Observed heterogeneity in the capture-recapture problem

Eugene Nkosinathi ZWANE

ISBN 90-393-3837-X

Observed heterogeneity in the capture-recapture problem

Geobserveerde heterogeniteit in vangst-hervangst methoden

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor
aan de Universiteit Utrecht
op het gezag van de Rector Magnificus, Prof. dr. W. H. Gispen,
ingevolge het besluit van het College voor Promoties
in het openbaar te verdedigen
op donderdag 28 october 2004 om 10.30 uur

door

Eugene Nkosinathi Zwane

geboren op 11 juni 1972, te Manzini, Swaziland

Promotoren: Prof. dr. P.G.M. van der Heijden
Faculteit der Sociale Wetenschappen
Universiteit Utrecht

Dit proefschrift werd mede mogelijk gemaakt met financiële steun van de Nederlandse Organisatie voor Wetenschappelijk Onderzoek.

Contents

Acknowledgements

It is a pleasure to thank many people who made this possible.

It is difficult to overstate my gratitude to my PhD supervisor and employer Peter van der Heijden for his enthusiasm, inspiration, and effort to explain things clearly and simply. Throughout my thesis-writing period he provided encouragement, sound advice, good teaching and lots of good ideas. Most importantly I would like to thank him for giving me a chance to explore my research capabilities.

I would also like to thank the rest of the academic and support staff of the Department of Methodology and Statistics (Faculty of Social Sciences) at Utrecht University for making my stay pleasant. In particular, I am grateful to Ardo van den Hout for assistance in the Dutch language.

Warm thanks are also due to my friends and "former" colleagues at the Central Statistical Office in Swaziland for their encouragement and words of support.

Finally, I am indebted to my entire extended family for their understanding, endless patience and encouragement when it was most required. To my mother, I do say to myself, you did a fine job raising me. To my daughter, Nkosephayo, you are the best thing that ever happened to me.

Chapter 1

Introduction

Suppose you would like to estimate the prevalence of a disease, say neural tube defects. For this purpose, you need a complete count of the number of children born or delivered with (a) neural tube defect(s). However no epidemiological registration is infallible, and thus capture-recapture methods are now accepted as a methods that can be used to arrive at a valid estimate of the population size. This thesis is concerned with efficient modelling of observed heterogeneity in capture-recapture studies. Although the emphasis in this thesis is on epidemiological applications, the capture-recapture technique can be used in other situations, such as, estimating the number of drug users, the number of people with infectious diseases, and the number of prostitutes.

1.1 Epidemiology

Epidemiology is the study of how often diseases occur in different groups of people and why (see Rothman and Greenland, 1998, Chapter 3). A central task in epidemiological research is to quantify the occurrence of disease in populations. This epidemiological information is then used to plan and evaluate strategies to prevent illness and as a guide to the management of patients in whom disease has already developed.

There are several measures of disease occurrence but the most common is the *prevalence rate*. Prevalence may be defined as the proportion of a population that has a disease at a specific point (see Rothman and Greenland, 1998, p. 42). To get accurate estimates of prevalence, accurate information on the subset of the population with disease is required.

One of the major problems in estimating the prevalence of a disease is under-ascertainment or under-counting. Even when the methodology and resources for creating a specialist registry aimed at registering all cases in a population are well established, it is unlikely that every case will be registered (see Tilling and

Sterne, 1999, p. 392). Thus accurate information on these quantities is often unavailable to health planners and epidemiologists. As a result, some estimate of the coverage of the registries and consequently the true number of diseased individuals is required.

In the next Section we present the capture-recapture methodology as a tool that can be used to estimate the number of diseased individuals in a population using several incomplete but overlapping registries.

1.2 The capture-recapture methodology

In this Section we first present the history of the capture-recapture methodology and then discuss how the methodology can be applied in human populations, with emphasis on epidemiology.

1.2.1 Historical development

The capture-recapture methodology was initially developed to estimate the size of wildlife populations. The animals are trapped, marked, and released on a number of occasions, and the individual trapping histories are then used to estimate the size of the population (Seber, 1982). The simplest capture-recapture model is the two sample problem, and can be set out as in table 1.1.

Table 1.1: Two sample problem

	Sample 2	
Sample 1	Not captured	Captured
Not captured	$n_{00} = ?$	n_{01}
Captured	n_{10}	n_{11}

The goal is to estimate the numbers not included in both samples, n_{00}. This number can be estimated using the information on the numbers captured in both samples and the numbers captured in only one sample, thus providing the total population size N. This estimate is valid under certain assumptions:

1. There is no change to the population during the period under investigation (population is closed), that is, there are no births, no deaths and no migrants. This implicitly implies that each individual in the population has a non-zero probability of being observed in all the samples.

2. Individuals can be matched without error or alternatively that the animals do not lose their tags (individuals can be matched from capture to recapture).

3. For each sample, each individual has the same chance of being included in the list (homogeneity of inclusion probabilities). If assumption 1 does not hold, then this assumption will not hold as cases which stay in the population are clearly likely to have higher catchability than those who migrate (or die).

4. The two samples are independent. This assumption actually follows from assumption 3 since the latter implies that marked and unmarked have the same probability of being caught in the second sample so that capture in first sample does not affect capture in second sample: Samples are independent. It is convenient, however, to list this assumption separately (see International Working Group for Disease Monitoring and Forecasting, 1995a, p. 1048).

If these assumptions hold then the estimated numbers missed is

$$\hat{n}_{00} = \frac{n_{10} \times n_{01}}{n_{11}}, \tag{1.1}$$

and the resulting estimate of the population size is given by $\hat{N} = n_{10} + n_{01} + n_{11} + \hat{n}_{00}$.

In wildlife the first two assumptions usually hold, because the experimenter actively samples the population to form the sample, and tries to do so in a way which approximates a random sample of the whole population (see Cormack, 1999, p. 909).

If certain groups of animals are more likely to be captured than other groups, or there is heterogeneity of the capture probabilities then (1.1) will be biased. Note that 'variable catchability' (or heterogeneity of the capture probabilities) does not refer to the difference among samples of the cases captured, but to the difference (defined by covariates) within the population in susceptibility to capture by a sample or samples (see Hook and Regal, 1999, p. 918). As an example consider the following hypothetical animal population, where large animals are more likely to be caught compared to smaller animals (see table 1.2). In table 1.2

Table 1.2: Heterogeneity in two sample problem

Animal		Sample 2	
Size	Sample 1	Not captured	Captured
Small	Not captured	20	16
	Captured	50	40
Large	Not captured	7	14
	Captured	50	100

at each level of animal size the capture probabilities are independent. Ignoring animal size violates assumption 3, making the samples to be dependent (see discussion of assumption 4), and consequently (1.1) results in a biased estimate of the population size. In this case it is preferable to estimate the population size for each group separately and then pool the estimates to arrive at the estimate of the population size. Instead of stratifying based on observable covariates one can model heterogeneity in some way, for example by using a log-linear model.

With two samples only independence between the samples can be entertained, as there is no evidence about the 'two-sample' interaction in the data. With more samples dependencies between the samples can be entertained except the 'all-sample' interaction. Analogous to the two sample problem, there is no evidence at all in the data about the 'all-sample' interaction from the numbers analyzed in any study (see Cormack, 1999, p. 913). Fienberg (1972) showed how dependencies between samples can be modelled by placing the capture-recapture problem with multiple samples in the standard framework of log-linear models. An advantage of the log-linear modelling framework is that both heterogeneity and dependencies can be modelled.

In the presence of continuous covariates, the log-linear modelling framework cannot be used directly. Alho (1990) and Huggins (1989) independently proposed a logistic regression approach to modelling heterogeneity due to continuous covariates. This approach assumes that the lists are independent given the covariates, or alternatively, the lists operate independently at the individual level (see Alho, 1990, p. 625).

1.2.2 Application in human populations

Capture-recapture methods in epidemiology are attempts to estimate or adjust for the extent of incomplete ascertainment using information from overlapping lists of cases from distinct sources (see International Working Group for Disease Monitoring and Forecasting, 1995a,b). The technique was first used in human populations by Sekar and Deming (1949). In this case, "being captured by sample i" is replaced by "being included in list i". Capture-recapture techniques have been used in estimating the prevalence of drug users and the number infected with the human immunodeficiency virus, see for example Davies et al. (1999). Other areas of application include the estimation of deaths due to traffic accidents (see Razzak and Luby, 1998) and the estimation of the prevalence of infectious and non-communicable diseases (for details, see International Working Group for Disease Monitoring and Forecasting, 1995b).

In contrast to animals studies, epidemiological capture-recapture data uses preconstructed lists, and thus there is no random selection nor common protocol (see Cormack, 1999, p. 909). The lists can be constructed from a variety of sources, such as, hospital records, doctors' medical files, medical prescriptions, and so on (see International Working Group for Disease Monitoring and Forecasting, 1995a, p. 1048). These lists are incomplete by their nature, and the problem is

to estimate those missing from both lists.

