Analisi di dati categorici con modelli marginali epressi tramite vincoli di uguaglianza e disuguaglianza

F. Bartolucci
Istituto di Scienze Economiche
Università di Urbino

Argomenti

- Modelli log-lineari per l'analisi di dati categorici

- Modelli marginali

- Classe di modelli proposta e metodi di inferenza connessi

- Estensioni

Modelli log-lineari

- Ampiamente utilizzati per l'analisi di tabelle di contingenza.

- Espressi tipicamente nella forma

$$\boldsymbol{K} \log(\boldsymbol{p}) = \boldsymbol{X}\boldsymbol{\beta}$$

$\boldsymbol{p}$ vettore delle probabilità congiunte $p(a, b, \ldots)$

$\boldsymbol{K}$ matrice di contrasti

$\boldsymbol{X}$ matrice del disegno

$\boldsymbol{\beta}$ vettore dei parametri log-lineari

- Consentono di parametrizzare parametrize contrasti tra logaritmi di probabilità come (con 3 variabili, A, B, C):

  - *local logits* (di $A|B = 1, C = 1$):

$$\beta_A(a) = \log \frac{p(a+1,1,1)}{p(a,1,1)}$$

  - *local log-odds ratios* (tra $A, B|C = 1$):

$$\begin{aligned} \beta_{AB}(a,b) &= \beta_{A|B}(a|b+1) - \beta_{A|B}(a|b) = \\ &= \log \frac{p(a+1,b+1,1)p(a,1,1)}{p(1,b+1,1)p(a+1,1,1)} \end{aligned}$$

  - *contrasti tra log-odds ratios* (interazioni di ordine superiore al secondo)

$$\beta_{ABC}(a,b,c) = \beta_{AB|C}(ab|c+1) - \beta_{AB|C}(ab|c)$$

Vantaggi dei modelli log-lineari

- Stime di massima verosimiglianza calcolabili facilmente (Fisher-scoring, Iterative proportional fitting).

- Metodi asintotici ampiamente sviluppati per test e confronto tra modelli basati su rapporto di verosimiglianza (devianza).

- Ampia disponibilità di software (GLIM, S-plus).

Svantaggi dei modelli log-lineari

- Non è possibile modellare probabilità marginali tramite (con 3 variables, A, B, C):

  – *local logits* (di A):

$$\eta_A(a) = \log \frac{p(A = a + 1)}{p(A = a)}$$

  – *local log-odds ratios* (between A, B):

$$\eta_{AB}(a, b) = \log \frac{p(A = a + 1, B = b + 1)p(A = a, B = b)}{p(A = a, B = b + 1)p(A = a + 1, B = b)}$$

  – *Interazioni di ordine superiore al secondo*

- Non è possibile utilizzare logit diversi da quelli local come (con 3 variables, A, B, C):

  – *global logits* (di A):

$$\eta_A^g(a) = \log \frac{p(A > a)}{p(A \leq a)}$$

  – *global log-odds ratios* (between A, B):

$$\eta_{AB}^{gg}(a, b) = \log \frac{p(A > a, B > b)p(A \leq a, B \leq b)}{p(A \leq a, B > b)p(A > a, B \leq b)}$$

- Non è possibile utilizzare vincoli di disuguaglianza sui parametri che permettono di esprimere particolari forme di confordanza *ordinamenti stocastici*.

Mobilità sociale

- Si indaga su come si evolvono le classi sociali da una generazione all'altra.

	y_1	y_2	y_3	y_4	y_5	y_6	y_7
x_1	1	0	3	7	2	1	0
x_2	4	20	9	22	16	2	0
x_3	4	12	5	7	5	2	0
x_4	12	54	32	143	50	10	4
x_5	5	29	34	116	67	22	6
x_6	0	8	8	50	18	12	0
x_7	4	4	6	14	11	4	2

Father (X) and son (Y) social class
for a sample of 847 Hertfordshire men

- Problemi di interesse:
  - Miglioramento da una generalizzazione all'altra (confronto tra la distribuzione marginale di A e di B).
  - Miglioramento da una generalizzazione all'altra (utilizzo di log-odds ratios di tipo global).

