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

Francesco Bartolucci

Istituto di Scienze Economiche, Unversità di Urbino,

Via Saffi, 42, 61029 Urbino, Italy

email: Francesco.Bartolucci@uniurb.it

Antonio Forcina

Dipartimento di Scienze Statistiche, Università di Perugia,

Via A. Pascoli, 06100 Perugia, Italy

email: forcina@stat.unipg.it

Abstract

We introduce a new family of latent class models for the analysis of capture-recapture data where continuous covariates are available. The present approach exploits recent advances in marginal parameterizations to model simultaneously, and conditionally on individual covariates, the size of the latent classes, the marginal probabilities of being captured by each list given the latent and possible higher order marginal interactions among lists conditionally on the latent. An EM algorithm for maximum likelihood estimation is described and an expression for the expected information matrix is derived. In addition, a new method for computing confidence intervals for the size of the population having given covariate configurations is proposed and its asymptotic properties are derived. Applications to data on HIV+ patients in the Region of Veneto (Italy) and to new cases of cancer in Tuscany are discussed.

KEY WORDS: Conditional inference; Latent class model; Marginal parameterization; Profile confidence intervals; Rasch model.

1 Introduction

Statistical methods for the estimation of animal abundance from data collected by multiple capture surveys have a long history: see Schwarz and Seber (1999) for an accurate review of the literature and Pollock (2000) for a concise discussion of the main lines of approach. Similar methods have found application in epidemiology, where multiple systems (usually called *lists*) are often operating simultaneously, and sometimes independently, to keep records of individuals who suffer from a certain disease. One can exploit the analogy between capture occasions and lists, though the proportion of people appearing on a list is usually higher than the proportion of animals captured in a single experiment. It is well known that even censuses may suffer from undercount as described by Darroch et al. (1993).

The earliest approaches for the analysis of data collected within *closed populations* (i.e. when migration, birth and death rates may be considered negligible) were based on simple models for contingency tables (see for example Bishop et al. 1975), where each list is treated as a binary variable and these variables are assumed to be independent; this is equivalent to assuming that the probability of appearing on a given list is constant across the whole population and that lists operate independently. More sophisticated models differ in the way they deal with observed and unobserved heterogeneity within the population of interest. Observed covariates may be discretized and used to define disjoint strata as, for example, in Darroch et al. (1993) or Plante et al. (1998). Alternatively, one may model individual capture probability as a function of continuous covariates: Alho (1990) applied a logistic model to a two-list model under independence while Zwane and van der Hijden (2005) extended this approach to multiple lists by allowing for possible dependence between lists. Unobserved heterogeneity has been modelled by assuming that capture probabilities depend upon a continuous latent trait having a given parametric distribution across the population; see for example Coull and Agresti (1999) or Dorazio and Royle (2003), who propose a beta-binomial model which they compare with a latent class approach. The use of latent class models are advocated by Pledger (2000), among others; these models have the advantage that the distribution of the latent is unconstrained; in addition, Lindsay et al. (1991) showed that a finite mixture with only a few latent classes provides the same fit of a conditional Rasch model.

A basic assumption of latent class models is that lists operate independently within homogeneous subjects, so that marginal association is due entirely to unobserved heterogeneity. This assumption may be violated; for instance, if a case of cancer has been detected by a pathology department, it is unlikely that the same case will also be detected at the hospital level. Stanghellini and van der Hijden (2004) allowed bivariate log-linear interactions conditionally on the latent. The same kind of bivariate dependence may be described by marginal models as in Bartolucci and Forcina (2001); computationally this is more demanding but it allows the univariate marginals and the bivariate associations to be modelled separately. A more detailed discussion of the marginal approach versus the log-linear approach is given in 2.2.

In this paper we propose a new class of models for capture-recapture data allowing for observed and unobserved heterogeneity. In particular, we extend the work of Bandeen-Roche et al. (1997) and Huang and Bandeen-Roche (2004) on latent class models with continuous covariates to the context of capture-recapture data, and in addition we allow for conditional dependence among lists by exploiting recent developments on marginal modelling (Bergsma and Rudas, 2002). When only discrete covariates are present, the models studied in this paper are only slightly more flexible than those in Bartolucci and Forcina (2001); however, the presence of continuous covariates poses new conceptual as well as computational challenges, the solution of which constitutes one of the main merits of this work. We also extend Cormack (1992)'s approach for computing confidence intervals for the true size of the population to the case of subpopulations corresponding to selected covariate configurations and derive its asymptotic properties. This extends an asymptotically equivalent procedure described by Stanghellini and van der Hijden (2004) to the presence of continuous covariates.

After describing a dataset on new cases of HIV infection in the Region of Veneto, in Section 2 we describe a family of marginal latent class models whose probability structure may depend on continuous individual covariates as in a generalized linear model. An EM algorithm for maximum likelihood estimation of the regression coefficients is described in Section 3, where we also indicate how to compute the expected information and asymptotic standard errors. Computation of confidence intervals for partial or overall undercounts is

discussed in Section 3.3. A set of MATLAB functions designed to perform all these tasks is available from the website `http://www.econ.uniurb.it/bartolucci/index.htm`. Their application to the HIV dataset is discussed in Section 4.

1.1 The data

The data we are going to analyze in Section 4 are about $n = 3079$ new cases of HIV infection detected among the residents in the Region of Veneto in the 1997-2003 period. The data were produced by linking $J = 3$ lists as described in Pezzotti et al. (2003):

- HIV (H): this is a list of individuals voluntarily took an HIV test at a Local Health Unit and were found to be positive.
- AIDS (A): this is a list of individuals diagnosed with AIDS; the list is managed by the National AIDS Center.
- DRLH (D): lists individuals who appear HIV positive according to a discharge report from a public hospital or from a private hospital when a reimbursement has been requested from the Regional Health System.

The frequencies for the $k = 7$ possible capture configurations are shown in Table 1, which indicates that there is not much overlapping between lists: only 21.3% of the subjects appear simultaneously in more than one list.

$\boldsymbol{r}$	D	A	AD	H	HD	HA	HAD
$y_{\boldsymbol{r}}$	1459	54	152	909	397	18	90

Table 1: *New cases of HIV infection in Veneto from 1997 to 2003 by capture configuration*

The *year* of first appearance in one of the lists, the corresponding *age* and the *sex* are also available. The marginal distribution of cases according to *year* is given in Table 2: the drop in number of cases from 1999 to 2000 is due mainly to the introduction of a new therapy which did not require hospitalization. The joint distribution by age and sex is summarized in Table 3; females appear to be younger and their proportion is much smaller.

year	1997	1998	1999	2000	2001	2002	2003
%	20.07	20.53	20.20	9.42	9.55	11.33	8.90

Table 2: *Marginal distribution of cases according to the year of first detection*

		age				
sex	%	Mean	Variance	1st quartile	Median	3rd quartile
Male	68.14	37.5	173.54	31.0	36.5	43.0
Female	31.86	33.1	179.98	27.0	32.0	38.0
Total	100.00	36.1	179.71	30.0	41.0	35.0

