Population estimation using the multiple system estimator in the presence of continuous covariates

Eugene Zwane and Peter van der Heijden
Department of Methodology and Statistics, Utrecht University, Utrecht, The Netherlands

Abstract: In the presence of continuous covariates, standard capture–recapture methods assume either that the registrations operate independently at the individual level or that the covariates can be stratified and log-linear models fitted, permitting the modelling of dependence between data sources. This article introduces an approach where direct dependence between registrations is modelled leaving the continuous covariates in their measurement scale. Simulations show that not accounting for possible dependence between registrations results in biased estimation of both the population size and standard error. The proposed method is applied to Dutch neural tube defect registration data.

Key words: capture–recapture; multinomial logit model; multiple system estimator; population size estimation

Data and software link available from: http://stat.uibk.ac.at/SMIJ
Received November 2003; **revised** September 2004; **accepted** December 2004

1 Introduction

One way to estimate the size of a closed population is to use capture–recapture models. These models have received considerable attention in epidemiology (International Working Group for Disease Monitoring and Forecasting, 1995a;b). The appeal of these models is that the investigator may use existing, overlapping, incomplete lists of diseased people (Hook and Regal, 2000), which may include, among others, hospital records or patient group records. The old fashioned approach assumes that all individuals have the same probability of being ascertained by a registration, implying that the registrations are independent. Any dependence among the registrations leads to a bias of the estimate derived under independence (Darroch *et al.*, 1993: 1145).

The most prevalent method for analysing such data uses log-linear models (Cormack, 1989; Fienberg, 1972). In epidemiology, list dependence and heterogeneity (the behaviour component) are the norm, and log-linear models are particularly useful in modelling these phenomena (Schwarz and Seber, 1999: 438–439). Direct dependence between lists is incorporated by introducing interaction terms in the models, and ‘observable’ heterogeneity is usually handled using stratification based on covariate information.

Address for correspondence: Eugene Zwane, Department of Methodology and Statistics, Utrecht University, P.O. Box 80.140, 3508 TC Utrecht, The Netherlands. E-mail: e.zwane@imperial.ac.uk

10.1191/1471082X05st086oa

In the presence of continuous covariates, the standard approach independently proposed by Alho (1990) and Huggins (1989) conditions on the captured individuals and then uses a generalized Horvitz–Thompson estimator to estimate population size (Pollock, 2002). This approach assumes that the lists are independent given the covariates, or alternatively, the lists operate independently at the individual level (Alho, 1990: 625). This is not plausible because most epidemiologic registrations are likely to be dependent even after controlling for observed variables.

To minimize this deficiency, other authors, for example, Darroch *et al.* (1993, 1145) proposed to stratify the observable continuous covariates of heterogenous catchability and then fit models, such as the Rasch model (Agresti, 1994; Coull and Agresti, 1999; Fienberg *et al.*, 1999), to accommodate possible further heterogeneity within each stratum. As the stratification is subjective, there is the possibility that for the same data, researchers using different stratification routines might arrive at different estimates of population size. Furthermore, there is possible loss of information and in some cases an increased number of parameters to be estimated.

In this article, we propose a new methodology for capture–recapture models with continuous covariates whereby list dependence is also modelled. It makes use of the multinomial logit model proposed by (Bock 1975), which integrates log-linear modelling with the multinomial logit approach. When there are more than two lists, this approach enables us to model the dependence between lists without stratifying the observable covariates of heterogenous catchability. Compared to other formulations of the multinomial logit model (Agresti, 2000; Haberman, 1979), Bock's approach explicitly formulates a design matrix for the columns of the data matrix under study, on top of the usual covariate matrix. When the design matrix for the columns is saturated, Bock's model is equivalent to other formulations, save for the interpretation of the logits.

In Section 2, we describe the data set we use to illustrate the approach. The multinomial logit model as proposed by Bock (1975) is discussed in Section 3. In Section 4, we show how the model can be used in the multiple system estimation. A simulation is presented in Section 5. In Section 6, we analyse the data set in detail, and conclude with a discussion in Section 7.

2 Data set

Neural tube defects (NTDs) are among the most frequent birth defects contributing to infant mortality and serious disability (Van der Pal *et al.*, 2003: 33). The most common NTDs are anencephaly and spina bifida. A child with anencephaly cannot survive and dies after birth, whereas a child with spina bifida can survive but often has serious functional impairments.

In the Netherlands, newborns/deliveries with NTDs are registered in several databases. The data include live births, fetal deaths and induced abortions. A problem is that none of these databases are complete, and thus we propose to use capture–recapture methods to estimate the numbers delivered with an NTD. As an illustration, we will use data from three of these incomplete registrations in the year 2000. We describe the registrations briefly

1) *Dutch perinatal database I* (LVR_1): This is a pregnancy and birth registry of low risk pregnancies and births, even if care only relates to a part pregnancy or delivery. In the Netherlands, the midwife is responsible for low risk pregnancies and births (primary care).
2) *Dutch perinatal database II* (LVR_2): This list registers anonymous data concerning the birth of a child in secondary care. If a woman is referred from primary care to secondary care (mostly high risk pregnancies), she can be registered in both LVR_1 and LVR_2.
3) *National neonate database* (LNR): This list contains anonymous information about all admissions and readmissions of newborns to paediatric departments within the first 28 days of life.

