrix list vc
```

```
symmetric vc[4,4]
      Ilwd_1      age      IlXage_1      _cons
Ilwd_1      2.974949
age          .03526621      .00157088
IlXage_1     -.12760349      -.00157088      .00573022
_cons        -.82827277      -.03526621      .03526621      .82827277
```

Calcoliamo ora l'OR per lwd controllato per age

```
b1= -1.944
b2=  0.080
b3=  0.132
```

$\ln(\text{OR}) = -1.944 + 0.132 \text{ age}$

$\text{var}(\ln(\text{OR})) = 2.975 + 0.00573 \text{ age}^2 + 2(-0.1276)\text{age}$

La tabella riporta i valori degli OR con i rispettivi intervalli di confidenza al 95% per lwd controllati per age

EtB	15	20	25	30	35	40	45
$\ln(\text{OR})$	0.04	0.7	1.36	2.02	2.68	3.34	4
$\text{var}(\ln(\text{OR}))$	0.42	0.15	0.16	0.45	1.03	1.90	3.06
OR	1.04	2.01	3.9	7.55	14.6	28.3	54.9
ORI	0.3	0.95	1.8	2	2	1.9	1.78
ORU	3.7	4.27	8.5	28.2	107	423	1690

Il notevole incremento degli intervalli di confidenza per un'etB superiore a 30 anni, suggerisce che le stime degli OR devono essere valutate con estrema cautela nel caso di soggetti con etB superiore a 30 anni

APPLICAZIONE: esempio n. 6

OR stimato con Mantel-Haenszel e tramite i coefficienti della logistica

Riprendiamo il data set giB considerato nell'esempio precedente.

. tabulate low smoke

birth weight<2500 g	smoked during pregnancy		Total
	0	1	
0	86	44	130
1	29	30	59
Total	115	74	189

L'OR grezzo vale:

$$OR_G = \frac{3086}{4429} = 2.021944$$

. sort race

. cs low smoke, by(race) or (vedi epitab [5s] pag.314 vol.2)

race	OR	[95% Conf. Interval]		M-H Weight
white	5.757576	1.850317	17.70502	1.375 (Cornfield)
black	3.3	.6673603	16.36071	.7692308 (Cornfield)
other	1.25	.3682685	4.278368	2.089552 (Cornfield)
Crude	2.021944	1.084529	3.770602	
M-H combined	3.086381	1.49074	6.389949	

Test for heterogeneity (M-H) chi2(2) = 3.031 Pr>chi2 = 0.2197

Test that combined OR = 1:

Mantel-Haenszel chi2(1) = 9.41
Pr>chi2 = 0.0022

N.B.: Test per l'omogeneità **H0: OR1=OR2=...ORn**

L'ipotesi nulla, nel nostro caso, non può venire rifiutata (Pr>chi2=0.2197), quindi **P** lecito tentare di combinare gli OR parziali in un unico OR globale.

L'ipotesi nulla che l'OR globale sia pari a 1 (Pr>chi2=0.0022) viene invece rifiutata

. sort race

. by race: tabulate low smoke

-> race= **white**

birth weight<2500 g	smoked during pregnancy		Total
	0	1	
0	40	33	73
1	4	19	23
Total	44	52	96

```
-> race= black
```

birth weight<2500 g	smoked during pregnancy		Total
	0	1	
0	11	4	15
1	5	6	11
Total	16	10	26

```
-> race= other
```

birth weight<2500 g	smoked during pregnancy		Total
	0	1	
0	35	7	42
1	20	5	25
Total	55	12	67

$$OR_{MH} = \frac{\sum \frac{a_i d_i}{N_i}}{\sum \frac{b_i c_i}{N_i}}$$

$$OR_{MH} = \frac{\frac{19 \cdot 40}{96} + \frac{6 \cdot 11}{26} + \frac{5 \cdot 35}{67}}{\frac{4 \cdot 33}{96} + \frac{5 \cdot 4}{26} + \frac{20 \cdot 7}{67}} = \frac{13.067}{4.2338} = 3.086$$

Oltre che secondo Mantel-Haenszel l'OR globale può ricavarsi come media pesata degli OR parziali con pesi pari al reciproco della varianza dei log(OR)

	White	Black	Other
OR	5.758	3.3	1.25
ln(OR)	1.751	1.194	0.223
var[log(OR)]	0.358	0.708	0.421
w	2.794	1.413	2.375

$$OR_L = \exp\left(\frac{\sum w_i \ln(OR_i)}{\sum w_i}\right)$$

ORL = 2.95 (leggermente minore dello stimatore di Mantel-Haenzel)

Utilizzando il modello logistico, si può agevolmente calcolare lo stimatore di max verosimiglianza dell'OR globale

Modello 1

. xi: logit low smoke

Iteration 0: Log Likelihood = -117.336
Iteration 1: Log Likelihood = -114.9123
Iteration 2: Log Likelihood = -114.9023

Logit Estimates

Number of obs = 189
chi2(1) = 4.87
Prob > chi2 = 0.0274
Pseudo R2 = 0.0207

Log Likelihood = -114.9023

low	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
smoke	.7040592	.3196386	2.203	0.028	.0775791	1.330539
_cons	-1.087051	.2147299	-5.062	0.000	-1.507914	-.6661886

Modello 2

. xi: logit low smoke i.race

i.race Irace_1-3 (naturally coded; Irace_1 omitted)

Iteration 0: Log Likelihood = -117.336
Iteration 1: Log Likelihood = -110.10441
Iteration 2: Log Likelihood = -109.98749
Iteration 3: Log Likelihood = -109.98736

Logit Estimates

Number of obs = 189
chi2(3) = 14.70
Prob > chi2 = 0.0021
Pseudo R2 = 0.0626

Log Likelihood = -109.98736

low	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
smoke	1.116004	.3692258	3.023	0.003	.3923346	1.839673
Irace_2	1.084088	.4899845	2.212	0.027	.1237362	2.04444
Irace_3	1.108563	.4003054	2.769	0.006	.3239787	1.893147
_cons	-1.840539	.3528633	-5.216	0.000	-2.532138	-1.148939

Modello 3

. xi: logit low smoke i.race i.race*smoke

i.race Irace_1-3 (naturally coded; Irace_1 omitted)

i.race*smoke IrXsmo_# (coded as above)

Note: Irace_2 dropped due to collinearity.
Note: Irace_3 dropped due to collinearity.
Note: smoke dropped due to collinearity.
Iteration 0: Log Likelihood = -117.336
Iteration 1: Log Likelihood = -108.9189
Iteration 2: Log Likelihood = -108.42021
Iteration 3: Log Likelihood = -108.4089
Iteration 4: Log Likelihood = -108.40889

Logit Estimates

Number of obs = 189
chi2(5) = 17.85
Prob > chi2 = 0.0031
Pseudo R2 = 0.0761

Log Likelihood = -108.40889

low	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
smoke	1.750517	.5982759	2.926	0.003	.5779173	2.923116
ltrace_2	1.514128	.7522689	2.013	0.044	.0397077	2.988548
ltrace_3	1.742969	.5946183	2.931	0.003	.5775389	2.9084
lrxsmo_2	-.556594	1.032235	-0.539	0.590	-2.579738	1.46655
lrxsmo_3	-1.527373	.8828152	-1.730	0.084	-3.257659	.202913
_cons	-2.302585	.5244039	-4.391	0.000	-3.330398	-1.274772

Abbiamo costruito tre modelli:

Modello 1 SMOKE
Modello 2 SMOKE RACE
Modello 3 SMOKE RACE RACE*SMOKE

Model	SMOKE	Log-Likelihood	G	df	p
1	0.704	-114.90			
2	1.116	-109.99	9.83	2	0.007
3	1.751	-108.41	3.16	2	0.206

OR grezzo = $\exp(0.704) = 2.02$

ORML = $\exp(1.116) = 3.05$ (molto prossimo allo stimatore di MH 3.09)

L'omogeneit**B** degli OR parziali tra i vari strati viene valutata tramite la quantificazione del contributo alla massima verosimiglianza fornita dall'interazione.

Si deve perci**b** impostare un likelihood ratio test del mod. 2 versus il mod. 3. Nel nostro esempio avremo un valore di G pari a 3.16 con 2 gradi di libert**B**; ne consegue un $p = 0.206$ che non ci consente di rifiutare l'ipotesi nulla di omogeneit**B** degli OR parziali lungo gli strati.

Come inserire e/o categorizzare le variabili

Molte volte possiamo chiederci se il legame tra il logit e la variabile indipendente sia veramente di tipo lineare oppure se quest'ultima debba essere eventualmente inserita nel modello dopo essere stata opportunamente trasformata per linearizzarne l'andamento in funzione del logit.

Si può tentare di risolvere il problema in due modi:

1) sia x la variabile su cui si vuole testare l'andamento lineare nei confronti del logit, si inserisce nel modello una nuova variabile $x \ln(x)$ e se ne valuta il contributo. Se la variabile $x \ln(x)$ non dà un contributo significativo al modello non vi è ragione di dubitare del legame lineare tra il logit e la variabile x .

2) si discretizza la variabile x in quartili e si valuta l'andamento dell'OR in funzione del punto medio del quartile corrispondente.

APPLICAZIONE: esempio n. 6

Consideriamo ancora il data set lbw e valutiamo se la variabile age possa essere inserita nel modello senza necessitare di alcuna trasformazione

