Analisi di dati cattura-ricattura tramite una versione estesa del modello di Rasch

F. Bartolucci

Università di Urbino

Francesco.Bartolucci@uniurb.it

A. Forcina

Università di Perugia

forcina@stat.unipg.it

Metodo Cattura-Ricattura

- Utilizzato per stimare la dimensione di una certa popolazione (N) sulla base di una serie di (J) "catture" effettuate in istanti di tempo diversi.

- A ogni soggetto "catturato" almeno una volta viene associata una configurazione di risposta:

$$\boldsymbol{r} = \begin{pmatrix} r_1 & \cdots & r_J \end{pmatrix}$$

con

$$r_j = \begin{cases} 1 & \text{se il soggetto è stato catturato all'occasione } j \\ 0 & \text{altrimenti} \end{cases}$$

- I dati possono essere raccolti in una tabella con $2^J - 1$ celle le cui frequenze sono $\boldsymbol{y} = \{y_{\boldsymbol{r}},\ \boldsymbol{r} \neq \mathbf{0}\}$; la cella mancante corrisponde al numero (incognito) di soggetti mai catturati, $y_{\mathbf{0}}$.

	$r_1 = 0$		$r_1 = 1$	
	$r_2 = 0$	$r_2 = 1$	$r_2 = 0$	$r_2 = 1$
$r_3 = 0$	y_{000} (?)	y_{010}	y_{100}	y_{110}
$r_3 = 1$	y_{001}	y_{011}	y_{101}	y_{111}

- Se $y_{\mathbf{0}}$ fosse noto, anche N lo sarebbe in quanto

$$N = y_{\mathbf{0}} + \sum_{\boldsymbol{r} \neq \mathbf{0}} y_{\boldsymbol{r}} = y_{\mathbf{0}} + n,$$

con n pari al numero complessivo delle catture.

Esempio

- Per stimare il numero di *diabetici residenti in Casale Monferrato* (Bruno *et al.*, 1994) sono state utilizzate $J = 4$ liste (catture):
  - ▷ cliniche private o medici di famiglia;
  - ▷ ospedali pubblici;
  - ▷ archivio pubblico dei casi di diabete;
  - ▷ pazienti che hanno richiesto rimborso dell'insulina.

$\boldsymbol{r}$	$y_{\boldsymbol{r}}$	$\boldsymbol{r}$	$y_{\boldsymbol{r}}$	$\boldsymbol{r}$	$y_{\boldsymbol{r}}$	$\boldsymbol{r}$	$y_{\boldsymbol{r}}$
0000	?	0100	74	1000	709	1100	104
0001	10	0101	7	1001	12	1101	18
0010	182	0110	20	1010	650	1110	157
0011	8	0111	14	1011	46	1111	58

Analisi di dati cattura-ricattura

- Modelli più utilizzati:
  - ▷ Modelli log-lineari (Fienberg, 1972, Cormack, 1989);
  - ▷ Modello a classi latenti (Cowan e Malec, 1986);
  - ▷ Modello di Rasch nella versione a classi latenti (Darroch, 1993, Agresti, 1994).
- Di particolare interesse solo il modello a classi e quello di Rasch (nella versione a classi latenti) in quanto tengono conto esplicitamente della eterogeneità tra i soggetti.

Modello a classi latenti

- Assunzioni di base:
  - ▷ la popolazione è divisa in k classi omogenee per quanto riguarda la tendenza ad essere catturati;
  - ▷ data la classe latente, le catture sono indipendenti tra loro (*locale indipendenza*, LI).

- *Probabilità condizionata* di osservare la configurazione $\boldsymbol{r}$ per un soggetto appartenente alla classe c

$$p_{\boldsymbol{r}|c} = \mathrm{pr}(\boldsymbol{r}|c) = \prod_j \lambda_{j|c}^{r_j}(1-\lambda_{j|c})^{1-r_j},$$

con $\lambda_{j|c} = \mathrm{pr}(r_j = 1|c)$.