In epidemiology it is likely that the assumptions of capture-recapture methods do not hold, for a critical discussion see Cormack (1999). A remedy for assumption 1, is to set up the experiments such that the population is closed, that is, such that the birth, death and migrations rates are negligible. This is usually accomplished by selecting the sources such that all the individuals have an opportunity to be identified by each of the sources, and that the sources have the same geographic coverage, time frame and apply the same case definition to recruit cases (see Chang et al., 1999, p. 927). Assumption 2 depends on the quality of the patients records and uniqueness of the patients matching codes. In some countries, matching can be a problem.

In this thesis we assume that the assumption 1 and 2 hold, and thus focus on the approaches which relax assumptions 3 and 4. This is due to that in epidemiology list dependence and heterogeneity of inclusion probabilities are a norm (see Schwarz and Seber, 1999). As mentioned earlier the effects of heterogeneity can be reduced by stratification. The estimates of the population size from each strata are then added up to arrive at an estimate of the population size. Alternatively log-linear models can be used to model both dependencies between lists and heterogeneity of inclusion probabilities. In the presence of continuous covariates the method of Alho (1990) and Huggins (1989) can be used. To be able to form strata the covariates used have to be measured by all lists. A difficulty in using the covariates in epidemiological capture-recapture data is that the lists are usually started for different purposes and as such are likely to cover different populations or have different covariate sets.

Dependence between lists is a serious issue in epidemiological studies. For example, if doctors refer patients to hospitals, then doctors records and hospital admission will not give two independent lists (see International Working Group for Disease Monitoring and Forecasting, 1995a, p. 1048). With only 2 lists, the dependence between the lists cannot be modelled. With more than 2 lists dependence between lists can be entertained except the 'all list' interaction. In the presence of continuous covariates, the standard method of Alho (1990) and Huggins (1989) assumes the lists are independent at the individual level, which might be incorrect due to lack of sufficient covariates which define the heterogeneous catchability.

1.3 Contents of this thesis

This thesis is based on six papers, both published and submitted for publication. As a result, these chapters can be read independently and in arbitrary order.

The main goal of the thesis is to develop methods for modelling observed heterogeneity in capture-recapture data. We start by elaborating the logistic regression methods of Alho (1990) and Huggins (1989) so that dependencies between registrations can be modelled in the presence of continuous covariates in

Chapter 2. The methods of Alho (1990) and Huggins (1989) assume that the lists are independent at the individuals level. This restriction is inappropriate as some heterogeneity may still remain due to other unobservable variables, and this may in-turn induce some dependence between some of the lists. Another reason is that some registrations are inherently dependent and inclusion of covariates will not eliminate the dependence. The modelling of dependencies is accomplished through the use of the multinomial logit model by Bock (1975).

In Chapter 3, we develop a method to relax the linear in the logit assumption of the multinomial logit model. The logistic function is often criticized as having an implicit shape unsuitable for capture-recapture studies (see Borchers et al., 1998). Here we essentially elaborate the multinomial logit model presented in Chapter 2 so that the covariates are not necessarily linear in the logits.

In Chapter 4, we focus on implementing the parametric bootstrap in the presence of continuous covariates. It has already been discussed at length that the symmetric (or asymptotic) confidence intervals are inappropriate for capture-recapture studies. The International Working Group for Disease Monitoring and Forecasting (1995a) noted that, for virtually all models proposed in capture-recapture literature the distribution of the estimate of the population size is skewed. In literature several authors have used the non-parametric bootstrap in the presence of continuous covariates (see Huggins, 1989; Tilling and Sterne, 1999), but as noted by Norris and Pollock (1996) this bootstrap method results in a variance estimate which is likely to be smaller than the true variance, because it conditions on being observed.

The next set of chapters focus on methods for analyzing defective capture-recapture data. In Chapter 5, we develop an approach for analyzing capture-recapture data when some registrations do not collect data on a subset of the population. Notice that if some registrations do not measure certain individuals then those individuals have a zero probability of being in all lists, implying the usual capture-recapture approaches fail.

We then work on approaches of modelling capture-recapture data when some of the registration do not measure some covariates of heterogenous catchability. The goal is to utilize all variables of heterogenous catchability available as in some cases dropping them may induce an 'all-list' interaction and thus invalidate results. In Chapter 6, we present the EM algorithm (see Little and Rubin, 1987, Chapter 9) for analyzing such data when there are only categorical covariates. In Chapter 7 we propose multiple imputation (Schafer, 1997) for problems with missing continuous (and categorical) covariates.

References

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

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

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

Chang, Y., LaPorte, R., Aaron, D., and Songer, T. (1999). The importance of source selection and pilot study in the capture-recapture application. *Journal of Clinical Epidemiology*, 52:927–928.

Cormack, R. (1999). Problems with using capture-recapture in epidemiology: an example of a measles epidemic. *Journal of Clinical Epidemiology*, 52:909–914.

Davies, A., Cormack, R., and Richardson, A. (1999). Estimation of injecting drug users in the City of Edinburgh, Scotland, and the number infected with human immunodeficiency virus. *International Journal of Epidemiology*, 28:117–121.

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

Hook, E. and Regal, R. (1999). Recommendations for presentation and evaluation of capture-recapture estimates in epidemiology. *Journal of Clinical Epidemiology*, 52:917–926.

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

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*, 142:1047–1058.

International Working Group for Disease Monitoring and Forecasting (1995b). Capture-recapture and multiple record systems estimation 2: applications. *American Journal of Epidemiology*, 142:1059–1068.

Little, R. and Rubin, D. (1987). *Statistical analysis with missing data*. J. Wiley & Sons, New York.

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

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

Rothman, K. and Greenland, S. (1998). *Modern Epidemiology.* Lippincott Willams & Wilkins, Philadelphia, USA.

Schafer, J. (1997). *Analysis of Incomplete Multivariate Data.* Chapman & Hall/CRC, New York.

Schwarz, C. and Seber, G. (1999). A review of estimating animal abundance III. *Statistical Science*, 14:427–456.

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

Sekar, C. and Deming, W. (1949). On a method of estimating birth and death rates and the extent of registration. *Journal of the Americal Statistical Association*, 44:101–115.

Tilling, K. and Sterne, J. (1999). Capture-recapture models including covariate effects. *American Journal of Epidemiology*, 149:392–400.

Chapter 2

Continuous covariates in capture-recapture models

In the presence of continuous covariates, standard capture-recapture methods either assume that the registrations operate independently at the individual level, or 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.

2.1 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 (see 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), that may include among others, hospital records, or patient group records. The traditional 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, p. 1145).

The most prevalent method for analyzing such data uses log-linear models (see Fienberg, 1972; Cormack, 1989). 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, p. 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.

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 (see Alho, 1990, p. 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, p. 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), that accommodates 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 (Haberman, 1979; Agresti, 2000), 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 2.2, we describe the data set we use to illustrate the approach. The multinomial logit model as proposed by Bock (1975) is discussed in Section 2.3. In Section 2.4, we show how the model can be used in the multiple system estimation. A simulation is presented in Section 2.5. In Section 2.6, we analyze the data set in detail, and conclude with a discussion in Section 2.7.

2.2 Data set

Neural tube defects (NTD's) are amongst the most frequent birth defects contributing to infant mortality and serious disability (see Van der Pal et al., 2003, p. 33). The most common NTD's are anencephaly and spina bifida. A child with anencephaly cannot survive and dies after birth whilst a child with spina bifida can survive but often has serious functional impairments.

In the Netherlands, newborns/deliveries with NTD's are registered in several databases. The data includes 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 re-admissions 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, see 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, caesarean section in prior pregnancy. During pregnancy the midwife decides, based on a list of criteria, whether the women 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 LVR_1 due 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. The data used in this analysis are shown in Table 2.1. Table 2.1 shows that deliveries listed in LNR tend to have normal delivery

Table 2.1: Overlap information for delivered children

	Ascertainment history †							
	[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

†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).

weight whilst cases with low delivery weight are frequently listed in both LVR_1 and LVR_2. The reason is that deliveries with a 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 for a child listed in "LVR_1 only" which we replaced by the median of delivery weight for children with the same ascertainment profile.

2.3 Bock's multinomial logit model

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

$$\pi_{k|i} = \frac{e^{z_{k|i}}}{\sum_{r=1}^{K} e^{z_{r|i}}}. \tag{2.1}$$

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 $(h = 1, 2, ..., H)$ 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 $\mathbf{\Lambda}$. The multinomial logits are decomposed as $\mathbf{Z} = \mathbf{X}\mathbf{\Lambda}\mathbf{Y}$. Let the elements of $\mathbf{Y}$ be y_{jk}, with $j = 1, 2, ..., J$, the elements of $\mathbf{X}$ be x_{ih}, and the elements of $\mathbf{\Lambda}$ be λ_{hj}. The category probabilities are then given by,

$$\pi_{k|i} = \frac{\exp(\sum_{h=1}^{H}\sum_{j=1}^{J} x_{ih}\lambda_{hj}y_{jk})}{\sum_{r=1}^{K}\exp(\sum_{h=1}^{H}\sum_{j=1}^{J} x_{ih}\lambda_{hj}y_{jr})}. \tag{2.2}$$

Notice that without $\mathbf{Y}$ (2.2) is equivalent to the standard multinomial logit model (see 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, but for the capture-recapture model presented in Section 2.4 the redundancy is eliminated by defining $\mathbf{Y}$ appropriately.

