

**A class of latent marginal models  
for capture-recapture data  
with continuous covariates**

F. Bartolucci

*Università di Urbino*

Francesco.Bartolucci@uniurb.it

A. Forcina

*Università di Perugia*

forcina@stat.unipg.it

# Outline

- Capture-recapture method and structure of a capture-recapture dataset
- Existing models for the analysis of capture-recapture data
- Proposed class of models
- Estimation of parameters and population size through the EM algorithm
- Computation of confidence intervals for the population size
- Application to a dataset provided by the Tuscany Cancer Registry

# Capture-recapture method

- The method was born for the estimation of the size of wild populations and now is commonly used for the estimation of the size of human populations when it is very difficult to directly count individuals (people who break a law or suffer of a certain disease)
- It is based on a series of  $(J)$  *trapping experiments* (or *lists*) so that to any subject captured at least once we can associate a *capture configuration*  $\mathbf{r} = (r_1, \dots, r_J)$  with
  - ▷  $r_j = 1$  if the subject has been captured at the  $j$ -th occasion
  - ▷  $r_j = 0$  otherwise
- We deal with closed populations when also a vector of covariates  $\mathbf{z} = (z_1, \dots, z_m)$  is available for any subject in the sample

# Data organization

- Let  $s$  be the number of distinct covariate configurations  $(z_1, \dots, z_s)$  anyone of which identifies a stratum
- For  $i = 1, \dots, s$  we denote by:
  - ▷  $n_i$  the number of subjects in the sample with covariate configuration  $z_i$
  - ▷  $y_{i,r}$  the number of these subjects with capture configuration  $r$
  - ▷  $\mathbf{y}_i$  the  $(2^J - 1)$ -dimensional vector of capture frequencies  $y_{i,r}$  for any  $r \neq \mathbf{0}$
- When  $n_i = 1$  (only one subject has covariate configuration  $z_i$ ), all the elements of  $\mathbf{y}_i$  are equal to 0 apart from an element equal to 1 corresponding to the capture configuration of this subject



# Analysis of capture-recapture data

- Most used models:
  - ▷ Log-linear models (Fienberg, 1972, Cormack, 1989)
  - ▷ Latent class (LC) model (Cowan & Malec, 1986)
  - ▷ LC version of the Rasch model (Darroch, 1993, Agresti, 1994)
- The most interesting approaches are those based on the LC model since this model is easily interpretable and takes explicitly into account the heterogeneity between subjects

# Latent class model

- Basic assumptions (Goodman, 1974):

- 1) The population is divided into  $c$  *latent classes* so that the distribution of the capture configuration is the same for all the subjects in the same class
- 2) Given the latent class, the probability of being captured at a certain occasion does not depend on the the event of being captured at other occasions (*local independence*, LI). This implies that

$$p(\mathbf{r}|h) = \prod_j \lambda_{j|h}^{r_j} (1 - \lambda_{j|h})^{1-r_j}, \quad h = 1, \dots, c$$

- ▷  $\lambda_{j|h} = p(r_j = 1|h)$ : probability that a subject in latent class  $h$  is captured at the  $j$ -th occasion

# Rasch model (in the latent class version)

- It may be seen as a constrained version of the LC model (Rasch, 1961, Lindsay *et al.*, 1991) in which

$$\log \frac{\lambda_{j|h}}{1 - \lambda_{j|h}} = \phi_h + \psi_j, \quad h = 1, \dots, c$$

- ▷  $\phi_h$ : tendency to be captured of the subjects in latent class  $h$
- ▷  $\psi_j$ : effectiveness of the list  $j$  in capturing subjects
- This constraint implies *unidimensional monotony* since we may always relabel the latent classes so that

$$\lambda_{j|1} \leq \lambda_{j|2} \leq \dots \leq \lambda_{j|c}, \quad j = 1, \dots, J;$$

this makes easier the interpretation of the latent structure

# Extensions of the Latent Class model

- Main limitations of the LC model in the capture-recapture context:
  - ▷ the assumption of LI may be too restrictive (e.g. *behavioral effect*: consequence of a capture on the behavior of a subject)
  - ▷ the model does not take into account available covariates
- To overcome the first limitation the LC model has been extended in several directions:
  - ▷ taking into account only the behavioral effect (Pledger, 2000)
  - ▷ allowing *marginal interactions* between capture occasions given the latent class (Bartolucci & Forcina, 2001)
  - ▷ allowing *log-linear interactions* between capture occasions given the latent class (Stanghellini & van der Heijden, 2004)

