The Analysis of Capture-Recapture data with a Rasch-type Model allowing for Conditional Dependence and Multidimensionality

Analisi di dati cattura-ricattura con modelli di tipo Rasch che ammettono dipendenza condizionata e multidimensionalità

Francesco Bartolucci
Dipartimento di Scienze Statistiche
Università di Perugia
bart@stat.unipg.it

Antonio Forcina
Dipartimento di Scienze Statistiche
Università di Perugia
forcina@stat.unipg.it

Riassunto: Nel presente lavoro viene proposta una classe di modelli orientati all'analisi di dati di tipo cattura-ricattura. Questi modelli possono essere visti come una generalizzazione del modello di Rasch, nel senso che sono ottenuti indebolendo parzialmente le ipotesi di base di tale modello: *unidimensionalità* e *indipendenza locale*. L'approccio proposto può essere utilizzato anche in presenza di strati creati sulla base di variabili esplicative discrete. Per la stima dei parametri si propone un uso congiunto degli algoritmi EM e Fisher-scoring e vengono discusse delle tecniche, basate sulla verosimiglianza profilo, per ottenere degli intervalli di confidenza per la dimensione della popolazione.

Keywords: EM algorithm; Fisher-scoring algorithm; Latent class model; Marginal modeling; Profile likelihood.

1. Introduction

A well-known method for estimating the size of wild populations is based on a sequence of trapping experiments so that, to each subject marked at least once, it is possible to associate a *capture configuration* $\boldsymbol{u} = (\, u_1 \quad \cdots \quad u_J \,)$ where J is the number of capture occasions and u_j is equal to 1 if the subject has been *captured at the* j*-th occasion* and 0 otherwise. Once the data have been prepared, they can be summarized into a contingency table with $t = 2^J - 1$ cells, since the entry corresponding to individuals that were never captured, obviously, cannot be filled.

Statistical models for analyzing such data have a long history and they are also applied to various social contexts where it is difficult or expensive to directly count individuals with a certain feature (disease, social problem); for a review see Schwarz and Seber (1999). Darroch *et al.* (1993) and Agresti (1994) were among the first to propose the use of the Rasch model. This model, originally conceived to analyze the results of an aptitude test assigned to a group of subjects, is based essentially on two assumptions: *unidimensionality* and *local independence*. Though the Rasch model provides frequently an adequate fit and has the advantage that the parameters can be easily interpreted, its basic assumptions may be too restrictive.

In this paper, we review a class of models, introduced by Bartolucci and Forcina (2001), that may be seen as an extension of the Rasch model which allows for lack of unidimensionality and conditional association between pairs of lists. This class of model is described in the following Section whereas maximum likelihood estimation is discussed in Section 3. Finally, the construction of confidence intervals for the population size is described in Section 4.

2. The proposed class of models

Assume that subjects have been drawn at random from a population composed of C latent classes and let $\lambda_{c,\boldsymbol{u}}$ denote the probability that any subject in latent class c experiences the capture configuration $\boldsymbol{u}$. According to the Rasch model we have that

$$\lambda_{c,\boldsymbol{u}} = \prod_j \frac{e^{u_j(\phi_c+\psi_j)}}{1+e^{\phi_c+\psi_j}}, \tag{1}$$

where the *subject parameter* ϕ_c may be interpreted as a measure of the tendency to be captured for individuals belonging to class c, while the *list parameter* ψ_j may be seen as a measure of the effectiveness of list j. Moreover, the so-called *manifest distribution* is given by

$$q_{\boldsymbol{u}} = \Pr(\boldsymbol{u}) = \sum_c \pi_c \lambda_{c,\boldsymbol{u}}$$

where π_c is the probability of belonging to class c. So, we have two important restrictions: (i) the ordering of latent classes with respect to the catchability is the same for any list (*unidimensionality*); (ii) given the latent class, the events of being captured by different lists are independent (*local independence*). However, (i) might hold only within disjoint subsets of lists (see also Darroch *et al.*, 1993) or, even conditionally on the latent class, the probability of appearing in a given list might be larger or smaller if the subject has already appeared in a related list. In these cases, a version of the Rasch model in which restrictions (i) and/or (ii) are relaxed should be used. To implement a model of this kind we adopt a *marginal modeling* approach (e.g. Bergsma, 1997, Ch. 4) which is briefly described in the following.