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

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

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

$$\frac{\partial^2 \ell}{\partial \lambda_{st}\partial \lambda_{uv}} = -\sum_{i=1}^{n}\sum_{k=1}^{K} \pi_{k|i}x_{is}x_{iu}y_{tk}\left[y_{vk} - \sum_{r=1}^{K} y_{vr}\pi_{r|i} \right]. \tag{2.3}$$

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 (see Bock, 1975, p. 526).

If we collect the element of $\mathbf{Z}$ by column into a vector $\boldsymbol{z}$ and the elements of $\boldsymbol{\Lambda}$ by row into a vector $\boldsymbol{\lambda}$ then we can define Bock's model as the conditional logit model given by $\boldsymbol{z} = [\mathbf{X} \otimes \mathbf{Y}^T]\boldsymbol{\lambda}$. Using this presentation shows that our approach is general in that some columns of the matrix resulting from $[\mathbf{X} \otimes \mathbf{Y}^T]$ can be dropped or alternatively that some elements in $\boldsymbol{\Lambda}$ can be set to zero. This implies that it is possible to restrict the model such that the covariates do not relate to the capture probabilities of some lists. We prefer the Bock model specification due to the clear separation of the design matrix for the list effects (and direct dependencies between lists) and the covariate matrix.

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 (see Chen and Kuo, 2001). For capture-recapture models where only the all list interaction is absent, the model can be fitted using the baseline category logit model. 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 (see Aitkin and Francis, 1992).

2.4 Multiple system estimator

In this Section we show how the multinomial logit model proposed by Bock can be used to estimate population size. As an illustration we consider the estimation of the population size for a problem with three lists but the ideas can 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 $\boldsymbol{x}_i$ with elements x_{ih} $(h = 1, \cdots, H)$, 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 $\boldsymbol{n}_i = [n_{100|i}, n_{010|i}, n_{001|i}, n_{110|i}, n_{101|i}, n_{011|i}, n_{111|i}]$, 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

$$\begin{array}{l} \\ \text{list 1} \\ \text{list 2} \\ \text{list 3} \\ \text{list 1:list 2} \\ \text{list 2:list 3} \end{array} \begin{array}{c} \begin{array}{ccccccc} 100 & 010 & 001 & 110 & 101 & 011 & 111 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \end{pmatrix} \end{array}. \tag{2.4}$$

In (2.4) the labels for the columns are the capture profiles and for clarity the elements of $\mathbf{Y}$ will be denoted by $y_{j(abc)}$ rather than y_{jk}, where $y_{j(abc)}$ is the element in row j of $\mathbf{Y}$ corresponding to capture profile $[a, b, c]$.

The labels for the rows in (2.4) are the list and interaction effects. Thus, the first three rows of $\mathbf{Y}$ refer to the list effects and the fourth and fifth rows to the interactions of list 1 and 2, and list 2 and 3 respectively. It is evident that this matrix is equivalent to the design matrix of a log-linear model (for situations without covariates) assuming list 1 and 2, and list 2 and 3 are dependent without the first column (or intercept) transposed. This shows that the list effects and dependencies between lists in $\mathbf{Y}$ are defined in the same way as a log-linear model without the all list interaction. In principle, we can also assume that the lists are of the same kind at the individual level (or after controlling for covariates) and define $\mathbf{Y}$ similar to Darroch et al. (1993).

To estimate population size $\hat{N}$ we use the fitted probabilities given by $[\hat{\pi}_{100|i}, \hat{\pi}_{010|i}, \hat{\pi}_{001|i}, \hat{\pi}_{110|i}, \hat{\pi}_{101|i}, \hat{\pi}_{011|i}, \hat{\pi}_{111|i}]$. These probabilities add up to one for each individual. Using the fitted probabilities, the individual's contribution to the estimate of the numbers missed is

$$\hat{m}_{000|i} = \frac{\hat{\pi}_{100|i}\hat{\pi}_{010|i}\hat{\pi}_{001|i}\hat{\pi}_{111|i}}{\hat{\pi}_{110|i}\hat{\pi}_{101|i}\hat{\pi}_{011|i}}. \tag{2.5}$$

This contribution, like in cases without covariates (see Darroch et al., 1993, p. 1139) assumes that there is no three list interaction. Using (2.5) the estimated population size is

$$\hat{N} = \sum_{i=1}^{n} \left(1 + \hat{m}_{000|i}\right). \tag{2.6}$$

Notice that in the case where there are no covariates (2.6) is identical to the estimate suggested by Bishop et al. (1975) among others.

The estimate of N given in (2.6) can also be written as

$$\hat{N} = \sum_{i=1}^{n} \left(1 + \hat{m}_{000|i}\right) = \sum_{i=1}^{n} \left(\frac{1}{1 - \hat{\Pi}_{0|i}}\right), \tag{2.7}$$

where $\hat{\Pi}_{0|i} = \hat{m}_{000|i}/(1 + \hat{m}_{000|i})$ is the probability that individual i is missed by all lists. Thus our estimator is the same as the Horvitz and Thompson (1952) estimator proposed by Alho (1990, p. 625) and Huggins (1989, p. 136), with the only difference being in the estimation approaches, implying that any optimality properties of their estimators also hold for (2.7). Thus our estimator is also unbiased. If all the lists are independent such that the fourth and fifth row in (2.4) drop off, (2.7) results in an estimate of the population size identical to those of Alho (1990) and Huggins (1989).

2.4.1 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 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 $\mathbf{Y}$ is given by (2.4) 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

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

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

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

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

2.4.2 Variance Estimation

In this section we will derive an estimator of both the conditional and unconditional variance of N. Conditioning on the observed sample size, we have

$$\text{var}(\hat{N}) = E\left[\text{var}(\hat{N}|n)\right] + \text{var}(E\left[\hat{N}|n\right]). \tag{2.8}$$

For ease of notation, let $V_1 = E\left[\text{var}(\hat{N}|n)\right]$ and $V_2 = \text{var}\left(E\left[\hat{N}|n\right]\right)$. V_1 is the conditional variance estimator, which 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

$$\frac{\partial \hat{N}_i}{\partial \lambda_{st}} = x_{is} \left[\frac{\hat{\pi}_{100|i} \hat{\pi}_{010|i} \hat{\pi}_{001|i} \hat{\pi}_{111|i}}{\hat{\pi}_{110|i} \hat{\pi}_{101|i} \hat{\pi}_{011|i}} \right] \left(- \sum_{\substack{a,b,c \\ (a,b,c) \neq (0,0,0)}} y_{t(abc)} \ \hat{\pi}_{abc|i} \right), \tag{2.9}$$

and the second derivative of Bock's model with respect to lambda given in equation (2.3). Using this, the conditional variance estimator is given by

$$V_1 = \sum_{i=1}^{n} \left[\left(\frac{\partial \hat{N}_i}{\partial \lambda_{st}} \right)^T \left(\frac{\partial^2 \ell}{\partial \lambda_{st} \partial \lambda_{uv}} \right)^{-1} \left(\frac{\partial \hat{N}_i}{\partial \lambda_{uv}} \right) \right]. \tag{2.10}$$

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

$$V_2 = \sum_{i=1}^{n} \frac{\Pi_{0|i}}{(1 - \Pi_{0|i})^2}$$

can be used.

From the equations it is evident that when the inclusion probabilities are large (that is $\Pi_{0|i}$ is small for all i) V_2 will be small, and when the estimated parameters in the model are unreliable (have large standard errors) V_1 will be large.

The problems with asymptotic confidence intervals has been discussed at length in literature. The International Working Group for Disease Monitoring and Forecasting (1995a, p. 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 (see Chao, 1987).

2.5 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 objective of this section is to examine the impact of residual dependence between lists, by comparing the method assuming independence between lists to the method we have just sketched. 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 in 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 $c_{12}(\gamma)$, 100 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 $[1, 2, 3]_x$, and the method assuming list 1 and list 2 are correlated, denoted by $[12, 3]_x$. The coverage is the number of times

the 95% asymptotic confidence interval includes the true value of the population of 500.

