Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Computational Statistics & Data Analysis 47 (2004) 729–743
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Semiparametric models for capture–recapture studies with covariates

E.N. Zwane*, P.G.M. van der Heijden

Department of Methodology and Statistics, University of Utrecht, P.O. Box 80.140, Utrecht 3508 TC, The Netherlands

Received 16 July 2003; received in revised form 6 November 2003; accepted 8 November 2003

Abstract

A flexible method for modelling capture–recapture data with continuous covariates that describe heterogeneous catchability is developed. The well established generalized additive modelling framework is used. An estimator of population size is developed using this method. The performance of the method is demonstrated using neural tube defect capture–recapture data from the Netherlands, with the birth weight of a child as a covariate. The parametric bootstrap is used for variance estimation.

Keywords: Capture–recapture; Generalized additive model; Multinomial logit model; Nonparametric; Population size estimation

1. Introduction

The estimation of the population size in the presence of covariates is currently dominated by parametric approaches. These approaches assume a logistic function for the inclusion probabilities (see, for example, Alho, 1990; Huggins, 1989). The logistic functional form has been criticized as having an implicit shape unsuitable for mark recapture line transect analysis (see Borchers et al., 1998a). Chen and Lloyd (2002, p. 506) also state that plausible parametric models for the inclusion probabilities are seldom available in wildlife or public health contexts, and that the functions for the inclusion probabilities are not identifiable, thus assuming parametric models leads to

* Corresponding author. Tel.: +31-30-2537635; fax: +31-30-253-5797.
E-mail address: e.zwane@fss.uu.nl (E.N. Zwane).

0167-9473/$ - see front matter
doi:10.1016/j.csda.2003.11.010

highly model sensitive results. The nonparametric approach of Chen and Lloyd (2000, 2002) goes a long way in answering these concerns.

Both the current approaches, that is, the parametric and nonparametric approaches, have the implicit assumption that given the covariates the lists are independent, or alternatively, that the lists operate independently at the individual level. Chen and Lloyd (2000, pp. 645–646) recently noted that when there are unmeasured sources of heterogeneity, accounting only for the measured ones will not eliminate all sources of bias. In support, Pollock (2002, p. 88) comments that "although using individual covariates has the purpose of accounting for heterogeneity, some inherent heterogeneity may still remain due to other unobserved variables". This remaining heterogeneity may result in some registrations to be dependent even after controlling for the observed covariates.

This work is motivated by data gathered routinely on children born with a neural tube defects (NTDs) in the Netherlands. The data consist of three incomplete but overlapping registrations with delivery weight of a child as a covariate. In a previous analysis we introduced a quadratic term to capture the nonlinear relationship between the logits of the inclusion probabilities and the birth weight of a child (see Zwane and Van der Heijden, 2002). In the said analysis some of the registrations were dependent even after controlling for the delivery weight. The data are presented in detail in Section 2.

In this article, we present a semiparametric approach which relaxes the linear-in-parameters assumption of the standard approaches using the vector generalized additive model (VGAM) framework proposed by Yee and Wild (1996). VGAMs are an extension of generalized additive models (Hastie and Tibshirani, 1990) to include a class multivariate regression models. In this approach the logits of the inclusion probabilities are specified as sums of nonparametric functions for *specific* covariates. Furthermore, any dependence between the registrations after controlling for the covariates is modelled. Kim and Cohen (2003) and Peng (2003) present similar approaches for matched case–control studies and cure models, respectively.

The paper is structured as follows. In Section 2 we present the data set on neural tube defects from the Netherlands which will be used to illustrate the approach developed in the paper. For completeness we present the triple list capture–recapture problem without covariates in Section 3. In Section 4 we show how the AMNL model can be used in the triple list capture–recapture problem with continuous covariates. A novel graphical technique for evaluating the fit in capture–recapture studies with continuous covariates is presented in Section 5. We present this technique mainly because it has been stated that assessing the goodness of fit in using auxiliary covariates is an *Achilles* heel (White, 2002). In Section 6 we apply the method to the data set presented in Section 2. We conclude with a discussion in Section 7.

2. Data

In the Netherlands data on NTDs can be obtained from various national and regional databases. For this analysis, we will use data collected in 1995 by three national

databases, namely the:

- **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).
- **Dutch perinatal database II** (LVR_2): This list registers anonymous data concerning the birth of a child in obstetrics departments (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.
- **National neonate database** (LNR): This list contains anonymous information about all admissions and re-admissions of newborns to paediatric departments within the first 28 days of life.

In each of these registrations the covariates pregnancy duration and delivery weight in kilograms (DW) are recorded. LVR_1 and LVR_2 also have information on the age of the mother and parity of the child which are not used in this analysis. In this analysis only delivery weight of the child will be used. It should be noted that abortions due to an NTD cannot be reported to LNR thus we only utilize births with a pregnancy duration from 24 weeks (the legal limit for pregnancy termination in the Netherlands). For other details on these registrations, see Van der Pal et al. (2003).

