STATISTICA, anno LXXIX, n. 4, 2019

PARAMETERIZATION OF CONTINUOUS COVARIATES IN THE POISSON CAPTURE-RECAPTURE LOG LINEAR MODEL FOR CLOSED POPULATIONS

Giuseppe Rossi
Unità di Epidemiologia e Biostatistica, Istituto di Fisiologia Clinica, CNR, Pisa, Italia

Pasquale Pepe
Ocular Technology Group - International, London, United Kingdom

Olivia Curzio
Unità di Epidemiologia e Biostatistica, Istituto di Fisiologia Clinica, CNR, Pisa, Italia

Marco Marchi [1]
Dipartimento di Statistica, Informatica, Applicazioni "G. Parenti", Università di Firenze, Firenze, Italia

1. Introduction

The capture-recapture method is widely used to estimate the size of animal and human populations. In human populations "being captured at the k_{th} occasion" is replaced by "being included in the k_{th} list". In this case the estimation is based on the information acquired from the overlapping lists of cases (Sekar and Deming, 1949; International working group for disease monitoring and forecasting, 1995a,b).

In epidemiology, the capture-recapture technique is used to estimates the population size of hidden phenomena such as drug users (Brugal *et al.*, 1999, 2004; Hay, 2000; Comiskey and Barry, 2001; Frischer *et al.*, 2001; Buster *et al.*, 2001; Lange *et al.*, 2003; Gemmell *et al.*, 2004; Hickmann *et al.*, 2004; Platt *et al.*, 2004; Hope *et al.*, 2005), deaths due to traffic accidents (Razzak and Luby, 1998), prostitution (Roberts and Brewer, 2006), people infected with the Human Immunodeficiency Virus (Abeni *et al.*, 1994; Bartolucci and Forcina, 2006), and other diseases (Tilling *et al.*, 2001; Zwane and Heijden, 2005). A closed population and constant covariate values across captures are assumed in this paper. There are no births, deaths or migrations, so the size of the population under study is constant over the capture time. These assumptions are likely to be valid for surveys performed in a relatively short time.

[1] Corresponding Author. E-mail: marchi@ds.unifi.it

In the epidemiological application of capture-recapture methods, two main problems may affect the estimation: the dependence among lists and the heterogeneity of capture probabilities among individuals. The dependence is due to the association between lists within each individual (the capture in a list has a direct causal effect on the capture in another list).

The heterogeneity relates to the effect of the individual characteristics (such as gender, age, ethnicity, etc...) on capture probabilities. Hwang and Huggins (2005) have shown that ignoring the heterogeneity of capture probabilities may lead to the underestimation of the population size. The Poisson log-linear model (LLM) can be used to model both the dependence and the heterogeneity (Schwarz and Seber, 1999) but, in presence of continuous covariates, the maximum likelihood estimation (MLE) may be inconsistent or biased (Baker, 1994; Tilling and Sterne, 1999) because the number of parameters may become close to the number of observations. Usually the continuous covariates are categorized but stratification is subjective and different categorization may provide different estimates of the population size (Pollock *et al.*, 1984; Pollock, 2002).

By the multinomial conditional logit model (MCLM), continuous covariates can be modeled in their original measurement scale (Zwane and Heijden, 2005). For this model the mathematical derivation of the standard error does not exist and bootstrap inference is needed in order to obtain an estimate of the confidence interval. The International working group for disease monitoring and forecasting (1995b) noted that, for all the models proposed in capture-recapture literature, the distribution of the population size is skewed. In presence of continuous covariates the bootstrap method is commonly used to obtain non-parametric confidence intervals (Huggins, 1989; Tilling and Sterne, 1999) but the estimated variance is likely to be smaller than the true variance (Norris and Pollock, 1996). Schwarz and Seber (1999) showed the equivalence between the standard Poisson LLM and the MCLM only when a dummy variable for each value of the continuous covariate is used in the Poisson LLM.

In the present work a new parameterization of the Poisson LLM, allowing to handle continuous covariates in their original measurement scale, is shown and the analytic estimate of the asymmetric confidence interval of the population size is derived. The proposed method was finally compared with the MCLM and evaluated on simulated and real data sets.

2. Preliminaries and notations

The simplest capture-recapture method assumes the presence of two captures or two lists and can be set out in a contingency table. The goal is to estimate the number of units missed in both occasions (n_{00}). The estimation is performed by means of the Petersen dual-system estimator:

$$\hat{n_{00}} = \frac{n_{10} n_{01}}{n_{11}}. \tag{1}$$

It is a function of the number of individuals caught in both occasions (n_{11}) and the number of individuals caught in one occasion only (n_{10} and n_{01}). The total population size is finally given by: $\hat{N} = \hat{n_{00}} + n_{10} + n_{01} + n_{11}$. For the Petersen estimator the following assumptions holds:

- the population is closed;
- the chance of being caught or present in a list is constant across individuals or occasions;
- the two captures are independent.

Usually, the first assumption may be controlled by the researcher performing the survey in a relatively short time. In contrast, the second and third assumptions relates to intrinsic characteristics of the individuals belonging to the population and, if this is not taken into consideration, the estimator given by Equation (1) is biased.