The children are matched using a set of key variables that form a unique set of data, namely, the combination of mother's birthday, child's birthday, gender of child and postal code. For more details on these registrations; refer Van der Pal *et al.* (2003).

The Dutch obstetric system is based on risk selection, meaning that everyone can start at the midwife level (primary care, LVR_1) unless there is a primary indication, such as chronic disease of the mother or caesarean section in prior pregnancy. During pregnancy the midwife decides, on the basis of a list of criteria, whether the woman should be referred to the obstetrician (secondary care, LVR_2). Thus low risk is used to refer to deliveries where the safety of the mother and/or child is certain. High risk pregnancies are referred to obstetric departments. It is possible to deliver in hospital skipping the midwife, thus some low risk pregnancies can appear in LVR_2. It is also possible to appear in LVR_2 and not in LVR_1 owing to omissions.

In each of the three registries, delivery weight of the child is recorded. Abortions are recorded in LVR_1 and LVR_2, but not in LNR. In this analysis, we will concentrate on estimating the numbers of children who were delivered, that is, excluding aborted children. A summary of data used in this analysis is shown in Table 1. Table 1 shows that deliveries listed in LNR tend to have normal delivery weight, whereas cases with low delivery weight are frequently listed in both LVR_1 and LVR_2. The reason is that deliveries with low delivery weight are likely to be referred by the midwife (who reports to LVR_1) to the obstetrician who reports to LVR_2. Delivery weight had a missing value

Table 1 Overlap information for delivered children

	Ascertainment history[a]							
	[1,0,0]	[0,1,0]	[0,0,1]	[1,1,0]	[1,0,1]	[0,1,1]	[1,1,1]	Total
Observed	43	37	16	24	7	17	4	148
Delivery weight								
Mean	3.209	2.339	2.745	2.151	3.357	2.729	3.050	2.717
s.e.	0.109	0.170	0.258	0.252	0.167	0.196	0.228	0.083

[a] The first element of the ascertainment profile refers to LVR_1, the second to LVR_2 and the third to LNR (1 is present and 0 is absent).

for a child listed in 'LVR_1 only', which we replaced by the median of delivery weight for children with the same ascertainment profile.

3 Bock's multinomial logit model

Assume that an individual i ($i = 1, 2, \ldots, n$) is classified in one of K nominal categories indexed by k ($k = 1, 2, \ldots, K$), such that $n_{k|i} = 1$ if individual i falls in category k and 0 otherwise. The multinomial logit for individual i is $z'_{\mathrm{i}} = [z_{1|i}, z_{2|i}, \ldots, z_{K|i}]$, implying that the category probabilities for individual i are

$$\pi_{k|i} = \frac{\mathrm{e}^{z_{k|i}}}{\sum_{r=1}^{K} \mathrm{e}^{z_{r|i}}} \tag{3.1}$$

The reason we condition on i is to make it explicit that an individual denotes a stratum.

Now assume that for individual i, there are continuous or categorical variables coded into H columns, indexed by $h(h = 1, 2, \ldots, H)$ and collected in a matrix $\mathbf{X}$ of size $n \times H$. Bock (1975) relates the multinomial logits in $\mathbf{Z}$ to the covariate matrix $\mathbf{X}$ and a design matrix $\mathbf{Y}$ by a matrix of (regression) parameters $\boldsymbol{\Lambda}$. The multinomial logits are decomposed as $\mathbf{Z} = \mathbf{X}\Lambda\mathbf{Y}$. Let the elements of $\mathbf{Y}$ be y_{jk}, with $j = 1, 2, \ldots, J$, the elements of $\mathbf{X}$ be x_{ih} and the elements of Λ be λ_{hj}. Then, the category probabilities are given by,

$$\pi_{k|i} = \frac{\exp(\sum_{h=1}^{H} \sum_{j=1}^{J} x_{ih}\lambda_{hj}y_{jk})}{\sum_{r=1}^{K} \exp(\sum_{h=1}^{H} \sum_{j=1}^{J} x_{ih}\lambda_{hj}y_{jr})} \tag{3.2}$$

Notice that without $\mathbf{Y}$, Equation (3.2) is equivalent to the standard multinomial logit model (Agresti 2000). This implies that $\mathbf{Y}$ can be thought of as a matrix of constraints. As in typical multinomial logit models, some of the logits can be redundant. Here the redundancy is eliminated by defining $\mathbf{Y}$ appropriately.