- *Probabilità manifesta* di osservare la configurazione $\boldsymbol{r}$ (a prescindere dalla classe latente)

$$p_{\boldsymbol{r}} = \mathrm{pr}(\boldsymbol{r}) = \sum_c p_{\boldsymbol{r}|c}\pi_c,$$

con $\pi_c = p(c)$, peso della classe c.

- *Probabilità a posteriori* che un soggetto con la configurazione $\boldsymbol{r}$ appartenga alla classe c

$$\mathrm{pr}(c|\boldsymbol{r}) = \frac{p_{\boldsymbol{r}|c}\pi_c}{\sum_h p_{\boldsymbol{r}|h}\pi_h}.$$

Modello di Rasch (nella versione a classi latenti)

- Può essere visto come un caso particolare del modello a classi latenti in cui (assunzione di *unidimensionalità*, U)

$$\alpha_{j|c} = \log \frac{\lambda_{j|c}}{1 - \lambda_{j|c}} = \phi_c + \psi_j$$

ϕ_c tendenza ad essere catturati dei soggetti nella classe c

ψ_j efficacia nel catturare della lista j

- L'assunzione di *monotonia* (M) è automaticamente soddisfatta, nel senso che si possono sempre riordinare le classi in modo che

$$\lambda_{j|1} \leq \lambda_{j|2} \leq \cdots \leq \lambda_{j|k}, \quad \forall j.$$

- La differenza principale rispetto al modello di Rasch (nella sua formulazione classica) è che per il fattore latente (tendenza ad essere catturati) non si utilizza un parametro per ognuno degli n individui, ma se ne utilizza un numero limitato ($k \ll n$).

- Ciò è dovuto all'assunzione dell'esistenza di classi latenti all'interno delle quali si soggetti hanno la tendenza ad essere catturati; ciò equivale ad assumere che il fattore latente abbia una distribuzione discreta con k livelli.

- Assunzioni di questo tipo sono utilizzate nell'ambito del metodo di stima *Marginal Maximum Likelihood* MML (Lindsay, *et al.*, 1991) allo scopo di ridurre il numero dei parametri e quindi ottenere stime consistenti degli item parameter.

- La caratteristica fondamentale del modello di Rasch è la facile interpretazione dei parametri.

- Tuttavia, in ambito cattura-ricattura, le sue assunzioni possono essere troppo restrittive:

  - il fattore latente (tendenza ad essere catturati) può non essere lo stesso rispetto a tutte le liste (violazione di U; Darroch *et al.*, 1993); *esempio*: il fattore latente per le liste 1 e 2 è diverso da quello per le liste 3 e 4.

  - anche condizionatamente alla classe latente due liste possono essere associate (violazione di LI); *esempio*: coloro che appaiono nella lista 1 più difficilmente appaiono nella lista 2.

Classe di modelli proposta

- Si propone una estensione del modello di Rasch in cui le assunzioni LI e U possono essere parzialmente indebolite:
  - ▷ si permette *multidimensionalità* ammettendo l'esistenza di più di un fattore latente:

$$\alpha_{j|c} = \phi_c(j) + \psi_j$$

  - ▷ si può avere *locale dipendenza* permettendo che alcune coppie di liste (j_1, j_2) siano associate condizionatamente alla classe latente:

$$\beta_{j_1,j_2|c} = \log \frac{p(r_{j_1} = 0, r_{j_2} = 0|c)p(r_{j_1} = 1, r_{j_2} = 1|c)}{p(r_{j_1} = 0, r_{j_2} = 1|c)p(r_{j_1} = 1, r_{j_2} = 0|c)} \neq 0$$

