Università degli Studi di Padova

Facoltà di Scienze Statistiche

CORSO DI LAUREA TRIENNALE IN SCIENZE STATISTICHE E GESTIONE DELLE IMPRESE

ANALISI DELLE TABELLE DI CONTINGENZA

Relatore

Ch.ma Prof.Laura Ventura

Laureanda

Debora Turolla Turatti

Matricola

571712-GEI

Anno Accademico 2010-2011

Quando le regole della matematica si riferiscono alla realtà non sono certe
e quando sono certe non si riferiscono alla realtà.

Albert Einstein

Indice

Introduzione

Questa tesi presenta una rassegna dei metodi più comunemente usati per analizzare un insieme di dati sotto forma di tabella di contingenza.

Nel primo capitolo si parlerà di indici, coefficienti e statistiche test per saggiare ipotesi di indipendenza e associazioni tra le variabili che compongono una tabella.

Nel secondo capitolo si tratteranno alcuni modelli lineari generalizzati, che sono adatti per l'analisi, la stima di tabelle e la valutazione della bontà. Nel terzo capitolo si approfondiranno due elementi di valutazione della bontà di adattamento di tali modelli ai dati, quali l'indice $pseudo - R^2$ e la curva ROC.

Nel quarto capitolo, infine, verranno riportati i codici R per eseguire quanto esposto nella tesi.

CAPITOLO 1

Le Tabelle di Contingenza

1.1 Introduzione

Questo capitolo si focalizza sulle tabelle di contingenza, le loro caratteristiche e proprietà e le varie tecniche inferenziali associate.
Nella prima parte del capitolo verranno analizzate le più semplici tabelle 2x2, presentati dei ben noti test, come il test chi-quadrato e il test di Fisher. Verranno, inoltre, presentati i concetti di rischio relativo e di quote.
Nella seconda parte del capitolo ci si soffermerà sulle tabelle $r \times c$, il test chi-quadrato per il trend e altre misure di associazione tra variabili.

1.2 Tabelle di Contingenza

Quando un campione, estratto da una popolazione, è classificato in base alle modalità di una o più variabili, qualitative o quantitative, i dati possono essere organizzati nella forma di tabelle di frequenza. Tali tabelle riportano le frequenze assolute di ciascuna classe. Un esempio di tabella di frequenza, o contingenza, è la Tabella 1.1, in cui viene riportata la popolazione residente in Italia nel 2001.

Tabella 1.1 Popolazione residente per sesso e classe d' età in Italia nel 2001

CLASSI DI ETÀ	Maschi	Femmine	Totale
Fino a 5 anni	1.344.296	1.274.498	2.618.794
5-9	1.375.399	1.303.705	2.679.104
10-14	1.440.659	1.364.628	2.805.287
15-19	1.517.900	1.445.729	2.963.629
20-24	1.739.347	1.685.003	3.424.350
25-29	2.138.204	2.108.572	4.246.776
30-34	2.283.606	2.260.176	4.543.782
35-39	2.313.969	2.309.619	4.623.588
40-44	2.024.945	2.040.634	4.065.579
45-49	1.850.242	1.889.328	3.739.570
50-54	1.895.424	1.954.267	3.849.691
55-59	1.620.147	1.704.626	3.324.773
60-64	1.657.480	1.807.467	3.464.947
65-69	1.426.778	1.653.170	3.079.948
70-74	1.229.113	1.574.399	2.803.512
75-79	913.342	1.373.434	2.286.776
80-84	445.332	789.985	1.235.317
85-89	267.981	573.970	841.951
90-94	88.270	240.947	329.217
95-99	13.468	49.372	62.840
100 e oltre	1.080	5.233	6.313
Totale	**27.586.982**	**29.408.762**	**56.995.744**

Fonte: Istat. 14º Censimento generale della popolazione e delle abitazioni.

È possibile classificare le unità di una popolazione secondo variabili dicotomiche (due categorie) o variabili con più di due categorie.

È importante che la classificazione sia **esaustiva** (ovvero che possieda sufficienti categorie per allocare ogni elemento della popolazione) e che le categorie siano **mutuamente esclusive** (ogni unità può essere collocata in una categoria solamente).

Quando la popolazione è così classificata nelle diverse categorie, è possibile contare il numero degli individui presenti in ognuna; questo numero, la **frequenza,** è il dato presente nelle tabelle di contingenza.

La forma generale di una tabella $r \times c$ è illustrata dalla Tabella 1.2,

Tabella 1.2: Tabella di contingenza $(\boldsymbol{r} \times \boldsymbol{c})$

X \ Y	y_1	y_2	·	·	y_j	·	·	y_c	TOT
x_1	n_{11}	n_{12}			·				$n_{1.}$
x_2					·				·
·					·				
·					·				·
x_i	·	·	·	·	n_{ij}				$n_{i.}$
·									·
·									·
x_r								n_{rc}	$n_{r.}$
TOT	$n_{.1}$	$n_{.2}$	·	·	$n_{.j}$	·	·	$n_{.c}$	n

Nella Tabella 1.2 il campione di n unità è classificato rispetto a due variabili, X e Y, in cui X presenta modalità $x_1, x_2, \dots, x_r$ e Y modalità $y_1, y_2, \dots, y_c$.

La frequenza osservata sull'i-esima riga e la j-esima colonna è rappresentata da $n_{ij}, i = 1, \dots, r, j = 1, \dots, c$. Il numero totale delle osservazioni della i-esima modalità di X è $n_{i.} = \sum_{j=1}^{c} n_{ij}$, mentre il numero totale delle osservazioni della j-esima modalità di Y è $n_{.j} = \sum_{i=1}^{r} n_{ij}$ e tali quantità sono note come **totali marginali**.

La numerosità campionaria è

$$n = \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij} = \sum_{j=1}^{c} n_{.j} = \sum_{i=1}^{r} n_{i.} \tag{1.1}$$

Una delle questioni più importanti che riguarda le relazioni tra variabili è l'indipendenza. Assumiamo che nella popolazione di interesse, la probabilità di estrarre un'osservazione appartenente alla i-esima categoria di X e alla j-esima

categoria di Y sia p_{ij}. Di conseguenza, la frequenza F_{ij} attesa nella cella ij risulta uguale a

$$F_{ij} = n \cdot p_{ij} , \qquad i = 1, \dots, r, \;\; j = 1, \dots, c. \tag{1.2}$$

Siano, inoltre, $p_{i.}$ la probabilità di estrarre un'osservazione appartenente alla i-esima categoria di X e $p_{.j}$ quella appartenente alla j-esima categoria di Y. Allora dalla legge di moltiplicazione delle probabilità, l' indipendenza tra le due variabili implica che:

$$p_{ij} = p_{i.} \cdot p_{.j} . \tag{1.3}$$

In termini di frequenza:

$$F_{ij} = n \cdot p_{i.} \cdot p_{.j} . \tag{1.4}$$

Le probabilità vengono stimate a partire dalle frequenze osservate, ossia

$$\hat{p}_{i.} = \frac{n_{i.}}{n} \qquad \text{e} \qquad \hat{p}_{ij} = \frac{n_{.j}}{n} \tag{1.5}$$

$$\hat{F}_{ij} = \; n \cdot \hat{p}_{i.} \cdot \hat{p}_{ij} = n \cdot \frac{n_{i.}}{n} \cdot \frac{n_{.j}}{n} = \frac{n_{i.} \cdot n_{.j}}{n} , \tag{1.6}$$

dove con $\hat{F}_{ij}$ si indicano le frequenze attese nell'ipotesi di indipendenza.

In caso di indipendenza, dunque, le frequenze stimate (1.6) e le frequenze osservate n_{ij} dovrebbero differire poco.

È, quindi, utile usare un test per l'indipendenza basato sulla differenza tra **frequenze attese** $\hat{F}_{ij}$ (numero di soggetti che ci aspettiamo di osservare in una determinata cella) e **frequenze osservate** n_{ij} (numero dei soggetti del campione che cadono nelle diverse categorie).

1.3 Il Test Chi-quadrato

Per testare l'indipendenza si verifica l'ipotesi nulla $H_0: p_{ij} = p_{i.} \cdot p_{.j}$. Pearson (1904) suggerì il ben noto test per l'indipendenza dato da

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(n_{ij} - \hat{F}_{ij}\right)^2}{\hat{F}_{ij}} . \qquad (1.7)$$

L'entità di questa statistica dipende dal valore delle differenze tra valori osservati e attesi: X^2 sarà "piccolo" quando H_0 è vera. Bisogna, però, decidere quali valori di X^2 portano ad accettare l'ipotesi nulla e quali al suo rifiuto.

Il criterio secondo il quale stabilire se queste distanze sono "grandi" o "piccole" viene fornito dalla distribuzione approssimata chi-quadrato. Infatti, il valore osservato X^2 viene confrontato con il valore tabulato di una variabile χ^2 con $k = (r-1)(c-1)$ gradi di libertà. La regola di decisione è: rifiuto H_0 se X^2 è maggiore o uguale a ${\chi^2}_{k;1-\alpha}$, dove ${\chi^2}_{k;1-\alpha}$ indica il quantile di livello $(1-\alpha)$ di un ${\chi^2}_k$ e α è il **livello di significatività** fissato.

Spesso nell'applicazione del test chi-quadrato le frequenze attese possono essere piccole, anche inferiori a 1.