In LNR the child has to be taken to a paediatric department to be registered. Given that children with a very low birth weight and pregnancy duration are more likely to die, they are less likely to be taken to paediatric departments during the first 28 days of life. As a result, these children have a low probability of being included in LNR; we expect the probability to rise rapidly as it approaches the *normal range* of birth weight and pregnancy duration and then level off. The midwife is more likely to perform deliveries of children with *normal* birth weight, whilst it is the opposite for the obstetrician. Subsequently, children with very low birth weight are more likely to be referred by the midwife to obstetric departments resulting in these children having a higher probability of being included in both lists (i.e. LVR_1 and LVR_2) than children with a normal birth weight.

Table 1 shows the cases ascertained and mean delivery weight in kilograms by capture profile. An ascertainment profile of [0,1,0] implies that the delivery is listed in LVR_2 only. As expected most of the children are listed in LVR_1 (the midwife level) and LVR_2 (obstetric departments). A few children are listed in the paediatric registry (LNR) which might be due to high mortality for children with NTDs. Table 1 also shows that deliveries listed in LNR tend to have normal delivery weight whilst cases with low delivery weight (DW less than 2.5 kg) are frequently listed in both LVR_1 and LVR_2, mainly due to that these deliveries are likely to be referred from LVR_1 to LVR_2. Cases listed in LVR_1 only seem to have a normal delivery weight, due to less referrals of these cases to obstetric departments.

A common feature of most epidemiological registrations is that of missing values. These data were not different as there were some missing values in pregnancy duration and delivery weight which had to be imputed before selecting the above data set

Table 1
Overlap information in data set of delivered children, 1995

	Ascertainment profile[a]							Total
	[1,0,0]	[0,1,0]	[0,0,1]	[1,1,0]	[1,0,1]	[0,1,1]	[1,1,1]	
Observed	52	51	12	20	2	15	6	158
Delivery weight								
Mean	3.083	2.587	3.013	2.233	2.953	3.321	2.573	2.812
s.e.	0.118	0.150	0.196	0.255	0.897	0.108	0.318	0.078

[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, 0 is absent).

in Table 1. There were two cases with missing values on both variables, only two missings for "pregnancy duration only" and five for "birth weight only" out of 202 cases of which 158 were cases with pregnancy duration from 24 weeks (see Table 1). The imputation was performed in the statistical package SPSS (SPSS, 1997). Other imputation methods could have been used, but as the proportion missing is less than 0.05, the method of imputation is not very important (see Harrell, 2001, p. 49).

3. Triple records system estimation without covariates

The notation for three lists can be specified as shown in Table 2. Without covariates log-linear models can be used for the estimation of the numbers missed and corresponding estimate of the population size. For example, under independence the log-linear model can be specified as,

$$\log(m_{abc}) = \mathbf{M} \times \boldsymbol{\alpha}; \tag{1a}$$

$$\log \begin{bmatrix} m_{100} \\ m_{010} \\ m_{001} \\ m_{110} \\ m_{101} \\ m_{011} \\ m_{111} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \times \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix}, \tag{1b}$$

where the m_{abc} denote the expected frequencies for the cell probabilities. Columns 2–4 of $\mathbf{M}$ relate to list effects. The estimate of the numbers missed can be computed as $\hat{m}_{000} = \exp[\alpha_0]$. Interaction between lists can be added by multiplying the corresponding list effects. The maximal model does not have all the list interaction.

Table 2
Three list problem without covariates

List 1	List 3			
	Not included		Included	
	List 2			
	Not included	Included	Not included	Included
Not included	n_{000}=?	n_{010}	n_{001}	n_{011}
Included	n_{100}	n_{110}	n_{101}	n_{111}

Alternatively, one can use the fact that Table 2 can be divided into one complete 2×2 table and one incomplete 2×2 table, and assume that the cross product ratio of the complete table, that is, $[m_{001}m_{111}]/[m_{101}m_{011}]$ is the same as in the table involving the missing cell, that is, $[m_{100}m_{010}]/[m_{000}m_{110}]$ (see Darroch et al., 1993, p. 1139). In this instance the estimate of the numbers missed is given by

$$\hat{m}_{000} = \frac{\hat{m}_{100}\hat{m}_{010}\hat{m}_{001}\hat{m}_{111}}{\hat{m}_{110}\hat{m}_{101}\hat{m}_{011}} = n \times \frac{\hat{\pi}_{100}\hat{\pi}_{010}\hat{\pi}_{001}\hat{\pi}_{111}}{\hat{\pi}_{110}\hat{\pi}_{101}\hat{\pi}_{011}}, \tag{2}$$

where $\hat{\pi}_{abc}$ denotes the estimated probability of $[a, b, c]$ conditional on being observed and n is the observed number of cases. Using (2) simply implies that there is no three factor interaction, because if it was present the cross product ratios for the subtables would be different.

4. Triple records system estimation in the presence of continuous covariates

Below we present the additive multinomial logit model that we have adopted for the capture–recapture problem. This enables us to use the multinomial logit model (MNL) in the capture–recapture setting without the assumption of linearity, and further model any residual and/or inherent dependencies between lists. We will first present some notation that we will use for the theoretical development, before proceeding to the parametric MNL model and how it can be generalized to have an additive specification.