Table 3: *Summary statistics for the joint distribution of age and sex*

2 A class of latent marginal regression models

2.1 Data organization

Assume that data are available for n subjects which have been captured at least once. These may be grouped into s distinct covariate configurations, each determined by a vector $\boldsymbol{z}_i$ common to n_i subjects with $\sum_i n_i = n$. A special case is when $n_i = 1$, $i = 1, \ldots, s$, i.e. there is a single subject in each configuration. In any case, because empty configurations need not be considered, $s \leq n$. Assume that there are J different lists and order the $k = 2^J - 1$ possible capture configurations, $\boldsymbol{r} = (r_1, \ldots, r_J)$, by letting each list go from 0 (not captured) to 1 (captured) in lexicographic order; let $\boldsymbol{y}_i$ be the vector containing the corresponding capture frequencies. When only one subject with a given covariate structure has been captured, $\boldsymbol{y}_i$ is a vector of zeros except for the entry corresponding to the observed capture configuration, which is equal to 1. Let $p_{i,\boldsymbol{r}}$ be the probability that a subject with covariate configuration $\boldsymbol{z}_i$ experiences the capture configuration $\boldsymbol{r} \neq \boldsymbol{0}$; these probabilities may be arranged into the vector $\boldsymbol{p}_i$ with capture configurations ordered as in $\boldsymbol{y}_i$. Let also q_i denote the probability of being captured at least once, given $\boldsymbol{z}_i$, so that $1 - q_i$ denotes the probability of being never captured. On the basis of a latent regression model, which will be discussed later, both $\boldsymbol{p}_i$ and $q_i = \sum_{\boldsymbol{r} \neq \boldsymbol{0}} p_{i,\boldsymbol{r}}$ will be assumed to depend on a common vector of regression coefficients $\boldsymbol{\beta}$. Finally let t_i denote the total number of subjects with covariate $\boldsymbol{z}_i$ in the population, i.e. the sum of n_i plus the number of those never captured, and $N = \sum_i t_i$ denote the unknown population size.

2.2 A class of marginal link functions

Unobserved heterogeneity will be modelled by assuming that each subject belongs to one among c disjoint latent classes as in a finite mixture model. More precisely, let $\pi_{i,h,\boldsymbol{r}}$ denote the conditional probability that a subject having covariate $\boldsymbol{z}_i$ belongs to latent class h and experiences capture configuration $\boldsymbol{r}$. With $\boldsymbol{\pi}_i$ we denote the vector with entries $\pi_{i,h,\boldsymbol{r}}$, where, within each latent class h, $\boldsymbol{r}$ varies in lexicographic order within the full set of 2^J capture configurations, including the event of being never captured.

The class of models that we are going to propose depends upon an invertible transformation that links the vector $\boldsymbol{\pi}_i$ to a vector of marginal logits and higher order interaction parameters of the form $\boldsymbol{\eta}_i = \boldsymbol{X}_i\boldsymbol{\beta}$ where $\boldsymbol{X}_i$ is a matrix of known constants which depend on $\boldsymbol{z}_i$ and $\boldsymbol{\beta}$ is a vector of regression parameters. The vector $\boldsymbol{\eta}_i$ is defined by a marginal and *ordered decomposable parameterization* (odp). The elements of an odp link function are variation independent (Bergsma and Rudas, 2002, p. 144) so that, roughly speaking, the values assigned to the parameters of different marginal distributions are always compatible. To keep the model parsimonious, we assume that all the interactions which are not defined within a suitable marginal distribution are assumed to be identically 0 and thus are not included into $\boldsymbol{\eta}_i$. As we describe in the Appendix, a linear predictor having these features may be expressed as

$$\boldsymbol{\eta}_i = \boldsymbol{C}\log(\boldsymbol{M}\boldsymbol{\pi}_i) = \boldsymbol{X}_i\boldsymbol{\beta} \tag{1}$$

where the matrices $\boldsymbol{C}$ and $\boldsymbol{M}$ are determined by the specific marginal model adopted; in particular $\boldsymbol{C}$ is a matrix of row contrasts required to compute the corresponding parameters and $\boldsymbol{M}$ is a matrix of 0's and 1's that produces the required marginal distributions. Let v denote the size of any $\boldsymbol{\eta}_i$; Bergsma and Rudas (2002, Theorems 1 and 2) show that (1) defines a transformation which is invertible and twice differentiable between the vector of the marginal parameters $\boldsymbol{\eta}_i$ and the probability vectors $\boldsymbol{\pi}_i$, within the set of probability distributions which satisfy the restriction that all log-linear interactions of a higher order than those included in $\boldsymbol{\eta}_i$ are equal to 0. Note that there must be $k - v$ linearly independent restrictions.

Bartolucci and Forcina (2001) also used a marginal link function whose elements, however,

are not odp, so that there exist values of $\boldsymbol{\eta}_i$ which do not correspond to any joint distribution, a feature which may cause numerical difficulties in the estimation algorithms. The log-linear link function used by Stanghellini and van der Hijden (2004) is a valid alternative. The present approach aims at modelling the marginal distribution of the latent given the covariates directly, and then the marginal distribution of each list given the covariates and the latent. With the log-linear link instead, the distribution of the latent can be modelled conditionally on the lists being fixed to a reference category. It may be shown that the two approaches are equivalent when lists are assumed to be conditionally independent given the latent and a saturated model is assumed for the distribution of the latent. Instead, when the regression model for the marginal logits of the latent is not saturated (for instance, because continuous covariates are available) or if bivariate and higher order interactions are allowed for the conditional distribution of the lists given the latent, the marginal link function may differ substantially from the log-linear one where, for instance, the main effect of a list is also conditional to the other lists being fixed to a reference category. When bivariate interactions are allowed, models like Rasch's can be formulated only if the univariate marginals of each list can be accessed directly.

Typically $\boldsymbol{\eta}_i$ consists of the following elements:

1. $c-1$ logits of type local which determine the marginal distribution of the latent;
2. cJ logits describing the probability of being captured in list j conditionally on being in latent class h, with j running faster than h; note that these logits are marginal with respect to other lists;
3. for each of at most $J-1$ selected pairs of lists, the c log-odds ratios conditionally on the latent but marginally with respect to other lists;
4. when the number of lists is, say, greater than 5, c three-way interactions for some specific triplet of lists conditionally on the latent.

The following examples may clarify the common features of an odp. Let 0 denote the variable indexing latent classes and $1,\ldots,4$ denote 4 different lists. To construct an odp which produces the model of conditional independence we need to consider the following

sequence of marginals: $\{0\}$, $\{0,1\}$, $\{0,2\}$, $\{0,3\}$, $\{0,4\}$; all the interactions which cannot be defined within these marginals are constrained to 0. Now suppose instead that the capture mechanism is such that lists follow a natural order, so that the probability of being captured by a list (given the latent) depends on whether the same subject has been captured by the previous list; in this case we should add to the previous sequence the three marginals: $\{0,1,2\}$, $\{0,2,3\}$, $\{0,3,4\}$ within which the set of interactions for specifying the bivariate association between adjacent lists may be defined. When there is a specific list, e.g. the first, such that the event of being captured by this list may affect the probability of being captured by other lists, we should consider, in addition to the set of marginals required for the model of conditional independence, the set of marginals $\{0,1,2\}$, $\{0,1,3\}$, $\{0,1,4\}$ where the bivariate interaction between each list and the first (conditionally on the latent) may be defined.

2.3 The linear model

The matrix $\boldsymbol{X}_i$, which determines the linear model, will usually be block diagonal with a block for each component of the linear predictor: latent distribution, univariate marginals given the latent, bivariate interactions given the latent and so on. The specific structure of each component may vary depending on certain basic assumptions concerning the nature and interpretation of the latent classes. If we believe that the latent classes are well-defined populations to which each subject may or may not belong depending on individual covariates, these covariates may enter only in the univariate logits for the probability that an individual with given covariates belongs to the different latent populations. Thus the conditional probabilities of being captured would depend only on the latent. In this context it makes sense to consider an additive model for the marginal logits of each list given the latent, like in the Rasch (1961) model, so that the corresponding elements of $\boldsymbol{\beta}$ could be interpreted as the capture effectiveness of each list (difficulty parameters, using the terminology of the Item Response Theory) and receptiveness of each latent class (ability parameters).