```
.use lbw
xi: logit low age lwt i.race smoke ptl ht ui (Modello completo)
i.race          Itrace_1-3      (naturally coded; Itrace_1 omitted)
```

```
Iteration 0:  Log Likelihood = -117.336
Iteration 1:  Log Likelihood =-101.39666
Iteration 2:  Log Likelihood =-100.73153
Iteration 3:  Log Likelihood = -100.724
Iteration 4:  Log Likelihood = -100.724
```

```
Logit Estimates                                     Number of obs =    189
                                                    chi2(8)          =   33.22
                                                    Prob > chi2      =  0.0001
Log Likelihood =   -100.724                        Pseudo R2       =  0.1416
```

low	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
age	-.0271003	.0364504	-0.743	0.457	-.0985418	.0443412
lwt	-.0151508	.0069259	-2.188	0.029	-.0287253	-.0015763
Itrace_2	1.262647	.5264101	2.399	0.016	.2309024	2.294392
Itrace_3	.8620792	.4391531	1.963	0.050	.0013548	1.722804
smoke	.9233448	.4008266	2.304	0.021	.1377391	1.708951
ptl	.5418366	.346249	1.565	0.118	-.136799	1.220472
ht	1.832518	.6916292	2.650	0.008	.4769494	3.188086
ui	.7585135	.4593768	1.651	0.099	-.1418484	1.658875
_cons	.4612239	1.20459	0.383	0.702	-1.899729	2.822176

```
. xi: logit low lwt i.race smoke ptl ht ui (Modello ridotto)
i.race          Itrace_1-3      (naturally coded; Itrace_1 omitted)
```

```
Iteration 0: Log Likelihood = -117.336
Iteration 1: Log Likelihood =-101.59389
Iteration 2: Log Likelihood =-101.00914
Iteration 3: Log Likelihood =-101.00398
Iteration 4: Log Likelihood =-101.00398
```

```
Logit Estimates                                     Number of obs =    189
                                                    chi2(7)          =    32.66
                                                    Prob > chi2     = 0.0000
Log Likelihood = -101.00398                       Pseudo R2       = 0.1392
```

low	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
lwt	-.015874	.0068532	-2.316	0.021	-.0293061	-.002442
Itrace_2	1.32521	.5221907	2.538	0.011	.3017352	2.348685
Itrace_3	.8975541	.4338379	2.069	0.039	.0472474	1.747861
smoke	.9387372	.3986848	2.355	0.019	.1573293	1.720145
ptl	.5032501	.3412117	1.475	0.140	-.1655126	1.172013
ht	1.853887	.6949829	2.668	0.008	.4917457	3.216029
ui	.7856561	.4564234	1.721	0.085	-.1089173	1.680229
_cons	-.0903816	.9516345	-0.095	0.924	-1.955551	1.774788

```
. quietly xi: logit low age lwt i.race smoke ptl ht ui
. lrtest, saving(0)      [memorizzo il modello completo]
. quietly xi: logit low lwt i.race smoke ptl ht ui
. lrtest
```

```
Logit: likelihood-ratio test                                chi2(1) =    0.56
                                                            Prob > chi2 =    0.4543
```

Il contributo della variabile age non risulta significativo a livello del 95%
 Controlliamo che il legame tra age e logit non sia di tipo lineare secondo le
 indicazioni di cui al punto 1 precedente.
 Costruiamo allo scopo una nuova variabile

```
. generate ala = age*ln(age)
. quietly xi: logit low age ala lwt i.race smoke ptl ht ui
. lrtest, saving(0)      [memorizzo il modello completo]
. quietly xi: logit low age lwt i.race smoke ptl ht ui
. lrtest
```

```
Logit: likelihood-ratio test                                chi2(1) =    0.52
                                                            Prob > chi2 =    0.4712
```

Il contributo della variabile age*ln(age) non risulta significativo a livello del
 95%. Non abbiamo percib ragione di dubitare del legame lineare del logit con age.

Procediamo ora secondo le indicazioni di cui al punto 2
 Ricaviamo i limiti corrispondenti ai quartili

```
. centile age, c(25,50,75)
```

Variabile	Obs	Percentile	Centile	-- Binom. Interp. -- [95% Conf. Interval]	
age	189	25	19	19	20
		50	23	21.50878	24
		75	26	25	28

Codifichiamo age secondo il quartile di appartenenza tramite la variabile categorica agec

```
. generate agec=0
(189 real changes made)
. replace agec=1 if age < 19
(35 real changes made)
. replace agec=2 if age >=19 & age <23
(59 real changes made)
. replace agec=3 if age >=23 & age <26
(41 real changes made)
. replace agec=4 if age >=26
(54 real changes made)

. xi: logit low i.agec lwt i.race smoke ptl ht ui
i.race          Irace_1-3      (naturally coded; Irace_1 omitted)
agec not found
r(111);
```

```
. xi: logit low i.agec lwt i.race smoke ptl ht ui
i.agec          Iagec_1-4      (naturally coded; Iagec_1 omitted)
i.race          Irace_1-3      (naturally coded; Irace_1 omitted)
```

```
Iteration 0: Log Likelihood = -117.336
Iteration 1: Log Likelihood =-100.92817
Iteration 2: Log Likelihood = -100.3155
Iteration 3: Log Likelihood = -100.3098
Iteration 4: Log Likelihood = -100.3098
```

```
Logit Estimates                                     Number of obs =    189
                                                    chi2(10)         =    34.05
                                                    Prob > chi2      = 0.0002
Log Likelihood = -100.3098                       Pseudo R2        = 0.1451
```

low	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
Iagec_2	-.150963	.5115327	-0.295	0.768	-1.153549	.8516227
Iagec_3	.38978	.537139	0.726	0.468	-.6629932	1.442553
Iagec_4	-.1085019	.5438405	-0.200	0.842	-1.17441	.957406
lwt	-.0147081	.0070912	-2.074	0.038	-.0286066	-.0008095
Irace_2	1.273804	.5323252	2.393	0.017	.230466	2.317142
Irace_3	.8858588	.4345804	2.038	0.042	.0340969	1.737621
smoke	1.014545	.4052074	2.504	0.012	.2203527	1.808736
ptl	.4639473	.3518591	1.319	0.187	-.225684	1.153579
ht	1.851264	.702434	2.635	0.008	.4745189	3.22801
ui	.8634525	.4691586	1.840	0.066	-.0560815	1.782987
_cons	-.2758081	.9806852	-0.281	0.779	-2.197916	1.6463

```
. xi: logistic low i.agec lwt i.race smoke ptl ht ui
i.agec          Iagec_1-4      (naturally coded; Iagec_1 omitted)
i.race          Irace_1-3      (naturally coded; Irace_1 omitted)
```