4.1. Issues of notation

In the capture–recapture problem the cell probabilities and cell counts are usually denoted by subscripts, for example, in a capture–recapture problem with three lists the capture profiles are denoted by $[a, b, c]$ (where $a, b, c = 0, 1$) with the cell probabilities denoted by π_{abc} (see Section 3). Rather than use this notation we use an alternative for the theoretical development, that is, for a capture profile we will simply use one index denoted by k, but will return to the conventional notation for specific problems. The notation used for the theoretical development is adapted from Yee and Wild (1996).

Table 3
Three list problem with covariates

List 1	List 3			
	Not included		Included	
	List 2			
	Not included	Included	Not included	Included
Not included	$n_{0\|i}$=?	$n_{2\|i}$	$n_{3\|i}$	$n_{6\|i}$
Included	$n_{1\|i}$	$n_{4\|i}$	$n_{5\|i}$	$n_{7\|i}$

The application of the AMNL to the capture–recapture problem is similar to using the MNL model. This is mainly due to that the only difference is that the covariates are no longer linear in the logits, but are now smooth functions. The triple list capture–recapture problem with covariates can be set out as illustrated in Table 3. Note that instead of using the inclusion profile as an index we are now using $k = 1, 2, \ldots, 7$ as an index, but there is a direct relation between the two (cf. Tables 2 and 3). For each individual there is a vector y_i with elements $[n_{1|i}, n_{2|i}, n_{3|i}, n_{4|i}, n_{5|i}, n_{6|i}, n_{7|i}]$, where $n_{k|i} = 1$ if individual i falls in cell k of Table 3 and zero otherwise.

4.2. Multinomial logit model

Assume that an individual indexed by i $(i = 1, 2, \ldots, n)$ is classified into one of K nominal categories, indexed by k $(k = 1, 2, \ldots, K)$, such that $n_{i|k} = 1$ if individual i falls in category k and 0 otherwise. For the capture–recapture problem if there are S lists/registrations then $K = 2^S - 1$. Further assume that for individual i there is a covariate vector $\boldsymbol{x}_i$ of length $H + 1$ consisting of continuous and/or categorical variables indexed by h $(h = 0, 1, 2, \ldots, H)$ with the first element being 1. Denoting the multinomial logit for individual i as $\eta_i' = [\eta_1(\boldsymbol{x}_i), \eta_2(\boldsymbol{x}_i), \ldots, \eta_K(\boldsymbol{x}_i)]$, the category probabilities are then given by,

$$\pi_{k|i} = \exp[\eta_k(\boldsymbol{x}_i)] \Big/ \sum_{r=1}^{K} \exp[\eta_r(\boldsymbol{x}_i)], \tag{3}$$

where $\eta_k(\boldsymbol{x}_i) = \sum_{h=0}^{H} x_{ih}\gamma_{hk}$ (where the γ_{hk}'s denote the parameters of the model). For the model to be identified usually $\eta_1(\boldsymbol{x}_i) = 0$, and the resulting model is called the baseline category logit model (with category 1 being the baseline).

In the two lists problem, Tilling and Sterne (1999) showed that the baseline category logit model is simply a different parameterization of the Alho/Huggins model. Recently Tilling et al. (2001) used the baseline category logit model to estimate the incidence of stroke for data with three registrations and several continuous covariates. They assumed dependence between all pairs of sources. The baseline category logit model is readily

available in standard software, but it is not directly suitable for the capture–recapture problem. For example, in the data set collected by V. Reid and distributed with the CAPTURE program (see Otis et al., 1978), there are six capture periods ($K=2^6-1=63$) and only 38 observations captured at least once, implying the model is not necessarily identified. Thus some restrictions have to be imposed to this model. For this we use the conditional logit model (McFadden, 1973).

To use the conditional logit model for capture–recapture estimation the data have to be rearranged to a suitable format. The responses for all individuals have to be collected in a vector of length $[n.K]$, $\boldsymbol{y} = [y_1|y_2|\cdots|y_n]$ and a covariate matrix specified as,

$$\mathbf{C} = \mathscr{M} \otimes \mathbf{X}, \tag{4}$$

where $\mathscr{M}$ is the design matrix $\mathbf{M}$ (which denotes dependencies between the lists, see Section 3) without the columns of ones (or intercept) and $\otimes$ denotes the Kronecker product. If we let $j = 1,2,\ldots,J$ index the columns of $\mathscr{M}$, the dimension of $\mathbf{C}$ is $[n.K] \times [J.H]$. Note that $S \leqslant J \leqslant K$. To implement the approach of Alho (1990) and Huggins (1989), $\mathscr{M}$ will be given by the design matrix in Eq. (1b) without the first column. Dependencies in lists can also be coded into $\mathscr{M}$.