If instead we believe that individual behavior is not entirely explained by the fact of belonging to a given latent class, we would allow individual covariates to affect also the

capture probabilities of each list. Thus a more general class of models may be constructed by relating either or both the first two components of the linear predictor to individual covariates. In doing so, it seems reasonable to require that the regression coefficients of the marginal logits of each list given the latent do not depend on the latent, which can affect only the intercept. This is one of the sufficient conditions of Theorem 1 of Huang and Bandeen-Roche (2004) for the local identifiability of the corresponding latent class model; additional comments on identifiability are at the end of section 3.3. The third component, containing bivariate interactions conditionally on the latent, will usually be introduced to capture residual unexplained association, and it is unlikely that this will have to depend on covariates, although in principle this is possible.

3 Likelihood inference

3.1 The likelihood function

Under the assumption of multinomial sampling, following Sanathanan (1972), the log-likelihood describing the overall population may be factorized as follows. Let $\boldsymbol{t} = (\, t_1, \ldots, t_s \,)$; then we have

$$L_U(\boldsymbol{t}, \boldsymbol{\beta}) = L_B(\boldsymbol{t}, \boldsymbol{\beta}) + L_C(\boldsymbol{\beta}) \tag{2}$$

$$\text{where} \qquad L_B(\boldsymbol{t}, \boldsymbol{\beta}) = \sum_i \left[\log \frac{t_i!}{(t_i - n_i)!\, n_i!} + (t_i - n_i) \log(1 - q_i) + n_i \log(q_i) \right] \tag{3}$$

$$\text{and} \qquad L_C(\boldsymbol{\beta}) = \sum_i \left[\log \frac{n_i!}{\prod_{\boldsymbol{r} \neq \boldsymbol{0}} y_{i,\boldsymbol{r}}!} + \sum_{\boldsymbol{r} \neq \boldsymbol{0}} y_{i,\boldsymbol{r}} \log \left(\frac{p_{i,\boldsymbol{r}}}{q_i} \right) \right]. \tag{4}$$

Under the assumption that the latent regression model is such that the so-called manifest probabilities $\boldsymbol{p}_i$ are identified by $\boldsymbol{\beta}$, model selection will be based on the conditional log-likelihood $L_C(\boldsymbol{\beta})$ because it is relatively simpler to maximize. Once a proper model has been selected, the binomial component of the unconditional log-likelihood will be used to obtain a point estimate of $\boldsymbol{t}$.

3.2 Maximizing the conditional likelihood

This may be seen as a problem of incomplete data, where a number of subjects are never captured and the latent class of those captured cannot be observed. The complete data would contain the number of subjects with capture configuration $\boldsymbol{r}$ who belong to latent class h and have covariate configuration $\boldsymbol{z}_i$. These frequencies may be arranged into the $c \times 2^J$ vector $\boldsymbol{m}_i$, having entries corresponding to those of $\boldsymbol{\pi}_i$. Then the log-likelihood may be maximized by the following EM algorithm (Dempster et al., 1977, Baker, 1990):

E-step: on the basis of the available estimate of the vector of regression coefficients $\dot{\boldsymbol{\beta}}$, first compute $\dot{\boldsymbol{\eta}}_i$ and then reconstruct $\dot{\pi}_{i,h,\boldsymbol{r}}$, the estimated underlying multinomial probabilities; on this basis the estimate of the underlying frequencies conditionally on $\boldsymbol{y}_i$ has the form

$$\dot{m}_{i,h,\boldsymbol{r}} = \begin{cases} (\dot{t}_i - n_i)\dfrac{\dot{\pi}_{i,h,\boldsymbol{0}}}{\sum_h \dot{\pi}_{i,h,\boldsymbol{0}}} & \text{if } \boldsymbol{r} = \boldsymbol{0} \\ y_{i,\boldsymbol{r}}\dfrac{\dot{\pi}_{i,h,\boldsymbol{r}}}{\sum_h \dot{\pi}_{i,h,\boldsymbol{r}}} & \text{otherwise} \end{cases}$$

where $\dot{t}_i = n_i/\dot{q}_i$;

M-step: maximize the multinomial likelihood $\tilde{L}(\boldsymbol{\beta}) = \sum_i \boldsymbol{m}_i' \log(\boldsymbol{\pi}_i) + \text{constant}$, having replaced $\boldsymbol{m}_i$ with $\dot{\boldsymbol{m}}_i$.

The M-step may be performed by a Fisher-scoring algorithm as follows. Let $\boldsymbol{\theta}_i$ be the vector of canonical parameters for the multinomial distribution in exponential family form, having removed the elements which are constrained to 0, so that this vector has the same dimension v of the linear predictor $\boldsymbol{\eta}_i$. As described in the Appendix, a *design matrix* $\tilde{\boldsymbol{G}}$ of full rank v may be easily constructed so that we may write

$$\log(\boldsymbol{\pi}_i) = \tilde{\boldsymbol{G}}\boldsymbol{\theta}_i - \boldsymbol{1}\log[\boldsymbol{1}'\exp(\tilde{\boldsymbol{G}}\boldsymbol{\theta}_i)].$$

Using this parameterization, the score vector and the average information matrix have the simple form

$$\tilde{\boldsymbol{s}} = \sum_i \boldsymbol{X}_i'\boldsymbol{R}_i'\tilde{\boldsymbol{G}}'(\boldsymbol{m}_i - t_i\boldsymbol{\pi}_i) \quad \text{and} \quad \tilde{\boldsymbol{F}} = \frac{1}{N}\sum_i t_i\boldsymbol{X}_i'\boldsymbol{R}_i'\tilde{\boldsymbol{G}}'\tilde{\boldsymbol{\Omega}}_i\tilde{\boldsymbol{G}}\boldsymbol{R}_i\boldsymbol{X}_i,$$

where

$$\tilde{\boldsymbol{\Omega}}_i = \text{diag}(\boldsymbol{\pi}_i) - \boldsymbol{\pi}_i\boldsymbol{\pi}_i' \quad \text{and} \quad \boldsymbol{R}_i = [\boldsymbol{C}\text{diag}(\boldsymbol{M}\boldsymbol{\pi}_i)^{-1}\boldsymbol{M}\text{diag}(\boldsymbol{\pi}_i)\tilde{\boldsymbol{G}}]^{-1}$$

are, respectively, the kernel of the multinomial variance and the derivative of $\boldsymbol{\eta}_i$ with respect to $\boldsymbol{\theta}_i'$. From the results in Bergsma and Rudas (2002) it follows that this matrix of derivatives is of full rank v.

A direct Fisher scoring algorithm which maximizes the likelihood of the incomplete data could also be used. An expression for the expected information matrix is derived in the next section and the corresponding score vector is computed in the Appendix. However, in our experience, such an algorithm would be much more unstable, though a little faster near convergence.