Formulazione di un modello nella classe proposta

- Si introduce un vettore dei paramentri $\boldsymbol{\eta} = (\, \boldsymbol{\eta}_1' \quad \boldsymbol{\eta}_{2,1}' \quad \cdots \quad \boldsymbol{\eta}_{2,k} \,)'$
  - ▷ $\boldsymbol{\eta}_1$ contenente i logit consecutivi dei pesi delle classi

$$\log(\pi_{c+1}/\pi_c), \qquad c = 1, \ldots, k-1$$

  - ▷ $\boldsymbol{\eta}_{2,c}$, $c = 1, \ldots, k$, contiene i parametri della distribuzione condizionata di $\boldsymbol{r}$ data la classe latente

$$\alpha_{j|c}, \qquad j = 1, \ldots, J,$$

$$\beta_{j_1,j_2|c}, \qquad j_1 = 1, \ldots, J-1, \quad j_2 = j_1 + 1, \ldots, J,$$

mentre le interazioni di ordine più elevato al secondo sono pari a 0.

- Il vettore dei parametri $\boldsymbol{\eta}$ viene a sua volta modellato tramite la forma lineare

$$\boldsymbol{\eta} = \boldsymbol{X}\gamma, \quad \boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{X}_2 \end{pmatrix}, \quad \gamma = \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix}$$

dove

  - ▷ $\boldsymbol{X}_1 = \boldsymbol{I}$ (non ci sono vincoli su $\pi_1, \ldots, \pi_k$);
  - ▷ $\boldsymbol{X}_2$ ha righe pari a $\mathbf{0}'$ a parte quelle in corrispondenza dei logit di ogni lista ($\alpha_{j|c}$) e delle interazioni doppie ($\beta_{j_1,j_2|c}$) non vincolate ad essere nulle.

- La classe dei modelli che possono essere formulati è in realtà più ampia di quella descritta inizialmente.

Stima dei parametri

- $\boldsymbol{\gamma}$ è stimato massimizzando la *verosimiglianza multinomiale*, condizionata al numero delle catture (n), che ha logaritmo

$$l(\boldsymbol{\gamma}) = \sum_{\boldsymbol{r} \neq \mathbf{0}} y_{\boldsymbol{r}} \log \left(\frac{p_{\boldsymbol{r}}}{q} \right),$$

  con $q = \sum_{\boldsymbol{r} \neq \mathbf{0}} p_{\boldsymbol{r}}$ pari alla probabilità che un soggetto venga catturato almeno una volta.

- Per la massimizzazione di $l(\boldsymbol{\gamma})$ si propone un uso congiunto degli algoritmi Fisher-scoring e EM. Il secondo è utilizzato per fornire valori di partenza per il primo che è molto veloce ma usualmente instabile.

Stima della dimensione della popolazione

- Una volta calcolata la stima di $\boldsymbol{\gamma}$ $(\widehat{\boldsymbol{\gamma}})$ quella della dimensione della popolazione viene ottenuta come:

$$\hat{N} = \frac{n}{\hat{q}} > n.$$

con $\hat{q} = \sum_{\boldsymbol{u} \neq \mathbf{0}} \hat{p}_{\boldsymbol{r}}$ pari alla stima della probabilità che un soggetto venga catturato almeno una volta.

- Stima della frequenza non osservata:

$$\hat{y}_{\mathbf{0}} = \frac{n}{\hat{q}} - n = n\frac{1-\hat{q}}{\hat{q}}.$$

Algoritmo Fisher-scoring

- Algoritmo iterativo che, al passo $h+1$, consiste nell'aggiornare la stima di $\boldsymbol{\gamma}$ come

$$\boldsymbol{\gamma}^{h+1} = \boldsymbol{\gamma}^{h} + [\boldsymbol{F}(\boldsymbol{\gamma}^{h})]^{-1}\boldsymbol{g}(\boldsymbol{\gamma}^{h})$$

dove:

- ▷ $\boldsymbol{\gamma}^{h}$ stima al passo h;
- ▷ $\boldsymbol{g}(\boldsymbol{\gamma}) = \dfrac{\partial l(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}$ (*vettore score*);
- ▷ $\boldsymbol{F}(\boldsymbol{\gamma}) = -\mathrm{E}\Big(\dfrac{\partial l^2(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}\partial \boldsymbol{\gamma}'}\Big)$ (*matrice di informazione attesa*)

Algoritmo EM

- È basato sul concetto di *dati completi* che, quando si assume l'esistenza di classi latenti, corrispondono alle frequenze di ogni configurazione $\boldsymbol{r}$ e ogni classe c ($\boldsymbol{x} = \{x_{c,\boldsymbol{r}}\}$) mentre i *dati incompleti* corrispondono alle frequenze osservate ($\boldsymbol{y} = \{y_{\boldsymbol{r}}\}$).

- L'algoritmo consiste nell'alternare i due passi:

E: si calcola il valore attesto condizionato di $\boldsymbol{x}$ dato $\boldsymbol{y}$;

M: si aggiorna γ massimizzando la log-verosimiglianza dei dati completi,

$$l_C(\gamma) = \sum_c \sum_{\boldsymbol{r}} x_{c,\boldsymbol{r}} \log(\pi_c p_{\boldsymbol{r}|c}),$$

con $\boldsymbol{x}$ sostituito con il valore atteso calcolato al passo E.

Analisi dei dati sui diabetici

Modello ($k = 2$)	Devianza	d.f.	$\widehat{N}$
Classi latenti	54,240	5	2.295
Rasch	93,953	8	2.332
Rasch + locale dipendenza (*)	0,879	5	2.403

(*) Si ammette associazione (indipendente dalla classe latente) tra:

1. *cliniche* e *ospedali* (γ_{12});
2. *cliniche* e *rimborso* (γ_{14});
3. *ospedali* e *archivio pubblico* ($\gamma_{23} = \gamma_{12}$);
4. *archivio pubblico* e *rimborso* (γ_{34}).

Stime ottenute

Parametro	Stima	s. e.	Parametro	Stima	s. e.
$\log(\pi_2/\pi_1)$	-0,7856	0,1523	ϕ_2	2,3335	0,1336
ψ_1	0,5394	0,1927	γ_{12}	-1,5132	0,3365
ψ_2	-2,5610	0,2052	γ_{14}	-1,1733	0,2884
ψ_3	-0,7895	0,1796	γ_{23}	-1,5132	0,3365
ψ_4	-3,8303	0,2106	γ_{24}	1,0704	0,2041

- la lista 1 (cliniche private o medici di famiglia) è la più efficace (ψ_1);
- la seconda classe latente è quella dei soggetti più facili da catturare (ϕ_2) e ha peso minore ($\log(\pi_2/\pi_1)$);
- le quattro associazioni sono significative anche condizionatamente alla classe latente e sono prevalentemente negative.

Costruzione di intervalli di confidenza

- Un intervallo di confidenza per N può essere ottenuto sulla base della log-verosimiglianza profilo di N (Cormack, 1992):

$$l^*(N) = \max_{\boldsymbol{\beta}} \ \log \ \frac{N!}{\prod_{\boldsymbol{r}} y_{\boldsymbol{r}}!} \prod_{\boldsymbol{r}} p_{\boldsymbol{r}}^{y_{\boldsymbol{r}}}.$$

  - ▷ si individua il valore di N, $\widehat{N}_U$, che massimizza $l^*(N)$;
  - ▷ per ogni N in un certo intervallo di interi si calcola $l^*(N)$ e la devianza

$$D(N) = 2\{l^*(\widehat{N}_U) - l^*(N)\};$$

- dato che $D(N) \sim \chi^2_1$ (asintoticamente sotto l'ipotesi che il vero valore della popolazione è N), l'intervallo di confidenza al livello $100(1-\alpha)\%$ per N è dato da