4.3. Additive multinomial logit model

The additive multinomial logit model we will use was developed by Yee and Wild (1996) as part of vector generalized additive models. In the vector generalized additive model the linear specification of the MNL is replaced with an additive specification, resulting in

$$\eta_k(\boldsymbol{x}_i) = \beta_{j(0)} + \sum_{j=1}^{J}\sum_{h=1}^{H} f_{j(h)}(\boldsymbol{x}_i), \tag{5}$$

where $f_{j(h)}$'s are smooth functions of the predictors and $\beta_{j(0)}$ denotes an intercept term for list effect j. The smooth functions are unknown and usually estimated using some form of scatterplot smoother. In Eq. (5) only the intercept is modelled parametrically and the (continuous) covariates are modelled nonparametrically. In a general formulation, some covariates can be modelled parametrically and others nonparametrically.

Let the coefficients of the parametrically modelled covariates be denoted by $\boldsymbol{\beta}$. To estimate the unknowns $\boldsymbol{f}$ and $\boldsymbol{\beta}$ it is common to use the penalized log-likelihood,

$$pl(\boldsymbol{f},\boldsymbol{\beta}) = \sum_{i=1}^{n}\sum_{k=1}^{K} n_{k|i}\log[\pi_{k|i}] + \frac{1}{2}\sum_{j=1}^{J}\sum_{h=1}^{H}\lambda_{j(h)}\int [f''_{j(h)}(x)]^2\,\mathrm{d}x. \tag{6}$$

The quantity $[f''_{j(h)}(x)]^2$ is the *roughness penalty function* which increases roughness in $f_{j(h)}$ and $\lambda_{j(h)}$ is the smoothing parameter which regulates the smoothness of $f_{j(h)}$. Formal approaches for selecting the smoothing parameter $\lambda_{j(h)}$ include the generalized cross validation (GCV) statistic (see Green and Silverman, 1994, Chapter 3), the Aikake information criterion (see Aikake, 1973). In this analysis we will use both the AIC and the informal/adhoc methods. The AMNL is implemented in the VGAM

library, available from http://www.stat.auckland.ac.nz/~yee/. This library will be used for the analysis.

4.4. Estimation of the population size

Given the parameters above the vector generalized additive model can be used to estimate the population size $\hat{N}$. Let $\hat{\pi}_{abc|i}$ denote the fitted probabilities for individual i conditional on being observed. We can estimate an individual specific unobserved count $\hat{m}_{000|i}$ as,

$$\hat{m}_{000|i} = \frac{\hat{\pi}_{100|i}\hat{\pi}_{010|i}\hat{\pi}_{001|i}\hat{\pi}_{111|i}}{\hat{\pi}_{110|i}\hat{\pi}_{101|i}\hat{\pi}_{011|i}}, \tag{7}$$

and finally use,

$$\hat{N} = \sum_{i=1}^{n} (1 + \hat{m}_{000|i}),$$

to obtain an estimate of total population. Eq. (7) is the same as (2) except for the fact that (7) is stratified by individual. Zwane and Van der Heijden (2002) explicitly showed that in the two list case, using Eq. (7) results in the same estimator of the population size and its corresponding asymptotic variance estimator as in Alho (1990). In this analysis we will use the parametric bootstrap (Buckland and Garthwaite, 1991; Zwane and Van der Heijden, 2003) to compute the variance of the estimate of the population size.

5. Graphical exploration

Hosmer and Lemeshow (1989, Chapter 8) give some guidelines for checking whether the linear-in-the-logit assumption is suitable for the analysis of data using the multinomial logit model. These guidelines are based on performing a series of logistic regressions on the data. In our previous analysis (see Zwane and Van der Heijden, 2002) we checked whether the linear-in-the-logit assumption was suitable for each list separately, and we noted that the logit for *LNR* was nonlinear in delivery weight. This process involves a lot of trial and error and it is a bit cumbersome. Furthermore, if the logit of the probability of being included in a list is nonlinear in a univariate analysis, it does not necessarily imply that it will be nonlinear in a multivariate analysis.

The advantage of the AMNL approach is that it makes it possible to visualize the fits of several models. For the capture–recapture problem we can compare the plot of the fitted probabilities against the covariate under the model and the "empirical" probabilities against the same covariate. Let $\Pi_{1|i}$, $\Pi_{2|i}$, and $\Pi_{3|i}$ denote the inclusion probabilities for individual i to list 1, list 2, and list 3, respectively. These inclusion

probabilities can be computed as

$$\Pi_{1|i} = \frac{\pi_{110|i}}{\pi_{010|i} + \pi_{110|i}}, \quad \frac{\pi_{101|i}}{\pi_{001|i} + \pi_{101|i}}, \quad \frac{\pi_{111|i}}{\pi_{011|i} + \pi_{111|i}}, \tag{8a}$$

$$\Pi_{2|i} = \frac{\pi_{110|i}}{\pi_{100|i} + \pi_{110|i}}, \quad \frac{\pi_{011|i}}{\pi_{001|i} + \pi_{011|i}}, \quad \frac{\pi_{111|i}}{\pi_{101|i} + \pi_{111|i}}, \tag{8b}$$

$$\Pi_{3|i} = \frac{\pi_{011|i}}{\pi_{010|i} + \pi_{011|i}}, \quad \frac{\pi_{101|i}}{\pi_{100|i} + \pi_{101|i}}, \quad \frac{\pi_{111|i}}{\pi_{110|i} + \pi_{111|i}}. \tag{8c}$$

If the probability of being listed in any of the lists does not depend on whether the individual is listed in another list, the quantities for $\Pi_{1|i}$, $\Pi_{2|i}$, and $\Pi_{3|i}$ will be equal. For example,

$$\Pi_{1|i} = \frac{\pi_{110|i}}{\pi_{010|i} + \pi_{110|i}} = \frac{\pi_{101|i}}{\pi_{001|i} + \pi_{101|i}} = \frac{\pi_{111|i}}{\pi_{011|i} + \pi_{111|i}},$$

and this result hold for the other inclusion probabilities.