- Within the approaches of Bartolucci & Forcina (2001) and Stanghellini & van der Heijden (2004) it is also possible to take into account (only discrete) covariates
- We propose a general approach in which it is possible to:
  - ▷ relax in a flexible way the assumption LI by allowing marginal and/or log-linear interactions between capture occasions given the latent class
  - ▷ take into account also continuous covariate
- The approach is based on recent advances on marginal parametrizations (Bergsma & Rudas, 2002) and may be seen as an extension of the *latent regression* approach of Bandeen-Roche *et al.* (1997) and Huang & Bandeen-Roche (2004) in which LI is retained

# Basic assumptions of the proposed approach

- 1) The population is divided into  $c$  latent classes so that the distribution of the capture configuration is the same for all the subjects in the same class and with the same covariate configuration
- 2) The following regression model holds

$$\boldsymbol{\eta}_i = \mathbf{X}_i \boldsymbol{\beta}, \quad i = 1, \dots, s$$

- ▷  $\boldsymbol{\eta}_i$ : vector of parameters of interest for the joint distribution of capture configuration and latent class given the covariate configuration  $\mathbf{z}_i$
- ▷  $\mathbf{X}_i$ : design matrix depending on  $\mathbf{z}_i$
- ▷  $\boldsymbol{\beta}$ : vector of regression parameters

# Marginal parametrization

- We make use of a *marginal parametrization* of the type (Colombi and Forcina, 2001)

$$\eta_i = \mathbf{C} \log(\mathbf{M} \boldsymbol{\pi}_i)$$

- ▷  $\boldsymbol{\pi}_i$ : vector of dimension  $c 2^J$  with elements  $\pi_{i,h,r}$  equal to the probability that a subject with covariate configuration  $\mathbf{z}_i$  experiences capture configuration  $\mathbf{r}$  and is in latent class  $h$
  - ▷  $\mathbf{C}$ : matrix of contrasts
  - ▷  $\mathbf{M}$ : matrix of aggregation with elements 0 or 1
- We require that the parametrization is *ordered decomposable* (Bergsma & Rudas, 2002), so that it is variation independent with advantages in the estimation process

- Typically the elements of  $\eta_i$  are
  - ▷ univariate logits for the marginal distribution of the latent classes
  - ▷ a logit for any capture occasion conditional on the latent class but marginal with respect to other occasions
  - ▷ possible log-odds ratios (and higher order-interactions) between capture occasions conditional on the latent class when we want to relax the assumption of LI
- Through  $\eta_i = \mathbf{X}_i\boldsymbol{\beta}$ ,  $\forall i$ , we may assume that the covariates affect
  - ▷ only the distribution of the latent classes
  - ▷ only the conditional distribution of the response configuration given the latent class
  - ▷ both distributions
  - ▷ none of them (LC model, Rasch model in its LC version)

# Estimation of the regression parameters

- The vector of regression parameters  $\beta$  is estimated by maximizing the *conditional multinomial likelihood* (Sanathanan, 1972), given the number of captures for any  $z_i$ ,  $i = 1, \dots, s$ , whose logarithm is

$$\ell(\beta) = \sum_i \sum_{\mathbf{r} \neq \mathbf{0}} y_{i,\mathbf{r}} \log \left( \frac{p_{i,\mathbf{r}}}{q_i} \right)$$

- ▷  $p_{i,\mathbf{r}}$ : manifest probability of capture configuration  $\mathbf{r}$  for a subject with covariate configuration  $z_i$
  - ▷  $q_i = \sum_{\mathbf{r} \neq \mathbf{0}} p_{i,\mathbf{r}}$ : probability that a subject of this type is captured at least once
- For the maximization of  $\ell(\beta)$  we have implemented an EM algorithm whose M step is, in turn, based on a Fisher-scoring algorithm