Modelli marginali

- McCullagh & Nelder (1989), Lang & Agresti (1994, *JASA*), Lang (1996, *Annals of Statistics*), Glonek & McCullagh (1995, *JRSS-B*), Bergsma (1997)

- Vengono formulati come

$$\boldsymbol{K}\log(\boldsymbol{M}\boldsymbol{p}) = \boldsymbol{X}\boldsymbol{\gamma}$$

$\boldsymbol{p}$ vettore delle probabilità congiunte

$\boldsymbol{M}$ matrice di marginalizzazione

$\boldsymbol{K}$ matrice di contrasti

$\boldsymbol{X}$ matrice del disegno

$\boldsymbol{\gamma}$ vettore di parametri

Esempio di $\boldsymbol{M}$ e $\boldsymbol{K}$ modello logistico multivariato

- Nel caso di 2 variabili binarie (A, B):

$$\boldsymbol{K} = \begin{pmatrix} -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & -1 & 1 \end{pmatrix}$$

$$\boldsymbol{M} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

$$\boldsymbol{K} \log(\boldsymbol{M}\boldsymbol{p}) = \begin{pmatrix} \log \frac{p(A=2)}{p(A=1)} \\ \log \frac{p(B=2)}{p(B=1)} \\ \log \frac{p(A=2,B=2)p(A=1,B=1)}{p(A=1,B=2)p(A=2,B=1)} \end{pmatrix}$$

Approccio proposto

- Si propone una classe di modelli marginali in cui si possono utilizzare:
  - logit di tipo generalizzato (non solo local).
  - log-odds-ratios e interazioni di ordine più elevato di tipo generalizzato (non solo local).
  - il modello viene direttamente formulato tramite vincoli lineari di uguaglianza e disuguaglianza.

Logit generalizzati

- *local* :

$$\log \frac{p(A = a + 1)}{p(A = a)}$$

- *global* (variabili ordinaili):

$$\log \frac{p(A > a)}{p(A \leq a)}$$

- *continuation* (sopravvivenza):

$$\log \frac{p(A > a)}{p(A = a)}$$

- *reverse continuation* (come continuation ma con categorie in ordine inverso):

$$\log \frac{p(A = a + 1)}{p(A \leq a)}$$

Log-odds ratio generalizzati

- Contrasti tra logit sono utilizzati per modellare l'associazione tra due variabili $\Rightarrow$ *log-odds ratio generalizzati* (Douglas *et al.*, 1990)

- Example: *global log-odds ratios*

$$\log \frac{\Pr(A > a, B > b)\Pr(A \leq a, B \leq b)}{\Pr(A > a, B \leq b)\Pr(A \leq a, B > b)}$$

- Interazioni di ordine superiore sono definite come contrasti tra log-odds ratio generalizzati

Formulazione di un modello

- Si ha *vettore di parametri margianali* del tipo

$$\boldsymbol{\eta} = \boldsymbol{K} \log(\boldsymbol{M}\boldsymbol{p})$$

- Le matrici $\boldsymbol{K}$ e $\boldsymbol{M}$ possono essere generate tramite semplici prodotti di Kronecker.

- Possibili problemi con l'inversione di $\boldsymbol{\eta}$:
  - parametri non variation independence;
  - per un dato $\boldsymbol{\eta}$, $\boldsymbol{p}$ si ottiene tramite algoritmo di Newton.