When the lists are dependent these quantities are not the same but a plot of, for example, $\hat{\pi}_{110|i}/(\hat{\pi}_{010|i} + \hat{\pi}_{110|i})$ under the model plotted against the covariate (in our case delivery weight) and compared with the corresponding empirical (or LOWESS fit of the) probability of being captured by list 1 given that individual is captured by list 2 is informative. A formal goodness of fit test can be the Kolmogorov–Smirnov two sample test. When all two factor interactions are in the model, the probabilities of being ascertained depends on whether the individual is ascertained in other lists, but the plots are still useful. Note that, the probability of being listed in one list given that the individual is not listed in any other list can also be computed. This probability involves the estimated (or missing) cell, and thus does not have a corresponding empirical probability.

Problems with using the LOWESS fit or the empirical probability of being captured is that for each probability in (8a)–(8c) there are likely to be a few observations used and that the range of the covariate for the selected probability might not cover the full range of the covariate distribution. Furthermore, the LOWESS fit does not use information in the other categories and thus it might be more preferable to compare fitted models against the most complex model that the investigator can entertain. The most complex model that we will consider in our analysis will be the default model in the VGAM library, that is the model incorporating all dependencies between lists with 4 effective degrees of freedom for birth weight (AIC=505.1). We compared this model to the LOWESS fit (using the `plsor` function in the HMISC library available from `http://www.cran.r-project.org/`) and the results are shown in Fig. 1. In Fig. 1, for some panels the LOWESS fit does not cover the whole covariate range, which is the contrary for the most complex model, otherwise the two lines are basically identical.

It is clear that some panels in Fig. 1 need further smoothing and this will be done in the next section. What is evident though is that, the probability of being in a list seems to be related to the delivery weight. This shows that models excluding the covariate, for example, log-linear models, will fit the data poorly. Another observation is that