The log-likelihood for the multinomial logit model can be expressed as

$$\ell = \sum_{i=1}^{n} \sum_{k=1}^{K} n_{k|i} \log [\pi_{k|i}]$$

and thus the first order derivative of the log-likelihood with respect to λ_{st} ($s = 1, 2, \ldots, J$ and $t = 1, 2, \ldots, H$) is

$$\frac{\partial \ell}{\partial \lambda_{st}} = \sum_{i=1}^{n} \sum_{k=1}^{K} \frac{n_{k|i}}{\pi_{k|i}} \left[x_{is}\pi_{k|i} \left(y_{tk} - \sum_{r} y_{tr}\pi_{r|i} \right) \right]$$

The second order derivatives of the log-likelihood are

$$\frac{\partial^2 \ell}{\partial \lambda_{st}\, \partial \lambda_{uv}} = -\sum_{i=1}^{n} \sum_{k=1}^{K} \pi_{k|i} x_{is} x_{iu} y_{tk} \left[y_{vk} - \sum_{r=1}^{K} y_{vr} \pi_{r|i} \right] \tag{3.3}$$

The solution of the likelihood equations corresponds to the maximum of the likelihood. The Newton–Raphson algorithm can be used to arrive at the solution (Bock, 1975: 526).

If we collect the elements of $\mathbf{Z}$ by row into a vector z and the elements of $\mathbf{\Lambda}$ by row into a vector $\boldsymbol{\lambda}$, then we can define Bock's model as the conditional logit model given by $z = [\mathbf{X} \otimes \mathbf{Y}^{\mathrm{T}}]\boldsymbol{\lambda}$. This representation shows that some columns of the matrix resulting from $[\mathbf{X} \otimes \mathbf{Y}^{\mathrm{T}}]$ can be dropped or alternatively that some elements in $\mathbf{\Lambda}$ can be set to zero.

The model just presented can be fitted with available software by exploiting the similarity of the likelihood function with that of the stratified proportional hazards model (Chen and Kuo, 2001). For small data sets, though impractical in the presence of continuous covariates, the model can be fitted as a log-linear model with a nuisance parameter for each distinct value of the set of covariates (Aitkin and Francis, 1992).

4 Multiple system estimator

In this section, we show how the multinomial logit model proposed by Bock can be used to estimate the population size. We consider the estimation of the population size for a problem with three lists, but the ideas can easily be extended to accommodate cases with more than three lists. For two lists, our approach is identical to the approach detailed in Alho (1990).

Assume that for each individual there is a covariate vector x_i with elements x_{ih} $(h = 1, \ldots, H)$, with $x_{i1} = 1$. Each individual has a unique capture profile and the set of possible capture profiles is $\{100, 010, 001, 110, 101, 011, 111\}$. Using this, we define a vector $n_i = [\mathrm{n}_{100|i}, \mathrm{n}_{010|i}, \mathrm{n}_{001|i}, \mathrm{n}_{110|i}, n_{101|i}, n_{011|i}, n_{111|i}]$, where $n_{abc|i} = 1$ if individual i has capture profile $[a,b,c]$ and $n_{abc|i} = 0$ otherwise.

To illustrate how $\mathbf{Y}$ is defined, consider a model assuming that list 1 and 2 and list 2 and 3 are dependent in the presence of covariates. For this model, $\mathbf{Y}$ is given by

$$\begin{array}{rc} & \begin{matrix} 100 & 010 & 001 & 110 & 101 & 011 & 111 \end{matrix} \\ \begin{matrix} \text{list 1} \\ \text{list 2} \\ \text{list 3} \\ \text{list 1:list 2} \\ \text{list 2:list 3} \end{matrix} & \begin{pmatrix} 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix} \end{array} \tag{4.1}$$

In Equation (4.1), the labels for the columns are the capture profiles, and for clarity the elements of **Y** will be denoted by $y_{j(abc)}$ rather than y_{jk}, where $y_{j(abc)}$ is the element in row j of **Y** corresponding to capture profile $[a,b,c]$. The labels for the rows in Equation (4.1) are the main and interaction effects. This matrix is equivalent to the transposed design matrix of a log-linear model (for situations without covariates) assuming list 1 and 2 and list 2 and 3 are dependent, without an intercept. This shows that the main and interaction effects are defined in the same way as in the log-linear model. In principle, **Y** can be a more complex design matrix whose parameters have an interpretation in terms of unobserved heterogeneity (Darroch *et al.*, 1993).

Once **X** and **Y** are defined, the multinomial logit model is used to relate the characteristics of each person listed to their probability of being captured by each list. The fitted capture probabilities are conditional on being observed at least once. However, the resulting parameter estimates can be used to estimate the unconditional probability that the *ith* individual is never captured, denoted by $\hat{\Pi}_{0|i}$ as

$$\hat{\Pi}_{0|i} = \frac{1}{1 + \sum_r \exp(\sum_h \sum_j x_{ih}\lambda_{hj}y_{jr})}$$

Using these probabilities, the estimated population size is

$$\hat{N} = \sum_{i=1}^{n} \hat{N}_i = \sum_{i=1}^{n} \left(\frac{1}{1 - \hat{\Pi}_{0|i}} \right) \tag{4.2}$$

where $\hat{N}_i$ is the contribution of individual i to the estimate of the population size. Thus our estimator is the same as the Horvitz and Thompson (1952) estimator proposed by Alho (1990: 625) and Huggins (1989: 136), with the only difference being in the estimation approaches, implying that any optimality properties of their estimators also hold for Equation (4.2). Thus our estimator is also unbiased. In a model in which the lists are assumed to be independent so that Equation (4.1) does not include the final two rows, Equation (4.2) results in an estimate of the population size identical to those of Alho (1990) and Huggins (1989).