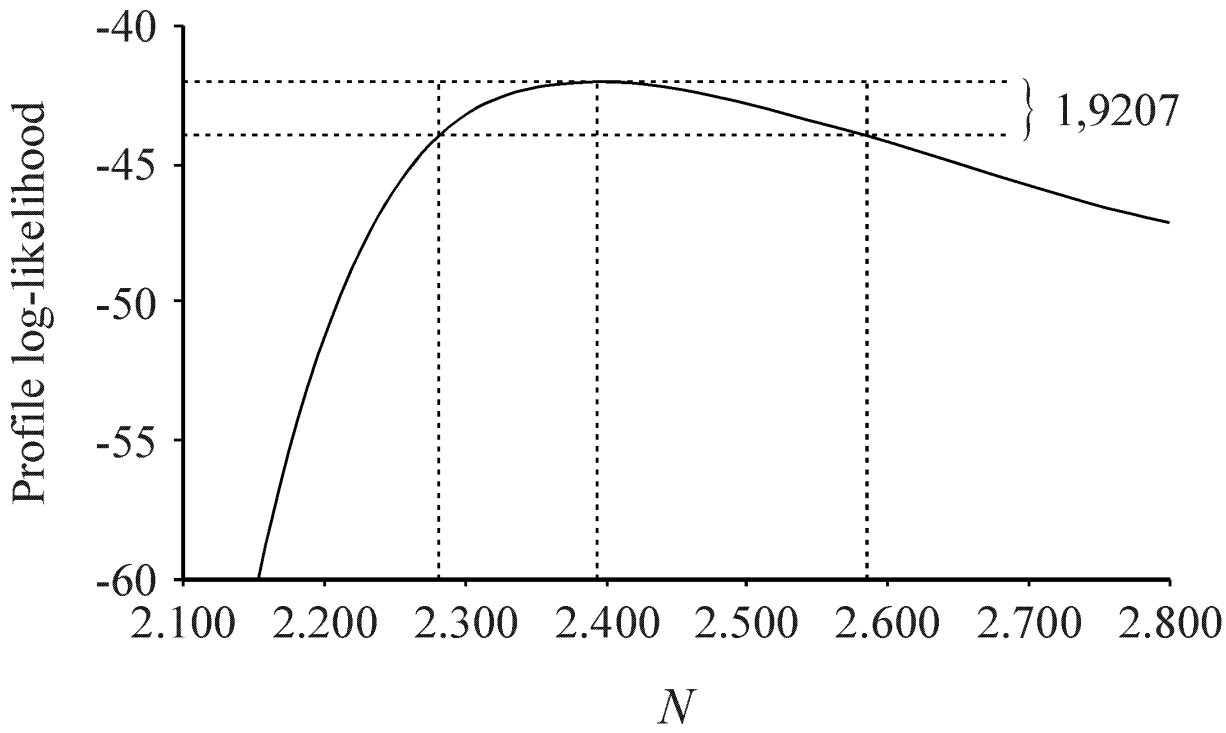
$$(N_1, N_2)$$

con

* N_1 più grande intero ($< \widehat{N}_U$) tale che $D(N_1) \geq \chi^2_{1,\alpha}$
* N_2 più piccolo intero ($> \widehat{N}_U$) tale che $D(N_2) \geq \chi^2_{1,\alpha}$

Dati su diabete

- Intervallo di confidenza per N: $(2.280,\ 2.585)$;



Inclusione di variabili esplicative discrete

- L'approccio proposto è stato esteso in modo da trattare dati stratificati secondo una o più variabili esplicative discrete.

- Si assume che la popolazione in ogni strato sia divisa in k classi latenti omogenee.

- Il modello può essere ancora formulato come $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\gamma}$; è possibile condividere alcuni parametri tra gli strati in modo da avere un modello più parsimonioso.

- Per la stima dei parametri e la costruzione degli intervalli di confidenza si possono utilizzare le stesse tecniche viste in precedenza.

Esempio

- Dati utilizzati da Darroch *et al.* (1993) per stimare il numero di persone di colore residenti in St. Luis (Missuri); vengono utilizzare $J = 3$ liste e due covariate: *età*, *tipo abitazione*.