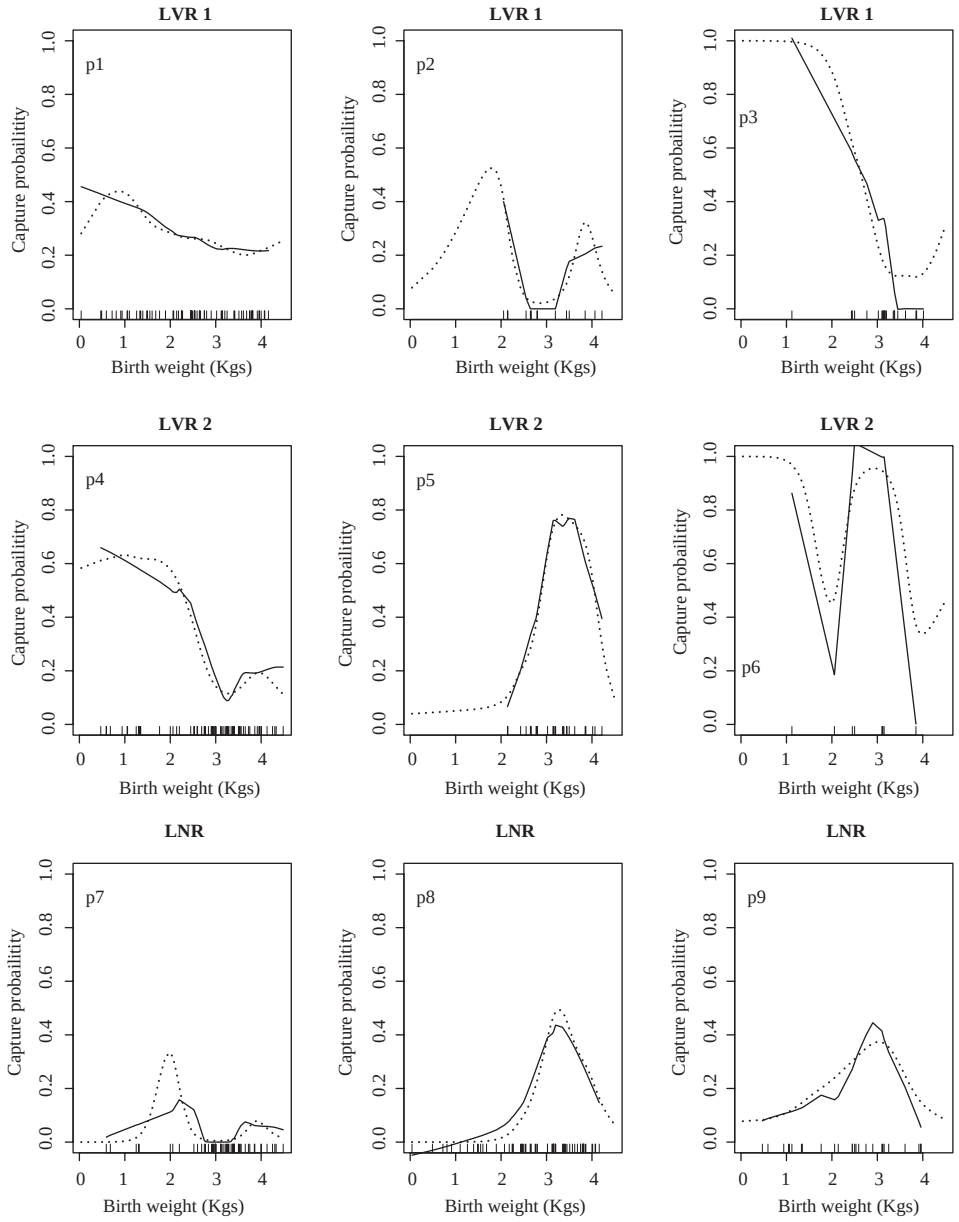

Fig. 1. Inclusion probabilities for LOWESS fit (solid line) and most complex model (dotted line).

the probability that a child is listed depends on whether the child has been listed in another list. This implies that the lists are dependent; as a result the standard method in the presence of covariates will be biased.

6. Application

In this section we will apply the method to the data presented in Section 2. Given that the effect of heterogeneity (which produces apparent dependence) can be reduced by stratification (International Working Group for Disease Monitoring and Forecasting, 1995) or by including continuous covariates, this has to be taken into account in the model search. Thus, we propose to first introduce the covariate in the model, and select the *best* smoothing before introducing dependencies.

In our model search we first considered log-linear models, and the independence model and best-fitting log-linear model are shown in Table 4. We then fitted parametric models incorporating delivery weight as a covariates, and afterwards the semiparametric models and the results are also shown in Table 4. To compute the confidence intervals we used the parametric bootstrap with 2000 replications. The results clearly show that there is a dependence between LVR_2 and LNR, that is, the secondary sources, and that the models without the covariate underestimate the population size. Among the models incorporating the covariate, the model assuming independence after controlling for the covariate (Alho/Huggins type model) also underestimates the population size. The other models (i.e., models M_4, M_5, and M_6) basically result in similar estimates of the population size.

A plot of models M_4 and M_6 using the approach in Section 5 is shown in Fig. 2. Fig. 2 shows that the curves for models M_4 and M_6 are similar, although the semiparametric models tend to be closer to the complex model than the linear fit. For M_4 and M_6, LVR_1 has no interaction for these models and thus the curves for LVR_1 are the same. As expected M_4 and M_6 tend to match the most complex model in places where there are more observations (as shown by rug plots in panels) and do not match in places with a few observations, due to oversmoothing by the complex model.

Fig. 2 shows that the probability of being listed in both LVR_2 when already listed in LVR_1 decreases rapidly with increasing delivery weight. This implies that there are more referrals of children delivered with a low delivery weight but these numbers reduce dramatically as the delivery weight becomes normal. The plot also shows that the inclusion probability to LVR_2 if listed in LNR is about 50% irregardless of whether the child is listed in LVR_1. Another important observation from Fig. 2 is that the

Table 4
Estimates of population size for different models

Model	Design matrix	Number of parameters	AIC	Point estimate	95% interval
M_1	[1,2,3]	4	521.9	252	[237,267]
M_2	[1,23]	5	512.2	303	[285,320]
M_3	$[1,2,3]\otimes DW$	6	517.7	254	[190,407]
M_4	$[1,23]\otimes DW$	8	501.5	346	[193,397]
M_5	$[1,23]\otimes s(DW,df=2)$	11.33	500.9	347	[194,394]
M_6	$[1,23]\otimes s(DW,df=3)$	14.99	501.4	349	[195,424]

