Mixtures of regressions for size estimation of heterogeneous populations

Misture di regressioni per la stima della numerosità di popolazioni eterogenee

Gianmarco Caruso

Abstract We propose a capture-recapture model which exploits finite mixtures of logistic regressions to account for latent heterogeneity between groups of individuals, in order to better understand their different propensities to the capture as well as different behavioral patterns. The additional behavioural variation in capture probabilities among individuals within a group is expressed by a suitable time-dependent covariate, which summarises the past individual experience [3]. A real data example and a simulation study illustrate how the proposed model performs.
Abstract *Si propone un modello cattura-ricattura che sfrutta le misture finite di regressioni logistiche per spiegare l'eterogeneità latente tra gruppi di individui, al fine di comprendere meglio le loro differenti propensioni alla cattura. La variabilità tra le probabilità di cattura di individui appartenenti ad uno stesso gruppo viene espressa mediante un'adeguata covariata tempo-dipendente, che riassume l'esperienza individuale passata [3]. Le potenzialità del modello proposto vengono illustrate attraverso un esempio basato su dati reali e uno studio di simulazione.*

Key words: capture-recapture, population size estimation, finite mixtures of GLM, logit regression.

1 Introduction

Capture-recapture methods are widely employed in estimating the size of elusive populations, whose units are subject to multiple captures across several occasions.

The main idea behind these techniques is to account for the number of unobserved individuals by suitably modelling and exploiting the capture histories of the observed units. One assumes that a closed population of unknown size N is sampled

Gianmarco Caruso
Dipartimento di Scienze Statistiche, La Sapienza Università di Roma, Italy
e-mail: gianmarco.caruso@uniroma1.it

t times, with independence between individuals. For example, in the common case of wildlife populations, animals that are captured for the first time are marked and then released, so that they can be recognizable in future trapping occasions. Supposing that M distinct individuals have been captured across t occasions, data are collected on a $M \times t$ matrix, $\boldsymbol{X} = [x_{ij}]$: in particular, $x_{ij} = 1$ if individual i is captured on occasion j, otherwise $x_{ij} = 0$. The i-th row of the matrix reports the capture history of the i-th individual. If there are N individuals in the population, then one can add $N - M$ rows of zeros to the matrix in order to include all the uncaptured individuals. In the following, one supposes to deal with closed populations, where there are no births, no deaths and no migrations: this assumption seems to be meaningful if the first and the last capture occasions are not too far in time and the range where the population lives is well bounded.

2 The model

One considers a model which allows capture probabilities to vary among individuals and across capture occasions. In addition, here one considers the presence of unobserved heterogeneity between groups of individuals, in the sense that different groups may exhibit different responses to captures. Finite mixtures of logistic regressions are thus exploited to account for latent heterogeneity and to better understand different responses by heterogeneous groups of individuals. The additional variation in capture probabilities among individuals within each group may be expressed by a suitable time-dependent covariate, which summarises the past individual experience [6, 3].

In the following, one considers a heterogeneous population of N individuals which can be partitioned in G subpopulations (or groups), $\mathscr{P}_1, \dots, \mathscr{P}_G$; namely, the N individuals are supposed to come from G different subpopulations of unknown proportions, $\pi_1, \dots, \pi_G$, which are non-negative and add up to 1. The proportion π_g represents the *a priori* probability for an individual to belong to the g-th subpopulation. The observed response x_{ij} is therefore supposed to be generated by a finite mixture of logistic regressions [11], where the mixture is assumed to be formed by G components: hence, each mixture component identifies a different group.

Conditional to the group g, the response at occasion j for individual i is given by

$$x_{ij}|p_{ij}^{(g)} \sim Bern\left(x_{ij}\middle|p_{ij}^{(g)}\right), \tag{1}$$

where $p_{ij}^{(g)}$ is the probability of being captured at occasion j for the i-th individual belonging to the g-th cluster ($i \in \mathscr{P}_g$).

If $\boldsymbol{\pi} = (\pi_1, \dots, \pi_G)$ is the vector of mixture weights, the unconditional probability distribution of x_{ij} is given by

$$h\left(x_{ij}\middle|\boldsymbol{\pi}, \{p_{ij}^{(g)}\}_{g=1,\dots,G}\right) = \sum_{g=1}^{G} \pi_g \, Bern\left(x_{ij}\middle|p_{ij}^{(g)}\right), \tag{2}$$

for $i = 1,\dots,N$ and $j = 1,\dots,t$.

The capture probability $p_{ij}^{(g)}$ depends on the group-specific regression parameters α_g and β_g and on the value of the covariate z_{ij}, according to a linear logistic model, namely

$$p_{ij}^{(g)} = \frac{\exp(\alpha_g + \beta_g z_{ij})}{1 + \exp(\alpha_g + \beta_g z_{ij})}\,, \tag{3}$$

$\forall i \in \mathscr{P}_g$, $g = 1,\dots,G$, $j = 1,\dots,t$ [9, 2]. The heterogeneity between groups of individuals is given by differences in the group-specific regression parameters which connect the covariate to the conditional expected value of the response: thus, same levels of the covariate affect the probabilities of recapture of individuals in distinct groups in different ways.

