Biometrics 60, 510–516
June 2004

A Multiple-Record Systems Estimation Method that Takes Observed and Unobserved Heterogeneity into Account

Elena Stanghellini

Dipartimento di Scienze Statistiche, Università di Perugia, 06100 Perugia, Italy
email: elena.stanghellini@stat.unipg.it

and

Peter G. M. van der Heijden

Department of Methodology and Statistics, Utrecht University, P.O. Box 80.140,
3508 TC Utrecht, The Netherlands
email: p.vanderheijden@fss.uu.nl

Summary. We present a model to estimate the size of an unknown population from a number of lists that applies when the assumptions of (a) homogeneity of capture probabilities of individuals and (b) marginal independence of lists are violated. This situation typically occurs in epidemiological studies, where the heterogeneity of individuals is severe and researchers cannot control the independence between sources of ascertainment. We discuss the situation when categorical covariates are available and the interest is not only in the total undercount, but also in the undercount within each stratum resulting from the cross-classification of the covariates. We also present several techniques for determining confidence intervals of the undercount within each stratum using the profile log likelihood, thereby extending the work of Cormack (1992, *Biometrics* **48,** 567–576).

Key words: Conditional independence model; Extended latent class model; Graphical models; Hierarchical log linear models; Identifiability; Observed information matrix; Profile log likelihood; Strata.

1. Introduction

In epidemiology multiple-record systems estimation methods are employed to estimate the size of an unknown population. Traditionally this was done using a log linear model. If there are two lists of individuals, it is necessary to assume independence of the two probabilities to be included in the lists. If there are more than two lists, and if other external information is not available, then it is possible to allow for interaction between the lists. An overview of this approach can be found in Bishop, Fienberg, and Holland (1975). In this traditional approach an assumption is that in each record system the inclusion probabilities are homogeneous over the individuals, i.e., individuals have the same probability of being caught (Chao et al., 2001). Violation of this homogeneity assumption will result in dependence between lists (see, for example, International Working Group for Disease Monitoring and Forecasting, 1995). Many strategies are proposed to go around this restrictive and often unrealistic assumption and the objective of this article is to extend one of these approaches.

Originally Bishop et al. (1975) proposed to tackle possible heterogeneity by stratifying the sample in such a way that in each stratum the homogeneity would be fulfilled. This option is possible if there is an observed categorical covariate that is closely related to the probability of being in each of the lists. Later Alho (1990) and Huggins (1991) generalized this approach by proposing a model where the capture probabilities of each list are a function of observed covariates that can be categorical or continuous. These two strategies have in common that they allow for observed heterogeneity, i.e., the heterogeneity of capture probabilities can be taken into account by observed covariates.

Alternatively, models are proposed for situations where such covariates are not available. One approach makes use of log linear models with homogeneous two-factor interactions (see International Working Group for Disease Monitoring and Forecasting, 1995); a second one makes use of latent class models, where it is assumed that the individuals can be classified in a small number of groups (the latent classes) with homogeneous capture probabilities and lists are independent conditional on the latent classes (see Agresti, 1994; Coull and Agresti, 1999); a third makes use of Rasch models, where it is assumed that the individuals differ on a continuous scale (see Darroch et al., 1993; Coull and Agresti, 1999; Fienberg, Johnson, and Junker, 1999). These three modeling strategies are closely related due to the relation that exists between the Rasch model, quasi-symmetry models, and latent class analysis (see Lindsey, Clogg, and Greco, 1991).

More recently, Biggeri et al. (1999) account for local dependence between a pair of lists in a latent class model. This approach is justified in epidemiological studies to estimate the size of a human population as there is unobserved heterogeneity between individuals and lists are set up for various purposes, so that the independence between lists is not guaranteed. A parallel approach, which makes use of the Rasch model, has been proposed by Bartolucci and Forcina (2001). Their model includes parameters for the associations between two lists, after conditioning on the unobserved heterogeneity and marginalizing with respect to the other lists.

In this article we extend the approach of Biggeri et al. (1999) by including a covariate in the model. We use a hierarchical log linear model with one categorical latent variable representing the unobserved heterogeneity. The model can be represented via a conditional independence graph (Whittaker, 1990; Edwards, 2000) with one unobserved node. As we are dealing with nonstandard latent variable models, we discuss identifiability by assessing the rank of the information matrix. We provide separate estimates for the undercount in each stratum as defined by the covariate and complement the estimates with confidence intervals evaluated by using the profile log likelihood, thereby extending the work of Cormack (1992).