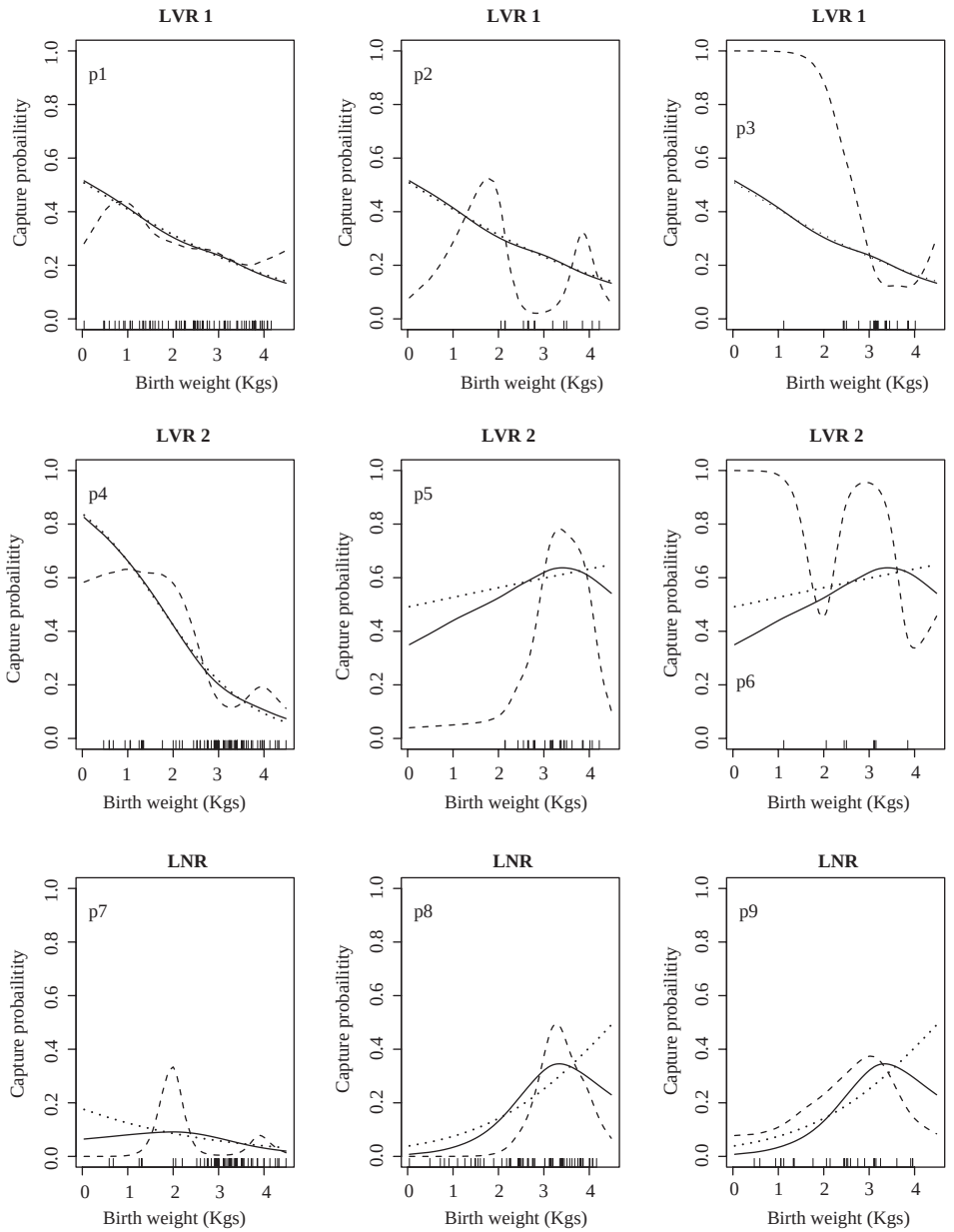

Fig. 2. Inclusion probabilities for different models, M_4 (dotted line), M_6 (solid line), and most complex model (dashed line).

inclusion probability to *LNR* if listed in LVR_1 is very low, around 5%, but inclusion to *LNR* if listed in LVR_2 increases with increasing delivery weight due to less deaths these children. At primary care (LVR_1) the inclusion probability is the same irregardless of

which lists the child is already listed in. This is sort of expected as the midwife is the *main* entry point for pregnancy related issues. One obvious deficiency of the parametric model is that it is rigid, for example the probability of being listed in *LNR* given that the child is listed in LVR_2 continues to increase even after the normal delivery weight range, whilst in the additive model it then decreases.

From the AICs and estimates of the population size in Table 4 and the plots in Fig. 2 it is clear that there is not much difference between the parametric model (M_4) and semiparametric models (M_5 and M_6). Thus we can conclude that in 1995 the true value of children born with an NTD in the Netherlands was not less than 190 and unlikely to have been more than 400 children.

7. Conclusions

We have shown how the additive multinomial logit model can be used in the capture–recapture problem. This model allows for modelling the covariates as smooth terms of the capture probabilities and also allows for dependencies in lists after controlling for the covariates. We also presented a graphical technique for evaluating the fit of multinomial logit models applied to the capture–recapture problem, though the graphs can be used in any multinomial logit model which has a structure (or structure can be devised). The plots we made are in the probability scale but using the logit scale will lead to the same conclusions.

In our example, the AMNL did not perform any better than a simple MNL model, but we envisage that in most other practical problems the AMNL will tend to fit much better than the MNL model. Thus our methods can be viewed as competitors to both the methods of Alho (1990) and Huggins (1989), and those of Chen and Lloyd (2000). Our methods are attractive because they allow for the modelling of residual dependence between lists whilst the other methods assume independence between lists given the covariates. We envisage that the approach can easily be incorporated to the full likelihood method of Borchers et al. (1998b). It would be interesting to compare the approach using the AMNL to the fully nonparametric approach of Chen and Lloyd (2000).

We did not concentrate on the model selection problem but we refer the reader to Stanley and Burnham (1998) for a comprehensive introduction specific to closed population capture–recapture models. If there is one disadvantage of our approach it will be the fact that model selection becomes a cumbersome task. On top of the selection of covariates and dependencies between lists, the value of the smoothing parameter has to be selected, though this can be circumvented by using an integrated smoothing parameter estimation (see Wood, 2000) which currently is not implemented for the AMNL. Chen and Lloyd (2002) stated that, "because the estimate of population size is an integrated quantity, we expect results to be quite insensitive to (a sensible) choice of the bandwidth". Within this GAM framework we also expect the smoothing parameter to also have little effect (if the choice is sensible), but it might have more effect on variability.