```
Logit Estimates                                     Number of obs =    189
                                                    chi2(10)         =    34.05
                                                    Prob > chi2      = 0.0002
Log Likelihood = -100.3098                       Pseudo R2        = 0.1451
```

low	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
Iagec_2	.8598795	.4398565	-0.295	0.768	.3155151 2.343447
Iagec_3	1.476656	.7931695	0.726	0.468	.5153066 4.231485
Iagec_4	.8971772	.4879213	-0.200	0.842	.3090013 2.604931
lwt	.9853996	.0069877	-2.074	0.038	.9717987 .9991908
Irace_2	3.574425	1.902756	2.393	0.017	1.259187 10.14664
Irace_3	2.425066	1.053886	2.038	0.042	1.034685 5.683804
smoke	2.758107	1.117605	2.504	0.012	1.246516 6.102732
ptl	1.590339	.5595754	1.319	0.187	.7979702 3.169515
ht	6.367865	4.473005	2.635	0.008	1.607241 25.22939
ui	2.371334	1.112532	1.840	0.066	.9454621 5.947592

Quartile	1	2	3	4
Midpoint	16.5	21	24.5	35.5
OR	1	0.86	1.48	0.89

L'OR mostra un andamento pendolare intorno a 1. Non appare per altro evidente un andamento a U che farebbe sospettare un legame di tipo quadratico tra la variabile in esame e il logit.

La non significatività del contributo della variabile age al modello considerato non impone tuttavia che tale variabile debba essere esclusa dal modello finale. Se una variabile **P** ritenuta biologicamente importante può essere inserita nel modello indipendentemente dal suo contributo misurato in termini statistici. Nell'esempio considerato i ricercatori, anche per valutare possibili interazioni con altre variabili, hanno ritenuto opportuno mantenere la variabile age nel modello

Non P detto che tutto ciò che sia statisticamente irrilevante debba anche essere trascurabile dal punto di vista clinico e/o biologico !!!

In fondo lo scopo di un modello **P** quello di prevedere e giustificare una realtà **B**, la statistica **P** solamente uno strumento, una lente attraverso cui esaminare i fenomeni e le loro frequenze. Se attraverso la lente non vediamo l'evidenza non per questo siamo autorizzati a negarla, più correttamente dovremo mettere in discussione la lente e la nostra capacità visiva!!!!

APPLICAZIONE: esempio n. 7

Introdurre una variabile continua o discreta?

Consideriamo il solito data set lbw
use lbw
(Hosmer & Lemeshow data)

```
. xi: logit low age lwt i.race smoke ptl ht ui  
i.race          Irace_1-3    (naturally coded; Irace_1 omitted)
```

```
Iteration 0:  Log Likelihood = -117.336  
Iteration 1:  Log Likelihood =-101.39666  
Iteration 2:  Log Likelihood =-100.73153  
Iteration 3:  Log Likelihood = -100.724  
Iteration 4:  Log Likelihood = -100.724
```

```
Logit Estimates                                     Number of obs =    189  
                                                    chi2(8)          =   33.22  
                                                    Prob > chi2     =  0.0001  
Log Likelihood =  -100.724                        Pseudo R2       =  0.1416
```

low	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
age	-.0271003	.0364504	-0.743	0.457	-.0985418	.0443412
lwt	-.0151508	.0069259	-2.188	0.029	-.0287253	-.0015763
Irace_2	1.262647	.5264101	2.399	0.016	.2309024	2.294392
Irace_3	.8620792	.4391531	1.963	0.050	.0013548	1.722804
smoke	.9233448	.4008266	2.304	0.021	.1377391	1.708951
ptl	.5418366	.346249	1.565	0.118	-.136799	1.220472
ht	1.832518	.6916292	2.650	0.008	.4769494	3.188086
ui	.7585135	.4593768	1.651	0.099	-.1418484	1.658875
_cons	.4612239	1.20459	0.383	0.702	-1.899729	2.822176

Possiamo valutare se inserire nel modello la variabile lwd come discreta anziché come continua. Proviamo a discretizzare lwt ricodificandola, in quartili, tramite lwd.

```
. centile lwt, c(25,50,75)
```

Variabile	Obs	Percentile	Centile	-- Binom. Interp. -- [95% Conf. Interval]	
lwt	189	25	110	105.3041	113
		50	121	120	128
		75	140.5	134	152.0876

```
. generate lwd=0  
. replace lwd=1 if lwt <110  
. replace lwd=2 if lwt >=110 & lwt<121  
. replace lwd=3 if lwt>=121 & lwt <140  
. replace lwd=4 if lwt >=140
```

```
. xi: logit low age i.lwd i.race smoke ptl ht ui  
i.lwd          Ilwd_1-4    (naturally coded; Ilwd_1 omitted)  
i.race          Irace_1-3    (naturally coded; Irace_1 omitted)
```

```
Iteration 0:  Log Likelihood = -117.336  
Iteration 1:  Log Likelihood =-101.59276  
Iteration 2:  Log Likelihood =-101.03944  
Iteration 3:  Log Likelihood =-101.03406  
Iteration 4:  Log Likelihood =-101.03406
```

```

Logit Estimates
Log Likelihood = -101.03406
Number of obs = 189
chi2(10) = 32.60
Prob > chi2 = 0.0003
Pseudo R2 = 0.1389

```

low	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
age	-.0339381	.0366729	-0.925	0.355	-.1058156	.0379395
Ilwd_2	-.8069816	.4764295	-1.694	0.090	-1.740766	.1268031
Ilwd_3	-.6935817	.4930435	-1.407	0.160	-1.659929	.2727659
Ilwd_4	-1.039099	.5306137	-1.958	0.050	-2.079083	.0008846
Irace_2	1.152482	.5148909	2.238	0.025	.143314	2.161649
Irace_3	.8531254	.4400472	1.939	0.053	-.0093512	1.715602
smoke	.8730467	.3983866	2.191	0.028	.0922232	1.65387
ptl	.5867308	.3495249	1.679	0.093	-.0983254	1.271787
ht	1.512056	.6605641	2.289	0.022	.2173738	2.806737
ui	.710506	.4603441	1.543	0.123	-.1917519	1.612764
_cons	-.6190517	.9773966	-0.633	0.526	-2.534714	1.296611

```

. xi: logistic low age i.lwd i.race smoke ptl ht ui
i.lwd          Ilwd_1-4      (naturally coded; Ilwd_1 omitted)
i.race         Irace_1-3     (naturally coded; Irace_1 omitted)

```

```

Logit Estimates
Log Likelihood = -101.03406
Number of obs = 189
chi2(10) = 32.60
Prob > chi2 = 0.0003
Pseudo R2 = 0.1389

```

low	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
age	.9666314	.0354492	-0.925	0.355	.8995905	1.038668
Ilwd_2	.4462029	.2125842	-1.694	0.090	.175386	1.135193
Ilwd_3	.4997828	.2464147	-1.407	0.160	.1901524	1.313593
Ilwd_4	.3537732	.1877169	-1.958	0.050	.1250449	1.000885
Irace_2	3.16604	1.630165	2.238	0.025	1.154092	8.685449
Irace_3	2.346971	1.032778	1.939	0.053	.9906924	5.560021
smoke	2.394194	.9538149	2.191	0.028	1.09661	5.22717
ptl	1.7981	.6284808	1.679	0.093	.9063539	3.567221
ht	4.536046	2.996349	2.289	0.022	1.242809	16.55582
ui	2.035021	.9368099	1.543	0.123	.8255116	5.016658

```

quartile      1          2          3          4
midpoint      95        115.5      129.5      195
number        42         50         47         50
b              0         -0.807     -0.649     -1.039
OR             1          0.44        0.49        0.35

```

Gli OR dei tre quartili terminali sono piuttosto omogenei, quindi **pub** essere conveniente discretizzare la variabile lwt in due categorie. In tal modo, tra l'altro, l'interpretazione degli OR sarebbe molto facilitata

```
lwd = 1 => lwt < 110          lwd = 0  altrimenti
```

```
. replace lwd=0
. replace lwd=1 if lwt < 110
```

```

. xi: logistic low age i.lwd i.race smoke ptl ht ui
i.lwd          Ilwd_0-1      (naturally coded; Ilwd_0 omitted)
i.race         Irace_1-3     (naturally coded; Irace_1 omitted)

```



```

. tabulate ptl
  premature|
    labor|
    history|
    (count)|
-----+-----
          0 |          159      84.13      84.13
          1 |           24      12.70      96.83
          2 |            5       2.65      99.47
          3 |             1       0.53     100.00
-----+-----
        Total |          189     100.00
Pare ragionevole codificare la variabile ptl su due livelli

ptd = 0  =>  ptl = 0                ptd = 1  altrimenti

```

```

. generate ptd=0
. replace ptd=1 if ptl >0

. xi: logit low age i.lwd i.race smoke i.ptd ht ui
i.lwd          Ilwd_0-1      (naturally coded; Ilwd_0 omitted)
i.race          Irace_1-3    (naturally coded; Irace_1 omitted)
i.ptd           Iptd_0-1    (naturally coded; Iptd_0 omitted)

```

```

Iteration 0:  Log Likelihood = -117.336
Iteration 1:  Log Likelihood =-99.431174
Iteration 2:  Log Likelihood =-98.785718
Iteration 3:  Log Likelihood =  -98.778
Iteration 4:  Log Likelihood =-98.777998

```