When we compare modeling each stratum separately with simultaneous modeling over strata, the advantage of the latter is that it allows to make the model more parsimonious by restricting certain parameters to be equal over the strata and to test for these equalities. We reanalyze the data on prevalence of diabetes in a town in northern Italy (Bruno et al., 1994) using a covariate related to the severity of the illness. The results are compared with other studies made on the same data set.

In Section 2 we present the general framework of log linear models with one latent categorical variable in the capture–recapture context, while in Section 3 we present the modeling strategy and address the issue of identifiability. In Section 4 we describe how confidence intervals for the undercount of each stratum can be derived and in Section 5 we report our analyses on the prevalence of diabetes in a small area of northern Italy. In Section 6 we draw our conclusions.

2. Generalities

Assume that there are R lists denoted by S_r, $r = 1, \dots, R$, with levels indexed by s_r, where $s_r = 1$ indicates being on list r and $s_r = 0$ indicates not being on list r. We denote a categorical latent variable by U with levels indexed by h, $h = 1, \dots, H$. The latent variable U accounts for unobserved heterogeneity of the individuals by assuming that individuals with the same level of h have identical probabilities of inclusion but individuals with different h may have different probabilities of inclusion. Assume also that categorical covariates are available. As we are not interested in modeling the interaction of the covariates, we will assume that we have a unique covariate C with levels indexed by j, $j = 1, \dots, J$, obtained by the cross-classification of the levels of the covariates. The interest is in estimating the undercount for each level j of C.

Let $\mathbf{s} = s_1, \dots, s_R$ denote a certain pattern of captures. Then let $X_{hj\mathbf{s}}$ refer to the counts with $U = h$, $C = j$ and a pattern of captures $\mathbf{s}$. We assume $X_{hj\mathbf{s}}$ to be independent Poisson random variables. We denote with X the vector of $X_{hj\mathbf{s}}$ with the convention that S_R is running fastest, and let $E(X) = m$.

In defining the likelihood for our model we should take into account that U is not observed. Let $t = 2^R$. The marginal entries $Y_{j\mathbf{s}} = \sum_{h=1}^{H} X_{hj\mathbf{s}}$ are also independent Poisson random variables, with expected values $E(Y_{j\mathbf{s}}) = \mu_{j\mathbf{s}}$. We denote by Y^* the vector with the counts in the marginal table, stacked in a way that $Y^* = L^*X$ with $L^* = \mathbf{1}'_H \otimes I_{J\times t}$. It then follows that $E(Y^*) = \mu^* = L^*m$. Moreover, as the entries in the cells where $\mathbf{s} = 0, \dots, 0$, are zero by construction, the first entry of each stratum j is not observable. We denote with Y the vector of the observable data obtained from Y^* by removing the entries $1,\ t+1, \dots, (J-1)t+1$. Note that $Y = LX$, with $L = \mathbf{1}'_H \otimes I_{J\times(t-1)}$. We will refer to X as the complete data vector and to Y as the incomplete data vector.

Let $n = \{n_j\}$ be the vector of the observed counts in stratum j and $N = \{N_j\}$ be the vector of the total counts in stratum j. The model considered in this article can be written as $\log(m) = Z\beta$ with Z a design matrix and β the vector of parameters. The unknown parameters of the models are N and β and their estimates should be derived by maximizing the log likelihood $l(y^* \mid M^*Y^* = N; \beta, N)$, with $M^* = I_J \otimes \mathbf{1}'_t$, where

$$P(Y^* \mid M^*Y^* = N)$$
$$= \prod_j \binom{N_j}{n_j} \left\{1 - p_j^*(\beta)\right\}^{N_j - n_j} p_j^*(\beta)^{n_j} n_j! \prod_{\mathbf{s}:\mathbf{s}\neq\mathbf{0}} q_{j\mathbf{s}}(\beta)^{y_{j\mathbf{s}}} \frac{1}{y_{j\mathbf{s}}!}$$

in which $p_{j\mathbf{s}} = \mu_{j\mathbf{s}}/\sum_{\mathbf{s}} \mu_{j\mathbf{s}}$, $p_j^* = \sum_{\mathbf{s}:\mathbf{s}\neq\mathbf{0}} p_{j\mathbf{s}}$, and $q_{j\mathbf{s}} = p_{j\mathbf{s}}/p_j^*$.