Tuttavia, l'approssimazione di X^2 con un χ^2 non è valida quando qualcuna delle frequenze attese è piccola. Cochran (1954) suggerisce che il test chi-quadro non dovrebbe essere usato, quando $n < 20$ o quando $20 < n < 40$ e c'è almeno una frequenza attesa inferiore a 5. Quando $n = 40$, una sola cella con frequenza attesa non più piccola di uno può essere tollerata.

1.4 Tabelle di contingenza 2x2

Le tabelle 2x2 sono usate molto spesso in ambito sociale e nella ricerca medica. In esse un campione di n unità viene classificato secondo due variabili dicotomiche. Alcune volte il numero degli individui in ogni categoria di una variabile è predeterminato e per ognuno di questi campioni è valutato il numero di individui della seconda variabile. Ad esempio, in un'analisi su un effetto collaterale (si veda la Tabella 1.3), il farmaco viene somministrato a 50 persone e alle restanti 50 viene dato un placebo. La seconda variabile registra, quindi, quanti individui hanno sofferto di effetto collaterale e quanti no.

Tabella 1.3 Effetto collaterale di un farmaco

		effetto collaterale		
		presente	assente	TOT
trattamento	farmaco	15	35	50
	placebo	4	46	50
	TOT	19	81	100

La formula per il calcolo della statistica X^2 è la (1.7) che nel caso 2x2 si semplifica in

$$X^2 = \frac{n \cdot (ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}, \tag{1.8}$$

con a, b, c e d definite come nella Tabella 1.4.

Tabella 1.4 Tabella di contingenza generale 2x2

		Y		
		Y_1	Y_2	TOT
	X_1	a	b	a+b
X				
	X_2	c	d	c+d
	TOT	a+c	b+d	n=a+b+c+d

Per la significatività del test si fa riferimento alla distribuzione ${\chi^2}_1$ e sono significativi contro l'ipotesi nulla valori grandi della statistica.

1.4.1 *Correzione di continuità di Yates*

Le frequenze osservate in una tabella di contingenza sono valori discreti e formano, quindi, una statistica X^2 discreta che viene però approssimata con la distribuzione continua χ^2.

Nel 1934 Yates suggerì una correzione per migliorare l'approssimazione per le tabelle 2x2, che consiste nel ridurre di ½ il valore assoluto della differenza tra

frequenze attese e frequenze osservate. La formula per il test chi-quadrato corretto per tabelle 2x2 diventa

$$X^2_{corr} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(|n_{ij} - \hat{F}_{ij}| - 0.5\right)^2}{\hat{F}_{ij}} = \frac{n \cdot (|ad - bc| - 0.5n)^2}{(a+b)(c+d)(a+c)(b+d)} \qquad (1.9)$$

Nelle regioni critiche tra 0.1 e 0.01 la correzione di Yates dà una buona approssimazione.

Sebbene questa correzione fu molto usata in passato, Conover (1968, 1974) mosse delle critiche e "alcuni autori ne sconsigliano l'uso" (Wayne, 2007). Inoltre, tale correzione è destinata a scomparire, dal momento in cui esiste un test esatto nelle tabelle 2x2 semplice da eseguire.

1.4.2 *Test esatto di Fisher*

A volte i dati possono provenire da campioni molto piccoli. Il test chi-quadrato non è un metodo idoneo di analisi se non è presente una dimensione minima per le frequenze attese. Un test che può essere usato, quando non ci sono le dimensioni minime è il test esatto di Fisher. È chiamato esatto poiché permette di calcolare la probabilità esatta di ottenere o il risultato osservato o i risultati maggiori o minori. È possibile far uso di questo test solo quando entrambi i totali marginali sono fissati. Ma, essendo una situazione poco frequente, viene usato anche in assenza di questa condizione.

Per usare questo test bisogna disporre i dati sotto forma di tabella di contingenza 2x2 come nella Tabella 1.4. L'ipotesi nulla è che la proporzione della caratteristica di interesse è uguale nelle due popolazioni, ovvero $p_1 = p_2$. La statistica test è b, ossia il numero di soggetti nel secondo campione con la caratteristica. Finney e Latscha (1963) hanno calcolato i valori critici di b per $(a + b) \leq 20$.

Per accettare o rifiutare H_0 bisogna trovare il valore critico di b cercando tra i valori di $(a+b)$, $(c+d)$ e a nell'opportuna tavola (Tavola j, Wayne, *Biostatistica*). Se il valore osservato di b è minore o uguale al numero intero di una data colonna, si rifiuta H_0 a un livello α pari al doppio di quello riportato sulla colonna.

1.5 Rischio relativo e Odds ratio

Nei paragrafi precedenti le statistiche test si intendono per dati provenienti da studi pianificati, nei quali almeno una variabile veniva manipolata. In campo medico-sanitario, però, i dati possono provenire da studi eseguiti sul campo.
Uno studio osservazionale è un'indagine nella quale sia i soggetti che le variabili di interesse non sono manipolate in nessun modo; in altre parole, è un'indagine non sperimentale. Il caso più semplice si verifica quando le variabili sono due, una indipendente chiamata *fattore di rischio* ed una dipendente chiamata *risposta*.
Quando queste variabili sono categoriche, i dati possono essere messi sotto forma di tabella (si veda la Tabella 1.4 con X=fattore di rischio e Y=malattia).
Il fattore di rischio X è, quindi, una variabile legata alla variabile risposta Y, in quanto causa sospetta di una particolare condizione della variabile risposta.
Ci sono due tipi di studi osservazionale:
-*Studio prospettico*. Vengono studiati due campioni, uno che presenta il fattore di rischio e uno che non lo presenta. I soggetti vengono seguiti nel tempo e registrati, dopo un certo periodo, nelle categorie della variabile risposta.
-*Studio retrospettivo*. I soggetti sono già registrati secondo la variabile risposta e il ricercatore, guardando indietro nel tempo, deve determinare quali dei soggetti avevano manifestato il fattore di rischio e quali no.

1.5.1 *Rischio Relativo*

A partire dai dati che risultano da uno studio prospettico, è possibile individuare il rischio di contrarre una malattia tra i soggetti che hanno manifestato il fattore e il rischio di contrarre la stessa malattia tra i soggetti che non hanno manifestato il fattore. Questi valgono rispettivamente $a/(a+b)$ e $c/(c+d)$, dove *a,b,c* e *d* sono definiti come nella Tabella 1.4.
Il rischio relativo del campione è definito come il rapporto tra le due quantità, ossia

$$\widehat{RR} = \frac{a/(a+b)}{c/(c+d)}. \tag{1.10}$$

Il rischio relativo $\widehat{RR}$ può essere usato come stima del rischio relativo *RR* della popolazione. Varia da zero a infinito; un valore paria a 1 significa che non c'è differenza tra il rischio di contrarre la malattia per i soggetti che presentano il fattore e quelli che non presentano il fattore. Un valore di $\widehat{RR}$>1 significa che sono più a rischio di contrarre la malattia i soggetti che presentano il fattore, mentre se $\widehat{RR}$<1 il rischio è più altro per coloro che non presentano il fattore.
È possibile costruire un intervallo di confidenza nel seguente modo

$$\widehat{RR}^{1\pm\left(z_{1-\alpha/2}/\sqrt{X^2}\right)}, \tag{1.11}$$

dove $z_{1-\alpha/2}$ è il quantile di livello $1-\alpha/2$ della distribuzione normale standard e X^2 è calcolato con la (1.8).
Se la malattia è relativamente rara nella popolazione (diciamo sotto il 5%), il Rischio Relativo può essere approssimato con l'odds ratio.

1.5.2 *Il rapporto di Odds o Odds Ratio*

Se i dati provengono da uno studio retrospettivo, dal campione dei soggetti con la malattia (casi) e dal campione di soggetti senza la malattia (controlli) è possibile determinare retrospettivamente la distribuzione del fattore rischio tra questi. I dati vengono inseriti in una tabella come la Tabella 1.4 con X=fattore di rischio e Y=gruppo.

In questo caso la misura più adeguata per confrontare i casi e i controlli è il *rapporto di odds (odds ratio)*. Se la probabilità di un certo evento è p, si definisce *odds* di quell'evento la quantità $o = p/(1-p)$, ovvero il rapporto tra la probabilità di successo e di insuccesso. L'utilizzo dell' *odds* è vantaggioso in alcuni tipi di analisi poiché varia da 0 a infinito, invece che da 0 a 1.

Spesso si lavora con il logaritmo dell'*odds*, detto *log odds* o *logit*

$$\log(o) = \log\left(\frac{p}{1-p}\right) \quad . \tag{1.12}$$

Questa quantità può variare da $-\infty$ a $+\infty$ e risulta utile nei modelli di regressione logistica che si tratteranno in seguito.

Facendo riferimento ad uno studio retrospettivo, l'*odds* di essere un caso (avere la malattia) rispetto all'essere un controllo (non avere la malattia) tra i soggetti con il fattore di rischio è:

$$o = {[a/(a+b)]}\Big/{[b/(a+b)]} = \boldsymbol{a}/\boldsymbol{b} \ . \tag{1.13}$$

L'odds di essere un caso rispetto all'essere un controllo tra i soggetti senza il fattore di rischio è:

$$o = {[c/(c+d)]}\Big/{[d/(c+d)]} = \boldsymbol{c}/\boldsymbol{d} \ . \tag{1.14}$$

A partire da queste quantità si definisce la stima dell' *odds ratio* $\widehat{OR}$ come:

$$\widehat{OR} = \frac{a/b}{c/d} = \frac{ad}{bc} \ . \tag{1.15}$$

L'*odds ratio* $\widehat{OR}$ calcolato sul campione può essere usato come stima dell'*odds ratio OR* della popolazione. Il rapporto di *odds* può assumere valori tra zero ed infinito. Un valore pari a 1 indica che non c'è associazione tra il fattore di rischio e la condizione della malattia. Un valore minore di 1 indica che ci sono quote ridotte di soggetti malati tra quelli con il fattore di rischio, mentre un valore maggiore di 1 indica che ci sono quote superiori di soggetti malati tra quelli con il fattore di rischio.