4.1 Interpretation of parameters

As in the models of Alho (1990) and Huggins (1989), the estimated parameters are the logits of the capture probabilities. Instead of having the number of different logits of the capture probabilities equal to the number of lists, if there are interactions between the lists the logits are dependent on whether an individual has been listed in another list. For example, assume that there is a single covariate x_i for individual i and that **Y** is given by Equation (4.1) and further let $\Pi_{j|i}$ denote the capture probability to list j

($j = 1{,}2{,}3$) for individual i. For this small example, $\mathbf{\Lambda}$ and $\mathbf{X}$ are given by

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_{11} & \lambda_{12} & \lambda_{13} & \lambda_{14} & \lambda_{15} \\ \lambda_{21} & \lambda_{22} & \lambda_{23} & \lambda_{24} & \lambda_{25} \end{bmatrix} \quad \text{and} \quad \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

Using this, the logits of the capture probabilities for list 1 are

$$\text{logit}(\Pi_{1|i}) = \begin{cases} \lambda_{11} + \lambda_{21}x_i, & \text{if not in 2} \\ (\lambda_{11} + \lambda_{14}) + (\lambda_{21} + \lambda_{24})x_i, & \text{if also in 2} \end{cases}$$

The equations for $\text{logit}(\Pi_{2|i})$ and $\text{logit}(\Pi_{3|i})$ can be derived in a similar way.

4.2 Variance estimation

In this section, we will derive an estimator of both the conditional and the unconditional variance of $\hat{N}$. Using standard results for conditional variances (Seber, 1982: 9) we have

$$\text{var}(\hat{N}) = E[\text{var}(\hat{N}|\hat{\Lambda})] + \text{var}(E[\hat{N}|\hat{\Lambda}]) \tag{4.3}$$

For ease of notation, let $V_1 = E[\text{var}(\hat{N}|\hat{\Lambda})]$ and $V_2 = \text{var}(E[\hat{N}|\hat{\Lambda}])$. V_1 reflects the sampling fluctuation in the multinomial distribution conditional on being observed and as such does not account for variability in the observed sample size. V_2 reflects variability in the observed sample size.

V_1 can be estimated using the delta method. To use the delta method, we need the first derivative of $\hat{N}_i$ with respect to λ_{hj} which is

$$\frac{\partial \hat{N}_i}{\partial \lambda_{st}} = x_{is}\left(\frac{\hat{\Pi}_{0|i}}{1 - \hat{\Pi}_{0|i}}\right)\left(-\sum_k y_{tk},\ \hat{\pi}_{k|i}\right) \tag{4.4}$$

and the second derivative of Bock's model with respect to lambda given in Equation (3.3). Using this, the conditional variance estimator is given by

$$\hat{V}_1 = \sum_{i=1}^{n}\left[\left(\frac{\partial \hat{N}_i}{\partial \lambda_{st}}\right)^{\mathrm{T}}\left(\frac{\partial^2 \ell}{\partial \lambda_{st}, \partial \lambda_{uv}}\right)^{-1}\left(\frac{\partial \hat{N}_i}{\partial \lambda_{uv}}\right)\right] \tag{4.5}$$

This variance only takes into account that the inclusion probabilities are estimated, but not the fact that the observed sample is drawn from a population.

To incorporate the variability in the observed sample, V_2 has to be added to the conditional variance estimate. Note that V_2 assumes that the estimated inclusion probabilities are fixed, and the usual Horvitz and Thompson (1952) variance estimator for independent observations given by

$$\hat{V}_2 = \sum_{i=1}^{n} \frac{\hat{\Pi}_{0|i}}{(1 - \hat{\Pi}_{0|i})^2}$$

can be used.

From these equations, it is evident that when the inclusion probabilities are large (that is $\hat{\Pi}_{0|i}$ is small for all i) $\hat{V}_2$ will be small, and when the estimated parameters in the model are unreliable (have large standard errors) $\hat{V}_1$ will be large.

The problems with asymptotic confidence intervals have been discussed at length in the literature. The International Working Group for Disease Monitoring and Forecasting (1995a: 1049) noted that for virtually all capture–recapture models, the distribution of the population size is skewed, and thus bootstrap and likelihood-based confidence intervals are preferred. An implementation of the parametric bootstrap for capture–recapture models with continuous covariates is discussed in Zwane and Van der Heijden (2003). Alternatively, the confidence intervals can be computed on the log scale (Chao, 1987).