	20-29		30-44	
$\boldsymbol{r}$	Owns	Rents	Owns	Rents
000	?	?	?	?
001	59	43	35	43
010	65	70	69	53
011	19	11	10	13
100	75	73	77	71
101	19	12	13	7
110	217	144	262	155
111	79	58	91	72

Modello ($C = 2$)	Devianza	d.f.	$\widehat{N}$
M1: Rasch	287.94	8	2.181
M2: M1 + violazione U (*)	3.71	5	4.153
M3: M2 + vincoli su strati (**)	6.84	13	3.089
M4: M3 + violazione LI (***)	1.37	9	3.579

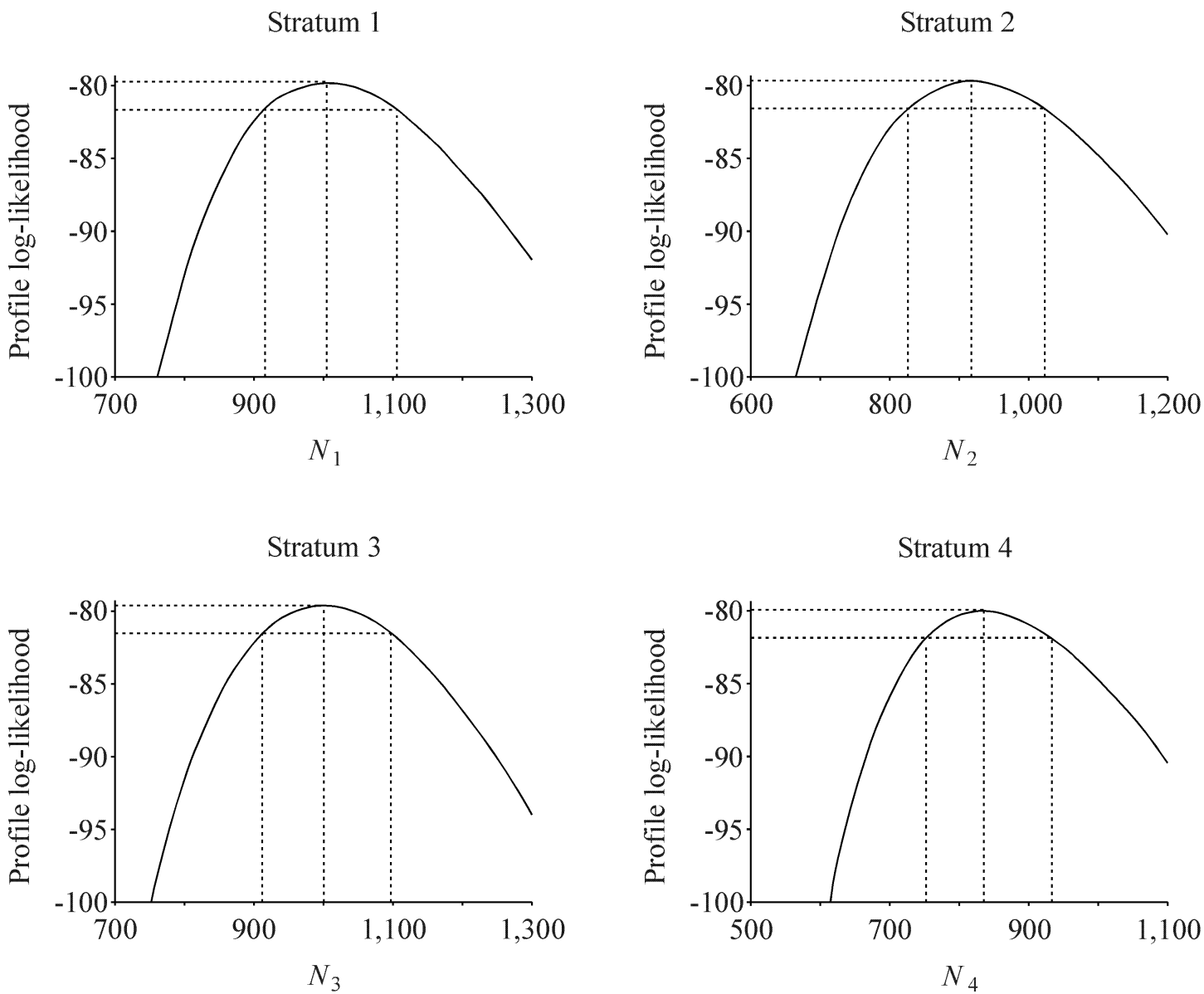
(*) La restrizione di logit additivo (U) viene rimossa per la terza lista.