However, maximum likelihood estimates for β and N are usually derived in two steps: (a) the maximum likelihood estimate of β is derived by maximizing the conditional log likelihood $l(y \mid MY = n; \beta)$ with $M = I_J \otimes \mathbf{1}'_{t-1}$; (b) the vector of the conditional estimates of N_j is derived as the integer part of $n_j/\hat{p}_j^*$ (see Bishop et al., 1975, p. 237).

Although point estimates of the undercount are derived conditionally on the n_j, confidence intervals will be evaluated by both considering n_j as fixed or as random variables.

3. The Model

As already noticed, when estimating the size of a human population by means of incomplete lists we cannot always assume the lists to act independently, even after controlling for the observed and unobserved heterogeneity. Furthermore, we do not exclude the covariate C to be conditionally associated with the lists, even after conditioning on all the other variables.

As there is no clear ordering between the variables, we consider all variables on an equal footing and we use a log linear model with one unobserved variable. We restrict to the class of models that are hierarchical. These models can be represented with an independence graph, that is a diagram in which the nodes correspond to variables. An edge between two nodes is missing whenever the two corresponding variables are independent given all the other variables. The conditioning set for this independence to hold can be reduced to its minimum by using the parallelism between conditional independence properties and the notion of separation (see Edwards, 2000).

However, not all log linear models with one latent variable are locally identifiable and the identification state should then be checked. Rothenberg (1971) shows that a way to assess local identifiability is by the rank of the observed information matrix evaluated at the maximum of the likelihood (see also Catchpole and Morgan, 1997, for a general discussion of local identifiability). Goodman (1974) discusses local identifiability of latent class models when the assumption of local independence between the observed variables holds.

We provide a separate estimate of the unobserved count in each stratum obtained from the cross-classification of the covariates. We complement the estimates with confidence intervals for the unobserved count in each stratum using both the profile log likelihood and the delta method.

Maximum likelihood estimation of the model can be performed with the EM algorithm. The EM algorithm alternates the following E and M steps until convergence:

E-step: The conditional expectation of the $E(X \mid LX = y)$ is computed given the current best parameter estimates of β.

M-step: New parameter estimates are computed by fitting a log linear model for the complete data X.

Due to the close similarity with the latent class model we refrain from describing the algorithm in detail but refer to Agresti and Lang (1993).

The observed information matrix can be obtained by differentiating twice $l(y; \beta)$ with respect to β. Alternatively we can use the results of Lang (1992) that showed that, for Poisson log linear models, Louis's formula (Louis, 1982) for the observed information matrix can be reformulated in a simplified way. Let the vector of the complete data X be Poisson distributed and the vector Y of the observed data be $Y = TX$ with T such that (i) each element is a '0' or a '1' and (ii) there is at most one '1' per column. Then the observed information matrix of the incomplete data $I(\beta \mid y = Tx)$ can be derived as $I(\beta \mid y = Tx) = Z'\{\mathrm{var}(X; \beta) - \mathrm{var}(X \mid TX = y; \beta)\}Z$. In our case, $T = L$ and conditions (i) and (ii) are fulfilled. Evaluation of the rank of the observed information matrix shows that all models considered in this article are locally identifiable.

Moreover, if we are not interested in modeling the marginal distribution of n_j but we regard them as fixed, we should evaluate the information matrix accordingly. This again can be done by differentiating twice the log of the conditional likelihood $l(y \mid MY = n; \beta)$ with respect to β. In doing so, we may note that $Y \mid MY$ is equally distributed as $X \mid TX$ with $T = \mathbf{1}'_H \otimes M$ with T satisfying the conditions (i) and (ii) above and therefore the observed information can be derived using the same procedure.

4. Confidence Intervals

After a suitable model is selected and the local identification is assessed, the profile log likelihood method can be used to derive confidence intervals for the undercount in each stratum. Here we extend the procedure proposed by Cormack (1992) to a situation with multiple strata. This will be done by externally assigning a value for the undercount in a given stratum r, and then deriving the maximum likelihood estimate for the model including that entry. The estimates of N_j, $j \neq r$ are then derived conditionally as described in Section 2. The procedure can be easily implemented in any software that allows to give external weights to the entries of a contingency table.