5 Simulations

The effect of dependence of inclusion probabilities on continuous covariates has been highlighted by several authors, but there is much less attention for remaining dependence between lists after accounting for covariates. The objectives of this section is to examine both the impact of residual dependence between lists (underfitting), by comparing the method assuming independence between lists to the method we have just sketched, and the effects of overfitting. To perform the simulation, a data set of 500 cases was generated, and for each case a single continuous covariate x_i was randomly drawn from the standard normal distribution. The probabilities of being seen in list 1 ($\Pi_{1|i}$), list 2 ($\Pi_{2|i}$) and list 3 ($\Pi_{3|i}$) were generated using $\text{logit}(\Pi_{1|i}) = 0.5 - 0.5x_i$, $\text{logit}(\Pi_{2|i}) = -0.5 - 0.5x_i$, and $\text{logit}(\Pi_{3|i}) = 1.0 - 0.25x_i$.

The data generation was performed using the method described by Oman and Zucker (2001). The data sets were created such that list 1 and list 2 are dependent. We varied $c_{12}(\gamma)$, which is a parametric normal model correlation matrix parameter for the correlation of list 1 and list 2, from 0 to 1. For each value of $c_{12}(\gamma)$, 1000 simulations were performed. For each data set, the population size was estimated using the method assuming independence between lists (the Alho–Huggins approach) denoted by $[1, 2, 3]_x$, a model assuming list 1 and list 2 are correlated (the true model), denoted by $[12, 3]_x$ and a model assuming list 1 and list 2 and list 1 and list 3 are correlated, denoted by $[12, 13]_x$. The coverage is the number of times the 95% asymptotic confidence interval includes the true value of the population of 500.

Table 2 shows that when the correlation between the lists is low ($c_{12}(\gamma) = 0.0$ or $c_{12}(\gamma) = 0.1$), the underfitting and true model perform well, but with increasing levels of correlation, the model $[1, 2, 3]_x$ fits poorly in terms of the AIC. When we compare the mean estimates, we find that $[12, 3]_x$ performs very well, whereas $[1, 2, 3]_x$

Table 2 Estimates of population size for varying degrees of dependence between the first and the second list

$c_{12}(\gamma)$	Mean Pearson's correlation $r_{12}(\gamma)$	Model AIC	Mean estimate	Minimum	Maximum	Mean standard deviation	Coverage
$[1, 2, 3]_x$							
0.0	0.06	1619	500.0	475.4	525.4	7.53	94.0
0.1	0.11	1608	497.1	472.8	519.1	7.37	90.7
0.2	0.17	1598	494.2	470.4	519.4	7.21	82.2
0.3	0.22	1589	491.3	467.6	516.7	6.98	72.0
0.4	0.28	1579	488.6	461.2	513.4	6.87	57.9
0.5	0.34	1569	485.8	458.7	506.5	6.71	45.7
1.0	0.62	1512	471.6	444.5	499.2	5.97	3.4
$[12, 3]_x$							
0.0	0.06	1621	500.2	474.2	527.3	7.92	94.4
0.1	0.11	1608	500.1	473.3	527.3	8.20	95.3
0.2	0.17	1594	499.9	472.2	525.9	8.48	94.2
0.3	0.22	1576	500.0	473.7	528.0	8.71	95.6
0.4	0.28	1554	500.4	470.0	536.1	9.10	94.3
0.5	0.34	1527	500.6	468.5	528.2	9.41	93.9
1.0	0.62	1298	500.3	467.5	538.6	11.00	94.8
$[12, 13]_x$							
0.0	0.06	1623	501.7	472.7	561.5	12.50	94.9
0.1	0.11	1610	502.3	467.7	584.6	13.70	92.6
0.2	0.17	1596	501.9	464.1	589.8	14.80	92.5
0.3	0.22	1578	502.5	463.3	579.3	16.32	94.0
0.4	0.28	1556	503.9	460.7	598.7	18.72	94.0
0.5	0.34	1529	505.2	457.5	636.1	21.90	93.8
1.0	0.62	1302	597.2	555.0	633.0	862729.30	100.0

underestimates the true population and the standard error. In terms of coverage model, $[12, 3]_x$ is superior, and this is more pronounced for values of $c_{12}(\gamma)$ greater than 0.1.

To evaluate the effects of another form of model mis-specification, we compare an overfitting model (i.e., one with more dependencies than the true model) with the true model. As shown in Table 2, the overfitting model is unbiased for low levels of dependence between the lists but is highly variable. In addition, as this model has more parameters it is penalized by the AIC resulting in it having a slightly higher AIC compared with the true model.

In conclusion, this simulation shows that in models with continuous covariates ignoring interaction between sources may severely bias the point estimates and the confidence interval. In addition, it is shown that fitting a more complex model than required results in a slight bias and increased variability in the estimate of the population size.