(**) Restrizioni:

1. struttura fattoriale delle variabili esplicative sui logit di appartenere alle varie classi;
2. efficacia delle liste non varia da strato a strato.

(***) Associazione doppia (condizionata) tra le liste 1-3 e 2-3.

- Viene costruito un intervallo di confidenza per la popolazione in ogni strato ottenendo $(916, 1106)$, $(826, 1023)$, $(912, 1097)$ and $(752, 933)$.



Inclusione di variabili esplicative continue

- Si sta tentando di estendere l'approccio al caso di covariate continue ($\boldsymbol{z}$) utilizzando una parametrizzazione del tipo

$$\log \frac{p(r_j = 1|\boldsymbol{z}, c)}{p(r_j = 0|\boldsymbol{z}, c)} = \phi_c + \psi_j + \boldsymbol{\delta}'\boldsymbol{z}$$

- Per la stima del parametri possono essere utilizzati degli algoritmi analoghi a quelli per il caso di variabili esplicative discrete.

- La costruzione di intervalli di confidenza risulta più complessa e richiede la messa a punto di algoritmi specifici.

Principali riferimenti bibliografici

Agresti, A. (1994). Simple capture-recapture models permitting unequal catchability and variable sampling effort. *Biometrics* **50**, 494-500.

Bartolucci, F. e Forcina, A. (2001), The Analysis of Capture-Recapture Data with a Rasch-type Model allowing for Conditional Dependence and Multidimensionality, *Biometrics*, **57**, pp. 714-719.

Biggeri, A., Stanghellini, E., Merletti, F. and Marchi, M. (1999). Latent class models for varying catchability and correlation among sources in capture-recapture estimation of the size of a human population. *Statistica Applicata* **11**, 563-586.

Bruno, G., Biggeri, A., Merletti, F., Laporte, R., McCarthy, D. and Pagano, G. (1994). Applications of capture-recapture to count diabetes. *Diabetes Care* **17**, 548-556.

Cormack, R. M. (1989). Log-linear models for capture-recapture. *Biometrics* **45**, 395-413.

Cormack, R. M. (1992). Interval estimation for mark-recapture studies of closed populations. *Biometrics* **48**, 567-576.

Cowan, C. D. and Malek, D. (1986). Capture-Recapture models when both sources have clustered observations. *Journal of the American Statistical Association* **81**, 461-466.

Darroch, J. N., Fienberg, S. E., Glonek, G. F. V. and Junker, B. W. (1993). A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability. *Journal of the American Statistical Association* **88**, 1137-1148.

Fienberg, S. E. (1972). The multiple recapture census for closed populations and incomplete 2^k contingency tables. *Biometrika* **59**, 591-603.

Green, P. J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

Green, P. J. and Mira, A. (2001), Delayed Rejection in Reversible Jump Metropolis-Hastings. *Biometrika* **88**, 1035-1053.

Lindsay, B., Clogg, C. and Grego, J. (1991). Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association* **86**, 96-107.

Sanathanan, L. (1972). Estimating the size of a multinomial population. *The Annals of Mathematical Statistics* **43**, 142-152.