```

Logit Estimates                                     Number of obs =    189
                                                    chi2(8)          =   37.12
                                                    Prob > chi2      =  0.0000
Log Likelihood = -98.777998                        Pseudo R2       =  0.1582

```

low	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
age	-.0464796	.0373888	-1.243	0.214	-.1197603	.0268011
Ilwd_1	.8420615	.4055338	2.076	0.038	.0472299	1.636893
Irace_2	1.073456	.5150752	2.084	0.037	.0639273	2.082985
Irace_3	.815367	.4452979	1.831	0.067	-.0574008	1.688135
smoke	.8071996	.404446	1.996	0.046	.0145001	1.599899
Iptd_1	1.281678	.4621157	2.774	0.006	.3759478	2.187408
ht	1.435227	.6482699	2.214	0.027	.1646415	2.705813
ui	.6576256	.4666192	1.409	0.159	-.2569313	1.572182
_cons	-1.216781	.9556797	-1.273	0.203	-3.089878	.656317

Quali interazioni inserire nel modello finale?

Ritenuto che il modello finale presentato nell'applicazione precedente sia comprensivo di tutti gli effetti principali, **P** giusto domandarsi quali interazioni possano eventualmente essere inserite per aumentarne la capacità **B** predittiva.

La scelta preliminare delle interazioni inseribili nel modello dovrà essere condotta principalmente in base alla plausibilità sul piano clinico-biologico. Ritornando all'ultimo modello presentato, **P** ben noto che et **B** (AGE), razza (RACE) e condizione di fumatore (SMOKE) hanno potenziali capacità **B** di interagire con numerosi fattori. Così come **P** abbastanza associato che peso e ipertensione possano sviluppare, tra loro, interazioni degne di nota.

Tra tutte le 42 interazioni attivabili soltanto le 17 sotto presentate paiono perciò, a priori, avere una certa attendibilità sul piano clinico-biologico

Interazione	Log-Likelihood	G	df	p-value
Main effect	-98.78			
AGE x RACE	-98.53	0.52	2	0.78
AGE x SMOKE	-98.51	0.54	1	0.46
AGE x HT	-98.39	0.78	1	0.38
AGE x UI	-98.76	0.04	1	0.84
AGE x LWD	-97.50	2.56	1	0.11
AGE x PTD	-98.36	0.84	1	0.36
RACE x SMOKE	-97.61	2.34	2	0.31
RACE x HT	-98.63	0.30	2	0.86
RACE x UI	-97.62	2.32	2	0.31
RACE x LWD	-97.08	3.40	2	0.18
RACE x PTD	-98.50	0.56	2	0.76
SMOKE x HT	-98.71	0.14	1	0.71
SMOKE x UI	-98.12	1.32	1	0.25
SMOKE x LWD	-97.61	2.34	1	0.13
SMOKE x PTD	-98.31	0.94	1	0.33
LWD x HT	-98.22	1.12	1	0.30
AGE x LWD x SMOKE x LWD	-96-01	5.54	2	0.06

Dalla tabella si deduce che le uniche interazioni che hanno una certa rilevanza statistica sono AGE x LWD e SMOKE x LWD.

Di seguito si riportano i coefficienti relativi al modello comprendente queste due interazioni.

```
xi: logit low age i.race smoke ht ui lwd ptd AxL SxL
i.race          Irace_1-3      (naturally coded; Irace_1 omitted)
```

```
Iteration 0: Log Likelihood = -117.336
Iteration 1: Log Likelihood = -97.135228
Iteration 2: Log Likelihood = -96.03855
Iteration 3: Log Likelihood = -96.006202
Iteration 4: Log Likelihood = -96.00616
```

```
Logit Estimates                                     Number of obs =    189
                                                    chi2(10)         =   42.66
                                                    Prob > chi2      =  0.0000
Log Likelihood = -96.00616                         Pseudo R2        =  0.1818
```

low	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
(b3)age	-.0839782	.0455663	-1.843	0.065	-.1732864	.0053301
Irace_2	1.083103	.5189153	2.087	0.037	.0660474	2.100158
Irace_3	.7596787	.4640335	1.637	0.102	-.1498103	1.669168
(b2)smoke	1.153131	.4584383	2.515	0.012	.2546084	2.051653
ht	1.359216	.661471	2.055	0.040	.062757	2.655676
ui	.7281685	.4794797	1.519	0.129	-.2115945	1.667932
(b1)lwd	-1.729949	1.868306	-0.926	0.354	-5.391762	1.931863
ptd	1.231578	.4713903	2.613	0.009	.3076701	2.155486
(b5)AxL	.1474112	.0828594	1.779	0.075	-.0149902	.3098127
(b4)SxL	-1.407375	.8186761	-1.719	0.086	-3.011951	.1972003
_cons	-.5117544	1.087536	-0.471	0.638	-2.643286	1.619777

```
. xi: logistic low age i.race smoke ht ui lwd ptd AxL SxL
i.race          Irace_1-3      (naturally coded; Irace_1 omitted)
```

```
Logit Estimates                                     Number of obs =    189
                                                    chi2(10)         =   42.66
                                                    Prob > chi2      =  0.0000
Log Likelihood = -96.00616                         Pseudo R2        =  0.1818
```

low	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
(b3)age	.9194513	.041896	-1.843	0.065	.8408967	1.005344
Irace_2	2.95383	1.532788	2.087	0.037	1.068277	8.167462
Irace_3	2.137589	.9919132	1.637	0.102	.8608713	5.307749
(b2)smoke	3.168096	1.452377	2.515	0.012	1.289956	7.780755
ht	3.893141	2.5752	2.055	0.040	1.064768	14.2346
ui	2.071284	.9931385	1.519	0.129	.8092928	5.301191
(b1)lwd	.1772934	.3312383	-0.926	0.354	.0045539	6.902359
ptd	3.426633	1.615282	2.613	0.009	1.360252	8.632086
(b5)AxL	1.15883	.09602	1.779	0.075	.9851216	1.36317
(b4)SxL	.2447849	.2003996	-1.719	0.086	.0491956	1.217988

Cerchiamo ora di valutare i differenti odds ratio relativi alle variabili SMOKE, LWD, AGE.

Potremmo ad esempio essere interessati a esaminare l'odds ratio relativo a LWD tra i fumatori e i non fumatori, controllando per l'età **B**; oppure potremmo avere interesse a quantificare l'odds ratio tra SMOKE e LWD controllando per l'età **B**..... In generale ciascun odds ratio può essere stimato dall'EXP di un'opportuna differenza di logit.

$$\mathcal{S} = b_0 + b_1 w + b_2 s + b_3 a + b_4 (ws) + b_5 (wa) + b'z$$

nel nostro caso avremo:

$b_1 = -1.730$ $b_2 = 1.153$ $b_3 = -0.084$ $b_4 = -1.407$
 $b_5 = 0.147$

Smoke	LWD=0	LWD=1	Logit diff
0	$b_0 + b_3 a + b'z$	$b_0 + b_1 + b_3 a + b_5 a + b'z$	$d_1 = b_1 + b_5 a$
1	$b_0 + b_2 + b_3 a + b'z$	$b_0 + b_1 + b_2 + b_3 a + b_4 + b_5 a + b'z$	$d_2 = b_1 + b_4 + b_5 a$
Logit diff	$d_3 = b_2$	$d_4 = b_2 + b_4$	$d_5 = b_1 + b_2 + b_4 + b_5 a$

I valori delle differenze fra i logit e gli odds ratio relativi a donne con età pari a 30 anni sono:

Effetti	Tra	Logit difference	OR
LWD=1	SMOKE=0	$-1.73 + 0.147 \times 30$	14.76
LWD=1	SMOKE=1	$-1.73 - 1.407 + 0.147 \times 30$	3.57
SMOKE=1	LWD=0	1.153	3.17
SMOKE=1	LWD=1	$1.153 - 1.407$	0.76
SMOKE+LWD=1		$-1.73 + 1.153 - 1.407 + 0.147 \times 30$	11.31

Per stimare gli intervalli di confidenza degli odds ratio dobbiamo ottenere le stime delle varianze per ciascuna combinazione lineare corrispondente alle differenze di logit.

$$\text{var}(d_1) = \text{var}(b_1) + a^2 \text{var}(b_5) + 2\text{cov}(b_1, b_5)$$

$$\text{var}(d_2) = \text{var}(b_1) + \text{var}(b_4) + a^2 \text{var}(b_5) + 2\text{cov}(b_1, b_4) + 2\text{cov}(b_1, b_5) + 2\text{cov}(b_4, b_5)$$

$$\text{var}(d_3) = \text{var}(b_2)$$

$$\text{var}(d_4) = \text{var}(b_2) + \text{var}(b_4) + 2\text{cov}(b_2, b_4)$$

$$\begin{aligned} \text{var}(d_5) = & \text{var}(b_1) + \text{var}(b_2) + \text{var}(b_4) + a^2 \text{var}(b_5) + 2\text{cov}(b_1, b_2) + 2\text{cov}(b_1, b_4) + 2\text{cov}(b_1, b_5) \\ & + 2\text{cov}(b_2, b_4) + 2\text{cov}(b_2, b_5) + 2\text{cov}(b_4, b_5) \end{aligned}$$

Dalla matrice di varianza e covarianza, relativa al modello in esame si ha:

```
. matrix varcov =get(VCE)
. matrix list varcov
symmetric varcov[11,11]
      age      Irace_2      Irace_3      smoke      ht
age      .00207628
Irace_2  .00088137      .26927313
Irace_3  -.00001522      .10078868      .21532713
smoke    -.00086319      .0526233      .07747354      .21016563
ht       -.00129261      -.00879836      .01206483      -.00288281      .43754388
ui       .00192378      .01181975      -.00187962      .01859802      .03272116
lwd      .04468759      -.03320748      -.15059091      .00625562      -.00152518
ptd      -.00449246      -.00014699      -.00607853      -.0298648      .0169116
AxL      -.00202808      .00256235      .00442796      .00284273      .00060774
SxL      .00077598      .00277675      .05217671      -.1653673      .02791049
_cons    -.04532396      -.14102394      -.12909522      -.11480406      -.02495821

      ui      lwd      ptd      AxL      SxL      _cons
ui       .22990082
lwd      .02640881      3.4905674
ptd      -.03175704      -.06091446      .22220884
AxL      -.00079498      -.14709336      .00259167      .00686568
SxL      -.05615267      -.11990997      .01582916      -.01000375      .6702306
_cons    -.08546587      -.97782026      .07718943      .04134667      .04648168      1.182735
```

Ricordiamo l'espressione dell'I.C. dell'OR

$$I.C.(OR_i) = \exp(b_i \pm z_{1-\alpha/2} \sqrt{\text{var}(b_i)})$$

L'I.C. dell'OR relativo ad un soggetto di 30 anni LWD=1 SMOKE=0 vale:

$$95\%I.C. (OR_{LWD=1;SMOKE=0;AGE=30;}) = \exp(2.692 \pm 1.96 \sqrt{\text{var}(d_1)})$$

$$95\%I.C. (OR_{LWD=1;SMOKE=0;AGE=30;}) = \exp(2.692 \pm 1.96 \sqrt{3.490 + (30^{20} \cdot 0.00687) - (60 \cdot 0.147)})$$

$$95\% I.C. (OR) = 2.41 - 90.2$$

In modo analogo si possono calcolare gli I.C. dei restanti OR

Regressione logistica negli studi caso-controllo

La regressione logistica può essere utilizzata non solo negli studi prospettici, ma anche negli studi caso-controllo. L'interpretazione dei coefficienti (escluso b_0^*)¹⁰ rimane immutata.

Negli studi caso-controllo con appaiamento tuttavia il modello richiede degli aggiustamenti per consentire di calcolare il contributo alla verosimiglianza in ogni strato. Si utilizza perciò il modello logistico condizionato oppure, in mancanza di un software specifico adeguato (ossia di un software che preveda l'implementazione della regressione logistica condizionata), si può procedere nel seguente modo:

- 1) si definisce un nuovo data set di numerosità pari al numero dei casi-controlli;
- 2) si utilizzano come covariate le differenze fra le coppie;
- 3) si esclude dal modello l'intercetta;
- 4) si utilizza il consueto modello di logistica non condizionata

Nell'applicazione seguente, con riferimento al data set riportato da Hosmer & Lemeshow (Appendix 3 pag.262) si presentano i risultati dell'analisi logistica condizionata (clogit command STATA) e dell'analisi non condizionata (logistic procedure SAS) effettuata su un data set manipolato secondo le indicazioni di cui ai punti precedenti.

¹⁰ dette J_1 e J_2 rispettivamente le frazioni di campionamento dei casi e dei controlli si ha:

$$b_0^* = \ln \frac{J_1}{J_2} + b_0$$

APPLICAZIONE: esempio n.8

use cc11 [Vedi Appendix 3 pag.262 Hosmer & Lemeshow]

```
pair low age lwt race smoke ptd ht ui
  1   0  14  135  1     0   0   0   0
  1   1  14  101  3     1   1   0   0
.....
  56  0  34  170  1     0   1   0   0
  56  1  34  187  2     1   0   1   0
```

. clogit low smoke ht ui ptd lwt, strata(pair) [STATA]

```
Iteration 0: Log Likelihood = -38.816242
Iteration 1: Log Likelihood = -27.643888
Iteration 2: Log Likelihood = -26.347492
Iteration 3: Log Likelihood = -26.238038
Iteration 4: Log Likelihood = -26.236872
Iteration 5: Log Likelihood = -26.236872
```

```
Conditional logistic regression                                Number of obs =    112
                                                            chi2(5)          =    25.16
                                                            Prob > chi2      =    0.0001
Log Likelihood = -26.236872                                Pseudo R2       =    0.3241
```

low	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
smoke	1.479564	.5620191	2.633	0.008	.3780272	2.581102
ht	2.329361	1.002549	2.323	0.020	.3644009	4.294322
ui	1.344895	.693843	1.938	0.053	-.0150127	2.704802
ptd	1.670594	.7468062	2.237	0.025	.2068811	3.134308
lwt	-.0150834	.0081465	-1.852	0.064	-.0310503	.0008834

. xi:clogit low i.race smoke ht ui ptd lwt, strata(pair) [STATA]
i.race Irace_1-3 (naturally coded; Irace_1 omitted)

```
Iteration 0: Log Likelihood = -38.816242
Iteration 1: Log Likelihood = -27.340328
Iteration 2: Log Likelihood = -25.927822
Iteration 3: Log Likelihood = -25.796042
Iteration 4: Log Likelihood = -25.794271
Iteration 5: Log Likelihood = -25.794271
```

```
Conditional logistic regression                                Number of obs =    112
                                                            chi2(7)          =    26.04
                                                            Prob > chi2      =    0.0005
Log Likelihood = -25.794271                                Pseudo R2       =    0.3355
```

low	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
Irace_2	.5713643	.6896449	0.828	0.407	-.7803149	1.923044
Irace_3	-.0253148	.6992043	-0.036	0.971	-1.39573	1.345101
smoke	1.400656	.6278396	2.231	0.026	.1701132	2.631199
ht	2.361152	1.086128	2.174	0.030	.2323797	4.489924
ui	1.401929	.6961584	2.014	0.044	.0374836	2.766375
ptd	1.808009	.7886502	2.293	0.022	.2622829	3.353735
lwt	-.0183757	.0100806	-1.823	0.068	-.0381333	.0013819

SOLUZIONE CON LOGISTICA NON CONDIZIONATA [SAS]

```
filename pippo 'c:\wstata\hosmel.txt';
data a;
infile pippo;
input pair low0 age0 lwt0 race0 smoke0 ptd0 ht0 ui0 race01 race02
low1 agel lwt1 racel smokel ptd1 ht1 uil racel1 racel2;
data a;
set a;
/* costruzione delle variabili DIFFERENZE */
razza1=racel1-race01;
razza2=racel2-race02;
lwt=lwt1-lwt0;
smoke=smoke1-smoke0;
ptd=ptd1-ptd0;
ht=ht1-ht0;
ui=uil-ui0;
/* logistic models with noint option */
proc logistic;
model low1=smoke ptd ht ui lwt/noint;
proc logistic;
model low1=razza1 razza2 smoke ptd ht ui lwt/noint;
run;
```

Struttura del file hosmel.txt