Si può stimare l'errore standard e l'intervallo di confidenza sfruttando il logaritmo dell'*odds ratio*. L'errore standard del logaritmo dell'*odds ratio* è:

$$SE\left(\log \widehat{OR}\right) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} \quad . \tag{1.16}$$

Se il campione è sufficientemente numeroso, si può costruire un intervallo di confidenza come

$$\log \widehat{OR} \pm z_{1-\alpha/2} \cdot SE\left(\log \widehat{OR}\right) . \tag{1.17}$$

1.5.3 La statistica di Mantel-Haenszel

Negli studi sopra descritti, la variabile malattia e la variabile fattore di rischio possono essere associate con un'altra variabile chiamata *variabile di confounding* (ovvero una variabile non considerata nello studio, che influenza le variabili che sono invece considerate nello studio) che, se ignorata, può falsare la relazione tra le prime due.

Per evitare che questo avvenga, Mantel e Haenszel (1959) proposero una tecnica: i soggetti, caso o controllo, vengono assegnati agli strati corrispondenti ai differenti valori della variabile di *confounding*. I dati sono poi analizzati all'interno dei singoli strati prima, e tra gli strati poi.

L'applicazione della procedura consiste nei seguenti passi:

1. Si formano *k* strati corrispondenti alle *k* categorie della variabile di *confounding* (vedi Tabella 1.5).

Tabella 1.5 Tabella dell'i-esimo $(i = 1, \dots, k)$ strato di una variabile di *confounding*

		campione		
		casi	controlli	TOT
fattore di rischio	presente	a_i	b_i	a_i+b_i
	assente	c_i	d_i	c_i+d_i
	TOT	a_i+c_i	b_i+d_i	n_i

2. Per ogni strato si calcola la frequenza attesa e_i relativa alla cella a sinistra della prima riga:

$$e_i = \frac{(a_i+b_i)(a_i+c_i)}{n_i} \quad . \tag{1.18}$$

3. Per ogni strato si calcola la quantità v_i:

$$v_i = \frac{(a_i+b_i)(c_i+d_i)(a_i+c_i)(b_i+d_i)}{n_i^2(n_i-1)} \quad . \tag{1.19}$$

4. Si calcola, quindi, la statistica test

$$X^2_{MH} = \frac{\left(\sum_{i=1}^k a_i - \sum_{i=1}^k e_i\right)^2}{\sum_{i=1}^k v_i} \quad . \tag{1.20}$$

5. Si rifiuta l'ipotesi nulla di associazione tra la malattia e il fattore di rischio se $X^2_{MH} \geq \chi^2_{1,1-\alpha}$.

È possibile, inoltre, stimare l'*odds ratio* comune a tutte le tabelle di contingenza con lo stimatore di Mantel-Haenszel, dato da

$$\widehat{OR}_{MH} = \frac{\sum_{i=1}^k (a_i d_i/n_i)}{\sum_{i=1}^k (b_i c_i/n_i)} \quad . \tag{1.21}$$

1.6 Tabelle di contingenza *r*x*c*

L'analisi delle tabelle di contingenza *r*x*c,* quando *r* o *c* o entrambe sono maggiori di due presenta alcune problematiche rispetto al caso più semplice 2x2. Formalmente il test χ^2 segue quello applicato nel caso più semplice. Per ogni cella si calcola la frequenza attesa $\hat{F}_{ij}$ con la formula (1.6) e si ottiene

l'indice X^2 dalla formula (1.7) sommando sulle rc celle. Sotto l'ipotesi nulla, la statistica X^2 segue la distribuzione χ^2_k, con $k = (r-1)(c-1)$ gradi di libertà. Se accade che $\hat{F}_{ij} < 5$, può essere opportuno accorpare le righe e/o le colonne prima di eseguire il test chi-quadrato.

1.6.1 Test chi-quadro per il trend

Consideriamo, ad esempio, la Tabella 1.6. Si può calcolare il test chi-quadro con la formula (1.7) per verificare l'ipotesi nulla di assenza di relazione tra le due variabili implicate.

Tabella 1.6 Episodi di tosse e fumo di sigaretta tra ragazzi di sesso maschile all'età di 12 anni (Bland, 1978)

	Fumo tra i ragazzi						Totale
	Non Fumatori		Occasionali		regolari		
Tosse	266	20,40%	395	28,80%	80	46,50%	714
No tosse	1037	79,60%	977	71,20%	92	53,50%	2106
Totale	1303	100%	1372	100%	172	100%	2847

La statistica X^2 vale 64.25, con 2 gradi di libertà (p-value <0.001). Dunque i dati non supportano l'ipotesi nulla. Avremmo ottenuto lo stesso risultato per qualsiasi ordine delle colonne, questo perché il test ignora il loro ordinamento. Tuttavia, ci si può aspettare che, se esiste una relazione tra le variabili, la prevalenza di episodi di tosse sia maggiore dove maggiore è il consumo di sigarette, ovvero che ci sia un trend nella prevalenza degli episodi di tosse da un capo all'altro della tabella. Un test che si usa in questi casi è il *Test chi-quadro per il trend*.

Innanzitutto, bisogna definire due variabili aleatorie, X e Y, i cui valori dipendono dalle categorie scelte per le variabili di riga e colonna. Ad esempio:

X =1 : non fumatori

X =2 : occasionali

X =3 : assidui

Y =1 : tosse

Y =2 : no tosse

Le categorie devono essere ordinate. Se ci sono n individui, si avranno n coppie di osservazioni (x_i, y_j). Se all'interno della tabella è presente un trend lineare, allora la pendenza della retta della regressione lineare di X su Y sarà diversa da 0. Sfruttando il modello di regressione lineare $Y = \alpha + \beta X$ e stimando α e β con il metodo dei minimi quadrati, si ottiene:

$$\hat{\beta} = \frac{\sum(y_i - \bar{y})(x_i - \bar{x})}{\sum(x_i - \bar{x})} \tag{1.22}$$

con standard error pari a

$$\widehat{SE}(\hat{\beta}) = \sqrt{\frac{s^2}{\sum(x_i - \bar{x})^2}}\ , \tag{1.23}$$

dove s^2 è la varianza di Y, data da

$$s^2 = \frac{1}{n}\sum(y_i - \bar{y})^2 \tag{1.24}$$

usando n al posto di $n - 1$ perché il test è condizionato ai totali marginali di riga e colonna.

Per n grande, $\widehat{SE}(\hat{\beta})$ è una buona stima della deviazione standard di $\hat{\beta}$. Quindi, se l'ipotesi nulla è vera, $\hat{\beta}/\widehat{SE}(\hat{\beta})$ è un'osservazione da una distribuzione Normale Standard e $\hat{\beta}^2/\widehat{SE}(\hat{\beta})^2$ ha distribuzione chi-quadro con un grado di libertà. A partire dalle formule (1.22), (1.23) e (1.24), $\hat{\beta}^2/\widehat{SE}(\hat{\beta})^2$ risulta essere:

$$\frac{\hat{\beta}^2}{\widehat{SE}(\hat{\beta})^2} = \left(\frac{\sum(y_i-\bar{y})(x_i-\bar{x})}{\sum(x_i-\bar{x})}\right)^2 \Big/ \frac{\sum(y_i-\bar{y})^2}{n\sum(x_i-\bar{x})^2} = \frac{n(\sum(y_i-\bar{y})(x_i-\bar{x}))^2}{\sum(x_i-\bar{x})^2\sum(y_i-\bar{y})^2} . \qquad (1.25)$$

Per facilitare il calcolo si può usare la forma alternativa della somma dei quadrati:

$$\chi_1^2 = n\left(\sum y_i x_i - \frac{(\sum y_i)(\sum x_i)}{n}\right)^2 \Big/ \left(\sum x_i^2 - \frac{\left(\sum x_i^2\right)^2}{n}\right)\left(\sum y_i^2 - \frac{(\sum y_i)^2}{n}\right). \quad (1.26)$$

Se l'ipotesi nulla è vera, χ_1^2 è un'osservazione da una distribuzione chi-quadro con 1 grado di libertà.

Vi sono alcune osservazioni da fare riguardo a questo test:

- Il trend può risultare significativo anche se il test chi-quadro per l'indipendenza non lo è; questo perché il test per il trend ha maggiore potenza nell'individuare la presenza di andamenti.
- Se l'ipotesi che si vuole verificare contempla l'ordine delle categorie è meglio usare il test per il trend; in caso contrario si usa il test chi-quadro per l'indipendenza.
- Se il numero delle osservazione è maggiore di 30, allora il test può essere ritenuto valido.