Table 2.2: Estimates of population size for varying degrees of dependence between the first and second list

$c_{12}(\gamma)$	Mean Pearson Correlation $r_{12}(\gamma)$	Model AIC	Mean estimate	Min	Max	Mean standard deviation	Coverage
$[12,3]_x$							
0.0	0.00	1631	499.4	481.4	523.8	7.64	95
0.1	0.06	1619	500.9	476.5	520.9	8.05	96
0.2	0.12	1604	500.5	481.8	516.8	8.32	94
0.3	0.18	1580	500.5	482.7	517.8	8.63	95
0.4	0.24	1561	500.5	479.7	522.3	8.91	92
0.5	0.30	1522	498.8	480.1	521.8	9.23	94
1.0	0.61	1282	500.0	473.6	534.8	10.88	93
$[1,2,3]_x$							
0.0	0.00	1629	499.3	481.1	517.9	7.15	93
0.1	0.06	1619	497.1	472.4	512.3	6.99	93
0.2	0.12	1609	493.8	478.2	508.9	6.81	74
0.3	0.18	1594	490.5	475.7	505.4	6.58	64
0.4	0.24	1588	487.4	468.6	504.4	6.38	47
0.5	0.30	1568	482.5	464.1	498.7	6.15	20
1.0	0.61	1516	468.8	451.6	488.4	5.34	3

Table 2.2 shows that when the correlation between the lists is low ($c_{12}(\gamma) = 0.0$, or $c_{12}(\gamma) = 0.1$), both models 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, whilst $[1,2,3]_x$ 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.

This simulation shows that in models with continuous covariates ignoring interaction between sources may severely bias the point estimates as well as the confidence interval.

2.6 Application to Neural Tube Defects Data

The purpose of this study is to estimate the number of children delivered with a neural tube defect (NTD). In table 2.3, 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 shown using simulations (for capture-recapture problems without covariates) that it tends to pick the data generating model more frequently (Stanley and Burnham, 1998, p. 492).

From these models it is clear that there is dependence between LVR_1 and LNR (see Table 2.3). The AIC's show that models incorporating delivery weight

Table 2.3: Estimates of population size for the covariates models

Model	Design Matrix †	Covariate Matrix	AIC	Estimated Population	$\hat{V}_1$	$\hat{V}_2$	95 % C.I. ‡
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]

†1 is LVR_1, 2 is LVR_2, and 3 is LNR

‡Confidence intervals computed on the log scale, see Chao (1987, p. 787). Using asymptotic confidence intervals results in some lower endpoints being less than the observed sample.

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 (see Hoeting et al., 1999, p. 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 AIC's 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, p. 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 2.4 presents the parameters for the "best" models, that is models M_{5b} and M_{7b}. Rather than interpreting the coefficients in Table 2.4, we will interpret

Table 2.4: Parameters for the models with the lowest AIC's

$\mathbf{\Lambda}$ 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

the equations for the logits of the inclusion probabilities, which are derived from these coefficients (see Section 2.4.1. For model M_{5b} the logits of the inclusion probabilities are

$$\text{logit } (LVR_1) = \begin{cases} -2.83 + 1.07D, & \text{if not in } LVR_2 \text{ and } LNR; \\ -0.15 - 0.11D, & \text{if in } LVR_2 \text{ but not in } LNR; \\ -4.59 + 1.24D, & \text{if not in } LVR_2 \text{ but in } LNR; \\ -1.91 + 0.07D, & \text{if in } LVR_2 \text{ and } LNR; \end{cases} \tag{2.11a}$$

$$\text{logit } (LVR_2) = \begin{cases} 0.11 - 0.02D, & \text{if not in } LVR_1; \\ 2.78 - 1.19D, & \text{if in } LVR_1; \end{cases} \tag{2.11b}$$

$$\text{logit } (LNR) = \begin{cases} -1.76 + 0.39D, & \text{if not in } LVR_1; \\ -3.52 + 0.56D, & \text{if in } LVR_1. \end{cases} \tag{2.11c}$$

From these equations, the probability of being ascertained by LVR_1 when neither listed in LVR_2 nor LNR increases with increasing delivery weight (estimate is 1.07). This also holds for inclusion to LVR_1 when listed in LNR but not in LVR_2 (estimate is 1.24). On the other hand, the probability of being ascertained by LVR_2 decreases with increasing delivery weight when listed in LVR_1 (estimate is 1.19). 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.

The logits of the inclusion probabilities for model M_{7b} show that the probability of being listed in LVR_2 tends to decrease with increasing delivery weight. This shows that the more abnormal the baby, in terms of the delivery weight, the more likely the child will be listed by the obstetrician registry (LVR_2). The probability of being ascertained by LVR_2 tends to increase with increasing delivery weight when listed by LNR, and vice versa. This might be due to the close proximity of obstetric and paediatric departments.

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

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

A summary of the resulting data is shown in Table 2.5. The log-linear model with the minimum AIC includes the interactions $LVR_1 : LVR_2$ and $LVR_1 : LNR$. The AIC for this model is 24.9 and 183 is the estimated number of deliveries with

Table 2.5: Overlap information for delivered children

Delivery	Ascertainment history †							
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's	4	17	4	14	0	4	1	44
≥ 2.5 kg's	39	20	12	10	7	13	3	104

†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).

an NTD. Although there is little difference between this model and model M_{5b} in table 2.3, some of the models with stratification, for example the saturated model, had large estimates of the population size. The quasi-symmetric model fitted poorly with an AIC of 34.0 and an estimate of 190 deliveries. Refinements of the Rasch model (see Darroch et al., 1993, p. 1143) did not fit better than the log-linear model with the lowest AIC. This implies that the lists are not of the same kind (see International Working Group for Disease Monitoring and Forecasting, 1995a, p. 1053).

The Rasch model can also be defined by assuming that the subjects are homogeneous within a set of latent classes (see Bartolucci and Forcina, 2001, p. 715). This model is equivalent to the quasi-symmetric latent class model (see Agresti, 1994, p. 498). A model of this form with 2 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. Due to lack of degrees of freedom we did not consider the ordinary latent class model.

2.7 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. 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 incorporation covariates (see Stanley and Burnham, 1998, p. 402).

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 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 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, p. 86). Our approach is different from the approaches of Darroch et al. (1993), Agresti (1994) and Bartolucci and Forcina (2001) among others, where both observed heterogeneity (due to only categorical covariates) and unobserved heterogeneity are taken into account. It would be interesting to develop models where unobserved and observed heterogeneity due to continuous covariates are taken into account. The approach just presented can be a good starting point, see section 2.4.

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*. Wiley, New York, USA.

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*, 46:623–635.

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

Bishop, Y., Fienberg, S., and Holland, P. (1975). *Discrete Multivariate Analysis: Theory and Practice*. McGraw-Hill, New York, USA.

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

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

Chao, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, 43:783–791.

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*, 88:1137–1148.

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*. Academic Press, New York.

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*, 152:771–778.

Horvitz, D. and Thompson, D. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47:663–685.

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

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*, 142:1047–1058.

International Working Group for Disease Monitoring and Forecasting (1995b). Capture-recapture and multiple record systems estimation 2: applications. *American Journal of Epidemiology*, 142:1059–1068.

Oman, J. and Zucker, D. (2001). Modelling and generating correlated binary variables. *Biometrika*, 88:287–290.

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*, 14:427–456.

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

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. *Eur. J. Obset. Gynecol. Reprod. 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*, 65:121–125.

Chapter 3

Semiparametric models for capture-recapture studies with covariates

Published as Zwane & van der Heijden (in press)
Computational Statistics & Data Analysis
©Elsevier. *Reprinted with permission*

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.

3.1 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 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, p. 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 (NTD's) 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 non-linear 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 3.2.

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). VGAM's are an extension of generalized additive models (se 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 non-parametric functions for *specific* covariates. Furthermore, any dependence between the registrations after controlling for the covariates is modelled. Kim and Cohen (ress) and Peng (2003) present similar approaches for matched case-control studies and cure models respectively.

The paper is structured as follows. In Section 3.2 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 3.3. In Section 3.4 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 3.5. 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 3.6 we apply the method to the dataset presented in Section 3.2. We conclude with a discussion in Section 3.7.

3.2 Data

In the Netherlands data on NTD's 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 mid-wife 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 3.1 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

Table 3.1: Overlap information in data set of delivered children: 1995

	Ascertainment profile[a]							
	[1,0,0]	[0,1,0]	[0,0,1]	[1,1,0]	[1,0,1]	[0,1,1]	[1,1,1]	Total
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).

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 NTD's. Table 3.1 also shows that deliveries listed in LNR tend to have normal delivery weight whilst cases with low delivery weight (DW less than 2.5 kilograms) 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, likely due to less referrals of these cases to obstetric departments.

A common feature of most epidemiological registrations is that of missing values. This data was no different as they were some missing values in pregnancy duration and delivery weight which had to be imputed before selecting the above data set. There were 2 cases with missing values on both variables, only 2 missings

for "pregnancy duration only" and 5 for "birth weight only" out of 202 cases of which 158 were cases with pregnancy duration from 24 weeks (see Table 3.1). 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).

3.3 Triple records system estimation without covariates

The notation for three lists can be specified as shown in Table 3.2. Without covari-

Table 3.2: Three list problem without covariates

	List 3			
	Not Included		Included	
	List 2			
List 1	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}

ates 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,

$$\log(m_{abc}) = \mathbf{M} \times \boldsymbol{\alpha}; \tag{3.1a}$$

$$\log \begin{bmatrix} m_{100} \\ m_{010} \\ m_{001} \\ m_{110} \\ m_{101} \\ m_{011} \\ m_{111} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \times \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix}; \tag{3.1b}$$

where the m_{abc} denote the expected frequencies for the cell probabilities. Columns 2 to 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 the all list interaction.