When data presents only two capture occasions and some covariates, which may affect the inclusion probabilities, a commonly used approach is to stratify the population and estimate the missing individuals in each stratum by the estimator (1). More generally, when the survey is performed in more than two occasions, it is possible to handle the dependence between occasions and modelling the heterogeneity of inclusion probabilities using generalized linear models. This class of models has been extensively used in the epidemiological applications of capture-recapture method.

3. The Poisson log linear model

Capture-recapture data with multiple captures and heterogeneity of inclusion probabilities can be handled in the standard framework of the Poisson LLM (Fienberg, 2002; Cormack, 1989). Here we describe the case of three captures (Table 1, see Appendix A) while the generalization to several occasions is straightforward. Data can be set out in a 3-way contingency table: 000, 100, 010, 001, 110, 101, 011, 111 where 1 means that the individual has been observed. For example, 010 means that the individual was captured only in the second occasion, therefore $s_1 = 0$, $s_2 = 1$ and $s_3 = 0$. Obviously, the unknown quantity to estimate is n_{000}, i.e. the number of individuals lost at any of the three captures.

The Poisson LLM allows to model the logarithm of the expected value of the number of individuals observed in each capture profile through the following linear equation:

$$\log\left[E\left(n_{s_1 s_2 s_3}\right)\right] = \beta_0 + \beta_1 s_1 + \beta_2 s_2 + \beta_3 s_3 + \beta_{12} s_{12} + \beta_{13} s_{13} + \beta_{23} s_{23}, \qquad (2)$$

where s_1, s_2 and s_3 are the indicator variables of the three captures, while s_{12}, s_{13}, and s_{23} are the two-way interaction terms.

As outlined by Chao (2001), in a capture-recapture experiment performed in three occasions, the observed profiles are 7 and there is no three-way interaction term, i.e.,

$s_{123} = 0$. Under this model the expected value of n_{000} is given by $E(n_{000}) = \exp(\beta_0)$. When information about individual characteristics is present, in the simplest case of a dummy covariate X, the interaction terms between the covariate and each capture profiles must be added to the above and the resulting model become:

$$\log\left[E\left(n_{s_1 s_2 s_3}\right)\right] = \beta_0 + \beta_1 s_1 + \beta_2 s_2 + \beta_3 s_3 + \beta_{12} s_{12} + \beta_{13} s_{13} + \beta_{23} s_{23} + \beta_x x + \\ + \beta_{1x} s_{1x} + \beta_{2x} s_{2x} + \beta_{3x} s_{3x} + \beta_{12x} s_{12x} + \beta_{13x} s_{13x} + \beta_{23x} s_{23x}, \quad (3)$$

where s_{1x}, s_{2x}, s_{3x} and $s_{12x}, s_{13x}, s_{23x}$ are all the two and three-way interactions between sources and the covariate, respectively.

Conditional on the two levels (0 and 1) of the dummy covariate X, the expected number of individuals missed at any occasion turns out to be $E(n_{0000}) = \exp(\beta_0)$ and $E(n_{0001}) = \exp(\beta_0 + \beta_x)$, where β_0 and β_x are the intercepts related to the covariate levels. Therefore, the total expected number of missed individuals is $E(n_{0000}) + E(n_{0001})$, and the total population size is obviously obtained adding the total expected number of missed individuals to the total number of observed individuals. The extension to several dichotomous covariates is straightforward. Continuous covariates can be instead modelled only using a set of dummy variables for each of them and for each single continuous value otherwise the MLE of the Poisson LLM is biased (Baker, 1994; Tilling and Sterne, 1999).

4. The multinomial conditional logit model

The MCLM can handle continuous covariates in their original measurement scale overcoming problems related to the MLE. The MCLM or the Bock's multinomial logit model are two alternative parameterizations of the same model. They extend the logistic approach of Huggins (1989) and Alho (2000) from two independent occasions to several dependent occasions (Zwane and Heijden, 2005). As outlined by Zwane and Heijden (2005), an individual i ($i = 1, 2, ..., n$) is classified in one of K capture profiles indexed by k ($k = 1, 2, ..., K$), such that the indicator variable $I_{k|i} = 1$ if individual i falls in the capture profile k and 0 otherwise. The multinomial logit for individual i is $z_i = [z_{1|i}, z_{2|i}, ..., z_{K|i}]$, imply that the category probabilities for the individual i are

$$\pi_{k|i} = \frac{e^{z_{k|i}}}{\sum_{k=1}^{K} e^{z_{k|i}}}. \quad (4)$$

When continuous or categorical variables, indexed by h ($h = 1, 2, .., H$) are present and collected in a matrix $\mathbf{X}$, the multinomial logits in $\mathbf{Z}$ are related to the covariate matrix $\mathbf{X}$ and a design matrix $\mathbf{Y}$ by a matrix of regression parameters $\mathbf{\Lambda}$. The multinomial logits are decomposed as $\mathbf{Z} = \mathbf{X\Lambda Y}$. Let the elements of $\mathbf{Y}$ be y_{jk}, with $j = 1, 2, ..., J$, the elements of X be x_{ih}, and the elements of $\mathbf{\Lambda}$ be λ_{hj}, where J is the number of lists and interaction effects. For example, with three lists, each individual has a unique capture