1.6.2 Misure di associazione per una tabelle di contingenza

Molte misure di associazione per tabelle di contingenza sono basate sulla statistica X^2 la quale, però, dipende dalla grandezza del campione e, quindi, aumenta con l'aumentare di n. La via più semplice per ovviare a questo problema è dividere il valore di X^2 per la numerosità campionaria ottenendo così un indice conosciuto come *Indice di contingenza quadratica media di Pearson:*

$$\phi^2 = \frac{X^2}{n} \quad . \tag{1.27}$$

Nel caso di una tabella 2x2 l'indice ϕ^2 varia da 0 (indipendenza) a 1 (associazione). Nel caso, però, di una tabella $r \times c$, l'indice non ha un limite superiore. A causa di ciò Pearson (1904) propone un altro coefficiente chiamato *Coefficiente di contingenza* e definito come

$$P = \sqrt{\frac{\frac{X^2}{n}}{1+\frac{X^2}{n}}} \quad . \tag{1.28}$$

A differenza del precedente indice, P varia tra 0 e 1. In generale, però, non raggiunge mai il suo limite superiore. Kendall e Stuart (1980) mostrano che, anche in caso di completa associazione, il valore di P dipende dal numero di righe e di colonne della tabella. L'indice da loro proposto è:

$$T = \frac{\frac{X^2}{n}}{\sqrt{[(r-1)(c-1)]}} \quad . \tag{1.29}$$

Ancora una volta il valore 0 indica la completa indipendenza, mentre il valore 1 indica completa associazione quando $r = c$, ma non è conclusivo se r è diverso da c.

Un'ulteriore modifica proposta da Cramer (1928) permette di dare un risultato per ogni valore di r e c:

$$V = \sqrt{\frac{\frac{X^2}{n}}{\min(r-1, c-1)}} \quad . \tag{1.30}$$

Il maggiore problema che accomuna tutti questi indici è che non possiedono un'interpretazione probabilistica.

I risultati ottenuti con i metodi precedenti, tutti fondati sula statistica X^2, restano difficili da interpretare, anche dopo le trasformazioni proposte, cioè mediante indici che tengono in considerazione della numerosità del campione e delle dimensioni della tabella. In particolare, quando i valori sono distanti da zero, quindi non si ha indipendenza tra le due variabili, non è chiaro il tipo di associazione. Per renderlo più evidente, Goodman e Kruskal (1954) hanno introdotto il concetto di *Riduzione Proporzionale nell'Errore*, abbreviato in *PRE*, nel predire Y conoscendo X rispetto all'errore che si farebbe se non si conoscesse il valore di X. Questa misura consente, a partire da X, di prevedere il valore assunto da Y con una percentuale di errore inferiore a quella che si otterrebbe non conoscendo X.

Sono, però, misure di associazione asimmetriche perché assumono valore differente a seconda di quale delle due variabili viene considerata indipendente: proprio per questo motivo vengono chiamate misure di associazione asimmetriche.

Il coefficiente di incertezza è una di queste misure. Si assuma che una variabile Y dipenda da una variabile X e che la loro dipendenza sia misurabile con l'entropia (una"versione nominale" della varianza). I dati vengono estratti da una tabella di contingenza *rxc* con frequenze relative (p_{ij}) nelle celle.

Si ha

$$H_{y.x} = \frac{H_y + H_x - H_{xy}}{H_y}\,, \tag{1.31}$$

dove

$$H_y = -\sum_{i=1}^{r} p_i \ln p_i \quad , \tag{1.32}$$

$$H_x = -\sum_{j=1}^{c} p_j \ln p_j \quad , \tag{1.33}$$

$$H_{xy} = -\sum_{i=1}^{r}\sum_{j=1}^{c} p_{ij} \ln p_{ij} \quad . \tag{1.34}$$

Goodman e Kruskal (1972) forniscono, inoltre, una formula per la varianza di $H_{y.x}$ per verificarne la significatività, data da

$$Var\left(\widehat{H}_{y.x}\right) = -\frac{1}{nH_y^4}\sum_{i=1}^{r}\sum_{j=1}^{c}\left\{H_y \ln\frac{p_{ij}}{p_{i.}} + \left(H_x - H_{xy}\right)\ln p_{.j}\right\} . \tag{1.35}$$

Il coefficiente $\widehat{H}_{y.x}$ misura la riduzione di errore di previsione di Y conoscendo X e varia tra 0 e 1.

L'altra misura è il λ di Goodman e Kruskal (1954) e valuta la riduzione proporzionale nell'errore, sulla base della relazione

$$\lambda_y = \frac{\sum_{i=1}^{r} max_j(n_{ij}) - max_j(n_{.j})}{n - max_j(n_{.j})}\,, \tag{1.36}$$

dove $max_j(n_{ij})$ è la frequenza maggiore in ogni riga e $max_j(n_{.j})$ è il totale per colonna maggiore. I dati provengono dalla Tabella 1.2.

L'indice λ_y stima la diminuzione relativa della probabilità d'errore nell'indovinare la categoria *Y*, utilizzando anche la classificazione di X.

Il valore di λ_y varia sempre tra 0 e 1. Il valore 0, che si ottiene quando le frequenze entro ogni cella sono distribuite casualmente, indica che la variabile indipendente non aggiunge informazioni nella previsione della variabile dipendente e che pertanto non può essere utile nella sua classificazione. Un valore uguale a 1 indica che esiste corrispondenza perfetta e, quindi, che la variabile dipendente è classificata correttamente anche dalla variabile indipendente.

Non esiste, però, corrispondenza biunivoca tra il valore 0 di λ_y e l'associazione tra le due variabili: quando le due variabili sono indipendenti λ_y è uguale a 0; ma quando λ_y risulta uguale a 0, non sempre si ha indipendenza. L'indice λ_y deve essere usato solo in condizioni particolari di analisi dell'associazione, ovvero quando i valori di una variabile qualitativa sono utilizzati per prevedere quelli dell'altra variabile. Come è possibile usare X per prevedere Y, nello stesso modo è possibile utilizzare la variabile Y per prevedere X. E', quindi, possibile calcolare un altro valore di λ_y scambiando le righe con le colonne, cioè il previsore con la variabile predetta. Salvo casi fortuiti, di norma, i diversi approcci danno risultati differenti. L'indice λ_y presentato è asimmetrico: è quindi importante scegliere la variabile dipendente adatta.

In vari casi non è possibile, o semplice, distinguere tra variabile dipendente ed indipendente. Viene quindi utilizzato un indice simmetrico, in cui le variabili di riga e di colonna hanno le stesse frequenze.

Innanzitutto si calcola λ_y con la formula (1.41) e λ_x con la seguente formula:

$$\lambda_x = \frac{\sum_{j=1}^{c} max_i(n_{ij}) - max_i(n_{i.})}{n - max_i(n_{i.})} \ . \qquad (1.37)$$

Da λ_y e λ_x si ricava, quindi, la formula di λ simmetrico come

$$\lambda = \frac{\sum_{i=1}^{r} max_j(n_{ij}) - max_j(n_{.j}) + \sum_{j=1}^{c} max_i(n_{ij}) - max_i(n_{i.})}{2n - max_j(n_{.j}) - max_i(n_{i.})} \ . \qquad (1.38)$$

CAPITOLO 2

Modelli Lineari Generalizzati per tabelle di contingenza

2.1 Introduzione

Il capitolo precedente trattava le tecniche per testare le ipotesi di indipendenza e associazione tra le variabili. In questo capitolo verrà esposto un diverso approccio: la creazione del modello e la stima dei parametri.

I modelli più comuni sono quelli lineari nei quali i valori attesi delle osservazioni sono dati da una combinazione lineare dei parametri. Per la stima di questi si utilizza la tecnica della massima verosimiglianza.

Il più grande vantaggio nell'usare tecniche di stima dei modelli è quello di riuscire a individuare effetti tra le variabili di tabelle di contingenza anche molto complesse.

2.2 Modelli log-lineari

Per modellare dati di conteggio, o frequenze, spesso si fa uso della distribuzione di Poisson (Dobson, 2008).

Se *Y* è il numero di eventi, la sua distribuzione di probabilità è

$$f(y) = \frac{\mu^y e^{-\mu}}{y!} \ , \qquad (2.1)$$

dove μ è il numero medio di eventi, $E(y) = \mu$ e $Var(y) = \mu$.

Il parametro μ è, spesso, un tasso. Ad esempio: il "numero medio di clienti che comprano un determinato oggetto su 100 clienti che entrano nel negozio" o il"numero medio di incidenti su 1000 patenti". I numeri 100 e 1000 sono l'esposizione all'evento. L'effetto delle variabili esplicative sulla risposta *Y* è modellato attraverso il parametro μ.

Siano Y_i , $i = 1, \dots, n$, variabili casuali indipendenti e sia Y_i è il numero di eventi osservati su un'esposizione n_i . Il valore atteso di Y_i è

$$E(Y_i) = \mu_i = n_i\theta_i \qquad i = 1, \dots, n. \tag{2.2}$$

Se suppongo, ad esempio, Y_i= "numero di denuncie assicurative per un particolare modello di moto", questo numero dipenderà dal numero n_i delle macchine di questo tipo assicurate e da altre variabili che influiscono su θ_i (età della macchina, revisione, ecc…).

Nei modelli log-lineari la relazione tra θ_i e le variabili esplicative è

$$\theta_i = e^{x_i^T\beta} \quad i = 1, \dots, n. \tag{2.3}$$

Di conseguenza il modello log-lineare diventa:

$$E(Y_i) = \mu_i = n_i e^{x_i^T\beta}, \;\; Y_i \sim Poisson(\mu_i), \quad i = 1, \dots, n. \tag{2.4}$$

Il legame canonico è la funzione logaritmo

$$\log \mu_i = \log n_i + x_i^T\beta \quad i = 1, \dots, n. \tag{2.5}$$

Il termine $\log n_i$ prende il nome di *offset*; è una costante nota incorporata nella procedura di stima del modello, x_i^T è l'i-esima riga della matrice X e β è il vettore dei coefficienti.