Alternatively, one can use the fact that Table 3.2 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,

$$\hat{m}_{000} = \frac{\hat{m}_{100}\hat{m}_{010}\hat{m}_{001}\hat{m}_{111}}{\hat{m}_{110}\hat{m}_{101}\hat{m}_{011}} = n \times \frac{\hat{\pi}_{100}\hat{\pi}_{010}\hat{\pi}_{001}\hat{\pi}_{111}}{\hat{\pi}_{110}\hat{\pi}_{101}\hat{\pi}_{011}}, \tag{3.2}$$

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 (3.2) simply implies that there is no three factor interaction, because if it was present the cross product ratio's for the subtables would be different.

3.4 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.

3.4.1 Issues of notation

In the capture-recapture problem the cell probabilities and cell counts are usually denote 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 3.3). 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).

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 3.3. Note that instead of using the inclusion profile as an index we are now using $k = 1, 2, \cdots, 7$ as an index, but there is a direct relation between the two (compare Table 3.2 and Table 3.3). 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 3.3 and zero otherwise.

Table 3.3: Three list problem with covariates

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

3.4.2 Multinomial logit model

Assume that an individual indexed by i ($i = 1, 2, ..., n$) is classified into one of K nominal categories, indexed by k ($k = 1, 2, .., K$), such that $n_{ik} = 1$ 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 ($h = 0, 1, 2, ..., H$) with the first element being 1. Denoting the multinomial logit for individual i as $\eta_i' = [\eta_1(\boldsymbol{x}_i), \eta_2(\boldsymbol{x}_i), ..., \eta_K(\boldsymbol{x}_i)]$, the category probabilities are then given by,

$$\pi_{k|i} = \exp[\eta_k(\boldsymbol{x}_i)] / \sum_{r=1}^{K} \exp[\eta_r(\boldsymbol{x}_i)], \tag{3.3}$$

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 3 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 6 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,

$$\mathbf{C} = \mathcal{M} \otimes \mathbf{X}; \tag{3.4}$$

where $\mathcal{M}$ is the design matrix $\mathbf{M}$ (which denotes dependencies between the lists, see Section 3.3) without the columns of ones (or intercept) and $\otimes$ denotes the Kronecker product. If we let $j = 1, 2, \cdots, J$ index the columns of $\mathcal{M}$, the dimension of $\mathbf{C}$ is $[n.K] \times [J.H]$. Note that $S \leq J \leq K$. To implement the approach of Alho (1990) and Huggins (1989), $\mathcal{M}$ will be given by the design matrix in equation (3.1b) without the first column. Dependencies in lists can also be coded into $\mathcal{M}$.

3.4.3 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

$$\eta_k(\boldsymbol{x}_i) = \beta_{j(0)} + \sum_{j=1}^{J} \sum_{h=1}^{H} f_{j(h)}(\boldsymbol{x}_i), \tag{3.5}$$

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 equation (3.5) 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 for these covariates be denoted by $\boldsymbol{\beta}$. To estimate the unknowns $\boldsymbol{f}$ and $\boldsymbol{\beta}$ it is common to use the penalized log-likelihood,

$$pl(\boldsymbol{f}, \boldsymbol{\beta}) = \sum_{i=1}^{n} \sum_{k=1}^{K} n_{k|i} \log [\pi_{k|i}] + \frac{1}{2} \sum_{j=1}^{J} \sum_{h=1}^{H} \lambda_{j(h)} \int [f''_{j(h)}(x)]^2 dx. \tag{3.6}$$

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.

3.4.4 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,

$$\hat{m}_{000|i} = \frac{\hat{\pi}_{100|i}\hat{\pi}_{010|i}\hat{\pi}_{001|i}\hat{\pi}_{111|i}}{\hat{\pi}_{110|i}\hat{\pi}_{101|i}\hat{\pi}_{011|i}}, \tag{3.7}$$

and finally use,

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

to obtain an estimate of total population. Equation (3.7) is the same as (3.2) except for the fact that (3.7) is stratified by individual. Zwane and Van der Heijden (2002) explicitly showed that in the two list case using equation (3.7) 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.

3.5 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 doesn't 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

$$\Pi_{1|i} = \frac{\pi_{110|i}}{\pi_{010|i}+\pi_{110|i}}; \frac{\pi_{101|i}}{\pi_{001|i}+\pi_{101|i}}; \frac{\pi_{111|i}}{\pi_{011|i}+\pi_{111|i}}; \tag{3.8a}$$

$$\Pi_{2|i} = \frac{\pi_{110|i}}{\pi_{100|i}+\pi_{110|i}}; \frac{\pi_{011|i}}{\pi_{001|i}+\pi_{011|i}}; \frac{\pi_{111|i}}{\pi_{101|i}+\pi_{111|i}}; \tag{3.8b}$$

$$\Pi_{3|i} = \frac{\pi_{011|i}}{\pi_{010|i}+\pi_{011|i}}; \frac{\pi_{101|i}}{\pi_{100|i}+\pi_{101|i}}; \frac{\pi_{111|i}}{\pi_{110|i}+\pi_{111|i}}. \tag{3.8c}$$

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 (3.8a)-(3.8c) 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 and *edf* = 4 (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 Figure 3.1. In Figure 3.1, 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 clearly that some panels in Figure 3.1 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

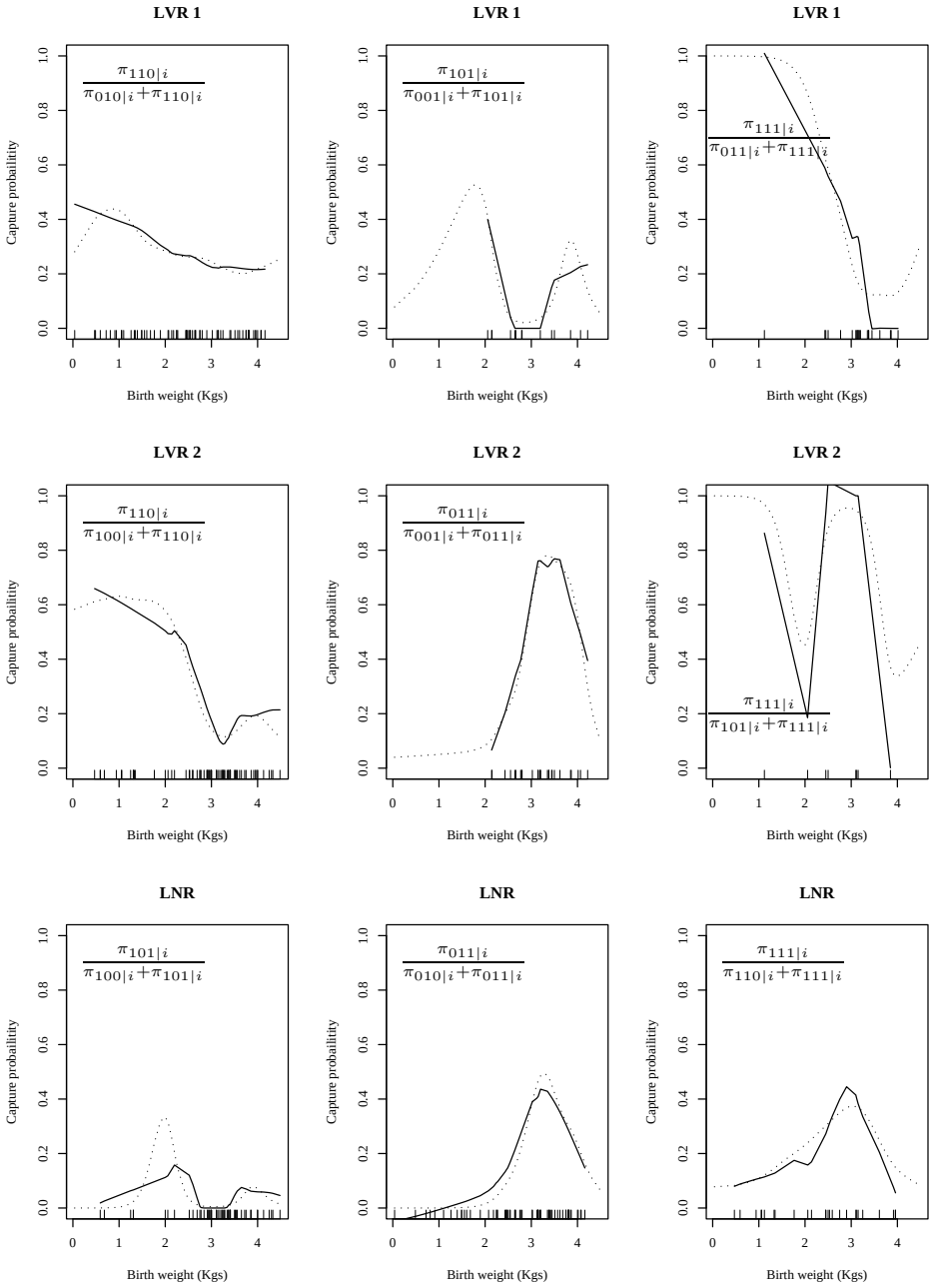

Figure 3.1: Inclusion probabilities for LOWESS fit (solid line) and most complex model (dotted line).

models excluding the covariate, for example, log-linear models, will fit the data poorly. Another observation is that 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.