With $\hat{N}(\beta)$ we indicate the conditional estimates of N obtained as a function of $\hat{\beta}$. Let $V_r = v_r$ be the random variable associated with the first entry on stratum r and $Y_r = y_r$ the vector of the random variables Y augmented by V_r. With v we denote the vector obtained from n by substituting the rth element with $n_r + v_r$. Let $\hat{\beta}_r$ denote the value of β that maximizes $l(y_r \mid M_r Y_r = v; \beta)$ where M_r is a matrix obtained from M by inserting a column in position $r(t-1)$. The column to be inserted is e_{Jr}, that is the rth column of I_J. Let $\hat{N}(\beta_r)$ be the vector with $n_r + v_r$ as the rth element and the conditional estimate $\hat{N}_j(\beta_r)$ as the jth element, $j \neq r$. The unconditional Poisson profile log likelihood for v_r is then obtained as $P_P(v_r) = 2\{l(y^*; \hat{\beta}) - l(y^*; \hat{\beta}_r)\}$.

Analogously, the unconditional multinomial profile log likelihood for v_r is obtained as $P_M = 2[l\{y^* \mid M^*Y^* = \hat{N}(\beta)\} - l\{y^* \mid M^*Y^* = \hat{N}(\beta_r)\}]$. By a parallel argument of Cormack (1992), Theorem 2, if $v_r = \hat{N}_r(\beta) - n_r$ then $\hat{\beta}_r = \hat{\beta}$. However, in general, $\hat{N}_j(\beta) \neq \hat{N}_j(\beta_r)$. We indicate with $\mathrm{Dev}_p(Y_r; \beta_r)$ the deviance of the Poisson model on Y_r with parameter β_r and with $\mathrm{Dev}_p(Y; \beta)$ the deviance of the Poisson model on Y with parameter β. The unconditional profile log likelihood can be obtained as a function of their difference. As an example, the unconditional multinomial profile log likelihood can be written as

$$\begin{aligned} P_M(v_r) = \mathrm{Diff} + 2\Bigg[& \log \frac{\hat{N}_r(\beta)!}{(n_r + v_r)!} - \log \frac{\{\hat{N}_r(\beta) - n_r\}!}{v_r!} \\ & + \{N_r(\beta) - n_r\} \log\{1 - p_r^*(\beta)\} \\ & - v_r \log\{1 - p_r^*(\beta_r)\} + \sum_{j \neq r} \log \frac{\hat{N}_j(\beta)!}{\hat{N}_j(\beta_r)!} \\ & - \log \frac{\{\hat{N}_j(\beta) - n_j\}!}{\{\hat{N}_j(\beta_r) - n_j\}!} + \{\hat{N}_j(\beta) - n_j\} \\ & \times \log\{1 - p_j^*(\hat{\beta})\} - \{\hat{N}_j(\beta_r) - n_j\} \\ & \times \log\{1 - p_j^*(\hat{\beta}_r)\} + \sum_j n_j \log \frac{p_j^*(\beta)}{p_j^*(\beta_r)} \Bigg], \end{aligned}$$

where $\mathrm{Diff} = \mathrm{Dev}_p(Y_r; \beta_r) - \mathrm{Dev}_p(Y; \beta)$. Note that the correction is summed over all strata. Note also that, as $\hat{N}_j(\beta)$ are not full maximum likelihood estimates, in the neighborhood of the unconditional estimates there may be terms taking negative values.

If we consider the n_j as fixed by the design, the conditional profile log likelihood can be derived as $P_C = 2\{l(y_r \mid M_r Y_r = \hat{n}; \hat{\beta}) - l(y_r \mid M_r Y_r = v; \hat{\beta}_r)\}$, where $\hat{n}$ is obtained from n by substituting the rth element with $\hat{N}_r(\beta)$.

A $(1 - \alpha)\%$ confidence interval is given by (n_{r1}, n_{r2}) where $n_{r1}(n_{r2})$ is the largest (smallest) integer smaller (larger) than $\hat{n}$ such that $P(n_{r1}) \geq \chi^2_{1,\alpha}\{P(n_{r2}) \geq \chi^2_{1,\alpha}\}$ and $\chi^2_{1,\alpha}$ is the $100\alpha\%$ critical value of the χ^2 with one degree of freedom. Note that this procedure does not extend easily to the evaluation of the confidence interval for the total undercount, as this one is obtained by summing over the undercounts in each stratum. We feel that other modeling strategies should be followed to derive the profile log likelihood for that entity.