```
1 0 14 135 1 0 0 0 0 0 0 1 14 101 3 1 1 0 0 0 1
.....
```

The LOGISTIC Procedure

Data Set: WORK.A
Response Variable: LOW0
Response Levels: 1
Number of Observations: 56
Link Function: Logit

Response Profile

Ordered Value	LOW0	Count
1	0	56

Model Fitting Information and Testing Global Null Hypothesis BETA=0

Criterion	Without Covariates	With Covariates	Chi-Square for Covariates
AIC	77.632	62.474	.
SC	77.632	72.601	.
-2 LOG L	77.632	52.474	25.159 with 5 DF (p=0.0001)
Score	.	.	19.785 with 5 DF (p=0.0014)

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
SMOKE	1	1.4796	0.5620	6.9305	0.0085	0.566221	4.391
PTD	1	1.6706	0.7468	5.0041	0.0253	0.509639	5.315
HT	1	2.3294	1.0025	5.3984	0.0202	0.539724	10.271
UI	1	1.3449	0.6938	3.7571	0.0526	0.385377	3.838
LWT	1	-0.0151	0.00815	3.4281	0.0641	-0.386762	0.985

NOTE: Since there is only one response level, measures of association between the observed and predicted values were not calculated.

The LOGISTIC Procedure

Data Set: WORK.A
 Response Variable: LOW0
 Response Levels: 1
 Number of Observations: 56
 Link Function: Logit

Response Profile

Ordered Value	LOW0	Count
1	0	56

Model Fitting Information and Testing Global Null Hypothesis BETA=0

Criterion	Without Covariates	With Covariates	Chi-Square for Covariates
AIC	77.632	65.589	.
SC	77.632	79.766	.
-2 LOG L Score	77.632	51.589	26.044 with 7 DF (p=0.0005)
	.	.	20.267 with 7 DF (p=0.0050)

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
RAZZA1	1	0.5714	0.6896	0.6864	0.4074	0.185061	1.771
RAZZA2	1	-0.0253	0.6992	0.0013	0.9711	-0.010808	0.975
SMOKE	1	1.4007	0.6278	4.9770	0.0257	0.536024	4.058
PTD	1	1.8080	0.7887	5.2557	0.0219	0.551559	6.098
HT	1	2.3612	1.0861	4.7259	0.0297	0.547090	10.603
UI	1	1.4019	0.6962	4.0554	0.0440	0.401720	4.063
LWT	1	-0.0184	0.0101	3.3229	0.0683	-0.471180	0.982

NOTE: Since there is only one response level, measures of association between the observed and predicted values were not calculated.

Valutazione dell'affidabilità del modello

Il test di Hosmer-Lemeshow¹¹

Il test Hosmer-Lemeshow può essere così schematizzato:

- 1) si suddividono i rischi predicted in percentili¹²;
- 2) per ogni decile si determinano i valori osservati e si calcolano i valori attesi¹³ della variabile risposta;
- 3) si determina la statistica C sotto definita

$$C = \sum \frac{(O-E)^2}{E}$$

La statistica C si distribuisce, asintoticamente, come una χ^2 con gradi di libertà $(n-2)$, con n pari al numero di percentili in cui si è categorizzata la risposta.

Se il valore di C^{14} supera il corrispondente valore di χ^2 il modello, con il livello di significatività specificato, non può essere rifiutato.

¹¹ Ricordiamo che il test di Hosmer & Lemeshow non può essere applicato nel caso di un modello di regressione logistica condizionata (Lee pag.311)

¹² In genere si suole suddividere la variabile risposta in dieci categorie (decili)

¹³ I valori attesi della variabile risposta si calcolano come somma dei valori predicted nell'ambito del percentile considerato

¹⁴ Valori elevati di C sono indici di elevati scostamenti tra i valori osservati e quelli attesi

APPLICAZIONE: esempio n.9

Applicazione del test di Hosmer & Lemeshow

Riportiamo un'applicazione del test relativa al file **lbw** con effetti principali age race smoke ht ui lwd ptd e interazioni AGExLWD e SMOKExLWD

```
xi: logistic low age i.race smoke ht ui lwd ptd axl sxl
i.race          Itrace_1-3      (naturally coded; Itrace_1 omitted)
```

```
Logit Estimates                                Number of obs =    189
                                                chi2(10)         =   42.66
                                                Prob > chi2      =   0.0000
Log Likelihood =  -96.00616                    Pseudo R2       =   0.1818
```

	low	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
age		.9194513	.041896	-1.843	0.065	.8408967	1.005344
Itrace_2		2.95383	1.532788	2.087	0.037	1.068277	8.167462
Itrace_3		2.137589	.9919132	1.637	0.102	.8608713	5.307749
smoke		3.168096	1.452377	2.515	0.012	1.289956	7.780755
ht		3.893141	2.5752	2.055	0.040	1.064768	14.2346
ui		2.071284	.9931385	1.519	0.129	.8092928	5.301191
lwd		.1772934	.3312383	-0.926	0.354	.0045539	6.902359
ptd		3.426633	1.615282	2.613	0.009	1.360252	8.632086
axl		1.15883	.09602	1.779	0.075	.9851216	1.36317
sxlwd		.2447849	.2003996	-1.719	0.086	.0491956	1.217988

```
. lpredict p
```

```
. centile p,c(10 20 30 40 50 60 70 80 90)
```

Variable	Obs	Percentile	Centile	-- Binom. Interp. -- [95% Conf. Interval]	
p	189	10	.0721003	.0498727	.0799334
		20	.0935177	.0799334	.1383023
		30	.1561116	.1092088	.1928434
		40	.206253	.1762366	.2616363
		50	.2780392	.2307341	.3159813
		60	.3313082	.2969172	.392711
		70	.4265189	.3592756	.4484282
		80	.493357	.4468605	.5600019
		90	.623149	.5597141	.702737

```
. lfit,group(10)
```

```
Logistic model for low, goodness-of-fit test  
(Table collapsed on quantiles of estimated probabilities)
```

```
number of observations =      189  
number of groups =      10  
Hosmer-Lemeshow chi2(8) =      4.39  
Prob > chi2 =      0.8201
```

```
. lfit, group(10) table
```

```
Logistic model for low, goodness-of-fit test  
(Table collapsed on quantiles of estimated probabilities)
```

_Group	_Prob	_Obs_1	_Exp_1	_Obs_0	_Exp_0	_Total
1	0.0721	0	0.9	19	18.1	19
2	0.0935	1	1.6	18	17.4	19
3	0.1561	4	2.4	15	16.6	19
4	0.2063	2	3.5	17	15.5	19
5	0.2780	6	5.0	14	15.0	20
6	0.3313	6	5.6	12	12.4	18
7	0.4265	6	7.2	13	11.8	19
8	0.4934	10	8.6	9	10.4	19
9	0.6231	10	10.6	9	8.4	19
10	0.9367	14	13.5	4	4.5	18

```
number of observations =      189  
number of groups =      10  
Hosmer-Lemeshow chi2(8) =      4.39  
Prob > chi2 =      0.8201
```

ALTRI TEST

Il test di Hosmer & Lemeshow presentato al paragrafo precedente ha lo scopo di valutare, nella sua globalità, l'affidabilità di un modello logistico.

Vi sono tuttavia altri test che consentono di valutare il contributo alla validità del modello fornito da un particolare set di covariate. In pratica si tratta di esaminare gli andamenti di alcune variabili in funzione dell'outcome predicted.

In seguito considereremo questi tre test:

- 1) ΔD_j versus p_j (p = predicted outcome) [fig.1]
- 2) ΔX^2_j versus p_j [fig.2]
- 3) $\Delta \beta_j$ versus p_j [fig.3]

ΔD_j misura la variazione della **devianza** conseguente alla eliminazione del j -esimo set di covariate

ΔX^2_j misura la variazione della statistica **chi-square di Pearson** conseguente alla eliminazione del j -esimo set di covariate

$\Delta \beta_j$ misura la differenza standardizzata del vettore dei **coefficienti stimati β** , conseguente alla eliminazione del j -esimo set di covariate

Virtualmente tutti i valori ΔX^2 e ΔD dovrebbero essere inferiori a 4¹⁵, valori superiori sono dovrebbero essere valutati con particolare cautela

Molto interessante risulta essere l'esame del grafico in cui i punti, con ordinata ΔX^2 , vengono tracciati con dei cerchi di diametro proporzionale al valore di $\Delta \beta$ [fig.4]. I punti con elevata ordinata e con rilevanti diametri sono espressione di punti che non 'fittano' in modo adeguato e contemporaneamente hanno un alto 'leverage'.

¹⁵ ΔX^2 e ΔD si distribuiscono approssimativamente come una χ^2 con un grado di libertà B . [$\chi^2_{0.95}(1)=3.84 \sim 4$]

APPLICAZIONE: esempio 10

In questa applicazione cercheremo di illustrare le analisi tendenti a valutare l'affidabilit  del modello logistico in corrispondenza di un determinato set di covariate.

Allo scopo utilizzeremo il solito data set lwb con i seguenti effetti principali e interazioni:

age i.race smoke ht lwd ptd agexlwd smokexlwd

```
use lbwf
(Hosmer & Lemeshow data)
```

```
. xi: logistic low age i.race smoke ht ui lwd ptd axl sxl
i.race          Itrace_1-3      (naturally coded; Itrace_1 omitted)
```