I valori stimati dal modello sono:

$$\hat{Y}_i = \hat{\mu}_i = n_i e^{x_i^T \hat{\beta}} \quad i = 1, \dots, n. \tag{2.6}$$

Questi valori vengono anche indicati con e_i , $i = 1, \dots, n$.
Considerando che $E(Y_i) = Var(Y_i)$ per le distribuzioni di Poisson, l'errore standard di Y_i è stimato dalla quantità $\sqrt{e_i}$. I residui di Pearson sono, quindi, definiti come

$$e_i^P = \frac{o_i - e_i}{\sqrt{e_i}}, \quad i = 1, \dots, n\ , \tag{2.7}$$

dove o_i sono i valori osservati di Y_i.
Per la distribuzione di Poisson, i residui (2.7) sono in relazione con il test X^2 per la bontà dell'adattamento. Infatti:

$$X^2 = \sum_{i=1}^{n} \left(e_i^P\right)^2 = \sum_{i=1}^{n} \frac{(o_i - e_i)^2}{e_i} \quad i = 1, \dots, n. \tag{2.8}$$

La devianza residua per un modello di Poisson è

$$D = 2 \sum_{i=1}^{n} \left[o_i \log\left(\frac{o_i}{e_i}\right) - (o_i - e_i) \right], \quad i = 1, \dots, n. \tag{2.9}$$

Se $\sum o_i = \sum e_i$, la (2.9) può essere semplificata in

$$D = 2\sum_{i=1}^{n}\left[o_i \log\left(\frac{o_i}{e_i}\right)\right] \quad i = 1, \dots, n. \tag{2.10}$$

I residui di devianza sono le componenti di D nella (2.9), ossia

$$d_i = e_i^D = sign(o_i - e_i)\sqrt{2\left[o_i \log\left(\frac{o_i}{e_i}\right) - (o_i - e_i)\right]}, \quad i = 1, \dots, n\,; \tag{2.11}$$

quindi $D = \sum d_i^2$.

La statistica X^2 e la devianza D sono in relazione. Infatti, approssimando la devianza con le formule di Taylor otteniamo $D \cong X^2$. Per misurare la bontà di adattamento, entrambe vanno comparate con una distribuzione χ^2_{n-p}, dove p è il numero di parametri stimato.

Altri test per la bontà sono:

-la statistica log-rapporto di verosimiglianza, che confronta il valore massimo della funzione di log-verosimiglianza del modello nella (2.5) con *p* parametri con il massimo valore della funzione di log-verosimiglianza del modello con la sola intercetta. Si calcola come $C = 2\left[l(\hat{\theta}) - l(\tilde{\theta})\right]$ e si distribuisce come una χ^2_{p-1} se i p parametri sono tutti uguali a 0 tranne l'intercetta.

-la statistica pseudo-R^2 che dà una misura generale della bontà di adattamento; è un indice compreso tra 0 e 1, dove 1 è il massimo adattamento del modello ai dati (questa statistica verrà trattata più nello specifico nel capitolo 3).

2.3 Modelli Logit

Nel modello log-lineare, tutti i tipi di variabile vengono trattati allo stesso modo. Nel modello di regressione logistica, invece, occorre differenziare il caso in cui la variabile risposta è binaria dal caso in cui la variabile risposta si compone di più categorie.

2.3.1 Modello logit con variabile risposta binaria

Si definisce la variabile casuale binaria *Z* che determina successo o insuccesso nel seguente modo:

$$Z = \begin{cases} 1 \; se \text{ è } un \; successo \\ 0 \; se \text{ è } un \; insuccesso \end{cases},$$

con probabilità $Pr(Z = 1) = \pi$ e $\Pr(Z = 0) = 1 - \pi$, ovvero Z segue una distribuzione di Bernoulli $B(\pi)$. Se ci sono n variabili così definite $Z_1, \dots, Z_n$ indipendenti con $\Pr(Z_i = 1) = \pi_i$ e se i π_i sono tutti uguali allora

$$Y = \sum_{i=1}^{n} Z_i \;, \tag{2.12}$$

dove Y è il numero di successi su n prove. La variabile casuale Y si distribuisce come una $Bin(n, \pi)$, ossia

$$\Pr(Y = y) = \binom{n}{y} \pi^y (1 - \pi)^{n-y}, \qquad y = 1, \dots, n. \tag{2.13}$$

Consideriamo il caso generale di N variabili casuali indipendenti $Y_1, \dots, Y_N$ corrispondenti ai numeri di successi in N sottogruppi o strati (Tabella 2.4), $Y_i \sim Bin(n_i, \pi_i), \; i = 1, \dots, N$.

Tabella 2.4: frequenze per N distribuzioni Binomiali

	Sottogruppi			
	1	2	…	N
successi	Y_1	Y_2	…	Y_N
insuccessi	$n_1 - Y_1$	$n_2 - Y_2$	…	$n_{N-}Y_N$
TOT	n_1	n_2	…	n_N

Sia $P_i = Y_i/n_i$ la proporzione di successi in ogni sottogruppo. Si ha $E(Y_i) = n_i\pi_i$ e $E(P_i) = \pi_i$. Si modella la probabilità π_i come:

$$g(\pi_i) = x_i^T\beta \ , \tag{2.14}$$

dove x_i è il vettore delle variabili esplicative e β è il vettore dei parametri e $g(\cdot)$ è una funzione opportuna di legame.

Il caso più semplice è il modello lineare $\pi = x^T\beta$. È usato in alcune applicazioni pratiche ma ha lo svantaggio che, sebbene π sia una probabilità, i valori stimati $x^T\hat{\beta}$ possono essere inferiori a 0 o superiori a 1. Per assicurarsi che π rientri nell'intervallo [0,1], si ricorre all'uso della distribuzione di probabilità cumulata:

$$\pi = \int_{-\infty}^{t} f(s)ds \ , \tag{2.15}$$

con $f(s) \geq 0$ e $\int_{-\infty}^{\infty} f(s)ds = 1$. La funzione di densità $f(s)$ è chiamata *tolerance distribution*.

Un'applicazione di quanto appena descritto è l'esempio "Modello dose-risposta".

ESEMPIO: Modello dose-risposta

La variabile risposta rappresenta la percentuale di successi. Lo scopo del modello è descrivere la probabilità π in funzione della dose, ad esempio $g(\pi) = \beta_1 + \beta_2 x$. Se $f(s)$ è la distribuzione uniforme nell'intervallo $[c_1, c_2]$ allora

$$f(s) = \begin{cases} \frac{1}{c_2 - c_1} & se\ c_1 \leq s \leq c_2 \\ 0 & altrimenti \end{cases} , \tag{2.16}$$

$$\pi = \int_{c_1}^{x} f(s)ds = \frac{x - c_1}{c_2 - c_1} , \quad se\ c_1 \leq x \leq c_2 . \tag{2.17}$$

Questa equazione ha la forma $\pi = \beta_1 + \beta_2 x$, dove $\beta_1 = \frac{-c_1}{c_2 - c_1}$ e $\beta_2 = \frac{1}{c_2 - c_1}$.

Questo modello è scarsamente usato. Al suo posto si preferisce usare il modello logit. La *tolerance distribution* e π sono date, rispettivamente, da

$$f(s) = \frac{\beta_2 exp(\beta_1 + \beta_2 s)}{[1 + exp(\beta_1 + \beta_2 s)]^2} , \tag{2.18}$$

$$\pi = \int_{-\infty}^{x} f(s)ds = \frac{exp(\beta_1 + \beta_2 x)}{1 + exp(\beta_1 + \beta_2 x)} . \tag{2.19}$$

Di conseguenza la funzione legame è

$$\log\left(\frac{\pi}{1 - \pi}\right) = \beta_1 + \beta_2 x . \tag{2.20}$$

Il termine $\log\left[\frac{\pi}{(1-\pi)}\right]$ prende il nome di *funzione logit.*

Definendo il modello di regressione logistica in termini più generali abbiamo che:

$$logit\,\pi_i = \log\left(\frac{\pi_i}{1-\pi_i}\right) = x_i^T\beta \ , \quad i = 1, \dots n, \tag{2.21}$$

dove x_i è il vettore delle variabili esplicative e β è il vettore dei parametri.
La devianza è

$$D = 2\sum_{i=1}^{n}\left[y_i \log\left(\frac{y_i}{\hat{y}_i}\right) + (n_i + y_i)\log\left(\frac{n_i - y_i}{n_i - \hat{y}_i}\right)\right] \tag{2.22}$$

e può essere usata per valutare la bontà del modello comparandola con il valore ${\chi^2}_{n-p,1-\alpha}$.
Sia Y_i il numero di successi, n_i il numero di prove e $\hat{\pi}_i$ la probabilità di successo stimata dal modello.
I residui di Pearson sono:

$$e_i^P = \frac{(y_i - n_i\hat{\pi}_i)}{\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i)}} \tag{2.23}$$

e la statistica di Pearson è:

$$X^2 = \sum_{i=1}^{n}\left(e_i^P\right)^2 \quad , \tag{2.24}$$

che va confrontata col il valore $\chi^2_{n-p,1-\alpha}$.

I residui di devianza, invece, sono:

$$d_i = sign(y_i - n_i\hat{\pi}_i)\left\{2\left[y_i \log\left(\frac{y_i}{n_i\hat{\pi}_i}\right) + (n_i + y_i)\log\left(\frac{n_i - y_i}{n_i - n_i\hat{\pi}_i}\right)\right]\right\}^{1/2}, \quad (2.25)$$

con $\sum_{i=1}^{n} d_i^2 = D$, e $D \sim \chi^2_{n-p}$.