Confidence intervals based on the profile log likelihood are approximate, and their validity relies on asymptotic results. A parallel procedure is to derive confidence intervals from the asymptotic distribution of $\hat{\mu}$ with covariance matrix evaluated using the delta method applied to both the unconditional and conditional information matrix. However, this procedure is not to be recommended, as it does not adequately correct for the asymmetry in the log likelihood (see Cormack, 1992).

5. Prevalence of Diabetes

We discuss the prevalence of diabetes in a town in northern Italy (see Bruno et al., 1994 for a detailed description). There are four sources (Table 1), namely the diabetic clinic and/or family physicians data source (S_1), the hospital discharges data source (S_2), the insulin and oral hypoglycerin data source (S_3), and the reagent strips and insulin syringes data source (S_4). There is a categorical covariate C, namely treatment, having three levels: (1) diet, (2) hypoglycemic agents, and (3) insulin.

Preliminary analyses of the interaction between the sources within each stratum show that there is a strong positive interaction between S_2 and S_4 and between S_3 and S_4 in all strata. There is a negative interaction between S_1 and S_2 and between S_1 and S_3 in stratum 1 (diet) and stratum 2 (hypoglycemic); and also a negative interaction between S_1 and S_4 in stratum 2 (hypoglycemic) and stratum 3 (insulin).

Model search is a difficult issue in general, and in particular in the area of multiple system records estimation, where the primary interest goes out into the estimation of the population size. For the moment we use an informal model search procedure making use of forward selection of interaction terms, but it is clear that the topic of model search needs further study, including approaches that make use of model averaging.

In fitting the model with the latent variable, we restricted the model search to the class of hierarchical interaction models. We used the software GLIM4. We started with two latent classes. Thus we fit models to a table of 3 (levels of C), times 2 (levels of latent variable) times ($2^4 - 1$) cells. The sample size is 2047.

Table 1
Data from prevalent cases of known diabetes for residents of Casale Monferrato, Piedmont, Italy

	s	$y_{\mathbf{s}}$	**s**	$y_{\mathbf{s}}$	**s**	$y_{\mathbf{s}}$	**s**	$y_{\mathbf{s}}$
Total								
	0000	–	0100	73	1000	701	1100	101
	0001	10	0101	7	1001	12	1101	18
	0010	180	0110	20	1010	650	1110	156
	0011	8	0111	14	1011	46	1111	58
Diet								
	0000	–	0100	13	1000	150	1100	11
	0001	0	0101	0	1001	0	1101	1
	0010	11	0110	4	1010	11	1110	1
	0011	1	0111	1	1011	1	1111	0
Hypoglycemic agents								
	0000	–	0100	58	1000	509	1100	68
	0001	4	0101	1	1001	5	1101	3
	0010	163	0110	13	1010	571	1110	98
	0011	3	0111	3	1011	9	1111	6
Insulin								
	0000	–	0100	2	1000	42	1100	22
	0001	6	0101	6	1001	7	1101	14
	0010	6	0110	3	1010	62	1110	57
	0011	4	0111	10	1011	36	1111	51

We started with the model $[CU][US_1][US_2][US_3][US_4]$. In a second step we add conditional association terms for those interactions that are not adequately explained by the latent variable. In a third step we take two ways to complicate the model further: First, we allow the conditional associations to be different in different strata, and second, we allow the conditional associations to be different for different levels of the latent variable.

Table 2 shows the results of our model search. Model $[CU][US_1][US_2][US_3][US_4]$ shows a deviance of 252.50 with 31 degrees of freedom. The most significant drop in the deviance is for the model where conditional association between S_1, S_3 is added with a deviance of 207.3 with 30 degrees of freedom. The next most significant drop in the deviance is for the model where the conditional association between S_1, S_4 is added, with a deviance of 191.60 with 29 degrees of freedom.

As the model does not show an acceptable fit, we now allow conditional associations to vary across the levels of the latent variable, but this did not lead to an acceptable model (models 16–18; the best model shows a deviance of 189.67 with 27 degrees of freedom). This is in line with our expectation, as the heterogeneity accounts for the difference between subjects of the probability of ascertainment, whereas we expect conditional associations between lists to be related to the severity of the illness. The second way, allowing conditional associations to vary across the strata, is investigated in models 13–15, and led to the following reasonable model: $[CU][US_1][US_2][US_3][US_4][CS_1S_3][S_1S_4]$ with a deviance of 30.70 and 23 degrees of freedom. This model essentially states that the conditional association between S_1 and S_3 varies across the strata but is constant when varying the latent group. The conditional association between S_1 and S_4 is constant within each combination of stratum and latent group. This model is not decomposable. The corresponding conditional independence graph is shown in Figure 1.