6 Application to NTD data

The purpose of this study is to estimate the number of children delivered with a NTD. In Table 3, we present two sets of analyses. In the first set, we ignore the delivery weight and in the second it is incorporated. For model selection, we use the crude AIC, as it has been

Table 3 Estimates of population size for the covariates models

Model	Design matrix[a]	Covariate matrix	AIC	Estimated population	$\hat{V}_1$	$\hat{V}_2$	95% C.I.[b]
Without delivery weight							
M_{1a}	[1, 2, 3]	1	524.6	217	197	101	[190, 260]
M_{2a}	[12, 3]	1	526.3	207	357	83	[178, 264]
M_{3a}	[13, 2]	1	522.9	202	170	74	[179, 242]
M_{4a}	[1, 23]	1	523.8	234	439	136	[198, 295]
M_{5a}	[12, 13]	1	522.6	183	172	43	[164, 225]
M_{6a}	[12, 23]	1	525.6	246	2143	164	[188, 391]
M_{7a}	[13, 23]	1	523.3	214	374	96	[183, 272]
M_{8a}	[12, 13, 23]	1	524.6	184	762	44	[157, 289]
With delivery weight (D)							
M_{1b}	[1, 2, 3]	$1 + D$	510.1	212	200	95	[187, 256]
M_{2b}	[12, 3]	$1 + D$	507.6	211	459	91	[179, 276]
M_{3b}	[13, 2]	$1 + D$	508.0	199	194	70	[175, 241]
M_{4b}	[1, 23]	$1 + D$	505.8	236	599	154	[196, 308]
M_{5b}	[12, 13]	$1 + D$	503.9	183	192	44	[163, 227]
M_{6b}	[12, 23]	$1 + D$	506.2	274	9835	245	[180, 646]
M_{7b}	[13, 23]	$1 + D$	503.9	226	921	133	[184, 319]
M_{8b}	[12, 13, 23]	$1 + D$	506.8	193	1437	59	[158, 341]

[a] 1 is LVR_1, 2 is LVR_2 and 3 is LNR.
[b] Confidence intervals computed on the log scale; refer Chao (1987: 787). Using asymptotic confidence intervals results in some lower endpoints being less than the observed sample.

shown using simulations (for capture–recapture problems without covariates) that it tends to pick the data generating model more frequently (Stanley and Burnham, 1998: 492).

Both sets of analyses show that there is dependence between LVR_1 and LNR (Table 3). The AICs show that models incorporating delivery weight fit much better than the log-linear models. After the inclusion of delivery weight the standard errors are larger, implying that there is extra uncertainty in the estimate of the population size. The best models (M_{5b} and M_{7b}) show that even after the inclusion of covariates the dependence between LVR_1 and LNR persists.

A cause for concern is that the estimated population sizes for the models are different even though the model fits are similar. In this situation, basing inferences on M_{5b} or M_{7b} alone is risky (Hoeting *et al.*, 1999: 383). To overcome this concern, we propose to incorporate model uncertainty into our estimates using the model averaging approach detailed in Burnham and Anderson (2002). This approach allows for model selection uncertainty to be incorporated into the standard errors and reduces bias in the parameter estimates in cases when there are a number of models with similar AICs with (substantially) different estimates of the population size and/or their standard errors. All the models with covariates had support from the data using the rough guide from Burnham and Anderson (2002: 171). The model averaged estimate of the population size computed from the models incorporating covariates is 214 deliveries with a standard error of 41.94 deliveries, implying the log-based confidence interval is [168, 367].

Table 4 presents the parameters for the best models, that is, models M_{5b} and M_{7b}. To interpret the parameters in Table 4, we focus on the significant parameter estimates for delivery weight $\hat{\lambda}_{2j}$, as the parameter estimates $\hat{\lambda}_{1j}$ play the role of intercepts. The parameter estimate $\hat{\lambda}_{24}$ in model M_{5b} shows that the probability of being listed in both LVR_1 and

Table 4 Parameters for the models with the lowest AICs

Λ entry	Parameter	Estimate	s.e.	*z-value*	*P-value*
Model M_{5b}					
λ_{11}	LVR_1	−2.826	1.386	−2.039	0.041
λ_{12}	LVR_2	0.108	1.075	0.101	0.920
λ_{13}	LNR	−1.760	0.849	−2.074	0.038
λ_{14}	$LVR_1 \times LVR_2$	2.675	1.397	1.915	0.056
λ_{15}	$LVR_1 \times LNR$	−1.760	1.643	−1.071	0.284
λ_{21}	$LVR_1{:}D$	1.066	0.471	2.262	0.024
λ_{22}	$LVR_2{:}D$	−0.017	0.372	−0.047	0.963
λ_{23}	$LNR{:}D$	0.387	0.303	1.276	0.202
λ_{24}	$[LVR_1 \times LVR_2]{:}D$	−1.172	0.477	−2.456	0.014
λ_{25}	$[LVR_1 \times LNR]{:}D$	0.174	0.522	0.333	0.739
Model M_{7b}					
λ_{11}	LVR_1	−0.064	0.607	−0.105	0.917
λ_{12}	LVR_2	2.957	0.936	3.159	0.002
λ_{13}	LNR	0.935	1.244	0.752	0.452
λ_{14}	$LVR_1 \times LNR$	−3.438	1.727	−1.991	0.046
λ_{15}	$LVR_2 \times LNR$	−2.556	1.360	−1.880	0.060
λ_{21}	$LVR_1{:}D$	−0.164	0.246	−0.670	0.503
λ_{22}	$LVR_2{:}D$	−1.282	0.320	−4.006	0.000
λ_{23}	$LNR{:}D$	−0.800	0.427	−1.873	0.061
λ_{24}	$[LVR_1 \times LNR]{:}D$	0.960	0.556	1.728	0.084
λ_{25}	$[LVR_2 \times LNR]{:}D$	1.110	0.458	2.424	0.015