profile and the set of all the possible capture profiles is 100, 010, 001, 110, 101, 011, 111 while the vector $I_i = [I_{100|i}, I_{010|i}, I_{001|i}, I_{110|i}, I_{101|i}, I_{011|i}, I_{111|i}]$, where $I_{abc|i} = 1$ if individual i has capture profile $[a, b, c]$ and $I_{abc|i} = 0$ otherwise. If we assume that list 1 and 2, and list 2 and 3 are dependent the $\mathbf{Y}$ matrix is given by:

$$\begin{pmatrix} \text{list}_1 : 1\,0\,0\,1\,1\,0\,1 \\ \text{list}_2 : 0\,1\,0\,1\,0\,1\,1 \\ \text{list}_3 : 0\,0\,1\,0\,1\,1\,1 \\ \text{list}_{12} : 0\,0\,0\,1\,0\,0\,1 \\ \text{list}_{23} : 0\,0\,0\,0\,0\,1\,1 \end{pmatrix}.$$

The elements of $\mathbf{Y}$ will be denoted by $y_{j(abc)}$ rather than y_{jk}, where $y_{j(abc)}$ is the element of $\mathbf{Y}$ in row j corresponding to capture profile $[a, b, c]$. The first three rows of $\mathbf{Y}$ refer to the single list effects and the fourth and fifth rows to the interactions of list 1 and 2, and list 2 and 3 respectively. The probabilities that each individual will belong to one of the 7 capture profiles can be estimated by the MCLM as a function of the covariate values and the overlapping captures. More precisely, the probability that the i_{th} individual belongs to the k_{th} capture profile, is given by:

$$\pi^{k|i} = \frac{e^{\sum_{h=1}^{H}\sum_{j=1}^{J} x_{ih}\lambda_{jh}y_{jk}}}{\sum_{r=1}^{K} e^{\sum_{h=1}^{H}\sum_{j=1}^{J} x_{ih}\lambda_{hj}y_{jr}}}. \tag{5}$$

The log-likelihood of the MCLM is given by $l_{mult} = \sum_{i=1}^{n}\sum_{k=1}^{K} I_{k|i}\log(\pi_{k|i})$ and the model can be fitted exploiting the similarity of the likelihood function with that of the stratified proportional hazards model (Zwane and Heijden, 2005; Chen and Lynn, 2001). In the case of three occasions, the individual contribution to the estimate of the missed numbers is:

$$\hat{m_{000|i}} = \frac{\hat{\pi_{100|i}}\,\hat{\pi_{010|i}}\,\hat{\pi_{001|i}}\,\hat{\pi_{111|i}}}{\hat{\pi_{110|i}}\,\hat{\pi_{101|i}}\,\hat{\pi_{011|i}}}. \tag{6}$$

Finally, the estimator of the population size is the following:

$$\hat{N} = \sum_{i=1}^{n}\left(1 + \hat{m_{000|i}}\right) = \sum_{i=1}^{n}\frac{1}{1 - \hat{\pi_{000|i}}}, \tag{7}$$

where $\hat{\pi_{000|i}} = \frac{\hat{m_{000|i}}}{1+\hat{m_{000|i}}}$ is the probability that individual i is missing in all lists.

5. A NEW PARAMETERIZATION OF THE POISSON CAPTURE-RECAPTURE LOG-LINEAR MODEL ALLOWING TO USE CONTINUOUS COVARIATES

The MLE of the Poisson LLM with continuous covariates in their original measurement scale and using the standard parameterization is biased (Baker, 1994; Tilling and Sterne,

1999). Therefore, when one or more continuous covariates are available, we propose to use a new parameterization to obtain an overall fit and an estimate of the unknown population size.

For each observed individual the capture profiles are defined using a dummy variable for each source and the value of the continuous variables in their original measurement scale. For example, in case of three lists (s_1, s_2, s_3) and one continuous covariate X, each individual can belong to a set of $K = 8$ possible capture profiles: 000, 100, 010, 001,110, 101, 011, 111, each with his own set of covariates (Table 2). The vector $I_i = [I_{000|i}, I_{100|i}, I_{010|i}, I_{001|i}, I_{110|i}, I_{101|i}, I_{011|i}, I_{111|i}]$, where $I_{s_1 s_2 s_3|i} = 1$ if individual i has capture profile $[s_1, s_2, s_3]$ and $I_{s_1 s_2 s_3|i} = 0$ otherwise, is defined. For the capture profile 001, $I_{s_1 s_2 s_3|i} = 1$ if individual i has the capture profile 001 and $I_{s_1 s_2 s_3|i} = 0$ otherwise. The capture indicator vector $I_{s_1 s_2 s_3|i}$ present a missing value for the no-capture profile 000 ($I_{000|i}$ value is missing). Finally, all the two-way interaction terms between sources (s_{12}, s_{13}, s_{23}), and all the two (s_{1x}, s_{2x}, s_{3x}) and three-way $(s_{12x}, s_{13x}, s_{23x})$ interactions between sources and the covariate, complete the data matrix. The resulting model is:

$$\log\left[E\left(I_{s_1 s_2 s_3|i}\right)\right] = \beta_0 + \beta_1 s_1 + \beta_2 s_2 + \beta_3 s_3 + \beta_{12} s_{12} + \beta_{13} s_{13} + \beta_{23} s_{23} + \beta_x x + \\ + \beta_{1x} s_{1x} + \beta_{2x} s_{2x} + \beta_{3x} s_{3x} + \beta_{12x} s_{12x} + \beta_{13x} s_{13x} + \beta_{23x} s_{23x}. \quad (8)$$

When three occasions are present, there are only seven available profiles for each subject in the MCLM while, in the proposed Poisson LLM, the profiles are eight since the unknown capture profile (000) is needed. Consequently, the Poisson LLM presents additional parameters compared to the MCLM, a constant and one parameter for each covariate. If three occasions and two covariates are present, the model is:

$$\begin{aligned}\log\left[E\left(I_{s_1 s_2 s_3|i}\right)\right] = {} & \beta_0 + \beta_1 s_1 + \beta_2 s_2 + \beta_3 s_3 + \beta_{12} s_{12} + \beta_{13} s_{13} + \beta_{23} s_{23} + \\ & + \beta_{x1} x_1 + \beta_{1x_1} s_{1x_1} + \beta_{2x_1} s_{2x_1} + \beta_{3x_1} s_{3x_1} + \\ & + \beta_{12x_1} s_{12x_1} + \beta_{13x_1} s_{13x_1} + \beta_{23x_1} s_{23x_1} + \\ & + \beta_{x2} x_2 + \beta_{1x_2} s_{1x_2} + \beta_{2x_2} s_{2x_2} + \beta_{3x_2} s_{3x_2} + \\ & + \beta_{12x_2} s_{12x_2} + \beta_{13x_2} s_{13x_2} + \beta_{23x_2} s_{23x_2} + \beta_{x_1 x_2} s_{x_1 x_2}.\end{aligned} \quad (9)$$

For the no-capture profile (000) of each subject i the model simplify to the logarithm of the expected number of missed units in i: $\log\left[E\left(I_{000|i}\right)\right] = \beta_0 + \beta_{x_1} x_1 + \beta_{x_2} x_2 + \beta_{x_1 x_2} x_1 x_2$, that is, to the sum of the constant, the covariates and the interactions effect between covariates. In this case the explicit expression of the variance of $\log\left[E\left(I_{000|i}\right)\right]$

is:

$$\begin{aligned}&\mathrm{Var}(\beta_0)+x_1^2\mathrm{Var}(\beta_{x_1})+x_2^2\mathrm{Var}(\beta_{x_2})+(x_1x_2)^2\mathrm{Var}(\beta_{x_1x_2})+\\&+2x_1\mathrm{Cov}(\beta_0,\beta_{x_1})+2x_2\mathrm{Cov}(\beta_0,\beta_{x_2})+2x_1x_2\mathrm{Cov}(\beta_0,\beta_{x_1x_2})+\\&+2x_1x_2\mathrm{Cov}(\beta_{x_1},\beta_{x_2})++2x_1(x_1x_2)\mathrm{Cov}(\beta_{x_1},\beta_{x_1x_2})+\\&2x_2(x_1x_2)\mathrm{Cov}(\beta_{x_2},\beta_{x_1x_2}).\end{aligned} \tag{10}$$

More generally, when H covariates are present, the variance of $\log\left[E\left(I_{000|i}\right)\right]$ is:

$$\mathrm{Var}\left[\log\left\{E\left(I_{000|i}\right)\right\}\right]=\sum_{h=1}^{H+1}x_h^2\mathrm{Var}(\beta_h)+\sum_{h\neq t}^{\binom{H+1}{2}}2x_hx_t\mathrm{Cov}(\beta_h,\beta_t), \tag{11}$$

where the covariates are indexed by h ($h=1,2,..,t,..,H+1$).

The confidence interval of $\log\left[E\left(I_{000|i}\right)\right]$ is then computed as:

$$\log\left[E\left(I_{000|i}\right)\right]+Z_\alpha\sqrt{\mathrm{Var}\left[\log\left\{E\left(I_{000|i}\right)\right\}\right]},$$

where, for the 95% confidence interval, $Z_\alpha=1.96$.

The expected number of missed units in i is $E\left(I_{000|i}\right)=e^{\log\left[E\left(I_{000|i}\right)\right]}$ and the total expected number of missed units is $\sum_{i=1}^{n}E\left(I_{000|i}\right)$. The asymmetric confidence interval of the total expected number of missed units ($\sum_{i=1}^{n}E\left(I_{000|i}\right)$) is computed as:

$$\sum_{i=1}^{n}e^{\left[\log\left\{E\left(I_{000|i}\right)\right\}+Z_\alpha\mathrm{Var}\left[\log\left\{E\left(I_{000|i}\right)\right\}\right]\right]^{\frac{1}{2}}}$$

for the upper limit and

$$\sum_{i=1}^{n}e^{\left[\log\left\{E\left(I_{000|i}\right)\right\}-Z_\alpha\mathrm{Var}\left[\log\left\{E\left(I_{000|i}\right)\right\}\right]\right]^{\frac{1}{2}}}$$

for the lower limit.