Table 3 reports the parameter estimates with their standard error, evaluated unconditionally on the n_j. All parameters are significant, apart from $C(2) \cdot U$ which indicates that

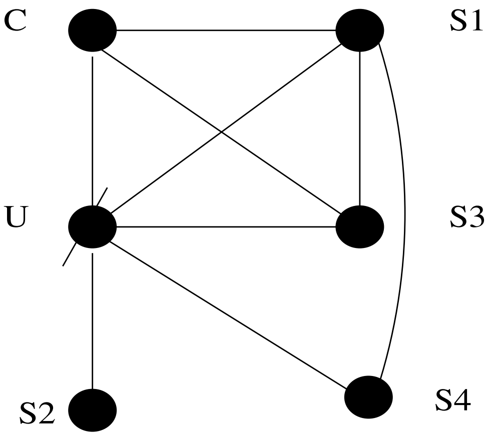

Figure 1. Conditional independence graph for model $[CU][US_1][US_2][US_3][US_4][CS_1S_3][S_1S_4]$.

Table 2
Models investigated. Observed in stratum 1 *is* 205, *in stratum* 2 *is* 1514, *and in stratum* 3 *is* 328 *individuals.*

	Models	Deviance	df	$\hat{N}$ in 1	$\hat{N}$ in 2	$\hat{N}$ in 3
1.	$C \cdot U + U \cdot (S_1 + S_2 + S_3 + S_4)$	252.5	31	233	1719	339
2.	Model 1 + $S_1 \cdot S_2$	243.4	30	230	1698	337
3.	Model 1 + $S_1 \cdot S_3$	207.3	30	268	1974	352
4.	Model 1 + $S_1 \cdot S_4$	233.2	30	233	1722	337
5.	Model 1 + $S_2 \cdot S_3$	251.4	30	233	1715	338
6.	Model 1 + $S_2 \cdot S_4$	251.2	30	233	1720	337
7.	Model 1 + $S_3 \cdot S_4$	252.3	30	233	1719	338
8.	Model 3 + $S_1 \cdot S_2$	206.2	29	279	2052	357
9.	Model 3 + $S_1 \cdot S_4$	191.6	29	269	1980	349
10.	Model 3 + $S_2 \cdot S_3$	220.4	29	271	1999	353
11.	Model 3 + $S_2 \cdot S_4$	207.1	29	268	1972	350
12.	Model 3 + $S_3 \cdot S_4$	207.0	29	268	1974	352
13.	Model 9 + $C \cdot S_1 \cdot S_3$	30.7	23	292	1948	342
14.	Model 9 + $C \cdot S_1 \cdot S_4$	170.4	23	292	2224	351
15.	Model 13 + $C \cdot S_1 \cdot S_4$	27.4	19	398	2928	339
16.	Model 9 + $U \cdot S_1 \cdot S_4$	190.3	28	268	1977	350
17.	Model 9 + $U \cdot S_1 \cdot S_3$	191.1	28	272	2001	349
18.	Model 16 + $U \cdot S_1 \cdot S_3$	190.0	27	271	1993	350

Table 3
Estimates of the parameters of model 13 *and their standard errors*

Parameter	Estimate	SE	Parameter	Estimate	SE
U_0	−5.7314	1.1402	$C(2)$	3.9905	0.0810
$C(3)$	0.7601	0.2235	U	4.4360	1.4410
S_1	1.5061	0.5983	S_2	2.0469	0.0077
S_3	2.5851	0.0851	S_4	6.0462	1.0152
$C(2) \cdot U$	−0.4388	0.4965	$C(3) \cdot U$	4.4419	0.4264
$C(2) \cdot S_1$	−1.6021	0.1431	$C(3) \cdot S_1$	−2.0938	0.3485
$U \cdot S_1$	−1.2105	0.8374	$U \cdot S(2)$	−2.3960	0.0305
$C(2) \cdot S_3$	−2.7005	0.0880	$C(3) \cdot S_3$	−2.2867	0.2017
$U \cdot S_3$	−1.5733	0.1748	$S_1 \cdot S_3$	−0.7703	0.2273
$U \cdot S_4$	−5.9273	0.9543	$S_1 \cdot S_4$	−1.2145	0.4399
$C(2) \cdot S_1 \cdot S_3$	1.9167	0.2462	$C(3) \cdot S_1 \cdot S_3$	1.7246	0.3735

the odds of being in the first latent class when in stratum 1 may not be different from the same odds when in stratum 2.