3.3 Covariance matrix of the estimator of the regression coefficients and local identifiability

The covariance matrix of the conditional estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ may be obtained from the expected information matrix of $L_C(\boldsymbol{\beta})$ written as an exponential family density. To this purpose let $\boldsymbol{\gamma}_i = \boldsymbol{H}\log(\boldsymbol{\rho}_i)$, with $\boldsymbol{\rho}_i = \boldsymbol{p}_i/q_i$, denote a vector of canonical parameters for the saturated model of the manifest multinomial distribution with parameters $(n_i, \boldsymbol{\rho}_i)$. Note that $\boldsymbol{H}$ may be any $(k-1)\times k$ matrix whose rows are linearly independent contrasts. Notice also that we can write $\boldsymbol{p}_i = \boldsymbol{A}\boldsymbol{\pi}_i$ and $q_i = \boldsymbol{a}'\boldsymbol{\pi}_i$, where $\boldsymbol{A} = \mathbf{1}_c' \otimes \bar{\boldsymbol{I}}$, $\boldsymbol{a}' = \mathbf{1}_k'\boldsymbol{A}$, $\bar{\boldsymbol{I}}$ is an identity matrix of size $k+1$ without the first row and $\otimes$ denotes the Kronecker product. It follows that the conditional log-likelihood may be written as

$$L_C(\boldsymbol{\beta}) = \sum_i \{\boldsymbol{y}_i'\boldsymbol{G}\boldsymbol{\gamma}_i - n_i \log[\mathbf{1}'\exp(\boldsymbol{G}\boldsymbol{\gamma}_i)]\} + \text{constant},$$

where $\boldsymbol{G}$ is the right inverse of $\boldsymbol{H}$. By applying the chain rule and after a few simplifications described in the Appendix, the average expected information matrix may be written as

$$\boldsymbol{F} = \frac{1}{n}\boldsymbol{X}'\boldsymbol{B}'\boldsymbol{O}\boldsymbol{B}\boldsymbol{X}, \tag{5}$$

where $\boldsymbol{X}$ is obtained by stacking the design matrices $\boldsymbol{X}_i$ one below the other, $\boldsymbol{B}$ is block diagonal with the ith block equal to $\boldsymbol{U}_i\boldsymbol{R}_i$, where

$$\boldsymbol{U}_i = \left\{\mathbf{1}_c' \otimes \left[\text{diag}(\boldsymbol{\rho}_i)^{-1} - \mathbf{1}_k\mathbf{1}_k'\right]\bar{\boldsymbol{I}}\right\}\tilde{\boldsymbol{\Omega}}_i\tilde{\boldsymbol{G}}/q_i,$$

and $\boldsymbol{O}$ is block diagonal with the ith block given by the conditional covariance matrix of $\boldsymbol{y}_i$, given n_i, which is equal to $n_i\boldsymbol{\Omega}_i$, with $\boldsymbol{\Omega}_i = \text{diag}(\boldsymbol{\rho}_i) - \boldsymbol{\rho}_i\boldsymbol{\rho}_i'$. On the basis of the information matrix above we can approximate the covariance matrix of $\hat{\boldsymbol{\beta}}$ with $(n\boldsymbol{F})^{-1}$.

The fact that $\boldsymbol{BX}$ is the Jacobian of the transformation from the manifest distribution to the model parameters $\boldsymbol{\beta}$ may be used for a heuristic assessment of local identifiability. Because the block diagonal matrix $\boldsymbol{O}$ is of full column rank $s(k-1)$, which is usually much bigger than the dimension of $\boldsymbol{\beta}$, full rank of $\boldsymbol{F}$ is at least a sufficient condition of local identifiability. Notice also that, unless there are cells of the conditional multinomials which have true probability equal to 0, a sufficient condition for local identifiability at the true value of the parameters is simply that $s(k-1)$ is bigger than the dimension of $\boldsymbol{\beta}$. This indicates that the presence of individual covariates may allow to fit a number of latent classes much larger than what would be possible with a single stratum.

3.4 Inference on the population size

The vector $\hat{\boldsymbol{t}}$ which maximizes the binomial likelihood $L_B(\boldsymbol{t}, \hat{\boldsymbol{\beta}})$ has elements

$$\hat{t}_i = \frac{n_i}{\hat{q}_i}, \quad i = 1, \ldots, s, \tag{6}$$

so that the overall population size is estimated on the basis of $\hat{\boldsymbol{\beta}}$ by $\hat{N} = \sum_i \hat{t}_i$. The properties of these estimators have been analyzed by Alho (1990), among others.

We now introduce a new statistic which extends the discrepancy measure studied by Cormack (1992) in the context of a single stratum to the context of continuous covariates; we also provide a formal argument for using the resulting confidence interval by deriving the asymptotic distribution of the statistic. When continuous covariates are present, it is unlikely that we are interested in the undercount for a given covariate configuration; most of the times it is the overall undercount to be of interest. However, if discrete covariates are also available, the size of the population with a given value of certain covariates might be of interest. The approach we are going to describe is very general, in the sense that we derive a confidence interval for the size of the population belonging to a chosen subset of strata as determined by an $s \times 1$ binary vector $\boldsymbol{u}$ whose entries are equal to 1 if the corresponding stratum is to be considered and are equal to 0 otherwise; we also denote with $N_{\boldsymbol{u}}$ the size of the corresponding

population. The approach proposed by Stanghellini and van der Heijden (2004) may be seen as a special case of the one we are going to describe when $\boldsymbol{u}$ has just one element different from 0 and an asymptotically equivalent statistics based on the unconditional estimator is used. Now let

$$G^2(N_{\boldsymbol{u}}) = \min_{\boldsymbol{t}'\boldsymbol{u}=N_{\boldsymbol{u}}} D(\boldsymbol{t}, \boldsymbol{\beta}) - \hat{D}, \tag{7}$$

where

$$D(\boldsymbol{t}, \boldsymbol{\beta}) = 2 \sum_i \left[(t_i - n_i) \log \left(\frac{t_i - n_i}{t_i p_{i,\mathbf{0}}} \right) + \sum_{\boldsymbol{r} \neq \mathbf{0}} y_{i,\boldsymbol{r}} \log \left(\frac{y_{i,\boldsymbol{r}}}{t_i p_{i,\boldsymbol{r}}} \right) \right] \tag{8}$$

is the hypothetical deviance of the assumed model under the assumption that t_i, the population size in any stratum i, is known and $\hat{D} = D(\hat{\boldsymbol{t}}, \hat{\boldsymbol{\beta}})$.

To compute the constrained minimum in (7) we use an algorithm that, starting from $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$, alternates the following steps until convergence:

- update $\boldsymbol{t}$, with $\boldsymbol{\beta}$ held fixed, by minimizing $D(\boldsymbol{t}, \boldsymbol{\beta})$ under the constraint $\boldsymbol{t}'\boldsymbol{u} = N_{\boldsymbol{u}}$ together with $t_i - n_i \geq 0$, $\forall i$. It may be shown that an algorithm that at each step minimizes a quadratic approximation of $D(\boldsymbol{t}, \boldsymbol{\beta})$ under the constraints at issue is equivalent to the following algorithm: (i) compute q_i, $\forall i$, from the available estimate of $\boldsymbol{\beta}$; (ii) solve with respect to ϕ the nonlinear equation

  $$\sum_i u_i n_i \frac{1 - q_i}{q_i[1 - \phi(1 - q_i)]} = N_{\boldsymbol{u}}$$

  by a simple Newton algorithm and (iii) update the estimate of $\boldsymbol{t}$ by letting $t_i = n_i(1 - q_i)/[q_i - \dot{\phi} q_i(1 - q_i)]$ if $u_i = 1$ and $t_i = n_i/q_i$ otherwise, where $\dot{\phi}$ is the solution of the equation above;

- update $\boldsymbol{\beta}$, with $\boldsymbol{t}$ held fixed, by maximizing the full multinomial likelihood on the complete table; this may be performed by an EM algorithm similar to the one described in Section 3.2; the only difference is that the E-step is slightly simpler because t_i is assumed to be known for any i.

In the special case where $N_{\boldsymbol{u}} = \hat{N}_{\boldsymbol{u}}$, the algorithm above stops after the first step, because, when we start from $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$, the conditional estimate $\hat{\boldsymbol{t}}$ satisfies the constraint $\hat{\boldsymbol{t}}'\boldsymbol{u} = \hat{N}_{\boldsymbol{u}}$ by construction and so $G^2(\hat{N}_{\boldsymbol{u}}) = 0$.