The total population size and its confidence interval can be finally calculated adding the number of observed units to the total expected number of missed units. The parameters of the proposed Poisson regression model can be estimated by the maximum likelihood using the Newton-Raphson method. The variance-covariance matrix of the parameters can be estimated by the observed information matrix (OIM), as suggested by Hardin and Hilbe (2001) in the case of canonical link.

6. SIMULATION STUDY

A simulation study was conducted to evaluate the performance of the proposed method to estimate the population size in a closed population and in presence of observed heterogeneity. Two datasets A and B with 1000 cases (Table 3) were generated from a population with an expected size of 4000 individuals. Each dataset was composed by three lists with different levels of dependence and heterogeneity. The dataset A was simulated to obtain independence between lists and a significant effect of the continuous covariate while the dataset B presents dependence between the first two lists and a significant effect of the continuous covariate. For each of the two simulated datasets, the number of individuals captured from only one source (s_1, s_2, s_3), two sources (s_{12}, s_{13}, s_{23}) and three sources (s_{123}) are reported in Table 3. More details regarding the construction of these datasets are reported in Rossi *et al.* (2010).

For each dataset, a comparison between the proposed Poisson LLM and the MCLM was made in terms of estimated total population (N) and deviance (D). Furthermore, the parametric confidence intervals obtained analytically from the proposed Poisson LLM were compared with the non-parametric confidence intervals obtained by the bootstrap method. From each of the two simulated dataset 1000 samples of size n=1000 were extracted with replacement.

For each of these samples the total population size was estimated using the model selected on the starting dataset by a backward procedure and according to the Akaike information criterion (AIC). This procedure led to a final distribution of the estimator of the total population, which was used to produce the non-parametric confidence intervals (Table 4).

For each of the two simulated data sets the proposed Poisson LLM, selected by the AIC, estimates exactly the simulated effects and converges to the MCLM along the variable selection process. The population size obtained by the two models were very close to the expected of 4000 individuals and the difference against the bootstrap confidence intervals ($\approx 1\%$) and deviance ($\approx 0.1\%$) appear to be negligible.

To investigate the ability of the proposed Poisson LLM to estimate correctly the population size (N) and the 95% confidence interval, we compared these estimates with those obtained using the standard Poisson LLM and the Wald method. First, we compared results obtained on datasets A and B without considering the effect of the continuous covariate in the model and for different sample size (1000, 10000, 50000). In this case, it is possible to estimate analytically the population size and the 95% C.I. also in the standard framework of Poisson LLM / Wald method. The equivalence between the two different types of parameterization of the Poisson LLM is shown in Table 5.

Furthermore, the population size (N) and the 95% C.I. obtained in closed form by the standard Poisson LLM and the proposed Poisson LLM were compared to the mean and the non-parametric 95% confidence interval of the distribution of N obtained by the bootstrap method. Both the population size and the non-parametric 95% confidence interval obtained by the Bootstrap method were only slightly different from those estimated analytically, both for the proposed Poisson LLM and the standard Poisson LLM.

However, such difference decreased significantly with increasing the sample size.

Finally, we compared the 95% C.I. obtained analytically by the proposed procedure with those obtained by the bootstrap method considering also the effect of the continuous covariate in the model. In this case, the bootstrap non-parametric 95% confidence intervals were only slightly different from those obtained by the proposed procedure and these difference decrease significantly increasing the sample size.

7. Application to real datasets

In this section, the proposed Poisson LLM was applied to two datasets regarding four notification lists of drug addiction users (opiate and cocaine) in the Liguria region:

- public services for drug addiction (s_1),
- operational unit for drug addiction at prefectures (s_2),
- therapeutic communities (s_3),
- hospital discharge records (s_4).

Gender and age were also present in the two datasets. The "captured" opiate and cocaine users were 4825 and 531 respectively. The estimated prevalences were obtained using all the above 4 sources of notification with sex and age (continuous scale) as covariates. Point and 95% confidence interval estimated by the proposed Poisson LLM are shown in Table 6. The estimated 15-64 years population prevalence of opiate and cocaine users of the Liguria region in the year 2002 (1002497 individuals), was about 1.3% (95% C.I.: 1.0-1.9) and 1.6% (95% C.I.: 0.7-3.6) respectively. According to similar studies, addressing the demanding problem of drug addiction within the European Union (Table 7), the proposed Poisson LLM seems to work well in obtaining a plausible quantitative evaluation of it.