# Estimation of the population size

- On the basis of the estimate of  $\beta$  ( $\hat{\beta}$ ), the number of subjects in the population with covariate configuration  $z_i$  is estimated as (Alho, 1990)

$$\hat{t}_i = \frac{n_i}{\hat{q}_i} > n_i$$

- ▷  $\hat{q}_i = \sum_{r \neq 0} \hat{p}_{i,r}$ : estimate of the probability that a subject with covariate configuration  $z_i$  is captured at least once
- The overall population size and the number of subjects never captured are then estimated as

$$\hat{N} = \sum_i \hat{t}_i \quad \text{and} \quad \hat{N} - n = \sum_i (\hat{t}_i - n_i)$$

# EM algorithm

- It is based on the concept of *complete data* corresponding, in this context, to the frequencies  $m_{i,h,r}$  of any capture configuration  $\mathbf{r}$ , any latent class  $h$  and any covariate configuration  $\mathbf{z}_i$
- The algorithm alternates two steps until convergence in  $\ell(\boldsymbol{\beta})$ :
  - E**: compute the conditional expected value of any  $m_{i,h,r}$  given the observed data  $y_{i,r}$  and the current value of  $\boldsymbol{\beta}$
  - M**: update  $\boldsymbol{\beta}$  by maximizing the log-likelihood of the complete data

$$\ell^*(\boldsymbol{\beta}) = \sum_i \sum_h \sum_{\mathbf{r}} m_{i,h,\mathbf{r}} \log(\pi_{i,h,\mathbf{r}})$$

with any  $m_{i,h,\mathbf{r}}$  substituted with the corresponding expected value computed during the E step

# Confidence interval for the population size

- The first approach that has been used for computing a  $100(1 - \alpha)\%$  confidence interval for  $N$  is based on the delta method

$$(\hat{N} - z_{\alpha/2} \hat{se}(\hat{N}); \quad \hat{N} + z_{\alpha/2} \hat{se}(\hat{N}))$$

- ▷  $\hat{se}(\hat{N})$ : estimate of the standard error of  $\hat{N}$  based on the Fisher information matrix of the parameters (Sanathanan, 1972)
- The resulting intervals are symmetric around  $\hat{N}$  and usually unnecessarily wide
- A better method, based on the *profile unconditional likelihood*, has been proposed by Cormack (1992) and extended by Stanghellini & van der Heijden (2004) to the case in which discrete covariates are present

# Proposed method for confidence intervals

- We make use of the statistic

$$D(\mathbf{t}, \boldsymbol{\beta}) = 2 \sum_i \left[ (t_i - n_i) \log \left( \frac{t_i - n_i}{t_i p_{i,0}} \right) + \sum_{r \neq 0} y_{i,r} \log \left( \frac{y_{i,r}}{t_i p_{i,r}} \right) \right]$$

corresponding to the *hypothetical* deviance of the selected model under the assumption that  $t_i$  is known for any stratum  $i$

- The minimum of  $D(\mathbf{t}, \boldsymbol{\beta})$  is reached at  $\mathbf{t} = \hat{\mathbf{t}}$  and  $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$
- We also consider the discrepancy statistic

$$G^2(N) = \min_{\mathbf{t}' \mathbf{1} = N} D(\mathbf{t}, \boldsymbol{\beta}) - D(\hat{\mathbf{t}}, \hat{\boldsymbol{\beta}})$$

that may be computed through a simple algorithm that alternates two steps (minimization with respect to  $\mathbf{t}$  and then with respect to  $\boldsymbol{\beta}$ )

- If  $N$  is equal to the true value of the population size

$$G^2(N) \sim \chi^2(1) \quad \text{as } N \rightarrow \infty$$

provided that the proportion of subjects in any stratum does not become negligible

- A confidence interval at level  $100(1 - \alpha)\%$  for  $N$  is then  $(N_1, N_2)$ 
  - ▷  $N_1$ : largest integer ( $< \hat{N}$ ) such that  $G^2(N_1) \geq \chi_\alpha^2(1)$
  - ▷  $N_2$ : smallest integer ( $> \hat{N}$ ) such that  $G^2(N_2) \geq \chi_\alpha^2(1)$
- A similar procedure may be used for computing confidence intervals for certain subpopulations (males, females)
- Asymptotically the approach is equivalent to that based on the profile likelihood, but it is simpler to implement