- Un modello viene formulato trami vincoli di uguglianza e disuguaglianza del tipo:

$$\boldsymbol{C}\boldsymbol{\eta} = \boldsymbol{0}$$

$$\boldsymbol{D}\boldsymbol{\eta} \geq \boldsymbol{0}$$

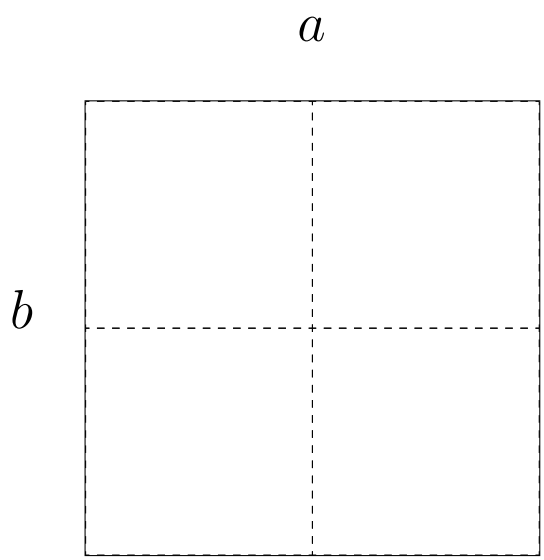
- Attraverso vincoli di disuguaglianza possiamo esprimere ipotesi di ordinamento stocastico come:

  - Una distribuzione marginale è più grande di un'altra

$$\log \frac{p(A > a)}{p(A \leq a)} \geq \log \frac{p(B > a)}{p(B \leq a)}, \ \forall a$$

  - Forme di concordanza tra due variabili (PQD)

$$\log \frac{p(A > a, B > b)p(A \leq a, B \leq b)}{p(A > a, B \leq b)p(A \leq a, B > b)} \geq 0, \ \forall a, b$$



Maximum Likelihood Estimation

- Let $y(a, b, \ldots)$ be the frequency of $A = a, B = b, \ldots$ in a sample of size n and $\boldsymbol{y}$ the vector with elements $y(a, b, \ldots)$ for any $a, b, \ldots$;

- *Multinomial log-likelihood*

$$l(\boldsymbol{\theta}) = \boldsymbol{y}' \log(\boldsymbol{p}) = \boldsymbol{y}' \boldsymbol{G}\boldsymbol{\theta} - n \log[\mathbf{1}' \exp(\boldsymbol{G}\boldsymbol{\theta})]$$

  $\boldsymbol{p}$ vector of joint probabilities $p(a, b, \ldots)$

  $\boldsymbol{\theta}$ vector of canonical parameters

  $\boldsymbol{G}$ design matrix

- $l(\boldsymbol{\theta})$ is maximized under $\boldsymbol{C}\boldsymbol{\eta} = \mathbf{0}, \boldsymbol{D}\boldsymbol{\eta} \geq \mathbf{0}$ through a generalization of the Aitchinson & Silvey algorithm (1958, *Annals of Statistics*)

Estimation algorithm

- At step $s+1$, a quadratic approximation of $l(\boldsymbol{\theta})$ is maximized under a linearized version of the constraints

$$\max_{\boldsymbol{\theta} \in \mathcal{C}_s} \; (\boldsymbol{v}_s - \boldsymbol{\theta}) \boldsymbol{F}_s (\boldsymbol{v}_s - \boldsymbol{\theta})$$

$\boldsymbol{F}_s$ information matrix (at step s)

$\boldsymbol{v}_s = \boldsymbol{\theta}_s + \boldsymbol{F}_s^{-1} \boldsymbol{s}_s$ working dependent variable

$\boldsymbol{s}_s$ score vector

$\mathcal{C}_s$ space generated by the linearized constraints

- Alternatively, we can use a constrained version of the Fisher-scoring algorithm that updates at any step the parameters $\boldsymbol{\eta}$ (possible problems with the transformation $\boldsymbol{\eta} \rightarrow \boldsymbol{p}$)

Hypothesis testing

- To compare two nested models, $M_1 \supset M_2$, we use the deviance

$$D = 2(\hat{l}_{M_1} - \hat{l}_{M_2})$$

$\hat{l}_M$ maximum log-likelihood under the model M

- When M_1 and M_2 are formulated by equality constraints, a p-value for D may be computed on the basis of the χ^2 distribution

- When M_1 is formulated by equality constraints and M_2 by equality and inequality constraints we have to use a mixture of χ^2's distributions known as $\bar{\chi}^2$ (Shapiro, 1985, *Biometrika*)