In the general case that subjects are divided into S strata according to one or more explanatory variables, let $\lambda_{s,c,\boldsymbol{u}}$, $\pi_{s,c}$ and $q_{s,\boldsymbol{u}}$ be the same quantity introduced above and referred to a certain stratum s. Then, denote by $\boldsymbol{\lambda}_{s,c}$ and $\boldsymbol{q}_s$ the $(t+1)\times 1$ vectors with elements $\lambda_{s,c,\boldsymbol{u}}$ and $q_{s,\boldsymbol{u}}$ for any $\boldsymbol{u}$, respectively; we follow the general convention that the elements of these vectors are ordered so that the entries of the binary vector $\boldsymbol{u}$ with a larger index run faster from 0 to 1. We now construct a vector $\boldsymbol{\eta}$ for specifying the manifest distribution $\{\boldsymbol{q}_s\}$. Let $\boldsymbol{\eta} = (\,\boldsymbol{\eta}_1' \quad \boldsymbol{\eta}_2'\,)'$ where $\boldsymbol{\eta}_1$ contains the logits of belonging to the different latent classes within each stratum and $\boldsymbol{\eta}_2$ contains the J univariate marginal (relative to other lists) logits, the $J(J-1)/2$ bivariate marginal log-odds ratios and higher order effects of any conditional distribution given s and c. In practice, $\boldsymbol{\eta}_2$ is obtained by stacking the vectors $\boldsymbol{\eta}_{2,s,c} = \boldsymbol{C}\log(\boldsymbol{M}\boldsymbol{\lambda}_{s,c})$ where the matrices $\boldsymbol{C}$ and $\boldsymbol{M}$ may be simply constructed (see the Appendix of Bartolucci and Forcina, 2001). The vector $\boldsymbol{\eta}$ is similar to the *linear predictor* of a generalized linear model in the sense that meaningful assumptions may be expressed by linear constraints, that is in the form $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$. In this way we may easily parametrize not only the Rasch model, but also a multidimensional Rasch model, a Rasch model compatible with certain bivariate associations specified conditionally on the latent classes or a full latent class model. Obviously, any of these models may be compared with the Rasch model through a likelihood ratio statistic that, under the null, has, as usual, asymptotic distribution of chi-square type with appropriate degrees of freedom. We may also implement the constraint that a subset of strata shares the same parameters so that, by modeling these simultaneously, a more parsimonious model may be found (see Bartolucci and Forcina, 2001, Sec. 3, for details).

3. Maximum likelihood estimation

Let $y_{s,\boldsymbol{u}}$ be the number of subjects in stratum s with capture configuration $\boldsymbol{u}$ and $n_s = \sum_{\boldsymbol{u}\neq\boldsymbol{0}} y_{s,\boldsymbol{u}}$ be the number of captures within stratum s. If we assume that the frequencies $\{y_{s,\boldsymbol{u}}\}$ in the hypothetical table including the missing cell follow a multinomial distribution, $\boldsymbol{\beta}$ may be estimated by maximizing the conditional likelihood, given $\{n_s\}$, whose logarithm is

$$l_y = \sum_s \sum_{\boldsymbol{u}\neq\boldsymbol{0}} y_{s,\boldsymbol{u}} \log(q_{s,\boldsymbol{u}}/r_s), \quad \text{with} \quad r_s = \sum_{\boldsymbol{u}\neq\boldsymbol{0}} q_{s,\boldsymbol{u}};$$

see Sanathanan (1972) and Darroch *et al.* (1993). The conditional estimator of N_s, where N_s denotes the size of the population in stratum s, is the integer part of $n_s/\widehat{r}_s$ which maximizes the binomial likelihood. This can be usefully complemented with a confidence interval derived on the basis of the unconditional profile likelihood (see the following Section). To maximize l_y with respect to $\boldsymbol{\beta}$ we propose to run the EM algorithm for a few steps and then use its estimates as starting values for the Fisher-scoring algorithm.