Acknowledgements

The authors thank the reviewer for helpful comments, which were very useful in revising the manuscript. We also gratefully acknowledge Karin van de Pal from TNO for preparing and providing us with the neural tube defects data set.

References

Aikake, H., 1973. Information theory and an extension of the maximum likelihood principle. In: Petran, B.N., Csaki, F. (Eds.), International Symposium on Information Theory. Akadmiai Kiadi, Budapest, Hungary, pp. 261–281.

Alho, J., 1990. Logistic regression in capture–recapture models. Biometrics 46, 623–635.

Borchers, D., Buckland, S., Goedhart, P., Clarke, E., Hedley, S., 1998a. Horvitz–Thompson estimators for double-platform line transect surveys. Biometrics 54, 1221–1237.

Borchers, D., Zucchini, W., Fewster, R., 1998b. Mark-recapture methods for line transect studies. Biometrics 54, 1207–1220.

Buckland, S., Garthwaite, P., 1991. Quantifying precision of mark-recapture estimates using the bootstrap and related methods. Biometrics 47, 255–268.

Chen, S., Lloyd, C., 2000. A non-parametric approach to the analysis of two stage mark-recapture experiments. Biometrika 88, 649–663.

Chen, S., Lloyd, C., 2002. Estimation of population size from biased samples using nonparametric binary regression. Statistica Sinica 12, 505–518.

Darroch, J., Fienberg, S., Glonek, G., Junker, B., 1993. A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability. J. Amer. Statist. Assoc. 88, 1137–1148.

Green, P., Silverman, B., 1994. Nonparametric Regression and Generalized Linear Models. Chapman & Hall, London.

Harrell, F., 2001. Regression Modelling Strategies: with Applications to Linear Models, Logistic Regression, and Survival Analysis. Springer, New York.

Hastie, T., Tibshirani, R., 1990. Generalized Additive Models.. Chapman & Hall, New York.

Hosmer, D., Lemeshow, S., 1989. Applied Logistic Regression. Wiley, New York.

Huggins, R., 1989. On the statistical analysis of capture experiments. Biometrika 76, 133–140.

International Working Group for Disease Monitoring and Forecasting, 1995. Capture–recapture and multiple record systems estimation 1: history and theoretical development. Amer. J. Epidemiology 142, 1047–1058.

Kim, I., Cohen, N., 2003. Semiparametric and nonparametric modeling for effect modification in matched studies. Comput. Statist. Data Anal., in press.

McFadden, D., 1973. Conditional logit analysis of qualitative choice behavior. In: Zarembka, P. (Ed.), Frontiers in Econometrics. Academic Press, New York, pp. 105–142.

Otis, D., Burnham, K., White, G., Anderson, D., 1978. Statistical inference from capture data on closed animal populations. Wildlife Monographs, No. 62.

Peng, Y., 2003. Fitting semiparametric cure models. Comput. Statist. Data Anal. 41, 481–490.

Pollock, K., 2002. The use of auxiliary variables in capture–recapture modeling : an overview. J. Appl. Statist. 27, 85–102.

SPSS, 1997. SPSS Missing Value Analysis 7.5. SPSS Inc., Chicago.

Stanley, T., Burnham, K., 1998. Information-theoretic model selection and model averaging for closed-population capture–recapture studies. Biometrical J. 40, 475–494.

Tilling, K., Sterne, J., 1999. Capture–recapture models including covariate effects. Amer. J. Epidemiol. 149, 392–400.

Tilling, K., Sterne, J., Wolfe, C., 2001. Estimation of incidence of stroke using a capture–recapture model including covariates. Internat. J. Epidemiol. 30, 1351–1359.

Van der Pal, K., Van der Heijden, P., Buitendijk, S., Den Ouden, A., 2003. Periconceptional folic acid use and the prevalence of neural tube defects in the Netherlands. Eur. J. Obset. Gynecol. Reprod. Biol. 108, 33–39.

White, G., 2002. Discussion comments on: the use of auxilliary covariates in capture–recapture modelling: an overview. J. Appl. Statist. 29, 103–106.

Wood, S.N., 2000. Modelling and smoothing parameter estimation with multiple quadratic penalties. J.R. Statist. Soc. B 62, 413–428.

Yee, T., Wild, C., 1996. Vector generalized additive models. J.R. Statist. Soc. B 58, 481–493.

Zwane, E., Van der Heijden, P., 2002. The multiple-system estimator in the presence of covariates. In: Stasinopoulos, M., Touloumi, G. (Eds.), 17th International Workshop on Statistical Modelling. Chania University, Chania, Greece, pp. 697–701.

Zwane, E., Van der Heijden, P., 2003. Implementing the parametric bootstrap in capture–recapture models with continuous covariates. Statist. Probab. Lett. 65, 121–125.