Modelli mistura per l'analisi di dati cattura-ricattura

F. Bartolucci

Istituto di Scienze Economiche
Università di Urbino

Metodo Cattura-Ricattura

- Metodo 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 la configurazione:

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

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

- I dati possono essere raccolti su una tabella $\{y_{\boldsymbol{r}}\}$ con $2^J - 1$ celle; la cella mancante corrisponde al numero (incognito) di soggetti mai catturati.

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

Esempio

- Diabetici residenti in Casale Monferrato (Bruno *et al.*, 1994, *Diabetes Care*).

- Sono state utilizzate $J = 4$ liste (catture):

  1. cliniche private o medici di famiglia;
  2. ospedali pubblici;
  3. archivio pubblico dei casi di diabete;
  4. 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

Modelli più noti per l'analisi di dati cattura-ricattura

- Modelli log-lineari (Fienberg, 1972, Cormack, 1989);
- Modello a classi latenti (Cowan e Malec, 1986);
- Modello di Rasch (Darroch, 1993, Agresti, 1994).

Modello a classi latenti

- Assunzioni di base:

  1. la popolazione è divisa in C classi omogenee (stessa tendenza ad essere catturati);

  2. data la classe latente, le catture sono indipendenti tra loro (*locale indipendenza*, LI).

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

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

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

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

$$q_{\boldsymbol{r}} = \Pr(\boldsymbol{r}) = \sum_c p_{\boldsymbol{r}|c}\pi_c,$$

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

Modello di Rasch

- Caso particolare del modello a classi latenti in cui (assunzione di *unidimensionalità*, U)

$$\lambda_{j|c} = \frac{e^{\phi_c + \psi_j}}{1 + e^{\phi_c + \psi_j}}$$

  θ_c tendenza ad essere catturati dei soggetti nella classe c

  β_j efficacia della lista j.

- Facile interpretazione dei parametri.

- Le assunzioni possono essere troppo restrittive:

  1. anche condizionatamente alla classe latente due liste possono essere non indipendenti (violazione di LI);

  2. la tendenza ad essere catturati può non essere la stessa per tutte le liste (violazione di U)

$$\lambda_{j|c} = \frac{e^{\phi_c(j) + \psi_j}}{1 + e^{\phi_c(j) + \psi_j}}.$$

Classe di modelli proposta

- Si propone una estensione del modello di Rasch in cui le assunzioni LI e U sono parzialmente indebolite.

- I pesi delle classi vengono parametrizzati tramite logits

$$\boldsymbol{\eta}_1 = \boldsymbol{G} \log(\boldsymbol{\pi})$$

con $\boldsymbol{\pi} = \{\pi_c\}$.

- Per ogni classe latente, si parametrizza $\boldsymbol{p}_c = \{p_{\boldsymbol{r}|c}\}$ come (*modello marginale*)

$$\boldsymbol{\eta}_{2,c} = \boldsymbol{C} \log(\boldsymbol{M}\boldsymbol{p}_c),$$

che contiene:

  - J logit (marginali);
  - $J(J-1)/2$ log-odds ratio (marginali);
  - parametri di interazione (marginali) di ordine superiore fino al J-mo.

Formulazione di un modello

- si utilizza una forma lineare

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$$

con

$$\boldsymbol{\eta} = \begin{pmatrix} \boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_{2,1} \\ \vdots \\ \boldsymbol{\eta}_{2,C} \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{X}_2 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}$$

- solitamente $\boldsymbol{X}_1 = \boldsymbol{I}$;

- $\boldsymbol{X}_2$ ha righe pari a $\mathbf{0}'$ in corrispondenza di tutte le interazioni di ordine triplo o superiore e in corrispondenza delle interazioni doppie che si vogliono nulle;

- $\boldsymbol{X}_2$ ha righe diverse da $\mathbf{0}'$ in corrispondenza dei logit marginali di ogni lista.

Inferenza classica

- $\boldsymbol{\beta}$ è stimato massimizzando la verosimiglianza condizionata, dato n, che ha logaritmo

$$l_{\boldsymbol{y}} = \sum_{\boldsymbol{r} \neq \boldsymbol{0}} y_{\boldsymbol{r}} \log\left(\frac{q_{\boldsymbol{r}}}{s}\right), \qquad s = \sum_{\boldsymbol{u} \neq \boldsymbol{0}} q_{\boldsymbol{r}}.$$

- 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 più veloce ma instabile.