LVR_2 decreases with increasing birth weight, whereas $\hat{\lambda}_{21}$ shows that the probability of being listed in LVR_1 increases with increasing birth weight. This simply confirms a feature of the databases, as high risk pregnancies (mostly with a low delivery weight) are more likely to be referred by the midwife than births with a normal delivery weight.

The parameters of model M_{7b} show another feature of our databases. For instance, $\hat{\lambda}_{25}$ shows that the probability of being listed both in LVR_2 and in LNR increases with increasing delivery weight. In the same model, $\hat{\lambda}_{22}$ shows that the probability of being listed in LVR_2 decreases with increasing delivery weight.

All these logits confirm our expectations, as children with a low delivery weight are more likely to be referred by the midwives (who report to LVR_1) to the obstetricians (reporting to LVR_2) than children with normal delivery weight, leading to a high probability to be in both lists for children with a low birth weight. Children with a normal delivery weight are also more likely to be taken to paediatric departments (who report to LNR) in the first 28 days of life. Children with abnormally low delivery weight are more likely to die or are still born.

To compare our estimates to other approaches proposed in the literature, for example the Rasch model (Agresti, 1994; Darroch *et al.*, 1993) we discretized the continuous variable delivery weight into a binary variable D^*, such that

$$D^* = \begin{cases} 0, & \text{if } D < 2.5\,\text{kg} \\ 1, & \text{if } D \geq 2.5\,\text{kg} \end{cases}$$

The resulting data are shown in Table 5. The log-linear model with the minimum AIC for the data in Table 5 includes the interactions $LVR_1{:}LVR_2$ and $LVR_1{:}LNR$. The AIC

Table 5 Overlap information for delivered children

	Ascertainment history[a]							
Delivery weight	[1, 0, 0]	[0, 1, 0]	[0, 0, 1]	[1, 1, 0]	[1, 0, 1]	[0, 1, 1]	[1, 1, 1]	Total
<2.5 kg	4	17	4	14	0	4	1	44
≥2.5 kg	39	20	12	10	7	13	3	104

[a]The first element of the ascertainment profile refers to LVR_1, the second to LVR_2 and the third to LNR (1 is present and 0 is absent).

for this model is 24.9, and 183 is the estimated number of deliveries with an NTD. Although there is little difference between this model and model M_{5b} in Table 3, some of the log-linear models with the discretized variable had an undefined estimate of the population size.

The Rasch model expressed as a log-linear model of quasi-symmetry (Darroch *et al.*, 1993) fitted poorly with an AIC of 34.0 and an estimate of 190 deliveries. Refinements of the Rasch model (Darroch *et al.*, 1993: 1143) did not fit better than the log-linear model with the lowest AIC. The Rasch model can also be defined by assuming that the subjects are homogeneous within a set of latent classes (Bartolucci and Forcina, 2001: 715). A model of this form with two latent classes resulted in an AIC of 47.8 and an estimate of the population size of 219. The algorithms used to fit this model are freely downloadable from http://www.stat.unipg.it/~bart/software.html. Owing to lack of degrees of freedom, we did not consider the ordinary latent class model.

In summary, as the Rasch model do not fit better than the log-linear model with the lowest AIC, we conclude that the lists are not of the same kind (International Working Group for Disease Monitoring and Forecasting, 1995a: 1053). In essence, this implies that some of the lists are connected, such that conditionally on a latent class, the probability of appearing in a given list is larger (or smaller), if the subject already appears in a related list (Bartolucci and Forcina, 2001: 715).

7 Conclusion and discussion

We extended the widely used conditional likelihood approach of Alho (1990) and Huggins (1989) by including residual dependence between lists when there are three or more sources for models incorporating continuous covariates. Thus rather than stratifying observable covariates and then fitting log-linear models, we have shown that it is possible to use the observable covariates in their measurement scale. This approach is particularly useful in epidemiology where minimal information is usually collected for each person by each registry resulting in omission of a number of variables that might explain the inclusion to each of the registries. In this instance, dependence will remain even after controlling for the observed variables.

Notice that although we use model averaging to arrive at one estimate of the population size, the performance of the model averaging approach has not been evaluated for models incorporating covariates (Stanley and Burnham, 1998: 402). This remains a topic of further study.