Analisi dei dati sulla mobilità sociale

- metti i vari risultati

Estensione al caso di strati con esempio

- metti i vari risultati

Estensione al tabelle multiple

- metti i vari risultati

Formulation of a marginal model

- A marginal model is formulated by choosing:

  – a set of marginal distributions (subsets of the set $\mathcal{V}$ of all the variables)

$$\mathcal{M}_1, \ldots, \mathcal{M}_r$$

  – a set of effects within any $\mathcal{M}_m, m = 1, \ldots, r$,

$$\mathcal{F}(\mathcal{M}_m) = \{\mathcal{L}_{m1}, \ldots, \mathcal{L}_{ms_m}\}$$

Bergsma & Rudas (2002) approach

- They deal with the *marginal log-linear parametrization* (local logits, local log-odds ratios,...)

- A marginal parametrization is *complete hierarchical* if:
  - all the possible effects are included with no replication of effects in different margins
  - there is an ordering of the $\mathcal{M}_m$'s such that

$$\mathcal{M}_i \not\subseteq \mathcal{M}_j, \forall i > j$$

  - for any m, all the effects $\mathcal{L} \subseteq \mathcal{M}_m$ belong to

$$\bigcup_{i \leq m} \mathcal{F}(\mathcal{M}_i)$$

- If the parametrization is hierarchical complete then it is *smooth*: Jacobian of full rank (Bardoff-Nielsen, 1978)

Example

Margin ($\mathcal{M}_m$)	Effect ($\mathcal{L}_{ml}$)	Description
A	A	local logit of A
AB	B	local logit of $B\|A=0$
	AB	log-odds ratio of AB
ABC	C	local-logit of $C\|A=0, B=0$
	AC	log-odds ratio $AC\|B=0$
	BC	log-odds ratio $BC\|A=0$
	ABC	contrast of log-odds ratios

An application

- Data set concerning allergies of a sample of $n = 997$ employees

- Response variables:

  I: infection to respiratory system (0=No, 1=Yes)

  A: asthma (0=No, 1=Moderate, 2=Serious)

  S: sneezes/lachrymation (0=No, 1=Moderate, 2=Serious)

- Data stratified according to:

  – G: gender (0=M, 1=F)

  – R: residence (0=Urban, 1=Suburban, 2=Countryside)

  – W: relatives with allergies (0=No, 1=Yes)

Data analysis

- For any variable we used logits of type *global*

- Parameters:
  - logit of I
  - logits of $A|I = i, \forall i$
  - logits of $S|I = i, \forall i$
  - log-odds ratios between $AS|I = i, \forall i$

- Model assumptions:
  - no effect of G, R, W on the marginal distribution of I
  - no effect of I on the joint distribution of AS
  - no effect of G, R, W on the association between A, S
  - no effect of G on $A|I$ and $S|I$
  - logits of $A|I$ and $S|I$ increase with W

- Deviance = 143.02 (175-223 d.f.) p-value = 1

References

Aitchison, J. and Silvey, S. D. (1958), Maximum likelihood estimation of parameters sunject to restraints, *Annals of Mathematical Statistics*, **29**, pp. 813-828.

Bergsma, W. P. (1997). *Marginal Models for Categorical Data*, Tilburg University Press.

Bergsma, W. P. and Rudas, T. (2002), Marginal models for categorical data, *Annals of Statistics*, **30**, .

Colombi, R., Forcina, A. (2001), Marginal regression models for the analysis of positive association of ordinal response variables, *Biometrika*, **88**, pp. 1007-1019.

Glonek, G. F. V. (1996), A class of regression models for multivariate categorical responses, *Biometrika*, **83**, pp. 15-28.

Glonek, G. J. N., and McCullagh, P. (1995). Multivariate Logistic Models, *Journal of the Royal Statistical Society* **B 57**, pp. 533-546.

Lang, J. B. and Agresti, A. (1994), Simultaneously modelling the joint and marginal distributions of multivariate categorical responses, *Journal of the American Statistical Association*, **89**, 626-632.

McCullagh P. & Nelder (1989), *Generalized linear models 2nd edition*, Chapman and Hall, London.