- Una volta stimati i parametri del modello, da dimensione della popolazione viene stimata come:

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

- Stima della frequenza non osservata:

$$\hat{y}_{\boldsymbol{0}} = \frac{n}{\hat{s}} - n = n\frac{1-\hat{s}}{\hat{s}}.$$

Algoritmo Fisher-scoring

- All'iterazione $h+1$ si aggiorna la stima di $\boldsymbol{\beta}$ come

$$\boldsymbol{\beta}^{h+1} = \boldsymbol{\beta}^{h} + (\boldsymbol{F}^{h})^{-1}\boldsymbol{g}^{h}$$

$\boldsymbol{\beta}^{h}$ stima al passo h

$\boldsymbol{g}^{h}$ vettore scre al passo h

$\boldsymbol{F}^{h}$ matrice di informazione attesa al passo h.

Algoritmo EM

- È basato sui due passi:

E: si calcola il valore attesto condizionato dei dati completi ($\{x_{c,\boldsymbol{r}}\}$) ai dati incompleti ($\{y_{\boldsymbol{u}}\}$);

M: si aggiorna $\boldsymbol{\beta}$ massimizzando la log-verosimiglianza (dei dati completi),

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

con $\{x_{c,\boldsymbol{r}}\}$ sostituiti con i corrispondenti valori attesti calcolati al passo E.

Analisi dei dati sui diabetici

Modello ($C = 2$)	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 un'associazione (condizionata) tra:

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

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

Costruzione di intervalli di confidenza

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

$$l_y(N) = \max_{\boldsymbol{\beta}} \ \log \ \frac{N!}{\Pi_{\boldsymbol{r}} \, y_{\boldsymbol{u}}!} \prod_{\boldsymbol{r}} q_{\boldsymbol{r}}^{y_{\boldsymbol{r}}}.$$

1. si individua il valore di N, $\widehat{N}_U$, che massimizza $l_y(N)$;

2. per ogni N in un certo intervallo di interi si calcola $l_y(N)$ e la devianza

$$D(N) = 2\{l_y(\widehat{N}_U) - l_y(N)\};$$

3. 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

$$(N_1, N_2)$$

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);

Inferenza Bayesiana
(per il modello a classi latenti)

- Distribuzioni a priori dei parametri:

  1. il numero di classi latenti, C, ha distribuzione uniforme nell'intervallo $[1, C_{max}]$;
  2. il vettore dei pesi, $\boldsymbol{\pi} = (\, \pi_1 \quad \ldots \quad \pi_C \,)'$, ha distribuzione Dirichlet con parametri $\boldsymbol{\nu} = (\, \nu_1 \quad \ldots \quad \nu_C \,)'$;
  3. per ogni j e c, $\lambda_{j|c}$ ha distribuzione Beta con parametri $\boldsymbol{\alpha} = (\, \alpha_1 \quad \alpha_2 \,)'$.

- Per avere una indicazione univoca delle classi latenti, questo sono in ordinate in modo crescente rispetto al peso:

$$\pi_1 < \pi_2 < \ldots < \pi_C.$$

- Per stimare la distribuzione congiunta a posteriori dei parametri $(\boldsymbol{\pi}, \boldsymbol{\lambda})$ e del numero di classi latenti (C) si ricorre al *Reversible Jump* (RJ) che permette di estrarre campioni da tale distribuzione:

$$(\boldsymbol{\pi}^{(t)}, \boldsymbol{\lambda}^{(t)}, C^{(t)}), \quad t = 1, \ldots, T.$$

Reversible Jump

- A ogni passo viene effettuata una delle due mosse:

  1. Aggiornamento dei parametri senza variare il numero di classi latenti;
  2. Aggiornamento del numero di classi latenti.

- La scelta della mossa da effettuare è casuale: ognuna è scelta con probabilità 1/2.

Aggiornamento dei parametri

- Condizionatamente al valore corrente di C si aggiornano separatamente $\boldsymbol{\pi}$ e $\boldsymbol{\lambda}$ tramite mosse di tipo Metropolis-Hastings.

- **Aggiornamento di $\boldsymbol{\pi}$**: Si propone un nuovo valore dei parametri, $\boldsymbol{\pi}^\star$, estratto dalla distribuzione

$$\text{Dirichlet}(\boldsymbol{\delta})$$

che viene accettato con probabilità

$$\min\left\{1, \frac{L(\boldsymbol{y}|\boldsymbol{\pi}^\star, \boldsymbol{\lambda})p(\boldsymbol{\pi}^\star)q(\boldsymbol{\pi}|\boldsymbol{\pi}^\star)}{L(\boldsymbol{y}|\boldsymbol{\pi}, \boldsymbol{\lambda})p(\boldsymbol{\pi})q(\boldsymbol{\pi}^\star|\boldsymbol{\pi})}\right\}.$$

$L(\cdot|\cdot)$ verosimiglianza

$p(\cdot)$ distribuzione a priori

$q(\cdot|\cdot)$

- **Aggiornamento di $\boldsymbol{\lambda}$**: Si propone un nuovo valore dei parametri, $\boldsymbol{\lambda}^\star$, estraendo, per ogni j e c, $\lambda^\star_{j|c}$ dalla distribuzione

$$\text{Beta}(\boldsymbol{\rho});$$

$\boldsymbol{\lambda}^\star$ viene accettato con probabilità

Aggiornamento del numero di classi latenti

- Viene effettuata casualmente una delle due seguenti mosse (scelta casuale con probabilità 1/2):

  1. si tenta di combinare due classi latenti (*conbine*);
  2. si tenta di dividere una classe latente in due (*split*).