From simulations, we observed that not accounting for dependencies in registrations (when they exist) after controlling for covariates leads to biased estimates of both the population size and the standard errors. This is more pronounced when the dependence between the lists is strong. Furthermore, with increasing number of lists, even at low levels of positive dependence the estimate of population size and its standard error are underestimated, resulting in very poor coverage levels.

It is worth mentioning that the method presented accounts only for observed heterogeneity, where the covariates defining heterogenous catchability are continuous (or a mixture of continuous and categorical variables). Using this approach enables parsimonious parameterization and thus the precision of all parameter estimates is increased (Pollock, 2002: 86). Our approach is different from the approaches of Bartolucci and Forcina (2001), Darroch *et al.*, (1993) and Stanghellini and Van der Heijden (2004) among others, where both observed heterogeneity (owing to only categorical covariates) and unobserved heterogeneity are taken into account. It would be interesting to develop models where unobserved and observed heterogeneity owing to continuous covariates are taken into account.

Acknowledgements

We thank the reviewers for providing helpful comments and suggestions to an earlier version. We also gratefully acknowledge Karin van de Pal from TNO for preparing and providing us with the NTD data set.

References

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

Agresti A (2000) *Categorical data analysis*. New York, USA: Wiley.

Aitkin M and Francis B (1992) Fitting the multinomial logit model with continuous covariates in GLIM. *Computational Statistics and Data Analysis* **14**, 89–97.

Alho J (1990) Logistic regression in capture–recapture models. *Biometrics* **46**, 623–35.

Bartolucci F and Forcina A (2001) Analysis of capture–recapture data with a Rasch-type model allowing for conditional dependence and multidimensionality. *Biometrics* **57**, 714–19.

Bock R (1975) *Multivariate statistical methods in behavioral research*. London: The MIT Press.

Burnham K and Anderson D (2002) *Model selection and multimodel inference: a practical information-theoretic approach*. New York: Springer.

Chao A (1987) Estimating the population size for capture–recapture data with unequal catchability. *Biometrics* **43**, 783–91.

Chen Z and Kuo L (2001) A note on the estimation of the multinomial logit model with random effects. *The American Statistician* 55, 89–95.

Cormack R (1989) Log-linear models for capture–recapture. *Biometrics* **45**, 395–413.

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

Darroch J, Fienberg S, Glonek G and Junker B (1993) A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability. *Journal of the American Statistical Association* **88**, 1137–48.

Fienberg S (1972) The multiple recapture census for closed populations and incomplete 2^k contingency tables. *Biometrika* **59**, 591–603.

Fienberg S, Johnson M and Junker B (1999) Classical multilevel and Bayesian approaches to population size estimation using multiple lists. *Journal of the Royal Statistical Society: Series A* **163**, 383–405.

Haberman S (1979) *Analysis of qualitative data*. New York: Academic Press.

Hoeting J, Madigan D, Raftery A and Volinsky C (1999) Bayesian model averaging: a tutorial. *Statistical Science* **14**, 382–417.

Hook E and Regal R (2000) Accuracy of alternative approaches to capture–recapture estimates of disease frequency: Internal validity of data from five sources. *American Journal of Epidemiology* **152**, 771–78.

Horvitz D and Thompson D (1952) A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47**, 663–85.

Huggins R (1989) On the statistical analysis of capture experiments. *Biometrika* **76**, 133–40.

International Working Group for Disease Monitoring and Forecasting (1995a) Capture–recapture and multiple record systems estimation 1: history and theoretical development. *American Journal of Epidemiology* **142**, 1047–58.

International Working Group for Disease Monitoring and Forecasting (1995b) Capturerecapture and multiple record systems estimation 2: applications. *American Journal of Epidemiology* **142**, 1059–68.

Oman J and Zucker D (2001) Modelling and generating correlated binary variables. *Biometrika* **88**, 287–90.

Pollock K (2002) The use of auxiliary variables in capture–recapture modeling: an overview. *Journal of Applied Statistics* **27**, 85–102.

Schwarz C and Seber G (1999) A review of estimating animal abundance III. *Statistical Science* **14**, 427–56.

Seber G (1982) *The estimation of animal abundance and related parameters*. New York: Macmillan.

Stanghellini E and Van der Heijden P (2004) A multiple-record systems estimation method that takes observed and unobserved heterogeneity into account. *Biometrics* **60**, 510–16.

Stanley T and Burnham K (1998) Information-theoretic model selection and model averaging for closed-population capture–recapture studies. *Biometrical Journal* **40**, 475–94.

Van der Pal K, Van der Heijden P, Buitendijk S and Den Ouden A (2003) Periconceptional folic acid use and the prevalence of neural tube defects in the Netherlands. *European Journal of Obstetrics Gynecology and Reproductive Biology* **108**, 33–39.

Zwane E and Van der Heijden P (2003) Implementing the parametric bootstrap in capture–recapture models with continuous covariates. *Statistics and Probability Letters* **65**, 121–25.