3.6 Application

In this section we will apply the method to the data presented in Section 3.2. 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 3.4. We then fitted parametric models incorporating delivery weight as a covariates, and afterwards the semiparametric models and the results are also shown in Table 3.4. 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

Table 3.4: 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]

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 3.5 is shown in Figure 3.2. Figure 3.2 shows that the curves for models M_4 and M_6 are similar,

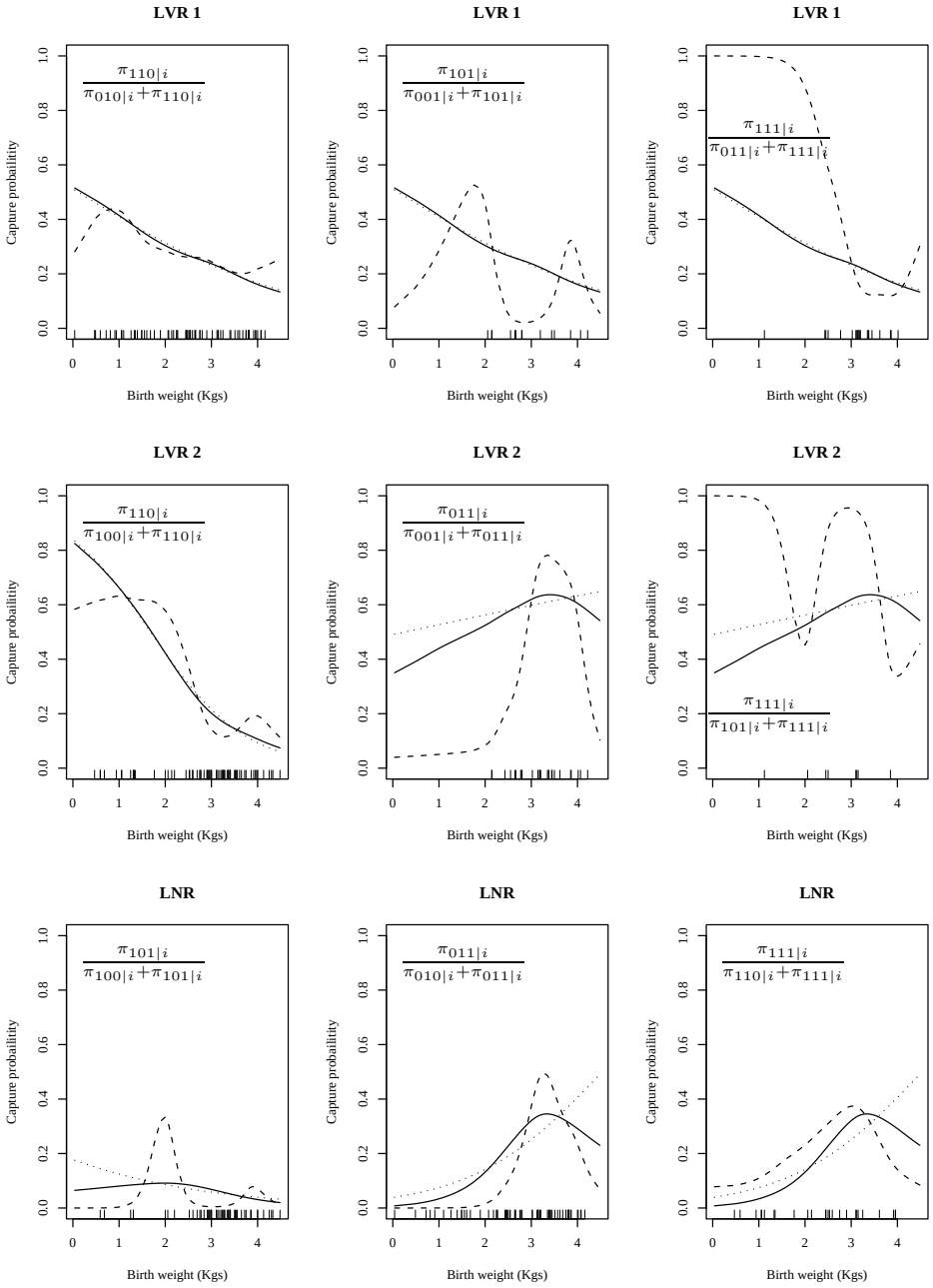

Figure 3.2: Inclusion probabilities for different models, M_4 (dotted line), M_6 (solid line), and most complex model (dashed line).

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 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.

Figure 3.2 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 percent irregardless of whether the child is listed in LVR_1. Another important observation from Figure 3.2 is that the inclusion probability to LNR if listed in LVR_1 is very low, around 5 percent, but inclusion to LNR if listed in LVR_2 increases with increasing delivery weight due to less deaths to 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 AIC's and estimates of the population size in Table 3.4 and the plots in Figure 3.2 it is clear that there is no 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 190 and unlikely to have been more than 400 children.

3.7 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 non-parametric 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.

References

Aikake, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petran B.N. and Csaki F., editor, *International symposium on information theory*, pages 261–281, Budapest, Hungary. Akadmiai Kiadi.

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

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

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

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

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

Chen, S. and Lloyd, C. (2002). Estimation of population size from biased samples using non-parametric binary regression. *Statistica Sinica*, 12:505–518.

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*, 88:1137–1148.

Green, P. and 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. and Tibshirani, R. (1990). *Generalized additive models*. Chapman and Hall, New York.

Hosmer, D. and 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. *American Journal of Epidemiology*, 142:1047–1058.

Kim, I. and Cohen, N. (in press). Semiparametric and nonparametric modeling for effect modification in matched studies. *Computational Statistics & Data Analysis*.

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

Otis, D., Burnham, K., White, G., and Anderson, D. (1978). *Statistical Inference from Capture Data on Closed Animal Populations*. Wildlife Monographs, No. 62.

Peng, Y. (2003). Fitting semiparametric cure models. *Computational Statistics & Data Analysis*, 41:481–490.

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

SPSS (1997). *SPSS Missing Value Analysis 7.5*. SPSS Inc., Chicago.

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

Tilling, K. and Sterne, J. (1999). Capture-recapture models including covariate effects. *American Journal of Epidemiology*, 149:392–400.

Tilling, K., Sterne, J., and Wolfe, C. (2001). Estimation of incidence of stroke using a capture-recapture model including covariates. *International Journal of Epidemiology*, 30:1351–1359.

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. *Eur. J. Obset. Gynecol. Reprod. Biology*, 108:33–39.

White, G. (2002). Discussion comments on: the use of auxilliary covariates in capture-recapture modelling: An overview. *Journal of Applied Statistics*, 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. and Wild, C. (1996). Vector generalized additive models. *J.R. Statist. Soc.*, B, 58:481–493.

Zwane, E. and Van der Heijden, P. (2002). The multiple-system estimator in the presence of covariates. In M. Stasinopoulos & G. Touloumi, editor, 17^{th} *International Workshop on Statistical Modelling*, pages 697–701, Chania, Greece. Chania University.

Zwane, E. and Van der Heijden, P. (2003). Implementing the parametric bootstrap in capture–recapture models with continuous covariates. *Statistics and Probability Letters*, 65:121–125.

Chapter 4

Implementing the parametric bootstrap in capture-recapture models with continuous covariates

Published as Zwane & van der Heijden (2003)
Statistics & Probability Letters, **65**, 121–125.
©Elsevier. *Reprinted with permission*

The parametric bootstrap is a method for variance estimation advocated by many researchers in multiple capture studies. Most applications thus far used the parametric bootstrap in log-linear modelling, that is, where there are possibly categorical covariates which relate to the probabilities of capture. In this article we present an algorithm for the parametric bootstrap that can be used when there are continuous covariates.

4.1 Introduction

The inappropriateness of symmetric confidence intervals in capture-recapture studies has been discussed by several authors, for example, the International Working Group for Disease Monitoring and Forecasting (1995) noted that, for virtually all models proposed in capture-recapture literature the distribution of the estimate of the population size is skewed. Proposals have been made to solve this problem, the first being to use a suitable transformation of the population size, mainly the logarithm, which would make the distribution look more like a normal random variable (see Chao, 1989). Another alternative is the profile likelihood confidence interval (see Evans et al., 1996). This approach is more appealing because it is *moderately* computer intensive (and in some instances can be performed by hand), but it is not readily applicable when there are continuous covariates. Finally, an approach which is now widely recognized as an adequate variance estimation method is the parametric bootstrap (Buckland and Garthwaite, 1991). In their paper, Buckland and Garthwaite (1991) presented both the non-parametric and parametric bootstrap for the capture-recapture problem without covariates.