- **Conbine**: Si scelgono casualmente due classi latenti consecutive $(c, c+1)$ che si tenta di sostituire con un'unica classe latente, c', i cui parametri sono scelti come:

$$\begin{aligned} \pi_{c'} &= \pi_c + \pi_{c+1}, \\ \lambda_{j|c'} &= \lambda_{j|c+1}, \quad j = 1, \ldots, J. \end{aligned}$$

- **Split**: Si scegli casualmente una classe, c, che si prova a dividere in due classi latenti (c_1, c_2), ponendo:

$$\begin{aligned} &\pi_{c_1} = \pi_{c'} g, \quad \pi_{c_2} = \pi_{c'}(1-g), \\ &\lambda_{j|c_1} = h_j, \quad \lambda_{j|c_2} = \lambda_{j|c'}, \quad j = 1, \ldots, J, \end{aligned}$$

con $g \sim \text{Beta}(2, 4)$ e $h_j \sim \text{Beta}(1, 1)$.

- La mossa proposta (combine o split) viene accetta con probabilità

$$\min\left\{1, \frac{L(\boldsymbol{y}|\boldsymbol{\pi}^\star, \boldsymbol{\lambda}^\star, C^\star)p(\boldsymbol{\pi}^\star, \boldsymbol{\lambda}^\star|C^\star)q(\boldsymbol{\pi}, \boldsymbol{\lambda}|\boldsymbol{\pi}^\star, \boldsymbol{\lambda}^\star, C^\star)}{L(\boldsymbol{y}|\boldsymbol{\pi}, \boldsymbol{\lambda}, C)p(\boldsymbol{\pi}, \boldsymbol{\lambda}|C)q(\boldsymbol{\pi}^\star, \boldsymbol{\lambda}^\star|\boldsymbol{\pi}, \boldsymbol{\lambda}, C)}\mathcal{J}\right\}.$$

Inferenza sui parametri del modello

- L'algoritmo MCMC produce un campione dalla distribuzione a posteriori dei parametri del modello. Indichiamo questo campione con $(\boldsymbol{\pi}^{(t)}, \boldsymbol{\lambda}^{(t)}, C^{(t)})$, $t = 1, \ldots, T$, dove T è il numero delle iterazioni.

- Una stima della probabilità a posteriori del numero di classi C è:

$$\hat{\Pr}(C|\boldsymbol{y}) = \frac{T_C}{T}.$$

- Una stima "across-model" della numerosità della popolazione è data da

$$\hat{N} = \frac{1}{T}\sum_{t=1}^{T}\frac{n}{u^{(t)}}.$$

- La probabilità λ_j di essere catturati dalla lista j, si può stimare come:

$$\hat{\lambda}_j = \frac{1}{T}\sum_{t=1}^{T}\sum_{c=1}^{C^{(t)}}\lambda_{j|c}^{(t)}\pi_c^{(t)}.$$

Analisi dei dati sui diabetici

Modello a classi latenti	$\hat{\Pr}(C\|\boldsymbol{y})$	$\hat{N}$	95% CI per N
Inferenza Bayesiana			
$C = 3$	0.463	2616.0	(2275.5, 3154.0)
$C = 4$	0.370	2543.2	(2274.1, 3233.3)
$C = 5$	0.132	2534.3	(2269.1, 3355.3)
Across-model		2575.5	(2273.6, 3204.2)
Inferenza Classica			
Biggeri *et al.* (1999)		2696	(2502, 2950)
Bartolucci and Forcina (2001)		2403	(2280, 2585)

	$j = 1$	$j = 2$	$j = 3$	$j = 4$
$\hat{\lambda}_j$	0.69	0.18	0.45	0.07
95% CI	(0.55,0.78)	(0.14,0.21)	(0.35,0.51)	(0.05,0.08)

Bibliografia

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.

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

Data with a Rasch-type Model allowing for Conditional Dependence and Multidimensionality, *Biometrics*, **57**, pp. 714-719.

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.

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.

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.

Darroch, J. N., Fienberg, S. E., Glonek, G. F. V. and Junker, B. W.

ulation 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.

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

Tierney, L. (1994) Markov chains for exploring posterior distributions. *Annals of Statistics*, **22**, 1701–1762.

Tierney, L. and Mira, A. (1999). Some adaptive Monte Carlo methods for Bayesian inference. *Statistics in Medicine* **18**, 2507-2515.