8. Conclusion

In this paper we proposed a new parameterization of the Poisson LLM. The estimation of the size of a closed population in presence of observed heterogeneity and the analytical asymmetric confidence interval were shown. Results obtained on simulated and real datasets suggested that the proposed parameterization of the Poisson LLM allows to treat continuous covariates in their original measurement scale This allowed us to overcome problems related to the choice of a correct categorization of continuous covariates or the introduction of a large number of dummy variables, which may lead to inconsistent or biased maximum likelihood estimates (Baker, 1994; Tilling and Sterne, 1999). The observed differences between the Poisson LLM and the MCLM in terms of goodness of fit and estimated population size appear negligible. The main advantage of the proposed Poisson LLM over MCLM is the analytical estimation of the confidence

interval. The bootstrap and the analytical confidence intervals are very similar and tend to be equivalent by increasing the sample size. The confidence bounds obtained by the proposed Poisson LLM are preferable to their bootstrap estimation because the latter can lead to unreasonable confidence limits in case of small datasets. Finally, the estimation obtained by the proposed Poisson LLM on a real dataset showed that the prevalence of opiate and cocaine users in the Liguria region seems to be consistent with that found in Inner London and in other European areas (Gemmell *et al.*, 2004; Hickmann *et al.*, 2004; Hope *et al.*, 2005).

Acknowledgements

The authors are very grateful to Professor Luigi Donato for his advice and encouragement and to the Liguria Region for making available data on drug addiction users.

La pubblicazione di questo articolo vuol essere un doveroso omaggio alla memoria di Giuseppe Rossi, amico e collega, prematuramente scomparso. A lui si deve l'idea alla base del lavoro la cui lettura auspichiamo serva a ricordare quale sia stata la perdita sia sotto il profilo umano che scientifico.

APPENDIX

A. TABLES

TABLE 1
Representation of a 3 sources capture–recapture data.

S_1	S_2	S_3	
		0	1
0	0	?	n_{001}
0	1	n_{010}	n_{011}
1	0	n_{100}	n_{101}
1	1	n_{110}	n_{111}

TABLE 2
New parameterization for the Poisson capture-recapture log linear model (LLM), referred to the i-th subject belonging to the capture profile 001 with covariate value x.

s_1	s_2	s_3	$I_{s_1 s_2 s_3}$	s_{12}	s_{13}	s_{23}	X	s_{1x}	s_{2x}	s_{3x}	s_{12x}	s_{13x}	s_{23x}
0	0	0	?	0	0	0	x	0	0	0	0	0	0
0	0	1	1	0	0	0	x	0	0	x	0	0	0
0	1	0	0	0	0	0	x	0	x	0	0	0	0
0	1	1	0	0	0	1	x	0	x	x	0	0	x
1	0	0	0	0	0	0	x	x	0	0	0	0	0
1	0	1	0	0	1	0	x	x	0	x	0	x	0
1	1	0	0	1	0	0	x	x	x	0	x	0	0
1	1	1	0	1	1	1	x	x	x	x	x	x	x

TABLE 3
Capture profiles of simulated datasets.

Dataset	s_0 (000)	s_1 (100)	s_2 (010)	s_3 (001)	s_{12} (110)	s_{13} (101)	s_{23} (011)	s_{123} (111)
A	?	314	307	312	40	35	42	11
B	?	233	321	312	120	29	41	18

TABLE 4

Mean deviance (D), analytically estimated population size (N) and 95% confidence interval (C.I.), bootstrap non-parametric 95% confidence interval (C.I.) obtained on simulated datasets for the proposed Poisson LLM and MCLM.

Dataset	AIC selected model	Type of model	D	N	Proposed 95% C.I.	Bootstrap 95% C.I.
A	$s_0\ s_1\ s_2\ s_3$ $s_{0x}\ s_{1x}\ s_{2x}\ s_{3x}$	LLM	3144	3824	3092-4879	3257-4707
	$s_1\ s_2\ s_3$ $s_{1x}\ s_{2x}\ s_{3x}$	MCLM	3143	3843	3092-4879	3250-4760
B	$s_0\ s_1\ s_2\ s_3$ $s_{0x}\ s_{1x}\ s_{2x}\ s_{3x}$	LLM	3378	3983	3170-5153	3353 - 4967
	$s_1\ s_2\ s_3$ $s_{1x}\ s_{2x}\ s_{3x}$	MCLM	3378	4001	3170-5153	3368 - 5009

LLM=Poisson log linear model, MCLM=multinomial conditional logit model, N=estimated population size on the starting datasets, D=mean of the deviance over the bootstrapped datasets.

TABLE 5

Analytical estimates of the population size (N) and 95% confidence interval (C.I.) obtained by the proposed Poisson LLM and the standard Poisson LLM compared to the non-parametric bootstrap estimates on simulated datasets with different sample sizes.