The time-varying covariate matrix $\boldsymbol{Z} = [z_{ij}]$ can be derived by exploiting the class of memory-related summaries introduced by [3], so that

$$z_{ij} = g_\lambda(x_{i1},\dots,x_{ij-1}) = \sum_{h=1}^{j-1} \frac{\lambda^{h-1}}{\sum_{k=1}^{j-1} \lambda^{k-1}}\, x_{ih}\,, \tag{4}$$

which takes values in $[0,1]$. Notice that $z_{ij} = 0$ for all partial capture histories such that $(x_{i1},\dots,x_{ij-1}) = (0,\dots,0)$ and, conventionally, for $j = 1$ (i.e. the first column of the matrix $\mathbf{Z}$ is composed by all zeros).

As discussed by [3], z_{ij} represents a weighted average of the past trapping experience for the individual i based on the first $j-1$ occasions. In particular, for $\lambda = 1$, all past captures has the same impact on the summary, while, for $\lambda > 1$, most recent captures have a greater impact on the summary. A positive value of β_g accounts for trap-happiness type of response to capture while a negative value accounts for trap-shyness.

3 Unconditional maximum likelihood estimation

Following [10], if $\boldsymbol{P} = \left[p_{ij}^{(g)}\right]$ is the matrix of capture probabilities, the unconditional likelihood for the model (2) is

$$L(N,\boldsymbol{P},\boldsymbol{\pi}) = \frac{N!}{(N-M)!} \prod_{i=1}^{N} \prod_{j=1}^{t} \sum_{g=1}^{G} \pi_g \left[p_{ij}^{(g)}\right]^{x_{ij}} \left[1 - p_{ij}^{(g)}\right]^{1-x_{ij}}. \tag{5}$$

Once the number of mixture components G is fixed, inference on N is made through iterative fitting of the mixture of logistic regressions for each $N \in \{M,\dots,N_{\max}\}$, where $N_{\max}$ is a high fixed upper bound for the population size [3]. The unconditional MLE (UMLE) for N is then the maximizer of the profile likelihood function

$$\hat{L}(N) = L\big(N, \hat{\boldsymbol{P}}(N), \hat{\boldsymbol{\pi}}(N)\big) = \sup_{\boldsymbol{\pi},\boldsymbol{P}} L(N,\boldsymbol{P},\boldsymbol{\pi})\,, \tag{6}$$

where the matrix $\boldsymbol{P}$ is function of the regression parameters $\alpha_1, \ldots, \alpha_G, \beta_1, \ldots, \beta_G$. Details about fitting of finite mixtures of GLMs are available in [7].

4 Illustration

A real data example and a simulation study are presented in the following, in order to show how the proposed model performs.

4.1 Real data example

One considers a data set coming from a survey in which snowshoe hares (*Lepus americanus*) were repeatedly captured during 6 consecutive days of trapping by using animal-baited traps. At the end of the sixth day, the number of observed individual hares was 68. The considered dataset has already been analysed by some authors (e.g. [1, 5]) and it is available in *R* package `Rcapture`.

The proposed model is fitted to hares' capture histories for different numbers of mixture components ($G = 1,2,3$) and for different values of λ (i.e. $\lambda = 1,2$). The choice of $\lambda = 1$ yields a time-dependent covariate which represents the relative frequency of the previous capture occurrences, while $\lambda = 2$ yields to a time-dependent covariate which enjoys a connection with Markov models [3]. For fixed G and λ, several finite mixtures of logit regressions are fitted for a set of candidate values of N, by using the functions in the R package `flexmix`: in particular, the function `initFlexmix` allows to repeat the EM algorithm with different starting values and chooses the solution which maximizes the likelihood.

The results displayed in Table 1 show that the models associated with the lowest values of the AIC are the ones corresponding to $G = 2$ components. This is somewhat expected since other authors - like [5] - have already shown the presence of groups of hares with different capture rates. The model with $G = 2$ and $\lambda = 1$ yields $\hat{\alpha}_1 = -1.45$, $\hat{\beta}_1 = 4.12$, $\hat{\alpha}_2 = -0.75$ (all of them associated to a p-value smaller than 7×10^{-3}) and $\hat{\beta}_2 = -0.75$, which appears not to be significantly different by 0 ($p = 0.28$). These results suggest that initial trap-happiness characterises the first group of hares, while for the second group no sufficient evidence of behavioural effects is provided. This indicates that a more parsimonious two-components mixture model with only one group manifesting behavioural effects could be further elaborated.