In Table 4 the point estimates within each stratum are reported, together with a 95% confidence interval constructed by using both the multinomial profile log likelihood and the delta method, evaluated unconditionally and conditionally on n_j. As expected, the confidence intervals evaluated conditionally are narrower. Also, the confidence intervals based on the profile are greater than the others, the only exception being the first stratum, where the lower limit is larger. This shows that the confidence intervals based on the delta method are overoptimistic.

The selected model shows that the latent variable has a monotone pattern of interactions with the sources which does not vary with the strata. Subjects in the third stratum have a higher probability of ascertainment in sources S_1 and S_3 with respect to the others. Also, conditioning on C, U, S_1, and S_3 accounts for the positive interaction between sources S_2 and S_4 that emerged in our preliminary analyses based on the marginal distribution. Similarly, the positive interaction between S_3 and S_4 that emerged in our preliminary analyses is accounted for after conditioning on the remaining variables. The negative interactions between sources S_1 and S_3, and S_1 and S_4 remain. In particular, the negative interaction between S_1 and S_3 varies in the strata. The other negative interactions that emerged from our preliminary analyses resulted in a nonsignificant improvement of the model fit.

We derive the interpretation of the latent variable from marginal estimated frequencies of the three strata by the two latent levels. The estimated frequencies for the observed cells show that, in the first stratum the estimated probability to fall in the second latent class is 0.027, for the second stratum 0.033, and for the third stratum 0.798. In each of the three strata the estimated probability of ascertainment is relatively low for the first level of the latent variable (stratum 1: 0.696;

Table 4
Point and 95% confidence interval estimates of the undercounts within each stratum

Stratum	Point estimate	Profile CI	Uncon. delta CI	Con. delta CI
1	87	45–145	36–138	37–136
2	434	313–580	397–471	404–463
3	14	4–32	8–20	8–20
Total	535	–	381–690	392–679

stratum 2: 0.771; stratum 3: 0.835) and high for the second level of the latent variable (stratum 1: 0.982; stratum 2: 0.992; stratum 3: 0.995). This leads to an interpretation of the latent variable as reflecting the severity of the illness, where the first level stands for a weak severity (ascertainment is low; a low probability to fall in this level for the third stratum) and the second level stands for a high severity (ascertainment is high; a high probability to fall in the third stratum).

Bruno et al. (1994) reported an estimated undercount of 10.42 (stratum 1), 501.53 (stratum 2), and 20.42 (stratum 3). There is a relevant distance between the estimates in the first stratum, which is where, according to our estimated model, the heterogeneity seems more severe. On the other hand, comparing with Biggeri et al. (1999; total undercount of 627, 95% confidence interval 433–881) we see that taking into account the observed component of the heterogeneity by introducing a covariate leads to a smaller total of undercount. Our result for the total N is also very close to the Bayesian Rasch model of Fienberg et al. (1999) including the interaction between the symmetry term and the reimbursements list S_4 (total $N = 2556$, 95% confidence interval 2446–2750) and is near to the upper bound provided by Bartolucci and Forcina (2001; total $N = 2403$, 95% confidence interval 2280–2585). This may be because their model finds more negative interactions between some of the lists.

6. Discussion

A range of models has been proposed in the literature to account for unobserved heterogeneity in capture–recapture context. When a genuine interaction between lists is present, a choice should be made on how to model this interaction. In this article we proposed a hierarchical log linear model where the interaction between lists is modeled after conditioning on all the other observed variables and a latent variable that takes unobserved heterogeneity into account. Another possibility is to model the induced interaction in the marginal distribution of the observable variables directly. However, the introduction of a latent variable that accounts for the unobservable heterogeneity leads to a model where the influence of the unobserved heterogeneity can be interpreted explicitly. Furthermore, if the latent variable model is true, the confidence intervals for the undercounts based on this marginal distribution are too optimistic.