```
Logit Estimates                                     Number of obs =    189
                                                    chi2(10)         =   42.66
                                                    Prob > chi2      =  0.0000
Log Likelihood = -96.00616                        Pseudo R2        =  0.1818
```

	low	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
age		.9194513	.041896	-1.843	0.065	.8408967 1.005344
Itrace_2		2.95383	1.532788	2.087	0.037	1.068277 8.167462
Itrace_3		2.137589	.9919132	1.637	0.102	.8608713 5.307749
smoke		3.168096	1.452377	2.515	0.012	1.289956 7.780755
ht		3.893141	2.5752	2.055	0.040	1.064768 14.2346
ui		2.071284	.9931385	1.519	0.129	.8092928 5.301191
lwd		.1772934	.3312383	-0.926	0.354	.0045539 6.902359
ptd		3.426633	1.615282	2.613	0.009	1.360252 8.632086
axl		1.15883	.09602	1.779	0.075	.9851216 1.36317
sxl		.2447849	.2003996	-1.719	0.086	.0491956 1.217988

```
. lpredict p
. lpredict dd, ddeviance
. lpredict dx2, dx2
. lpredict dbeta, dbeta
. sort p
. list low p dd dx2 dbeta
```

	low	p	dd	dx2	dbeta
1.	0	.0135093	.0276855	.0139373	.0002473
2.	0	.0283335	.1189681	.0603472	.0020983
.....					
187.	1	.8736002	.2877671	.1540583	.0099768
188.	0	.9153448	5.305643	11.61685	.8640447
189.	1	.936707	.138007	.0713094	.0039467