The 90% profile likelihood confidence intervals are built following [4], who highlights their advantages in a mark-recapture context. Notice that as the number of components increases, the confidence intervals tend to get wider, due to the flatter shape of the corresponding profile log-likelihood. This feature is probably due to the fact that the information provided by the data is insufficient to establish any upper bound on the number of animals, above all when a complex model is fitted on data coming from a relative low number of occasions [8].

Table 1: Unconditional maximum likelihood estimates for the population size, 90% confidence intervals and AIC index associated with alternative fitted models for different values of G and λ.

G	λ	$\hat{N}$	$(N_{\text{low}}, N_{\text{upp}})$	AIC
1	1	80	(73, 94)	81.53
	2	78	(72, 89)	83.20
2	1	79	(71, 197)	**75.72**
	2	75	(70, 111)	76.69
3	1	80	(72, 178)	81.39
	2	76	(70, 146)	82.09

4.2 Simulation study

Motivated by the results of the previous example, a simulation study is carried out in order to assess the ability of the proposed model in estimating the population size. Capture histories are generated for two subpopulations of individuals (thus $G = 2$) and collected binary entries matrix with $N = 100$ rows and t columns, where $N - M$ rows have zero entries. The probability that an individual belongs to the first group is $\pi_1 = 0.33$ and the regression parameters are set to $\alpha_1 = -3$, $\beta_1 = -2$, $\alpha_2 = -3$ and $\beta_2 = 4$. Since the probability of first capture is completely determined by the value of the intercept α, one is implicitly assuming that the first capture probability is the same for all the individuals of the population, regardless of the group they belong to. The replication of 20 simulated datasets has been carried out, for different time-dependent covariate specifications ($\lambda = 1, 2$) and for different number of occasions ($t = 15, 30$). For each data set, the true data-generating process is fitted to the data. From the results reported in 2, it appears that the the empirical confidence intervals coverage is consistent with its theoretical counterpart. The population size seems to be slightly overestimated, though the bias decreases with the number of occasions, as expected.

Table 2: Simulation study with 20 simulated data sets for several model specifications, determined by different numbers of occasions ($t = 15, 30$) and different values of λ. The table contains: average and median of the UML estimates of N (respectively, N_{ave} and N_{med}), root mean square error (*rmse*), percentage of 95% confidence intervals coverage (*CI coverage*), average length of the confidence intervals (l_{CI}).

t	λ	N_{ave}	N_{med}	l_{CI}	CI coverage	rmse
15	1	110.0	88.0	122.6	0.95	49.0
	2	113.9	120.0	71.5	0.95	31.7
30	1	104.7	104.5	56.2	0.90	14.5
	2	108.9	100.5	46.4	0.95	22.7

5 Final remarks and further developments

The proposed model appears a flexible extension of the one proposed in [3], allowing for the presence of latent heterogeneity between groups of individuals by means of group-specific regression parameters. Some possible further developments should involve a more in-depth study of the groups composition, along with a more flexible and parsimonious model which accounts for the possibility that some groups are not subject to behavioural effects, as suggested from the real data example. Moreover, a more extensive simulation study should be carried out, mainly in order to assess whether a model misspecification could be correctly identified when the population is composed by heterogeneous groups. Still through simulation studies, it can be interesting to investigate whether the better performances (in terms of AIC) of the proposed model on real data are indeed reliable; or whether, on the other hand, the AIC may tend to favour one model against the other. A Bayesian alternative might be proposed too, in order to overcome possible annoying problems due to the flatness of the profile likelihood when G is large.

References

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

[2] Alho, J. M. (1990). Logistic regression in capture-recapture models. *Biometrics*, pages 623–635.

[3] Alunni Fegatelli, D. and Tardella, L. (2016). Flexible behavioral capture–recapture modeling. *Biometrics*, 72(1):125–135.

[4] Cormack, R. M. (1992). Interval estimation for mark-recapture studies of closed populations. *Biometrics*, pages 567–576.

[5] Dorazio, R. M. and Andrew Royle, J. (2003). Mixture models for estimating the size of a closed population when capture rates vary among individuals. *Biometrics*, 59(2):351–364.

[6] Farcomeni, A. (2011). Recapture models under equality constraints for the conditional capture probabilities. *Biometrika*, 98(1):237–242.

[7] Grün, B. and Leisch, F. (2007). Fitting finite mixtures of generalized linear regressions in r. *Computational Statistics & Data Analysis*, 51(11):5247–5252.

[8] Hirst, D. (1994). An improved removal method for estimating animal abundance. *Biometrics*, pages 501–505.

[9] Huggins, R. (1989). On the statistical analysis of capture experiments. *Biometrika*, 76(1):133–140.

[10] Pledger, S. (2000). Unified maximum likelihood estimates for closed capture–recapture models using mixtures. *Biometrics*, 56(2):434–442.

[11] Wedel, M. and DeSarbo, W. S. (1995). A mixture likelihood approach for generalized linear models. *Journal of classification*, 12(1):21–55.