2.3.2 Modello logit con variabile risposta a più categorie

Se la variabile risposta ha più di due categorie occorre usare un metodo diverso da quello descritto al paragrafo precedente. Bisogna fare, però, un ulteriore differenziazione sul tipo di variabile risposta poiché il modello sarà differente a seconda che la suddetta variabile sia nominale oppure ordinale.

REGRESSIONE LOGISTICA CON VARIABILE RISPOSTA NOMINALE

I modelli di regressione logistica con variabile risposta nominale sono usati quando non c'è un ordine naturale tra le categorie. Perciò una categoria viene scelta come riferimento (ad esempio la prima) e di conseguenza, si definisce la funzione logit per le altre categorie come

$$logit(\pi_i) = \log\left(\frac{\pi_i}{\pi_1}\right) = x_i^T \beta_i\,, \qquad i = 2, \dots, r. \quad (2.262)$$

Le $(r-1)$ equazioni logit vengono usate simultaneamente per stimare i parametri β_i. Una volta ottenute le stime $\hat{\beta}_i$, si può calcolare il predittore lineare $x_i^T \hat{\beta}_i$. Passando all'esponenziale, la formula (2.26) diventa:

$$\hat{\pi}_i = \hat{\pi}_1 \exp\left(x_i^T \hat{\beta}_i\right), \qquad i = 2, \dots, r. \quad (2.27)$$

Ma considerando che $\hat{\pi}_1 + \hat{\pi}_2 + \cdots + \hat{\pi}_I = 1$ si ha che:

$$\hat{\pi}_1 = \frac{1}{1+\sum_{i=2}^{r} exp(x_i^T \hat{\beta}_i)} \tag{2.28}$$

e

$$\hat{\pi}_i = \frac{exp(x_i^T \hat{\beta}_i)}{1+\sum_{i=2}^{r} exp(x_i^T \hat{\beta}_i)} \ , \ \ i = 2, \dots, r. \tag{2.29}$$

I valori attesi si possono calcolare moltiplicando le probabilità stimate $\hat{\pi}_i$ per le frequenze. I residui di Pearson, la statistica chi-quadro per la bontà del modello, la devianza, i residui di devianza, il rapporto di verosimiglianza e l'indice pseudo-R^2 sono gli stessi visti nei modelli precedenti.

Spesso è più semplice interpretare gli effetti di fattori esplicativi in termini di rapporti di *odds* piuttosto che di parametri β. Si consideri una variabile risposta con r categorie e una variabile esplicativa x binaria, che indichi l'esposizione ad un fattore, $(x = 0)$ se è assente e $(x = 1)$ se è presente. Il rapporto di *odds* per la categoria i $(i = 2, \dots, r)$ in relazione alla categoria $i = 1$ è

$$OR_i = \frac{\pi_{ip}}{\pi_{ia}} \Big/ \frac{\pi_{1p}}{\pi_{1a}} \ , \tag{2.30}$$

dove π_{ip} e π_{ia} sono le probabilità della categoria i $(i = 1, \dots, r)$ a seconda che il fattore sia, rispettivamente, presente o assente. Il modello diventa, quindi:

$$\log\left(\frac{\pi_i}{\pi_1}\right) = \beta_{0i} + \beta_{1i} x \ , \ \ i = 2, \dots, r. \tag{2.31}$$

I logaritmi degli *odds* sono: $\log\left(\frac{\pi_{ia}}{\pi_{1a}}\right) = \beta_{0i}$ quando $(x = 0)$ e $\log\left(\frac{\pi_{ip}}{\pi_{1p}}\right) = \beta_{0i} + \beta_{1i}$ quando $(x = 1)$.

Di conseguenza il logaritmo del rapporto di *odds* si può scrivere come

$$\log OR_i = \log\left(\frac{\pi_{ip}}{\pi_{1p}}\right) - \log\left(\frac{\pi_{ia}}{\pi_{1a}}\right) = \beta_{1i} \,. \tag{2.32}$$

Quindi $OR_i = \exp(\beta_{1i})$ che si può stimare con $\exp(\hat{\beta}_{1i})$. Se $\hat{\beta}_{1i} = 0$ allora $OR_i = 1$ che corrisponde al caso in cui il fattore è assente. Inoltre, un intervallo di confidenza per OR_i è dato da $exp\left[\hat{\beta}_{1i} \pm z_{1-\frac{\alpha}{2}} \cdot s.e.(\hat{\beta}_{1i})\right]$. Gli intervalli di confidenza che non includono il valore 1 corrispondono a valori di β significativamente diversi da 0.

Per la regressione logistica con variabile risposta nominale la scelta della categoria di riferimento può avere effetto sul parametro $\hat{\beta}$ stimato, ma non sulle probabilità $\hat{\pi}$ stimate o sui valori attesi.

REGRESSIONE LOGISTICA CON VARIABILE RISPOSTA ORDINALE

Se esiste un evidente ordine naturale tra le categorie di una variabile risposta, allora tale ordine deve essere tenuto in considerazione nel procedimento di stima del modello.

Sia Z una variabile continua. Si definiscono, inoltre, per questa variabile, degli intervalli C_i , $i = 1, \dots I$. Ad esempio, i pazienti di un ospedale con valori piccoli della variabile Z vengono classificati come "non malati", quelli con valori medi "moderatamente malati" e quelli con valori alti "gravemente malati". Gli estremi degli intervalli $C_1, \dots, C_{I-1}$ definiscono le i categorie ordinali con associate probabilità $\pi_1, \dots, \pi_i$ (con $\sum_{i=1}^{I} \pi_i = 1$).

Esistono diversi metodi per trattare variabili di questo tipo:

1) *Modello logit cumulativo*

Gli *odds* cumulati per la i-esima categoria sono:

$$\frac{P(z \leq C_i)}{P(z > C_i)} = \frac{\pi_1 + \pi_2 + \cdots + \pi_i}{\pi_{i+1} + \cdots + \pi_I} \tag{2.33}$$

e il modello logit cumulativo è:

$$\log \frac{\pi_1 + \pi_2 + \cdots + \pi_i}{\pi_{i+1} + \cdots + \pi_I} = x_i^T \beta_i \tag{2.34}$$

2) *Modello con odds proporzionali*

Se il predittore lineare $x_i^T \beta_i$ comprende l'intercetta β_{0i} che dipende dalla categoria i, ma le altre variabili esplicative non dipendono da i, allora il modello diventa:

$$\log \frac{\pi_1 + \pi_2 + \cdots + \pi_i}{\pi_{i+1} + \cdots + \pi_I} = \beta_{0i} + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1} \, . \tag{2.35}$$

Questo modello con *odds* proporzionali si basa sull'ipotesi che gli effetti delle variabili esplicative sono gli stessi per tutte le categorie.

3) *Modello logit per categorie adiacenti*

Un alternativa al modello precedente è considerare il rapporto delle probabilità per categorie successive, ad esempio: $\frac{\pi_1}{\pi_2}, \frac{\pi_2}{\pi_3}, \ldots, \frac{\pi_{r-1}}{\pi_r}$.

Il modello logit per categorie adiacenti è:

$$\log\left(\frac{\pi_i}{\pi_{i+1}}\right) = x_i^T \beta_i \quad i = 1, \dots, r. \tag{2.36}$$

Se viene semplificato con

$$\log\left(\frac{\pi_i}{\pi_{i+1}}\right) = \beta_{0i} + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1} \ , \tag{2.37}$$

si assume che l'effetto delle variabili esplicative è lo stesso per tutte le coppie di categorie adiacenti. Si interpretano, solitamente, i parametri β_i come rapporti di *odds* tramite $OR = \exp(\beta_i)$.

4) *Modello logit dei rapporti continuativi*

Un ultima opzione è quella di modellare il rapporto delle probabilità

$$\frac{\pi_1}{\pi_2}, \frac{\pi_1 + \pi_2}{\pi_3}, \dots, \frac{\pi_1 + \cdots + \pi_{r-1}}{\pi_r}$$

oppure

$$\frac{\pi_1}{\pi_2 + \cdots + \pi_r}, \frac{\pi_2}{\pi_3 + \cdots + \pi_r}, \dots, \frac{\pi_{r-1}}{\pi_r}.$$

Il modello è

$$\log\left(\frac{\pi_i}{\pi_{i+1} + \cdots + \pi_r}\right) = x_i^T \beta_i \ . \tag{2.38}$$

In tutti i modelli presentati, la bontà di adattamento può essere testata tramite i residui, la devianza, il rapporto di verosimiglianza, le statistiche X^2 e pseudo-R^2 come per i modelli visti nei paragrafi precedenti.