# Analysis of cancer registry dataset

- We first considered a latent marginal model based on the assumptions
  - ▷ local independence
  - ▷ A and S affect the distribution of the latent classes
  - ▷ A and S affect the distribution of the capture configuration given the latent with regression coefficients constant across latent classes (to ensure identifiability; Huang & Bandeen-Roche, 2004)
- For this model we chose  $c = 3$  latent classes on the basis of BIC

$c$	$d(c)$	$\hat{\ell}_C(c)$	BIC
1	9	-9749.4	19579
2	15	-9320.1	18773
3	21	-9235.1	18657
4	27	-9225.5	18691

- Through a series of likelihood ratio tests we then selected the model in which (deviance decreases of 2.7 with 1 d.f.):
  - ▷ the logit for being in latent class 2 against latent class 1 does not depend on A
  - ▷ the regression coefficient for S on the conditional logits of P given the latent is the same of that for the logits of H
  - ▷ association between H and P conditional on the latent is allowed
- We have to reject the hypotheses:
  - ▷ A and S affect only the distribution of the latent (deviance with respect to the initial model 188.4 with 6 d.f.)
  - ▷ A and S affect only the conditional distribution of the capture configuration given the latent (deviance 211.6 with 4 d.f.)

## Parameter estimates

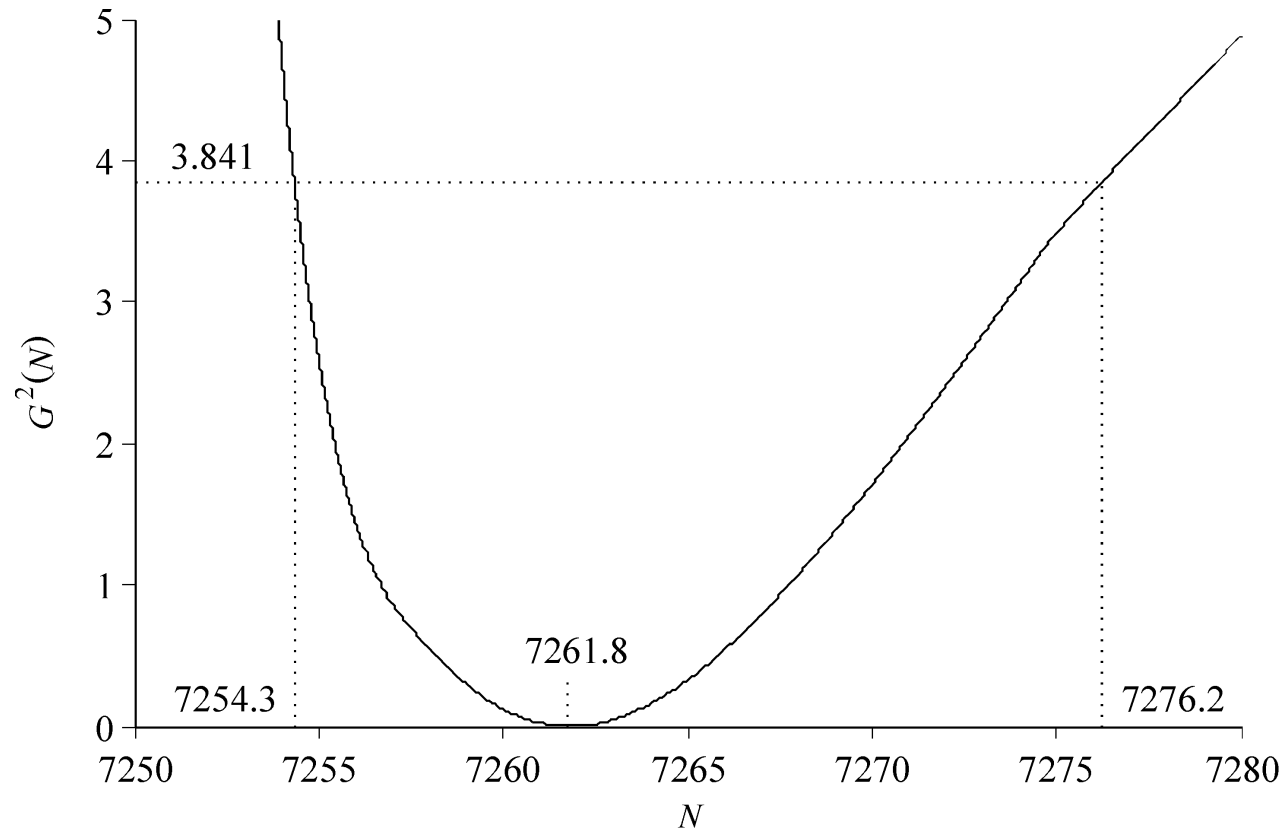
- For a "reference subject" with  $z_{i1} = 0$  and  $z_{i2} = 68.6$  we have