In this paper we discuss ways of implementing the parametric bootstrap in the presence of (continuous) covariates. Thus far, only the asymptotic variance estimator does not condition on the observed sample size in the presence of continuous covariates (see Alho, 1990; Zwane and Van der Heijden, 2002). Several authors have used the non-parametric bootstrap in the presence of continuous covariates, for example Huggins (1989), Tilling and Sterne (1999) and, Tilling et al. (2001). But as noted by Norris and Pollock (1996), the non-parametric bootstrap results in a variance estimate which is likely to be smaller than the true variance, because it conditions on being observed. This is in-line with a simulation conducted by Tilling and Sterne (1999), which showed that the non-parametric bootstrap has a coverage consistently lower than the nominal coverage.

The plan of the paper is as follows. In Section 4.2, we discuss two bootstrap methods which can be used to estimate variances in the capture-recapture problem, but we concentrate on the parametric bootstrap. Section 4.3 presents a simulation to evaluate the properties of these methods.

4.2 Variance estimation methods

Assume that the true population size is N and the individuals are indexed by i ($i = 1, 2, ..., N$) of which n are observed, and further assume that the individuals act independently. For ease of exposition, assume that we have two registrations (or lists), but the results can be easily extended to more than two lists, and for individual i there is a covariate vector $\boldsymbol{x}_i$. The inclusion profile for individual i (denoted by $\boldsymbol{w}_i$) is (1,0) if individual i is observed in list 1 only, (0,1) if observed in list 2 only, and (1,1) if observed in both lists. Individuals with an inclusion profile of (0,0) are unobserved and have to be estimated.

Let the probability that an individual is captured (registered or listed) at least once be denoted by ϕ_i (the estimated probability is denoted by $\hat{\phi}_i$). This probability is not necessarily the same for all individuals, as it is dependent on the individual level covariates. Using this quantity, the estimate of the population size is,

$$\hat{N} = \sum_{i=1}^{n} \hat{N}_i = \sum_{i=1}^{n} \frac{1}{\hat{\phi}_i},$$

where $\hat{N}_i$ is the contribution of individual i to the estimate of the population size (see Huggins, 1989). This quantity is useful in the parametric bootstrap variance estimator.

Below we present two bootstrap methods, which are analogous to those presented by Norris and Pollock (1996) and our presentation will basically follow the same scheme. The only difference of the methods presented here, is that we allow for individual level covariates (i.e. continuous covariates) whilst the methods given by Norris and Pollock (1996) do not. For ease of exposition we do not discuss the notion of model uncertainty, but it can be easily incorporated in the approaches presented.

4.2.1 Nonparametric bootstrap

This method samples with replacement from the ascertainment histories, that is, for each bootstrap sample n individuals are drawn with replacement and the estimate of the population size calculated. This method is equivalent to method 1 in Norris and Pollock (1996). Tilling et al. (2001) also give a discussion of how this method can be implemented in the capture-recapture setting with covariates. As has been discussed above, this bootstrap procedure results in a variance estimate which is usually smaller than the true variance. Therefore the coverage will usually be too low.

4.2.2 Parametric bootstrap

Assuming a good estimate of the probability model exists, then an estimate of the unconditional variance can be computed based on the fitted inclusion proba-

bilities. Unlike the non-parametric bootstrap, this approach provides a non-zero probability of being missed (see Norris and Pollock, 1996, p.238).

The method is equivalent to method 3 in Norris and Pollock (1996). The only difference is that the fitted inclusion probabilities are different for each individual (dependent on continuous covariates). For each bootstrap replication draw one or more ascertainment histories for each individual based on a multinomial probability model, using an estimate of the individual's contribution to the estimate of population size. In most cases $\hat{N}_i$ is not an integer, and as noted by Buckland and Garthwaite (1991, p.258), it is simplest to round off to the nearest integer. This approach, though suited for the capture-recapture problem without covariates (especially when $\hat{N}$ is large), it is not suited for the problem incorporating continuous covariates. Rounding individual values to the nearest integer might lead to an overestimation or underestimation of the true value of population size dependent on the proportion of $\hat{N}_i$'s rounded up or down.

To go around the above problem we propose to first randomly determine an integer estimate of each individual's contribution to the total population. We assume that the true N_i is either INT$[\hat{N}_i]$ or INT$[\hat{N}_i + 1]$ (where INT$[\hat{N}_i]$ is the integer part of $\hat{N}_i$). We give a higher probability to the integer close to $\hat{N}_i$ in the following way: The probability that the true N_i is INT$[\hat{N}_i]$ is $1 - d_i$ and the probability is d_i if the true N_i is INT$[\hat{N}_i + 1]$, where $d_i = \hat{N}_i - \text{INT}[\hat{N}_i]$. This approach results in each bootstrap sample (including individuals missed by all lists) to be close to the estimate of population size. In each bootstrap sample (after excluding individuals missed by all lists) the estimate of the population size is calculated.

4.3 Simulation

To better understand the properties of the methods described above, a simulation study was undertaken. The two methods were compared based on the percentage of times that the "unknown" population size fell within the simulated confidence intervals. Several two list capture-recapture experiments with a single standard normal covariate x were generated for different population sizes (50, 100, 250, 500). The probabilities of being ascertained by list 1 (Π_1) and list 2 (Π_2) were generated using $\text{logit}(\Pi_1) = A + 0.5x_i$ and $\text{logit}(\Pi_2) = 0.5 + x_i$. The values of A were (-1, -0.5, 0, 0.5, 1). For each population size and A, 200 data-sets were simulated, and for each data-set the estimate of population size was computed. We then used the bootstrap methods presented with 1000 replications to derive 95 percent quantile confidence intervals for each data-set. Using these intervals we ascertained the coverage proportions for each of the methods, and they are tabulated in Table 4.1. The table also shows several parameters of the estimate of the population size over the 200 samples.

The table shows that the coverage of the parametric bootstrap is approximately equal to the nominal coverage, whilst this is not true for the non-parametric

Table 4.1: Coverage of the bootstrap methods

						Coverage	
N	A	Mean	Median	Min	Max	Non Parametric	Parametric
50							
	-1.0	94.8	48.8	34.2	2224.1	89.0	96.0
	-0.5	57.5	51.2	30.2	212.1	91.5	95.0
	0.0	54.4	50.7	32.9	116.4	93.0	98.5
	0.5	54.3	51.0	39.1	123.6	90.0	95.5
	1.0	56.3	50.4	40.6	517.5	88.0	94.5
100							
	-1.0	108.2	99.4	73.4	458.1	91.5	95.0
	-0.5	108.2	101.9	78.0	250.8	90.0	95.0
	0.0	104.3	99.4	71.6	172.9	88.5	94.5
	0.5	105.7	101.4	80.1	250.3	91.0	96.5
	1.0	102.0	99.9	86.2	151.8	90.5	98.5
250							
	-1.0	258.0	250.5	194.2	500.4	91.5	93.5
	-0.5	253.8	249.8	207.5	361.6	92.5	95.0
	0.0	251.9	248.8	211.1	316.0	93.0	96.5
	0.5	252.4	250.1	221.4	299.7	85.5	96.0
	1.0	251.4	250.8	224.5	288.7	92.0	98.5
500							
	-1.0	505.8	505.8	398.4	631.7	91.5	96.5
	-0.5	506.2	504.7	437.4	618.9	91.5	95.5
	0.0	506.6	502.5	434.3	637.6	93.0	96.0
	0.5	497.9	496.0	441.9	568.2	91.0	96.0
	1.0	502.4	499.2	469.3	565.4	88.0	97.0

bootstrap. A comparison of the means and the medians in Table 4.1, shows that the distribution of the estimate of the population size is skewed, but as the two methods presented here do not assume asymptotic normality they are unaffected. We conclude that in the context of continuous covariates the parametric bootstrap works well.

References

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

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

Chao, A. (1989). Estimating population size from sparse data in capture-recapture experiments. *Biometrics*, 45:427–438.

Evans, M. A., Kim, H. M., and O'Brien, T. E. (1996). An application of profile-likelihood confidence interval to capture-recapture estimators. *Journal of Agricultural, Biological and Environmental Statistics*, 1:131–140.

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. *American Journal of Epidemiology*, 142:1047–1058.

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

Tilling, K. and Sterne, J. (1999). Capture-recapture models including covariate effects. *American Journal of Epidemiology*, 149:392–400.

Tilling, K., Sterne, J., and Wolfe, C. (2001). Estimation of incidence of stroke using a capture-recapture model including covariates. *International Journal of Epidemiology*, 30:1351–1359.

Zwane, E. and Van der Heijden, P. (2002). The multiple-system estimator in the presence of covariates. In M. Stasinopoulos & G. Touloumi, editor, 17^{th} *International Workshop on Statistical Modelling*, pages 697–701, Chania, Greece. Chania University.