Though the plot of $G^2(N_{\boldsymbol{u}})$ (or its asymptotic variants) as a function of $N_{\boldsymbol{u}}$ is commonly referred to as a profile likelihood, $G^2(N_{\boldsymbol{u}})$ is not a proper likelihood ratio because it is neither based on the unconditional nor on the conditional likelihood. To understand its nature let, N_0 denote the true value of the overall population size N, $N_{\boldsymbol{u}0}$ that of $N_{\boldsymbol{u}}$, $\boldsymbol{\alpha} = \boldsymbol{t}/N_0$ and $\boldsymbol{\lambda}' = (\boldsymbol{\alpha}', \boldsymbol{\beta}')$; let also $\hat{\boldsymbol{\lambda}}$ denote the value of $\boldsymbol{\lambda}$ obtained by replacing $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ with their conditional estimates and $\boldsymbol{\lambda}_0$ denote the true value of $\boldsymbol{\lambda}$. Finally define

$$F(\boldsymbol{\lambda}, \hat{\boldsymbol{\lambda}}) = [D(\boldsymbol{t}, \boldsymbol{\beta}) - \hat{D}]/N_0.$$

Lemma 1 *Let $\boldsymbol{V}_0$ denote the matrix of the second derivatives of $F(\boldsymbol{\lambda}, \hat{\boldsymbol{\lambda}})/2$ at $\boldsymbol{\lambda}_0$ and $\tau = N_{\boldsymbol{u}0}/N_0$; then*

$$\min_{(\boldsymbol{\alpha}'\boldsymbol{u}=\tau)} F(\boldsymbol{\lambda}, \hat{\boldsymbol{\lambda}}) = \min_{(\boldsymbol{\alpha}'\boldsymbol{u}=\tau)} (\boldsymbol{\lambda} - \hat{\boldsymbol{\lambda}})'\boldsymbol{V}_0(\boldsymbol{\lambda} - \hat{\boldsymbol{\lambda}}) + o(\|\boldsymbol{\lambda}_0 - \hat{\boldsymbol{\lambda}}\|^2),$$

PROOF. $F(\boldsymbol{\lambda}, \hat{\boldsymbol{\lambda}})$ satisfies the conditions given by Shapiro (1985, p.135) for a *discrepancy function*: (i) is non-negative; (ii) is equal to 0 only when $\boldsymbol{\lambda} = \hat{\boldsymbol{\lambda}}$; (iii) is twice continuously differentiable and (iv) for any a such that $\|\boldsymbol{\lambda} - \hat{\boldsymbol{\lambda}}\| \geq a$, exists $b > 0$ such that $F(\boldsymbol{\lambda}, \hat{\boldsymbol{\lambda}}) > b$. Then the result follows from Lemma 2.1 in Shapiro (1985).

Theorem 1 *Provided that*

$$\lim_{N_0 \to \infty} \alpha_{0i} = c_i > 0, \forall i, \tag{9}$$

$G^2(N_{\boldsymbol{u}})$ has asymptotic $\chi^2(1)$ distribution.

PROOF. Sanathanan (1972, Theorem 4) has shown that $\hat{\boldsymbol{\lambda}}$ is a consistent estimator of $\boldsymbol{\lambda}_0$ and that $\sqrt{N_0}(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0)$ has an asymptotic $N(\boldsymbol{0}, \boldsymbol{\Sigma}_0)$ distribution, where $\boldsymbol{\Sigma}_0$, as we show in the Appendix, is such that $\boldsymbol{V}_0 \to^p \boldsymbol{\Sigma}_0^{-1}$ and thus $G^2(\tau N_0) = N_0 \min_{(\boldsymbol{\alpha}'\boldsymbol{u}=\tau)} F(\boldsymbol{\lambda}, \hat{\boldsymbol{\lambda}})$, because of Lemma 1, has asymptotic $\chi^2(1)$ distribution (see also Shapiro, 1985, Lemma 2.2 for a detailed derivation).

Theorem 1 implies that, if a given $N_{\boldsymbol{u}}$ was equal to the true value $N_{\boldsymbol{u}0}$, the probability that the corresponding statistic $G^2(N_{\boldsymbol{u}})$ exceeds the critical value $\chi^2_\alpha(1)$ is approximately equal to α, where $\chi^2_\alpha(1)$ denotes the $100(1-\alpha)$-th percentile of the χ^2 distribution with 1 degree of freedom. Because of this, a confidence interval for $N_{\boldsymbol{u}}$, $(\hat{N}_{\boldsymbol{u},1}, \hat{N}_{\boldsymbol{u},2})$, may be

constructed as follows: (i) compute the value of $G^2(N_{\boldsymbol{u}})$ for a grid of values of $N_{\boldsymbol{u}}$ around $\hat{N}_{\boldsymbol{u}}$ so that a reasonable approximation of this function may be constructed and, possibly, plotted; (ii) compute $\hat{N}_{\boldsymbol{u},1}$ as the largest $N_{\boldsymbol{u}}$ such that $N_{\boldsymbol{u}} \leq \hat{N}_{\boldsymbol{u}}$ and $G^2(N_{\boldsymbol{u}}) \geq \chi^2_\alpha(1)$ and $\hat{N}_{\boldsymbol{u},2}$ as the smallest $N_{\boldsymbol{u}}$ such that $N_{\boldsymbol{u}} \geq \hat{N}_{\boldsymbol{u}}$ and $G^2(N_{\boldsymbol{u}}) \geq \chi^2_\alpha(1)$.

4 An application

In order to choose the appropriate number of latent classes, we fitted a preliminary model whose regression component was, in the light of prior information possibly over parameterized. Because the covariate *year* should affect the functioning of the lists and not that of the latent classes, we assumed that:

(i) the marginal logits of the latent have the form

$$\boldsymbol{\eta}_{i0} = \boldsymbol{\beta}_0 + x_{ia}\boldsymbol{\beta}_{0a} + x_{is}\boldsymbol{\beta}_{0s} + x_{ia}x_{is}\boldsymbol{\beta}_{0as}, \tag{10}$$

where x_{ia} is the age of the i-th subject and x_{is} equals -1 for male and 1 for female;

(ii) the logits of the capture probabilities given the latent are defined as follows:

$$\begin{aligned} \boldsymbol{\eta}_{i1} \;=\; & \boldsymbol{\beta}_1 + x_{iy}(\mathbf{1}_c \otimes \boldsymbol{\beta}_{1y}) + x_{ia}(\mathbf{1}_c \otimes \boldsymbol{\beta}_{1a}) + x_{is}(\mathbf{1}_c \otimes \boldsymbol{\beta}_{1s}) + \\ & + x_{iy}x_{ia}(\mathbf{1}_c \otimes \boldsymbol{\beta}_{1ya}) + x_{iy}x_{is}(\mathbf{1}_c \otimes \boldsymbol{\beta}_{1ys}) + x_{ia}x_{is}(\mathbf{1}_c \otimes \boldsymbol{\beta}_{1as}), \end{aligned} \tag{11}$$

where x_{iy} is equal to -1 or 1 depending on whether the year of first detection was before 2000 or not; note that regression coefficients are constant across latent classes;

notice that all bivariate and higher order interactions (given the latent) are, for the moment, constrained to 0.

The maximum log-likelihood ($\hat{L}_C(c)$) as a function of the number c of latent classes and the $\text{BIC}(c) = -2\hat{L}_C(c) + \log(n)d(c)$, where $d(c)$ denotes the number of parameters, are given in Table 3 below for values of c between 1 and 4.