CAPITOLO 3

Approfondimenti

3.1 La statistica Pseudo-R^2 nei modelli con regressione logistica

Il coefficiente di determinazione R^2 per la regressione lineare misura la proporzione di variabilità totale dei dati spiegata dal modello e varia tra 0 e 1, dove 0 indica un modello con nessun valore predittivo e 1 un modello che si adatta perfettamente ai dati. Si è cercato, negli ultimi trent'anni, un simile indicatore per i modelli di regressione logistica e sono state proposte diverse estensioni di R^2, alcune delle quali basate sull'entropia (Mittlböck e Schemper, 1996). Queste estensioni, chiamate statistiche pseudo-R^2, hanno trovato larga applicazione nelle scienze sociali (Maddala, 1983, Laitla, 1993 e Long, 1997).
McKelvey e Zavoina (1975) proposero una statistica pseudo-R^2 basata su una struttura di modello latente, dove la variabile risposta deriva da una variabile latente continua in relazione con il predittore lineare del modello. Questa statistica pseudo-R^2 è definita come

$$pseudo - R^2 = \frac{\widehat{Var}(\hat{y}^*)}{\widehat{Var}(\hat{y}^*)+Var(\varepsilon)} \ , \qquad (3.1)$$

dove $\hat{y}^*$ è la variabile latente e ε è l'errore. Poiché la variabile latente non è osservata, non è possibile calcolare $Var(\varepsilon)$, ma si assume che, nel modello logistico valga $Var(\varepsilon) = \pi^2/3$.

McFadden (1973) suggerì un indice che compara il modello con la sola intercetta e il modello completo, dato da

$$pseudo - R_F^2 = 1 - \frac{lnL(\hat{\theta})}{lnL(\tilde{\theta})} \ , \tag{3.2}$$

dove $L(\tilde{\theta})$ è la massima verosimiglianza per il modello ridotto e $L(\hat{\theta})$ è la massima verosimiglianza per il modello completo.

Maddala (1983) sviluppò un ulteriore indice $pseudo - R^2$ che può essere applicato a qualsiasi modello e che viene stimato con il metodo della massima verosimiglianza. Questo indice è:

$$pseudo - R_M^2 = 1 - \left(\frac{L(\tilde{\theta})}{L(\hat{\theta})}\right)^{\frac{2}{n}} \ . \tag{3.3}$$

Poichè la statistica rapporto di verosimiglianza è $\lambda = -2\log\left(\frac{L(\tilde{\theta})}{L(\hat{\theta})}\right)$, si può scrivere $pseudo - R_M^2 = 1 - e^{-\lambda/n}$. Maddala dimostrò che tale indice ha un limite superiore pari a $1 - \left(L(\tilde{\theta})\right)^{2/n}$ e, perciò, modificò il suo indice in:

$$pseudo - R_N^2 = \frac{1-\left(\frac{L(\tilde{\theta})}{L(\hat{\theta})}\right)^{\frac{2}{n}}}{1-\left(L(\tilde{\theta})\right)^{2/n}} \ . \tag{3.4}$$

L'indice R^2 per il modello lineare si interpreta, come detto in precedenza, come la quota di variabilità spiegata dai regressori. Tuttavia, non c'è una chiara interpretazione degli indici $pseudo - R^2$ in termini di variabilità nella risposta in un modello di regressione logistica.

3.2 La curva ROC

Uno strumento per valutare l'adeguatezza di un criterio di classificazione è fornito dalla curva ROC (*Receiver Operating Characteristic*).

Prendiamo in esame una tabella 2x2 che consenta di quantificare la proporzione di falsi positivi rispetto al totale di individui positivi e l'analoga proporzione di falsi negativi (Tabella 3.1). Questi valori sono determinati dal valore della soglia adottato che discrimina i valori del test positivi da quelli negativi.

Tabella 3.1: Tabella di classificazione

		Positivo (malato)	Negativo (non malato)
Test in esame	positivo	a	b
	negativo	c	d

Il confronto tra i risultati del test in esame e il vero stato del campione permette di stimare due parametri: la sensibilità ($Se = a/(a + c)$), ovvero la probabilità che un individuo malato risulti positivo al test e la specificità ($Sp = d/(d + b)$), ovvero la probabilità che un individuo malato risulti negativo al test.

L'analisi con la curva ROC viene effettuata attraverso lo studio della funzione che lega la sensibilità alla probabilità di ottenere un falso positivo nella classe dei sani (ossia $1 -$specificità). In altre parole, si studia il rapporto tra allarmi veri e allarmi falsi. La relazione tra questi due parametri si raffigura tramite una linea ottenuta riportando, in un sistema di assi cartesiani e per ogni valore della soglia adottato, la proporzione di veri positivi in ordinata e la proporzione di falsi positivi in ascissa. L'unione dei punti ottenuti riportando nel piano cartesiano ciascuna coppia (Se) e $(1 -$Sp) genera una curva spezzata con

andamento a scaletta (ROC *plot*). Interpolando questi punti si ottiene una curva (ROC *curve*) (Fig. 2).

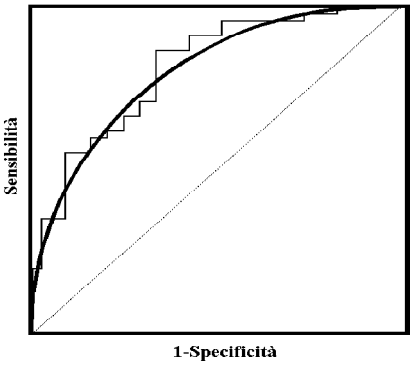

Fig. 2: Curva ROC prima e dopo l'interpolazione

Il valore della soglia ottimale è il punto sulla curva ROC più vicino all'angolo superiore sinistro.

La capacità discriminante di un test, ossia la sua capacità di separare correttamente il campione sano da quello malato è proporzionale all'area sottesa alla curva ROC (*Area Under Curve*, AUC) ed equivale alla probabilità che il risultato di un test su un individuo estratto a caso dal gruppo dei malati sia superiore a quello di un individuo estratto dal gruppo dei sani (Bamber, 1975; Zweig e Campbell, 1993).

L'area sottesa ad una curva ROC rappresenta un parametro fondamentale per la valutazione della capacità diagnostica di un test.

L'interpretazione del valore AUC può essere fatta attraverso questo schema proposto da Swets (1998):

- AUC=0.5 test non informativo
- 0.5<AUC≤0.7 test poco accurato
- 0.7<AUC≤0.9 test moderatamente accurato
- 0.9<AUC<1.0 test altamente accurato
- AUC=1 test perfetto

Un test per l'AUC può essere costruito nel seguente modo:

$$t = \frac{AUC - 0.5}{\sqrt{s.e.(AUC)}} \ , \tag{3.5}$$

dove $s.e.(AUC)$ è lo standard error di AUC.

Se il valore di t eccede il valore critico $z_{1-\alpha/2}$, si può affermare che il test diagnostico presenta una *performance* significativamente superiore a quella di un test non discriminante.

CAPITOLO 4

Codici R e un esempio

4.1 Alcuni codici R

Test chi-quadrato

```
chisq.test(dataset)
```

Test chi-quadrato con correzione di Yates

```
chisq.test(dataset,correct=T)
```

Test di Fisher

```
fisher.test(dataset)
```

Statistica di Mantel-Haenszel

```
mantelhaen.test(dataset)
```

Test chi-quadrato per il trend

```
prop.trend(x,n)
```

Modello log-lineare

```
model.glm<-glm(y~x,family=poisson,data=dataset)
```

Modello logit

```
Model.glm<-glm(y~x,family=binomial(link="logit"),data=dataset)
```

Curva ROC

```
library(ROCR)
data(ROCR.simple)
pred <- prediction(ROCR.simple$predictions, ROCR.simple$labels)
perf <- performance(pred,"tpr", "fpr")
plot(perf,colorize = TRUE)

library(verification)
roc.plot(ROCR.simple$labels,ROCR.simple$predictions, xlab = "False
positive rate",ylab = "True positive rate", main = NULL, CI = T,
n.boot = 100, plot = "both", binormal = TRUE)(auc <-
as.numeric(performance(pred, measure = "auc", x.measure =
"cutoff")@y.values))
```

4.2 Applicazione in R: analisi di una tabella 2x2

La Tabella 4.1 riporta i risultati di uno studio eseguito su 148 individui. Si vuole vedere se esiste un legame tra l'insorgenza di una malattia ostruttiva dell'arteria coronarica (OCAD) e il sesso dell'individuo.

Tabella 4.1: Malattia ostruttiva coronarica

		sesso		
		M	F	TOT
OCAD	si	92	15	107
	no	21	20	41
	TOT	113	35	148

Fonte: Prof. L. Ventura, Lucidi corso Modelli Statistici I, Università di Padova

Innanzitutto si importano i dati in R. Si crea una tabella a doppia entrata nel seguente modo:

```
> OCAD<-matrix(c(92,21,15,20), nrow=2)
> colnames(OCAD) <- c("M", "F")
> rownames(OCAD) <- c("SI", "NO")
> table <- as.table(OCAD)
```

```
> OCAD
    M  F
SI 92 15
NO 21 20
```

Si può eseguire il test chi-quadrato per saggiare l'indipendenza:

```
> chisq.test(OCAD, correct=F)

        Pearson's Chi-squared test

data:  OCAD
X-squared = 19.8375, df = 1, p-value = 8.431e-06

> chisq.test(OCAD, correct=T)

        Pearson's Chi-squared test with Yates' continuity correction

data:  OCAD
X-squared = 17.959, df = 1, p-value = 2.257e-05
```

Confronto i valori ottenuti nei due output con il percentile di livello $1 - \alpha = 0.95$ della distribuzione chi-quadrato:

```
>qchisq(0.950, 1)
[1] 3.841459
```

Si conclude che entrambi i valori `X-squared` sono più grandi di 3.84, pertanto rifiuto l'ipotesi nulla di indipendenza al livello 0.05.