The Fisher-scoring algorithm consists in updating the estimate at step $h+1$ as

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

where $\boldsymbol{\beta}^h$ is the estimate at step h and $\boldsymbol{g}^h$ and $\boldsymbol{F}^h$ are, respectively, the first derivative vector and the expected information matrix evaluated at step h.

The EM algorithm alternates the following steps until convergence:

- *E-step*: on the basis of the current value of the estimate of $\boldsymbol{\beta}$, compute the conditional expected value of $\{x_{s,c,\boldsymbol{u}}\}$, given $\{y_{s,\boldsymbol{u}}\}$, where $x_{s,c,\boldsymbol{u}}$ denotes the number of subjects within stratum s which belong to latent class c and experience configuration $\boldsymbol{u}$;
- *M-step*: update the estimate of $\boldsymbol{\beta}$ by maximizing the complete log-likelihood,

$$l_x = \sum_s \sum_c \sum_{\boldsymbol{u}} x_{s,c,\boldsymbol{u}} \log(\pi_{s,c}\lambda_{s,c,\boldsymbol{u}}),$$

  with the frequencies $\{x_{s,c,\boldsymbol{u}}\}$ replaced by the corresponding expected values computed at previous step; this can be performed again by a Fisher-scoring algorithm.

We follow this approach, based on the combined use of the two algorithms, since the Fisher-scoring is much faster than the EM but extremely sensitive to starting values. We experimented that 10 steps of the EM are usually enough to obtain good starting values for the Fisher-scoring.

4. Likelihood-based confidence intervals

With a single stratum, a confidence interval for the population size N may be determined from the profile unconditional log-likelihood as follows (see also Cormack, 1992):

1. for each value of N varying on a suitable interval of integers, evaluate the unconditional profile log-likelihood, $l_y(N)$, that is the maximum of

$$\log(N!) - \sum_{\boldsymbol{u}} \log(y_{\boldsymbol{u}}!) + \sum_{\boldsymbol{u}} y_{\boldsymbol{u}} \log(q_{\boldsymbol{u}}), \quad \text{with} \quad y_{\mathbf{0}} = N - n;$$

2. find the unconditional estimate of N, $\widehat{N}_U$, as the value of N that maximizes $l_y(N)$;
3. a confidence interval at level $100(1-\alpha)\%$ for N is given by (N_1, N_2) where N_1 and N_2 are chosen, respectively, as the largest integer smaller than $\widehat{N}_U$ and the smallest integer greater than $\widehat{N}_U$, so that $D(N_1) \geq \chi^2_{1,\alpha}$ and $D(N_2) \geq \chi^2_{1,\alpha}$, where $D(N) = 2\{l_y(\widehat{N}_U) - l_y(N)\}$ and $\chi^2_{1,\alpha}$ is the $100\alpha\%$ critical value on the χ^2_1 distribution.

When there are two or more strata we suggest to set up a confidence interval for any N_s by performing the procedure above with $l_y(N)$ replaced by $l_{y,s}(N_s)$, where $l_{y,s}(N_s)$ is the unconditional profile log-likelihood where each N_h, for $h \neq s$, has been replaced by the conditional estimate $n_h/\hat{r}_h$. A refinement of this procedure consists in replacing N_h, for $h \neq s$, with a more appropriate value than the conditional estimate, so that the nominal confidence level is surely respected. This new procedure and its performances are currently under investigation.

References

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

Bartolucci F. and Forcina A. (2001) The analysis of capture-recapture data with a rasch-type model allowing for conditional dependence and multidimensionality, *Biometrics*, 57, 714–719.

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

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.

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

Schwarz C.J. and Seber A.F. (1999) Estimating animal abundance: review iii, *Statistical Science*, 14, 427–456.