These results seem to indicate that three latent classes should be adequate to represent unobserved heterogeneity; we denote this model by M_0. Certain meaningful restrictions were tested in an attempt to simplify M_0. Taken as a whole, the hypothesis that covariates do not affect the marginal distribution of the latent weights and the effectiveness of the

c	$d(c)$	$\hat{L}_C(c)$	$\delta(c)$	BIC(c)
1	21	-3918.3	-	8005.2
2	28	-3794.8	17.6	7814.6
3	35	-3762.6	4.6	7806.2
4	42	-3749.0	1.9	7835.3

Table 4: *Fit of the preliminary model as a function of the number of latent classes:* $d(c)$ = *number of parameters,* $\delta(c) = [\hat{L}_C(c) - \hat{L}_C(c-1)]/[d(c) - d(c-1)]$

lists can be easily rejected as well as the more specific assumption of no interaction between covariates in the two components of the linear predictor. However, inspection of the estimates and the corresponding standard errors under M_0 suggested the following list of possible simplifications:

(i) there is no interaction between age and sex on the marginal logits of the latent: $\beta_{0as1} = \beta_{0as2} = 0$;

(ii) there is no interaction between year and age and between year and sex on the effectiveness of list HIV: $\beta_{1ya1} = \beta_{1ys1} = 0$;

(iii) the effectiveness of list AIDS does not depend on covariates: $\beta_{1y2} = \beta_{1a2} = \beta_{1s2} = \beta_{1ya2} = \beta_{1ys2} = \beta_{1as2} = 0$;

(iv) the effectiveness of list DRLH does not depend on sex: $\beta_{1s3} = \beta_{1ys3} = \beta_{as3} = 0$.

Model M_4, obtained by imposing the above restrictions upon M_0, has a deviance of 19.07 with 13 degrees of freedom and a p-value of 0.121. We finally tried to relax the assumption of conditional independence given the latent; however, no pair of lists seems to exhibit a significant association given the latent and the covariates and so M_4 is our final model. The analysis of the deviance leading to the selection of this model is displayed in Table 5 below.

Model	Description	Deviance	d.f.	p-value
M_0:	$c = 3$ with $\boldsymbol{\eta}_{i0}$ as in (10) and $\boldsymbol{\eta}_{i1}$ as in (11)	-	-	-
M_1:	M_0 with $\boldsymbol{\beta}_{0a} = \boldsymbol{\beta}_{0s} = \boldsymbol{\beta}_{0as} = \mathbf{0}$	78.94	6	$< 10^{-4}$
M_2:	M_0 with $\boldsymbol{\beta}_{1y} = \boldsymbol{\beta}_{1a} = \boldsymbol{\beta}_{1s} = \boldsymbol{\beta}_{1ya} = \boldsymbol{\beta}_{1ys} = \boldsymbol{\beta}_{1as} = \mathbf{0}$	488.36	18	$< 10^{-4}$
M_3:	M_0 with $\boldsymbol{\beta}_{0as} = \mathbf{0}$, $\boldsymbol{\beta}_{1ya} = \boldsymbol{\beta}_{1ys} = \boldsymbol{\beta}_{1as} = \mathbf{0}$	55.00	11	$< 10^{-4}$
M_4:	M_0 with restrictions (i), (ii), (iii) and (iv)	19.07	13	0.121

Table 5: *List of marginal regression models fitted on the HIV dataset*

Table 6 gives the estimates of the weights of the latent classes and the capture probability of each list (given the latent) for a "reference subject" with $x_{iy} = x_{is} = 0$ and $x_{ia} = 36.1$ (average age in the sample).

	Latent class		
	1	2	3
Class weight	0.8414	0.0146	0.1440
Conditional prob. HIV	0.2419	0.0031	0.5203
Conditional prob. AIDS	0.0111	0.0254	0.2231
Conditional prob. DRLH	0.0415	0.7707	0.9964

Table 6: *Estimates of class weights and conditional probability of appearing in the lists based on M_4, for a subject with covariate values equal to the sample average*

Parameter	Estimate	Standard error	p-value
Effect of sex on η_{i01} (β_{0s1})	3.7473	0.8513	$< 10^{-4}$
Effect of sex on η_{i02} (β_{0s2})	-3.7188	0.8516	$< 10^{-4}$
Effect of age on η_{i01} (β_{0a1})	0.1697	0.0203	$< 10^{-4}$
Effect of age on η_{i02} (β_{0a2})	-0.1186	0.0214	$< 10^{-4}$

Table 7: *Estimates of the regression coefficients for the marginal logits of the latent class weights*

Parameter	Estimate	Standard error	p-value
Effect of year on H (β_{1y1})	0.2845	0.0646	$< 10^{-4}$
Effect of sex on H (β_{1s1})	0.6608	0.1940	0.0007
Effect of age on H (β_{1a1})	0.0129	0.0095	0.1740
Interaction age.sex on H (β_{1as1})	0.0282	0.0080	0.0005
Effect of year on D (β_{1y3})	-1.7776	0.3438	$< 10^{-4}$
Effect of age on D (β_{1a3})	-0.0579	0.0109	$< 10^{-4}$
Interaction year.age on D (β_{1ya3})	-0.0351	0.0101	0.0005

Table 8: *Estimates of the regression coefficients for the conditional univariate logits of the lists*

Table 6 indicates that latent class 1 is the most common and contains subjects with the smallest chance of being detected by AIDS and DRLH. Latent class 2 is very small (but its size increases with age) and contains subjects for whom the probability of being detected by DRLH is quite high, while that of being detected by HIV or AIDS is rather small. Latent class 3 corresponds to cases with the highest probability of being detected by HIV, AIDS

and DRLH. Table 7 indicates that the probability of being in class 2 increases with age and is larger for females than for males. Finally, Table 8 indicates that the effectiveness of list HIV is larger in the second period (2000-2003) for females (compared to males) and increases with age. The probability of being detected by DRLH, instead, is much smaller in the second period than in the first period and decreases with age.

The estimate of N based on M_4 equals 7495.9, with the number of missing subjects being 4416.9 (58.92%). As illustrated in Figure 1, the 95% confidence interval for N computed on the basis of the statistic $G^2(N)$ described in Section 3.4 is $(6126.1, 9660.3)$, so that undercount should be between 49.7% and 68.1%. We also computed a 95% confidence interval for the number of males in the population, N_m, and for the number of females, N_f (see Figure 2). The interval for N_m is $(4442.9, 7374.7)$, which obviously contains the point estimate $\hat{N}_m = 5581.5$, while that for N_f is $(1582.7, 2466.2)$ around the point estimate $\hat{N}_f = 1914.3$.

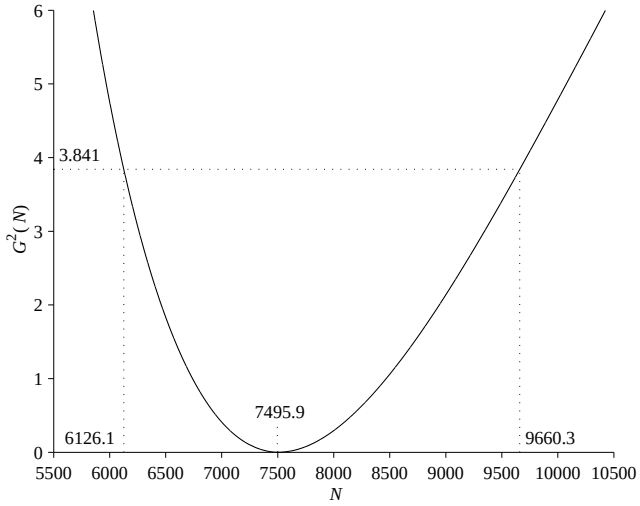

Figure 1: *Confidence interval for the population size based on model M_4*

4.1 Discussion

The size of the population estimated by Pezzotti et al. (2003) based on the same 3 lists and a saturated log-linear model equals 12682. There are two possible explanations for obtaining substantially different results. One is that their lists refer to periods of different length, with the list AIDS starting from 1983, HIV from 1988 and DRLH from 1997. In addition, we note that, when the same log-linear model is fitted to our data, an estimate of 9013 is obtained.