Type of dataset	Covariate adjusted	Selected model	Sample size	Proposed Poisson model estimates		Standard Poisson model estimates		Non - parametric bootstrap estimates		Bootstrap vs analytical estimates (% of variation)	
				N	95% C.I.	N	Wald 95% C.I.	N	95% C.I.	N	95% C.I.
			1000	3314	2903/3818	3314	2903/3818	3329	2920/3817	0.44	0.62 / -0.03
A		$s_0\ s_1\ s_2\ s_3$	10000	33143	31750/34628	33143	31750/34628	33130	31789/34600	-0.04	012 / -0.08
	not		50000	165716	162546/168977	165716	162546/168977	165750	162669/168887	0.02	0.08 / -0.05
			1000	3336	2842/3968	3336	2842/3968	3356	2851/3968	0.59	0.32 / 0.01
B		$s_0\ s_1\ s_2\ s_3\ s_{12}$	10000	33363	31677/35196	33363	31677/35196	33370	31748/35200	0.02	0.26 / 0.01
			50000	166813	162939/170824	166813	162939/170824	166900	163253/170790	0.05	0.19 / -0.02
		$s_0\ s_1\ s_2\ s_3\ s_{0x}$	1000	3823	3092/4879			3867	3257/4707	1.11	5.35 / -2.12
A		$s_{1x}\ s_{2x}\ s_{3x}$	10000	38245	35645/41162			38250	36319/40288	0.01	1.89 / -0.71
	yes		50000	191225	185225-197541			191250	186383/196132	0.01	0.63 / -3.61
		$s_0\ s_1\ s_2\ s_3\ s_{12}\ s_{0x}$	1000	3934	3027/5367			4000	3244/5060	1.69	5.74 / -1.49
B		$s_{1x}\ s_{2x}\ s_{3x}\ s_{12x}$	10000	39336	36028/43157			39370	36666/42332	0.09	1.46 / -0.68
			50000	196678	189024/204842			196700	190497/202962	0.01	0.78 / 0.92

TABLE 6
Analytically estimated population size (N) and 95% confidence interval (C.I.) obtained on the real dataset (Liguria region 2002) by the proposed Poisson LLM.

Type of drug	N	95% C.I.
Opiate	12758	9589 - 18049
Cocaine	16759	4391 - 69163

TABLE 7
Estimated prevalence (x 100) in other European countries / towns.

Type of drug	Opiate		Cocaine	
Country/Town	Prevalence	Years	Prevalence	Years
	World Drug Report 2006 (15-64 years)			
Italy	0.8	2004	1.02	2003
UK	1.09	2001	2.04	2003
Spain			2.07	2003
	Hope *et al.* Addiction 2005 (15-54 years)			
London	1.2 - 1.6	2000-2001	1.5 - 1.9	2000-2001
Brighton	2			
Liverpool	1.05			

REFERENCES

D. ABENI, G. BRANCATO, C. PERUCCI (1994). *Capture-recapture to estimate the size of the population with human immunodeficiency virus type 1 infection*. Epidemiology, 5, pp. 410–414.

J. ALHO (2000). *Unified maximum likelihood estimates for closed capture-recapture models using mixtures*. Biometrics, 56, pp. 443–450.

S. BAKER (1994). *The multinomial-Poisson transformation*. The Statistician, 43, no. 4, pp. 495–504.

F. BARTOLUCCI, A. FORCINA (2006). *A class of latent marginal models for capture-recapture data with continuous covariates*. Journal of the American Statistical Association, 101, pp. 786–794.

M. BRUGAL, A. DOMINGO-SALVANY, A. MAGUIRE, J. CAYLA, R. H. J.R. VILLALBI (1999). *A small area analysis estimating the prevalence of addiction to opioids in Barcelona 1993*. Journal of Epidemiology and Community Health 1999, 53, pp. 488–494.

M. BRUGAL, A. DOMINGO-SALVANY, E. D. D. QUIJANO, L. TORRALBA (2004). *Prevalence of problematic cocaine consumption in a city of Southern Europe, using capture-recapture with a single list*. Journal of Urban Health, 81, pp. 416–427.

M. BUSTER, G. V. BRUSSEL, W. V. D. BRINK (2001). *Estimating the number of opiate users in Amsterdam by capture-recapture: The importance of case definition*. European Journal of Epidemiology, 17, pp. 935–942.

A. CHAO (2001). *An overview of closed capture-recapture models*. Journal of Agricultural, Biological, and Environmental Statistics, 6, no. 2, pp. 158–175.

Z. CHEN, K. LYNN (2001). *A note on the estimation of the multinomial logit model with random effects*. The American Statistician, 55, no. 2, pp. 89–95.

C. COMISKEY, J. BARRY (2001). *A capture-recapture study of the prevalence and implications of opiate use in Dublin*. European Journal of Public Health, 11, pp. 198–200.

R. CORMACK (1989). *Log-linear models for capture-recapture*. Biometrics, 45, no. 2, pp. 395–413.

S. FIENBERG (2002). *The multiple recapture census for closed populations and incomplete 2k contingency tables*. Journal of Applied Statistics, 29, pp. 85–102.

M. FRISCHER, M. HICKMANN, L. KRAUS, F. MARIANI, L. WIESSING (2001). *A comparison of different methods for estimating the prevalence of problematic drug measure in Great Britain*. Addiction, 96, pp. 1465–1476.

I. GEMMELL, T. MILLAR, G. HAY (2004). *Capture-recapture estimates of problematic drug use and the use of simulation based confidence intervals in a stratified analysis*. Journal of Epidemiology and Community Health, 58, pp. 758–765.