For the proposed model we showed how to calculate the information matrix and to check for local identification. We also extended the procedure in Cormack (1992) to evaluate confidence intervals for the undercounts in each stratum of the covariates based on the profile log likelihood of multinomial and Poisson models. Farrington (2002) argued that in epidemiological capture–recapture studies confidence intervals should be based on the profile Poisson log likelihood and provided a simplified way to construct them for relevant parameters, in a situation with no covariates. The extension of this method to a multiple strata context may be usefully investigated. The comparison between the confidence intervals based on the asymptotic distribution of the estimates and the ones evaluated using the profile log likelihood can be used to detect situations of scarce information in the data about the unobserved counts within each stratum.

Acknowledgements

We thank an associate editor and an anonymous referee for helpful comments. Part of this article was completed while the first author was visiting the Department of Mathematics and Statistics of the University of Plymouth whose hospitality is gratefully acknowledged.

Résumé

Nous présentons un modèle permettant d'estimer l'effectif inconnu d'une population à partir de listes, même dans les cas où les hypothèses (a) d'homogénéité des probabilités de capture des individus, et (b) d'indépendance marginale entre listes, ne sont pas respectées. Cette situation est fréquente dans les études épidémiologiques, où l'hétérogénéité enter individus est forte, et où les chercheurs ne maîtrisent pas l'indépendance des sources de renseignements. Nous nous intéressons à la situation où des covariables qualitatives sont disponibles, et où l'intérêt de l'étude porte non seulement sur le comptage de la population sous-jacente totale, mais aussi sur l'estimation des effectifs de chaque strate résultant du croisement des covariables. Nous présentons aussi plusieurs techniques de construction d'intervalles de confiance pour les effectifs de chaque strate, en utilisant la vraisemblance projetée, ce qui prolonge le travail de Cormack (1992).

References

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

Agresti, A. and Lang, J. B. (1993). Quasi-symmetric latent class models, with application to rater agreement. *Biometrics* **49,** 131–139.

Alho, J. M. (1990). Logistic regression in capture-recapture models. *Biometrics* **46,** 623–635.

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

Biggeri, A., Stanghellini, E., Merletti, F., and Marchi, M. (1999). Latent class models for varying catchability and correlation among sources in capture–recapture estimation of the size of a human population. *Statistica Applicata* **11,** 563–576.

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

Bruno, G., Biggeri, A., LaPorte, R. E., McCarty, D., Merletti, F., and Pagano, G. (1994). Appplication of capture–recapture to count diabetes? *Diabetes Care* **17,** 548–556.

Catchpole, E. and Morgan, B. J. T. (1997). Detecting parameter redundancy. *Biometrika* **84,** 187–196.

Chao, A., Tsay, P. K., Lin, S.-H., Shau, W.-Y., and Chao, D. Y. (2001). The application of capture-recapture models to epidemiological data. *Statistics in Medicine* **20,** 3123–3157.

Cormack, R. M. (1992). Interval estimation for mark-recapture studies of closed populations. *Biometrics* **48,** 567–576.

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

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.

Edwards, D. (2000). *Introduction to Graphical Modeling*, 2nd edition. New York: Springer.

Farrington, C. P. (2002). Interval estimation for Poisson capture-recapture models in epidemiology. *Statistics in Medicine* **21,** 3079–3092.

Fienberg, S. E., Johnson, M. S., and Junker, B. W. (1999). Classical multilevel and Bayesian approaches to population size estimation using multiple lists. *Journal of the Royal Statistical Society, Series A* **162,** 383–405.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* **61,** 215–231.

Huggins, R. M. (1991). Some practical aspects of a conditional likelihood approach to capture experiments. *Biometrics* **47,** 725–732.

International Working Group for Disease Monitoring and Forecasting. (1995). Capture-recapture and multiple-record systems estimation 1: History and theoretical development. *American Journal of Epidemiology* **142,** 1047–1058.

Lang, J. B. (1992). Obtaining the observed information matrix for the Poisson log linear model with incomplete data. *Biometrika* **79,** 833–836.

Lindsay, B., Clogg, C., and Greco, J. (1991). Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association* **86,** 96–107.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B* **44,** 226–233.

Rothenberg, T. (1971). Identification in parametric model. *Econometrica* **39,** 577–591.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. New York: Wiley.

Received August 2002. *Revised November* 2003.
Accepted December 2003.