Parameter	Latent class		
	1	2	3
Class weight	0.5129	0.4178	0.0693
Conditional prob. H	0.8538	0.9786	0.9534
Conditional prob. P	0.9827	0.8734	0.2399
Conditional prob. D	0.0063	0.6817	0.8196

- Class 1 is the most common and contains cases with the smallest risk of D and the largest chance of being detected by P
- Subjects in latent class 2 are most likely to be detected by H and have a much bigger chance of D
- Latent class 3 corresponds to a very small number of cases which are unlikely to be detected by P and have the highest risk of D

# Estimate of the population size

- The estimated population size is  $\hat{N} = 7261.8$  with only 8.8 (0.12%) cases of cancer are missing from the Tuscany Registry
- The confidence interval for  $N$  is  $(7254.3, 7276.2)$



# References

- Agresti, A. (1994), Simple capture-recapture models permitting unequal catchability and variable sampling effort, *Biometrics*, **50**, pp. 494-500.
- Alho, J. M. (1990), Logistic Regression in Capture-Recapture Models, *Biometrics*, **46**, pp. 623-635.
- Bandeen-Roche, K. and Miglioretti, D. L., Zeger, S. L. and Rathouz, P. J. (1997), Latent variable regression for multiple discrete outcomes, *Journal of the American Statistical Association*, **92**, pp. 1375-1386.
- 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**, pp. 207-212.
- Bergsma, W. P. and Rudas, T. (2002), Marginal models for categorical data, *Annals of Statistics*, **30**, pp. 140-159.
- Colombi, R. and Forcina, A. (2001), Marginal regression models for the analysis of positive association of ordinal response variables, *Biometrika*, **88**, pp. 1007-1019.
- Cormack, R. M. (1989), Log-linear models for capture-recapture, *Biometrics*, **45**, pp. 395-413.
- Cormack, R. M. (1992), Interval estimation for mark-recapture studies of closed populations, *Biometrics*, **48**, pp. 567-578.
- Cowan, C. D. and Malek, D. (1986), Capture-Recapture models when both sources have clustered observations, *Journal of the American Statistical Association*, **81**, pp. 461-466.
- Crocetti, E., Miccinesi, G., Paci, E. and Zappa, M. (2001), An application of the two-source capture-recapture method to estimate the completeness of the Tuscany Cancer Registry, Italy, *European Journal of Cancer Prevention*, **10**, pp. 417-423.

- 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**, pp. 1137-1148.
- Fienberg, S. E. (1972), The multiple recapture census for closed populations and incomplete  $2^k$  contingency tables, *Biometrika*, **59**, pp. 591-603.
- Goodman, L. A. (1974), Exploratory latent structure analysis using both identifiable and unidentifiable models, *Biometrika*, **61**, pp. 215-231.
- Huang, G. and Bandeen-Roche, K. (2004), Building an identifiable latent class model, with covariate effects on underlying and measured variables, *Psychometrika*, **69**, pp. 5-32.
- Lindsay, B., Clogg, C. and Grego, J. (1991), Semiparametric estimation of the Rasch model and related exponential response models, including a simple latent class model for item analysis, *Journal of the American Statistical Association*, **86**, pp. 96-107.
- Pledger, S. (2000), Unified maximum likelihood estimation for closed capture-recapture models using mixtures, *Biometrics*, **56**, pp. 434-442.
- Rasch, G. (1961), On general laws and the meaning of measurement in psychology, *Proceedings of the IV Berkeley Symposium on Mathematical Statistics and Probability*, **4**, pp. 321-333.
- Sanathanan, L. (1972), Estimating the size of a multinomial population, *The Annals of Mathematical Statistics* **43**, 142-152.
- Stanghellini, E. and van der Heijden, P. G. M. (2004), A multiple-record systems estimation method that takes observed and unobserved heterogeneity into account, *Biometrics*, **60**, pp. 510-516.