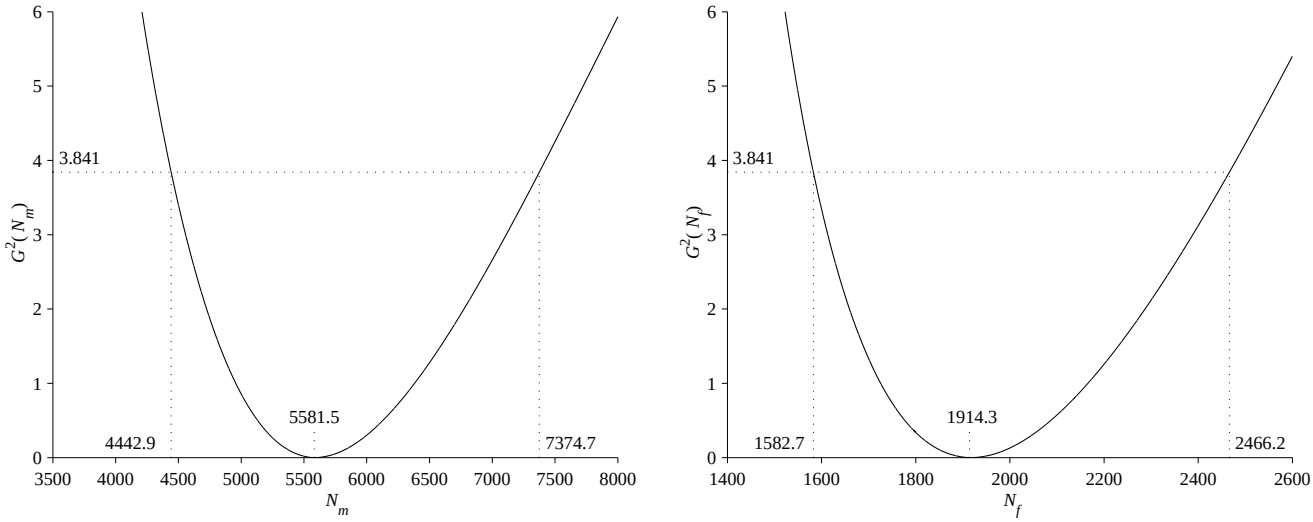

Figure 2: *Confidence intervals for the size of the population of males (N_m) and of females (N_f) based on model M_4*

This seems to suggest that, because of the structure of the data, ignoring observed and latent heterogeneity may lead to a substantial over-estimation of the population size.

The period of first appearance clearly poses a substantial problem. We could use all the available cases since 1983; there were however several reasons for not doing so. One is that, with such a long period, the assumption of closed population is untenable and it becomes less clear what population size we are really talking about. For instance, from the data provided by Pezzotti et al. (2003), the estimated number of deaths within the population of interest from 1983 to 2003 may be set at over 2500. In addition, because in certain periods some lists were not active, their capture probability during the same period would have to be constrained to 0 and this would have introduced technical difficulties.

A similar modelling approach was applied also to a dataset provided by the Tuscany Cancer Registry and concerning 7253 new cases of cancer detected during the year 2000. The data were produced by linking discharge reports from local hospitals, reports from the pathologic anatomy units in Tuscany and death certificates issued from 2000 to 2002 concerning new cases of cancer detected during the year 2000 by one of the previous lists or by a private doctor; see Crocetti et al. (2001) for a description of the sources and the criteria used for data collection. Although for brevity detailed results will not be described here in detail, there are two features which are worth mentioning. One is the very high coverage of this registry, which is over 99.8%. We also found a significant violation of conditional independence with

a significantly negative association between hospitals and pathology unit reports.

Acknowledgments

We would like to thank Dr. P. Pezzotti (Agenzia per la Salute, Regione Lazio) and Dr. C. Piovesan (Servizio di Epidemiologia e Sanità Pubblica, Regione Veneto) for providing the HIV dataset, describing the context and discussing the results. We are indebted to Dr. E. Crocetti and Dr. E. Paci (Centro per lo Studio e la Prevenzione Oncologica, Regione Toscana) for providing the Tuscany Cancer Registry dataset. Finally, we thank an Associate Editor and two Referees for stimulating comments and we acknowledge the financial support of the research project MIUR 2002 "Ordinamenti stocastici nell'analisi della dipendenza in tabelle multiple con applicazioni socio-sanitarie"

Appendix

Construction of the $\boldsymbol{C}$, $\boldsymbol{M}$ and $\tilde{\boldsymbol{G}}$ matrices

The matrix $\boldsymbol{C}$ is block diagonal with blocks of the form $(\,-1 \quad 1\,)$ for univariate logits, $(1\,-1\,-1\;1)$ for log-odds ratios and so on. Let $\boldsymbol{c}_h$ denote the hth row of $\boldsymbol{I}_c$; then the matrix $\boldsymbol{M}$ may be obtained by stacking one below the other blocks of rows of the form $\boldsymbol{M}_{\mathcal{I}} = \bigotimes_{i=1}^{J+1} \boldsymbol{M}_{\mathcal{I},i}$ where $\mathcal{I}$ may be:

- the hth marginal logit of the latent, then $\boldsymbol{M}_{\mathcal{I},i}$ is equal to $(\,\boldsymbol{c}'_{h-1} \quad \boldsymbol{c}'_h\,)'$ if $i = 1$ and to $\mathbf{1}'_2$ otherwise;
- the univariate logit of list j given the latent equal to h, then $\boldsymbol{M}_{\mathcal{I},i}$ is equal to $\boldsymbol{c}_h$ for $i = 1$, to $\boldsymbol{I}_2$ for $i = j + 1$ and to $\mathbf{1}'_2$ otherwise;
- the log-odds ratio between lists j_1, j_2 given the latent equal to h, then $\boldsymbol{M}_{\mathcal{I},i}$ is equal to $\boldsymbol{c}_h$ for $i = 1$, to $\boldsymbol{I}_2$ for $i = j_1 + 1$ or $i = j_2 + 1$ and to $\mathbf{1}'_2$ otherwise.

The matrix $\tilde{\boldsymbol{G}}$ has a block of columns for each effect; the block corresponding to the marginal logits of the latent is a Kroneker product of $\boldsymbol{I}_c$ without the first column and $\mathbf{1}_{k+1}$; the set

of univariate logits for list j correspond to the Kroneker product between $\boldsymbol{I}_c$ and $\bigotimes_{l=1}^{J}\tilde{\boldsymbol{G}}_l$ where $\tilde{\boldsymbol{G}}_l$ is equal to $\boldsymbol{I}_2$ if $l=j$ and to $\mathbf{1}_2$ otherwise.