Posso eseguire, inoltre, il test esatto di Fisher:

```
> fisher.test(OCAD)

        Fisher's Exact Test for Count Data

data:  OCAD
p-value = 2.345e-05
```

```
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  2.373501 14.381177
sample estimates:
odds ratio
  5.754167
```

Il p-value porta al rifiuto dell'ipotesi nulla che le popolazioni di origine dei due campioni abbiano la stessa suddivisione dicotomica e che le differenze osservate con i dati campionari siano dovute semplicemente al caso.

Per effettuare la regressione logistica si crea le matrici:

```
> y = c(rep(0,41),rep(1,107))
> x = c(rep(0,21),rep(1,20),rep(0,92),rep(1,15))
OCAD= data.frame(y,x)
```

dove y è la variabile risposta e x la variabile esplicativa.

Si effettua la regressione logistica:

```
> OCAD.glm<- glm(y~x,family=binomial(link="logit"))
> summary(OCAD.glm)

Call:
glm(formula = y ~ x, family = binomial(link = "logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8346  -1.0579   0.6412   0.6412   1.3018

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.4773     0.2418   6.108 1.01e-09 ***
x            -1.7649     0.4185  -4.217 2.47e-05 ***
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 174.68  on 147  degrees of freedom
Residual deviance: 156.31  on 146  degrees of freedom
AIC: 160.31

Number of Fisher Scoring iterations: 4
```

Creo la tabella di corretta classificazione

```
> table(y,fitted(OCAD.glm)>0.5)

y   FALSE TRUE
  0    20   21
  1    15   92
```

Una misura di adeguatezza della previsione è la frazione dei casi totali correttamente o erroneamente classificati:

```
> (20+92)/148
[1] 0.7567568
> (15+21)/148
[1] 0.2432432
```

Quindi la probabilità di corretta specificazione è del 75% e quella di errata classificazione è il restante 25%.

Calcolo l’indice $pseudo - R^2$:

```
>pseudoR2<-1-
((exp(0.5*deviance(OCAD.glm)))/(exp(0.5*(OCAD.glm$null.devianc
e))))
> pseudoR2
[1] 0.999897
```

Questa statistica suggerisce un buon adattamento ai dati.

BIBLIOGRAFIA

Armitage, P.e G.Berry (2003). *Statistica Medica*. McGrow-Hill.

Bamber, D. (1975). The Area above the Ordinal Dominance Graph and the Area below the Receiver Operatin Characteristics Graph. *J. Mah Psycol.*, **12**, 387-415.

Bland, M. (2009). *Statistica Medica.* Apogeo.

Bottarelli, E. e S. Padrodi (2003). Un approccio per la valutazione della validità dei test diagnostici: le curve ROC (*Receiver Operating Characteristic*). Ann. Fac Medic. Vet di Parma (Vol XXIII, 2003). 49-68. Disponibilità all'indirizzo: http://www.unipr.it/arpa/facvet/annali/2003/49.pdf

Dobson, A. J. e A. G. Barnett (2008). *An Introduction to Generalized Linear Models*. Chapman & Hall/CrC Press.

Everitt, B (1992). *The Analysis of Contingency Tables.* Chapman & Hall.

Hu, B., J. Shao e M. Palta (2006). Pseudo-R^2 in logistic regression model. *Statistica Sinica*. **16**, 847-860. University of Wisconsin-Madison.

Laitila, T. (1993). A pseudo-R^2 measure for limited and qualitative dependent variable models. *J. Econometrics.* **56**, 341-356.

Long, J. S. (1997). *Regression Models for Categorical and Qualitative Variables in Econometrics*. Sage Publications.

Maddala, G. S. (1983). *Limited-Dependen and Qaulitative Variables in Econometrics.* Cambridge University Press. Cambridge.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. In *Frontiers in Econometrics* (Edited by P. Zarembka), 105-152. Academy Press, New York.

Mittlböck, D. e M. Schemper (1996). Explained Variation for logistic regression. *Statist. Medicine*. **15**, 1987-1997.

Pace, L. e A. Salvan (2001). *Introduzione alla Statistica vol 2.* Cedam.

Rudas, T. (1998). *Odds RAtios in the Analysisi of Contingency Tables*. Sage Publications.

Scarpa, B. e A. Azzalini (2004). *Analisi dei dati e data mining*. Springer Verlag.

Soliani, L. Manuale di Statistica per la ricerca e la professione (Internet). (pubblicato: Aprile 2005; consultato: Marzo 2011). Disponibile all'indirizzo: http://www.dsa.unipr.it/soliani/soliani.html

Swets, J. A. (1998). Measuring the accuracy of diagnostic system. *Science*. **240**, 1289-1293.

Ventura, L. Lucidi corso di Modelli Statistici I. 2011. Università di Padova.

Wayne, W. D. (2005). *Biostatistica.* Edises.

Zweig, H. H. e G. Campbell (1993). Receiver Operating Characteristic (ROC) plots: a fundamental evolution tool in medicine. *Clin. Chem.*, **39**, 561-577.

Ringraziamenti

Ecco la parte che non volevo scrivere, non perché non ci tenga e ringraziare le persone, ma perché c'è sempre qualcuno che dimentichi o che non ringrazi nella maniera dovuta. Ma Letizia ha detto "scrivi!", perciò io scrivo.

Per prima cos voglio dire che, come in alcuni modelli statistici non è importante l'ordine in cui le variabili sono inserite, così in questa pagina non è importante chi viene per primo e chi per ultimo.

Un primo ringraziamento è dovuto alla professoressa Ventura che ha pazientemente corretto tutti i miei errori e mi ha dato l'opportunità di affrontare un argomento interessante come questo.

Ovviamente, poi, viene la famiglia. Quindi grazie mamma e grazie papà per il sostegno economico che mi ha permesso di frequentare questa università e per tutto quello che avete fatto per me dal giorno in cui mi avete messa al mondo.

Un ringraziamento particolare va alla nonna e al nonno. La prima lavora ancora alla sua veneranda età per aiutare "noi", il secondo mi leggeva le favole da piccola e si metteva la sveglia a mezzanotte per venirmi a prendere il sabato sera. Avete fatto molto di più di quello che i nonni fanno e, nonostante i nostri scontri dovuti alle diverse opinioni, al modo opposto in cui vediamo la vita e il mondo, al vostro orgoglio e al mio orgoglio, vi sono grata per tutto quello che avete fatto.

Completiamo la famiglia con la "sis"; spesso ti ho ignorata, mandata via, mai fatta avvicinare; sono consapevole del mio comportamento e mi dispiace; sei arrivata in un momento nel quale non era permesso a nessuno di entrare. Nonostante questo, per non so quale motivo, tu mi vuoi bene e quando ti ho

chiesto di esserci, per te “no” non era una risposta considerabile. Ti sono grata per amarmi anche se non lo merito perché so che di te avrò sempre bisogno, anche se mai l’ho dimostrato e, forse, mai riuscirò a dimostrartelo.
C’è una famiglia che non ti scegli, quella con cui condividi il DNA e c’è un’altra famiglia che ti scegli. Io non l’ho scelta, loro hanno scelto me e li ringrazio ogni giorno per averlo fatto. Se non fosse per i miei genitori non sarei a Padova, se non fosse per i miei amici non ci sarei affatto. Quindi ecco i miei fratelli:

- Berto. Sei stato il primo a dirmi di uscire con il gruppo e non l’ho mai dimenticato; forse tu non ti rendi conto di quello che hai fatto quel giorno, ma lo so io, quindi se mai un giorno avessi bisogno di qualcosa, anche un polmone, è tuo, fratello!! (non so in che condizioni lo troverai però).
- Cleme. Beh, “no Cleme no party” dice tutto di lui. Sei il fratello con cui posso parlare di tutto e che capisce quello che dico. Sei paziente e sopporti i miei difetti senza farmeli pagare troppo: grazie, continua così.
- Giolo. Il piccolo, anzi il “gavanel”. Nonostante sia il più piccolo del gruppo mi ha insegnato molte cose, mi ha stupito molte volte. Ti voglio bene fratellino!

Non voglio escludere nessuno: Giorgio, Nicholas, Alex, Erick, Bardellino e Silvia (anche se ultimamente abbiamo avuto i nostri problemi non posso dimenticare che sei stata importante).

Grazie alle mie compagne di università, Valentina e Chiara, per l’aiuto e la compagnia durante questi anni e ad Amos, senza il quale non avrei passato l’ultimo difficile esame

Un ringraziamento speciale prima di concludere va a Sandra. Se avessi un euro peri tutti i caffè, le sigarette e le chiacchierate che abbiamo fatto sarei

miliardaria. Voglio ringraziarti perché sei fondamentale, sei la voce fuori dal coro, quella che dice quello che non voglio sentire, ma di cui ho bisogno, quella che mi fa venire i dubbi che mi fanno cambiare, in meglio, spero. Grazie! Lo sai che sei come una madre per me!

Con un nome ho iniziato e con quel nome concludo. Grazie per avermi fatto scrivere questa pagine, ne avevo bisogno. Al contrario della Sandra, tu mi dici quello che ho bisogno di sentirmi dire, mi rassicuri sempre, mi dai fiducia e anche di questa ne ho molto bisogno. Grazie sorella!

E così concludo. Spero di non aver tralasciato o offeso nessuno e se l'ho fatto, non era mia intenzione.

"Ora non resta che darci buon tempo
in ufficiali feste di trionfo,
in gioiosi spettacoli di scena,
degni del gradimento della corte.
Suonin le trombe! Rullino i tamburi!
Pene, amarezze, turbolenze, addio!
Per noi da oggi ha inizio - com'io spero -
un'era lunga di felicità!"

William Shakespeare, Re Enrico VI

Un abbraccio a tutti