```
.graph dd p [fig.1]
.graph dx2 p [fig.2]
.graph db p [fig.3]
.graph dx2 p [w=db],border ylab xlab t1(Symbol size proporzional to dBeta")
[fig.4]
```

Nelle pagine successive sono rappresentati i grafici sopra definiti realizzati con SIGMA-PLOT.

- [Plot of \$\Delta D\$ Versus \$\pi\$](#)
- [Plot of \$\Delta X^2\$ Versus \$\pi\$](#)
- [Plot of \$\Delta \beta\$ Versus \$\pi\$](#)
- [Plot of \$\Delta X^2\$ Versus \$\pi\$ with the Plotting Symbol Proportional in Size to \$\Delta \beta\$](#)

Dall'esame dei grafici possiamo trarre le seguenti considerazioni:

1) dalle figure 1-2 notiamo che il modello 'fitta' abbastanza bene; infatti quasi tutti i pattern di covariate considerati presentano valori di D e X^2 inferiori al valore critico 4. In particolare dalla figura 2 vengono evidenziati 5 pattern critici che sono poi facilmente individuabili sulla figura 1 insieme ad altri due pattern critici situati nella 'coppa' definita dalle due curve similkadratiche.

2) dalla figura 3 si individuano cinque set di covariate con ordinate particolarmente elevate. Il valore di P espressione sia di elevati valori di X^2 sia di elevato 'leverage'¹⁶

¹⁶ I valori di leverage, di X^2 , di P possono anche essere espressi in funzione della probabilit  stimata

prob. stimata	0.0-0.1	0.1-0.3	0.3-0.7	0.7-0.9	0.9-1
X^2	L/S	M	M/S	M	L/S
P	S	L	M	L	S
h (leverage)	S	L	M/S	L	S

L large
 S small
 M moderate

Selezione delle variabili

Una prima selezione naturale delle variabili può essere effettuata tramite l'analisi univariata, prendendo in considerazione tutte le variabili ritenute a priori biologicamente importanti e quelle che presentino un **p-value < 0.25**

Praticamente ci si può affidare a una delle procedure di selezione automatica implementate dai vari package statistici. Anche in questo caso, tuttavia, sarebbe opportuno adottare come criterio di ingresso e di rimozione dei p-value p_{ij} valori di quanto normalmente codificato.

Hosmer e Lemeshow consigliano di adottare una $p_{\text{ENTRY}}=0.15$ e una $p_{\text{REMOVE}}=0.20$

I PUNTI ESSENZIALI

Il modello lineare non può essere usato quando la variabile dipendente è di tipo dicotomico (pagg.1-3)

Le variabili categoriche devono sempre essere inserite o escluse in toto dal modello (pag.5)

Tra ODDS RATIO (OR) e coefficienti della logistica valgono le seguenti relazioni (pag.9)

$$\begin{aligned}OR_i &= e^{b_i} \\OR_{iL} &= e^{[b_i - 1.96SE(b_i)]} \\OR_{iH} &= e^{[b_i + 1.96SE(b_i)]}\end{aligned}$$

Il valore di OR sovrastima sempre il Rischio Relativo (RR) (pag.11)

Quando il legame tra logit e una covariata continua non è lineare è preferibile discretizzare quest'ultima (pag.14)

Negli studi caso-controllo deve usarsi il modello logistico condizionato (pag.36)

Nell'uso di procedure di selezione automatica scegliere valori adeguati di p-value (pag.46)

BIBLIOGRAFIA

- Hosmer D.W, Lemeshow S. *Applied Logistic Regression* Wiley
- AA.VV. *Reference Manual* Stata Press
- Pagano M. *Biostatistica* Gnocchi

Plot of ΔD Versus π

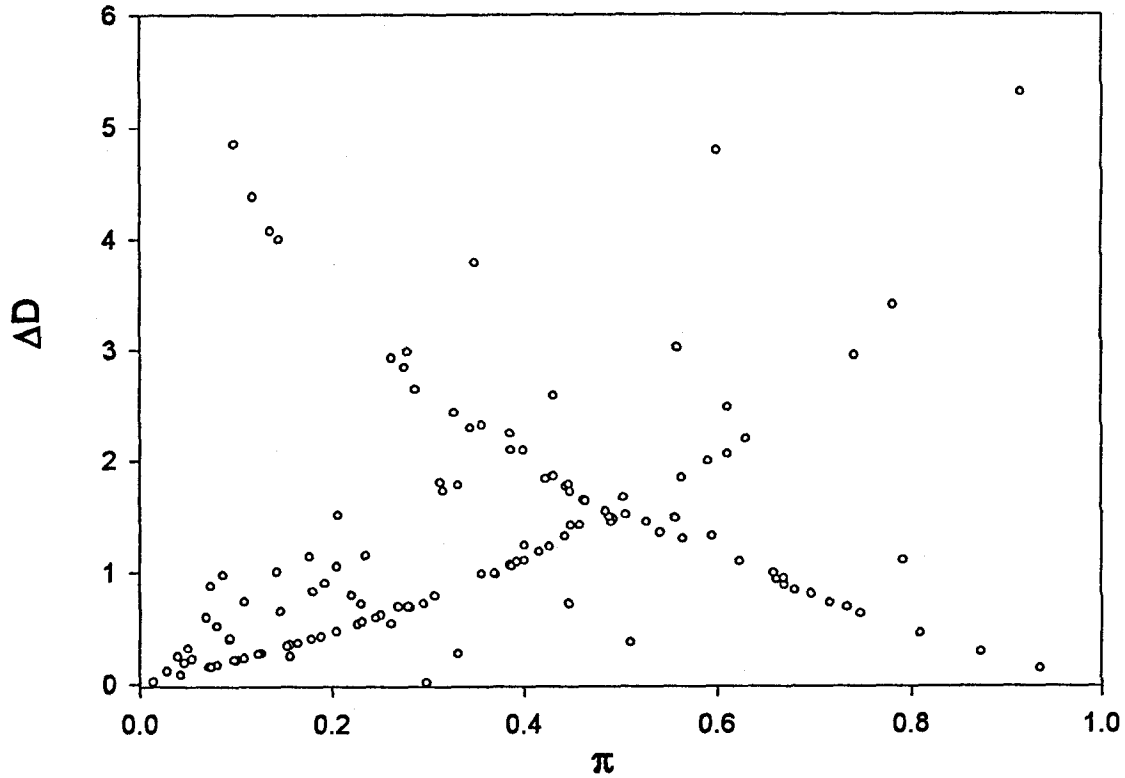


Fig. 1

Plot of ΔX^2 Versus π

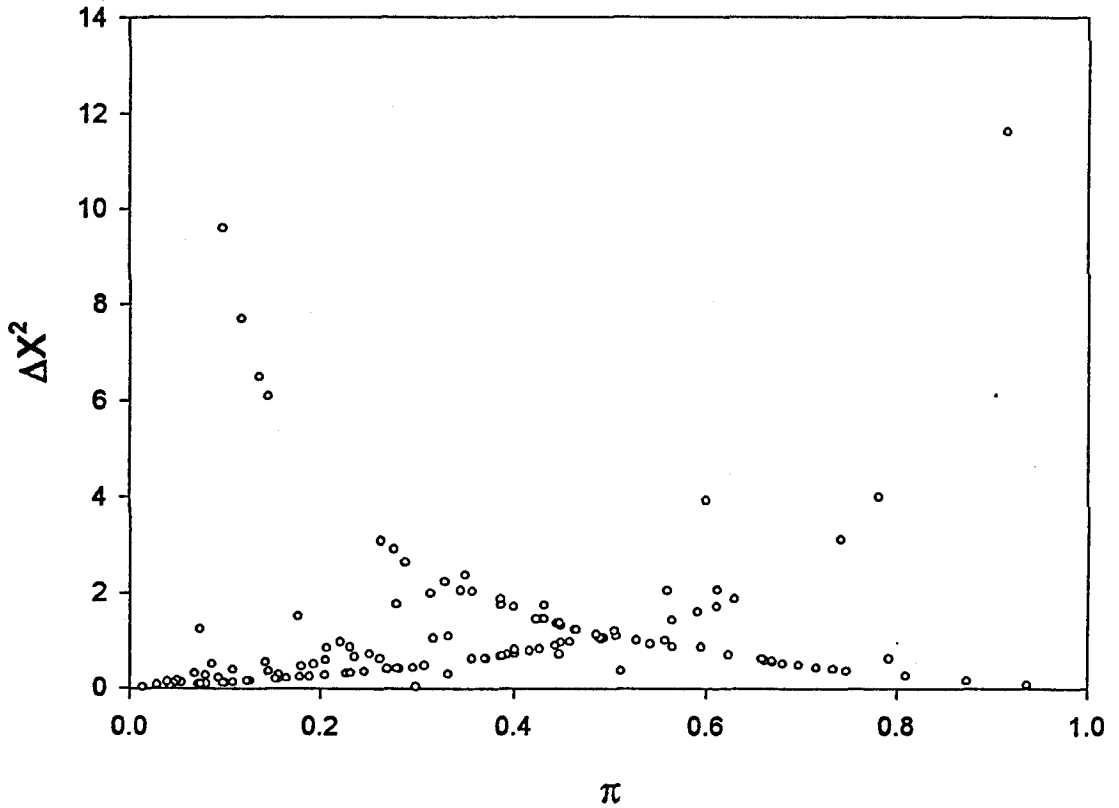


Fig. 2

Plot of $\Delta\beta$ Versus π

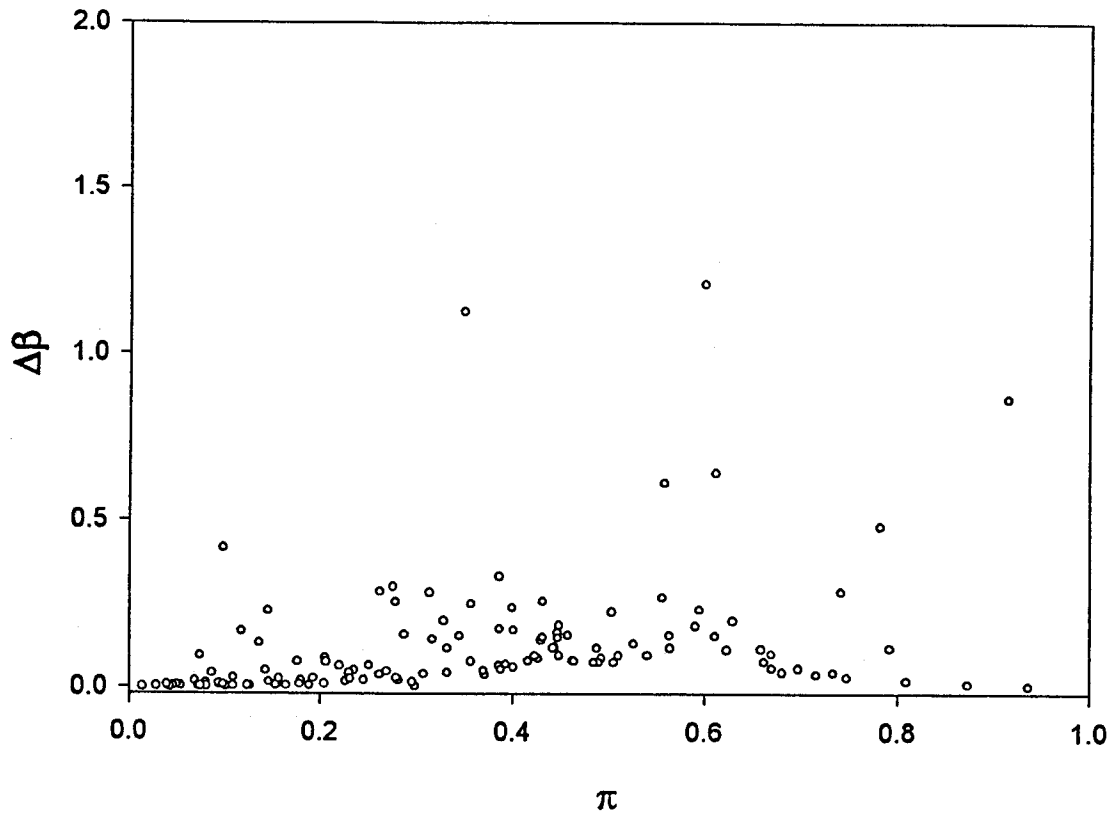


Fig. 3

Plot of ΔX^2 Versus π with
the Plotting Symbol Proportional in Size to $\Delta\beta$

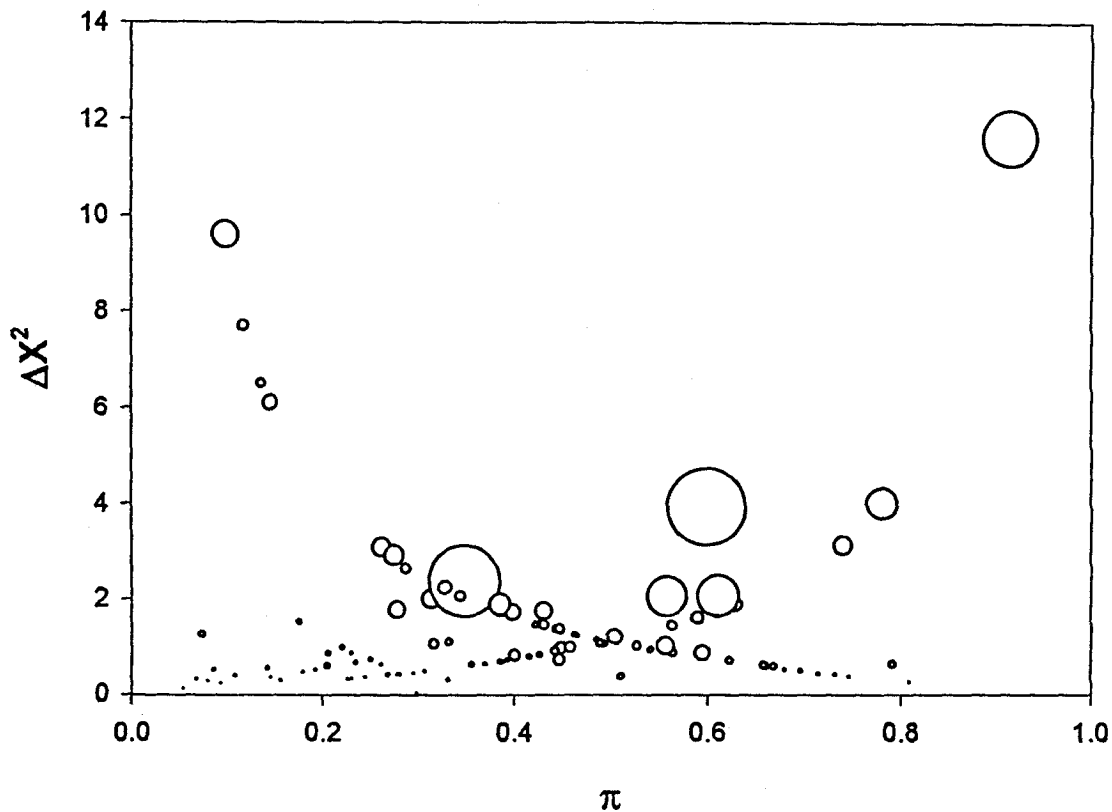


Fig. 4