The covariance matrix of $\hat{\boldsymbol{\beta}}$

Use the following chain rule to differentiate L_C:

$$\boldsymbol{s}' = \frac{\partial L_C}{\partial \boldsymbol{\beta}'} = \sum_i \frac{\partial L_C}{\partial \boldsymbol{\gamma}_i'}\frac{\partial \boldsymbol{\gamma}_i}{\partial \boldsymbol{\theta}_i'}\frac{\partial \boldsymbol{\theta}_i}{\partial \boldsymbol{\eta}_i'}\frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\beta}'} = \sum_i [(\boldsymbol{y}_i - n_i\boldsymbol{\rho}_i)'\boldsymbol{G}](\boldsymbol{H}\boldsymbol{U}_i)\boldsymbol{R}_i\boldsymbol{X}_i,$$

where

$$\begin{aligned}\boldsymbol{U}_i &= \frac{\partial \log(\boldsymbol{\rho}_i)}{\partial \boldsymbol{\theta}_i'} = \mathrm{diag}(\boldsymbol{\rho}_i)^{-1}\frac{q_i\boldsymbol{A} - \boldsymbol{p}_i\boldsymbol{a}'}{q_i^2}\tilde{\boldsymbol{\Omega}}_i\tilde{\boldsymbol{G}} = \\ &= \mathrm{diag}(\boldsymbol{\rho}_i)^{-1}(\boldsymbol{A} - \boldsymbol{p}_i\boldsymbol{a}'/q_i)\tilde{\boldsymbol{\Omega}}_i\tilde{\boldsymbol{G}}/q_i = \\ &= \{\mathbf{1}_c' \otimes [\mathrm{diag}(\boldsymbol{\rho}_i)^{-1} - \mathbf{1}_k\mathbf{1}_k']\bar{\boldsymbol{I}}\}\tilde{\boldsymbol{\Omega}}_i\tilde{\boldsymbol{G}}/q_i.\end{aligned}$$

The information matrix $\boldsymbol{F}$ may be obtained as the expected value of $\boldsymbol{s}\boldsymbol{s}'/n$ by noting that $E[(\boldsymbol{y}_i - n_i\boldsymbol{\rho}_i)(\boldsymbol{y}_i - n_i\boldsymbol{\rho}_i)'] = n_i\boldsymbol{\Omega}_i$. This simplifies to (5) because, irrespective of the specific form of $\boldsymbol{H}$, $\boldsymbol{G}\boldsymbol{H} = \boldsymbol{I}_k - \mathbf{1}_k\mathbf{1}_k'/k$ and so $\boldsymbol{H}'\boldsymbol{G}'\boldsymbol{\Omega}_i\boldsymbol{G}\boldsymbol{H} = \boldsymbol{\Omega}_i$.

The covariance matrix $\boldsymbol{\Sigma}_0$

The second derivative matrix of $F(\boldsymbol{\lambda})$, $\boldsymbol{V}(\boldsymbol{\lambda})$, has elements

$$\frac{\partial^2 F(\boldsymbol{\lambda})}{\partial\alpha_i\partial\alpha_j} = \frac{\delta_{ij}f_i}{\alpha_i(\alpha_i - f_i)}, \qquad \frac{\partial^2 F(\boldsymbol{\lambda})}{\partial\alpha_i\partial\beta_j} = -\frac{1}{p_{i,\mathbf{0}}}p_{i,\mathbf{0}}^{(j)},$$

$$\frac{\partial^2 F(\boldsymbol{\lambda})}{\partial\beta_j\partial\beta_l} = -\sum_i\left[\frac{\alpha_i - f_i}{p_{i,\mathbf{0}}}\left(p_{i,\mathbf{0}}^{(jl)} - \frac{p_{i,\mathbf{0}}^{(j)}p_{i,\mathbf{0}}^{(l)}}{p_{i,\mathbf{0}}}\right) + \sum_{\boldsymbol{r}\neq\mathbf{0}}\frac{f_{i,\boldsymbol{r}}}{p_{i,\boldsymbol{r}}}\left(p_{i,\boldsymbol{r}}^{(jl)} - \frac{p_{i,\boldsymbol{r}}^{(j)}p_{i,\boldsymbol{r}}^{(l)}}{p_{i,\boldsymbol{r}}}\right)\right]$$

where $\delta_{ij}=1$ if $i=j$ and 0 otherwise, $p_{i,\boldsymbol{r}}^{(j)}$ denotes the first derivative $p_{i,\boldsymbol{r}}$ with respect to β_j and $p_{i,\boldsymbol{r}}^{(jl)}$ denotes the second derivative with respect to β_j and β_l. Now consider the matrix $\boldsymbol{V}_0 = \boldsymbol{V}(\boldsymbol{\lambda}_0)$; condition (9) implies that, as $N_0\to\infty$, $f_i \to^p c_i(1-p_{i,\mathbf{0}})$ and $f_{i,\boldsymbol{r}} \to^p c_i p_{i,\boldsymbol{r}}$ $i=1,\ldots,k$. It follows that $\boldsymbol{V}_0 \to^p \boldsymbol{\Sigma}_0^{-1}$ the information matrix given by Sanathanan (1972), which has elements

$$s_0(\alpha_i,\alpha_j) = \frac{\delta_{ij}(1-p_{i,\mathbf{0}})}{c_i p_{i,\mathbf{0}}}, \qquad s_0(\alpha_i,\beta_j) = -\frac{1}{p_{i,\mathbf{0}}}p_{i,\mathbf{0}}^{(j)}, \qquad s_0(\beta_j,\beta_l) = \sum_i c_i\sum_{\boldsymbol{r}}\frac{p_{i,\boldsymbol{r}}^{(j)}p_{i,\boldsymbol{r}}^{(l)}}{p_{i,\boldsymbol{r}}}.$$

References

Alho, J. M. (1990), Logistic Regression in Capture-Recapture Models, *Biometrics*, **46**, pp. 623-635.

Baker, S. G. (1990), A simple EM algorithm for capture-recapture data with categorical covariates, *Biometrics*, **46**, pp. 1193-1200.

Bandeen-Roche, K., Miglioretti, D. L., Zeger, S. L. and Rathouz, P. J. (1997), Latent variable regression for 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.

Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. (1975), *Discrete Multivariate Analysis*, Cambridge, Massachussets, MIT Press.

Cormack, R. M. (1992), Interval estimation for mark-recapture studies of closed populations, *Biometrics*, **48**, pp. 567-578.

Coull, B. A. and Agresti, A. (1999), The use of mixed logits models to reflect heterogeneity in capture-recapture studies, *Biometrics*, **55**, pp. 294-301.

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.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977), Maximum Likelihood from Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society, Series B*, **39**, pp. 1-22.

Dorazio, R. M. and Royle, J. A. (2003), Mixture models for estimating the size of a closed population when capture rates vary among individuals, *Biometrics*, **59**, pp. 351-364.

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.

Pezzotti, P., Piovesan, C., Michieletto, F., Zanella, F., Rezza, G. and Gallo, G. (2003), Estimating the cumulative number of human immunodeficiency virus diagnoses by cross-linking from four different sources, *International Journal of Epidemiology*, **32**, pp. 778-783.

Plante, N. Rivest, L. P. and Tremblay, G. (1998), Stratified capture-recapture estimation of the size of a closed population, *Biometrics*, **54**, pp. 47-60.

Pledger, S. (2000), Unified maximum likelihood estimation for closed capture-recapture models using mixtures, *Biometrics*, **56**, pp. 434-442.

Pollock, K. H. (2000), Capture-recapture models, *Journal of the American Statistical Association*, **95**, pp. 293-296.

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.

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.

Schwarz, C. J. and Seber, G. A. F. (1999), Estimating animal aboundance: Review III, *Statistical Science*, **14**, pp. 427-456.

Zwane, E. and van der Heijden, P. (2005), Population estimation using the multiple system estimator in the presence of continuous covariates, *Statistical Modelling*, **5**, pp. 39-52.