J. HARDIN, J. HILBE (2001). *Generalized Linear Models and Extensions*. Stata Press, Texas, USA.

G. HAY (2000). *Capture-recapture estimates of drug measure in urban and non-urban settings in the north east of Scotland*. Addiction, 95, pp. 1795–1803.

M. HICKMANN, V. HIGGINS, V. HOPE, K. T. M. BELLIS, A. WALKER (2004). *Injecting drug use in Brighton, Liverpool, and London: best estimates of prevalence and coverage of public health indicators*. Journal of Epidemiology and Community Health, 58, pp. 766–771.

V. HOPE, M. HICKMANN, K. TILLING (2005). *Capturing crack cocaine use: estimating the prevalence of crack cocaine use in London using capture-recapture with covariates*. Addiction, 11, pp. 1701–1708.

R. HUGGINS (1989). *On the statistical analysis of capture experiments*. Biometrika, 76, pp. 133–140.

W. HWANG, R. HUGGINS (2005). *An examination of the effect of heterogeneity on the estimation of population size using capture-recapture data*. Biometrika, 92, pp. 229–233.

INTERNATIONAL WORKING GROUP FOR DISEASE MONITORING AND FORECASTING (1995a). *Capture-recapture and multiple-record estimation I: History and theoretical development*. American Journal of Epidemiology, 142, no. 10, pp. 1047–1058.

INTERNATIONAL WORKING GROUP FOR DISEASE MONITORING AND FORECASTING (1995b). *Capture-recapture and multiple-record estimation II: Applications in human diseases*. American Journal of Epidemiology, 142, no. 10, pp. 1059–1068.

J. LANGE, R. L. PORTE, E. TALBOTT, Y. CHANG (2003). *Capture-recapture method: the gold standard for incidence and prevalence*. Journal of the New Zealand Medical Association, 20, p. 116.

J. NORRIS, K. POLLOCK (1996). *Including model uncertainty in estimating variances in multiple capture studies*. Environmental and Ecological Statistics, 3, pp. 235–244.

L. PLATT, M. HICKMANN, T. RHODES, L. MIKHAILOVA, V. KARAVASHKIN, A. VLASOV, K. TILLING, V. HOPE, M. KHUTORKSOY, A. RENTON (2004). *The prevalence of injecting drug use in a Russian city: Implications for harm reduction and coverage*. Addiction, 99, pp. 1430–1438.

K. POLLOCK (2002). *The use of auxiliary variables in capture-recapture modelling: An overview*. Journal of Applied Statistics, 29, pp. 85–102.

K. POLLOCK, J. HINES, J. NICHOLS (1984). *The use of auxiliary variables in capture-recapture andremoval experiments*. Biometrics, 40, pp. 329–340.

J. RAZZAK, S. LUBY (1998). *Estimating deaths and injuries due to road traffic accidents in Karachi, Pakistan, through the capture-recapture method*. International Journal of Epidemiology, 27, pp. 866–870.

J. ROBERTS, D. BREWER (2006). *Estimating the prevalence of male clients of prostitute women in Vancouver with a simple capture-recapture method*. Journal of the Royal Statistical Society, series A, 169, no. 4, pp. 182–204.

G. ROSSI, P. PEPE, O. CURZIO, M. MARCHI (2010). *Generalized linear models and capture-recapture method in a closed population: Strengths and weaknesses*. Statistica, 70, no. 3, pp. 371–390.

C. SCHWARZ, G. SEBER (1999). *A review of estimating animal abundance III*. Statistical Science, 14, pp. 427–456.

C. SEKAR, W. DEMING (1949). *On a method of estimating birth and death rates and the extent of registration*. Journal of the American Statistical Association, 44, pp. 101–115.

K. TILLING, J. STERNE (1999). *Capture-recapture models including covariate effects*. American Journal of Epidemiology, 149, no. 4, pp. 392–400.

K. TILLING, J. STERNE, C. WOLFE (2001). *Estimation of the incidence of stroke using a capture-recapture model including covariates*. International Journal of Epidemiology, 30, no. 6, pp. 1351–1359.

E. ZWANE, P. V. D. HEIJDEN (2005). *Population estimation using the multiple system estimator in the presence of continuous covariates*. Statistical Modeling, 5, no. 1, pp. 39–52.

SUMMARY

The capture-recapture method is widely used by epidemiologists to estimate the size of hidden populations using incomplete and overlapping lists of subjects. Closed populations, heterogeneity of inclusion probabilities and dependence between lists are taken into consideration in this work. The main objective is to propose a new parameterization for the Poisson log linear odel (LLM) to treat continuous covariates in their original measurement scale. The analytic estimate of the confidence bounds of the hidden population is also provided. Proposed model was applied to simulated and real capture-recapture data and compared with the multinomial conditional logit model (MCLM). The proposed model is very similar to the MCLM in dealing with continuous covariates and the analytic confidence interval performs better than the bootstrap estimate in case of small sample size.

Keywords: Capture-recapture; Closed population; Continuous covariates; Poisson log-linear model; Multinomial conditional logit model.