Chapter 5

The multiple-record systems estimator when registrations refer to different but overlapping populations

Published as Zwane, van der Pal-de Bruin & van der Heijden (2004)
Statistics in Medicine, **23**, 2267–2281.
©Wiley International. *Reprinted with permission*

In multiple-record systems estimation it is usually assumed that all registrations relate to the same population. In this paper we develop a method which can be used when the registrations relate to different populations, in the sense that they cover, for example, different time periods or regions. We show that under certain conditions ignoring that the registrations relate to different populations results in correct estimates of population size. The EM algorithm is presented as a method that can be used for more general problems. The parametric bootstrap is used to construct a confidence interval. The proposed method is then applied to a data set with five registrations of neural tube defects, that cover different time periods.

5.1 Introduction

Capture-recapture analysis was developed by ecologists for assessing the size of animal populations in the wild (Seber, 1982). The population size is estimated from the degree of overlap between two or more samples obtained from the same population. In epidemiology, capture-recapture methods are used to estimate or adjust for the extent of incomplete ascertainment using information from overlapping cases from distinct sources (see International Working Group for Disease Monitoring and Forecasting, 1995a,b). The common labels for the methods in human populations are, *multiple-system*, *multiple-records systems*, and *multiple-record systems method* (see International Working Group for Disease Monitoring and Forecasting, 1995a).

For two overlapping samples (from the same population) the method is used to estimate the part of the population that is not observed (individuals in neither of the two samples). This estimation is accomplished under the assumption of independence of inclusion probabilities (International Working Group for Disease Monitoring and Forecasting, 1995a,b). Another assumption is homogeneity of inclusion probabilities over individuals. Although it has been long thought that the inclusion probabilities for both lists should be homogeneous, it has recently been shown that only one of the lists has to have homogeneous inclusion probabilities when the *joint* capture probability is positive (Alho, 1990; Chao et al., 2001).

Categorical covariates are frequently used to diminish heterogeneity of inclusion probabilities, and for this log-linear models are widely used as it is possible to incorporate stratification variables, and permit dependence between sources when there are more than two sources (Bishop et al., 1975; Fienberg, 1972; Cormack, 1989).

In this paper we deal with a special case of heterogeneous inclusion probabilities, namely the case where the populations from where the lists emanate partially overlap. This results in some individuals being systematically missed by one or more of the lists. Therefore the *joint* inclusion probabilities are zero for some individuals in the combined population. A first example is that lists do not cover the same region. For this example, a stratum is defined as a subregion, and not every list is observed in each subregion. A second example is when lists do not

cover the same time periods. Here the strata are defined as the subperiods of time. We approach the absence of observations in certain strata for certain lists as an incomplete data problem.

The EM algorithm is an iterative procedure for obtaining maximum likelihood estimates in incomplete data (Little and Rubin, 1987). In the standard capture-recapture problem the EM algorithm can be used to estimate part of the population missed by all sources (see Baker, 1990). As we have partially overlapping populations there are more entries missing in the contingency table than in the standard capture-recapture problem. When some lists are not operating in some strata, implying that there are several unobservable cells that are a result from non-operating lists, the EM algorithm can *still* be used to estimate these missing entries, and thus the population size.

Section 5.2 introduces the data set that will be used to illustrate the results. In Section 5.3 we present two simple capture-recapture models; the first with one list operating in a subperiod of another list, and the other where the two lists operate in different but partially overlapping years (time periods). We show under which conditions stratification by year can be ignored. We then present the EM algorithm in Section 5.4, and show how it can be used to estimate population size in partially overlapping populations, and further show that for simple models the results will be equivalent with using traditional methods. In Section 5.5 we analyze the data set on neural tube defects which motivated this article. Finally, Section 5.6 gives some conclusions.

5.2 Data

In this section we will introduce the data set on neural tube defects (NTD's) in the Netherlands that will be used to illustrate the procedure presented in the paper. In the Netherlands cases with NTD's are registered in several national databases. Furthermore the Dutch Association of Patients with a NTD also conducts its own surveys (see Van der Pal et al., 2003). In this analysis we will use five registrations which we describe briefly.

1. **Dutch Perinatal Database I (R1)**: This is an anonymous pregnancy and birth registry of low risk pregnancies and births, even if care only relates to a part pregnancy or delivery. Data over the period 1988 through 1998 are used.

2. **Dutch Perinatal Database II (R2)**: This list registers anonymous data concerning the birth of a child in secondary care. Data over the period 1988 through 1998 are used.

3. **National Neonate Database (R3)**: This list contains anonymous information about all admissions and re-admissions of newborns to paediatric departments within the first 28 days of life. Data was used for the period 1992-1998.

4. **Dutch Monitoring System of Child Health Care (R4)**: R4 registers live born infants with a NTD who visit a paediatrician for the first time. All paediatric departments participate. NTD's are registered since 1993.

5. **Dutch Association of Patients with a NTD (R5)**: A short questionnaire was sent to every member of R5 with a NTD affected child between 1988 and 1998.

Children were linked on date of birth, zip code, mother's date of birth and gender of child (Van der Pal et al., 2003). It should be noted that abortions are possible in R1 and R2, whereas they cannot appear in the other registrations. Therefore we consider only children with a pregnancy duration from 24 weeks (the legal limit for pregnancy termination in the Netherlands) .

None of these databases include all cases of neural tube defects because of, for instance, non-participation of health care professionals. Thus multiple-record systems estimation has to be used to estimate the size of babies born with NTD's. The usual approach is to fit log-linear models with a structural zero for the observations missed by all lists (International Working Group for Disease Monitoring and Forecasting, 1995b). In our situation this usual approach cannot be adopted as some of the registrations are not available for some years: for 1988-1991 only three registrations are available (R1,R2,R5) and in 1992 only four registrations are available (R1,R2,R3,R5). The frequencies for all years are given in Table 5.1.

In 1988-1991, observations with an inclusion profile of 01000 also include observations that could have been 01100, 01010, and 01110 had R3 and R4 been active. In 1992, observations with an inclusion profile of 01100 also include observations that could have been 01110 had R4 been active. Similar statements could be said for some other inclusion profiles.

In the next sections we show that the EM algorithm is a tool which can effectively analyze data of this form, by utilizing information on relations between registrations while stratifying by year. We start by showing for simple cases what can go wrong when one ignores the fact that the registrations do not come from the same population.

5.3 Capture-recapture methodology

In this section we discuss the problem of estimating population size in "dual record systems" when the registrations relate to different but overlapping populations, for example, the registrations may cover different but overlapping time periods. In particular we study what happens if this fact is ignored, that is, if it is assumed that both registrations refer to the same population. We show one example where the union of the two populations is estimated unbiased, and one where the resulting estimate is biased. This then serves as a motivation for a general solution discussed in Section 5.4.

Table 5.1: Numbers ascertained by inclusion profile for all years

	Ascertainment Profile [†,‡]																															
Year	00000	00001	00010	00011	00100	00101	00110	00111	01000	01001	01010	01011	01100	01101	01110	01111	10000	10001	10010	10011	10100	10101	10110	10111	11000	11001	11010	11011	11100	11101	11110	11111	Total
1988		9							101	5							4	0							24	2							145
1989		3							114	8							3	1							30	4							163
1990		7							105	5							3	3							43	4							170
1991		3							95	7							4	1							32	8							150
1992		10			15	0			80	3			12	2			9	1			0	0			27	3			7	3			172
1993		3	12	1	4	0	2	0	61	0	18	7	1	0	4	0	5	0	0	1	1	0	0	1	24	1	7	4	0	0	3	0	160
1994		3	15	1	6	0	5	0	34	1	18	1	1	0	5	0	24	0	7	3	1	1	4	1	13	1	9	0	1	1	2	4	162
1995		2	16	2	5	0	4	3	27	2	18	5	2	3	9	1	29	1	15	2	1	1	0	0	15	0	4	1	1	0	5	0	174
1996		5	9	5	10	0	5	1	26	0	11	2	1	0	4	2	26	0	6	6	1	0	0	2	11	0	9	1	1	0	5	4	153
1997		4	12	1	13	2	3	1	26	0	11	1	0	0	6	0	41	2	11	1	2	0	4	2	18	1	7	3	1	0	3	4	180
1998		1	8	0	13	0	6	0	25	0	14	1	2	0	4	1	27	0	7	0	0	0	11	0	20	0	3	2	1	0	7	1	154

[†]Denotes whether a child was included in a list (Included=1) or not (Not Included=0). 00001 implies the child was seen only in R5.

[‡]Zero frequencies stand for sampling zeros

5.3.1 Simple capture-recapture model

The simplest multiple-record systems consists of two lists. Let Π_1 and Π_2 be the probability of capture by list 1 and 2 respectively. The joint probabilities are denoted by π_{ij} $(i = 0, 1;\ j = 0, 1)$, where π_{10} is the probability to be in list 1 only, π_{01} is the probability to be in list 2 only and, π_{11} is the probability to be in both lists. The corresponding frequencies are shown in table 5.2. The probability π_{00} and frequency n_{00} are unknown and have to be estimated in order to compute an estimate of the unknown population size N. Furthermore $n_{ij} = N\pi_{ij}$.

Table 5.2: Dual-record system problem

	List 2	
List 1	Not included	Included
Not included	0 *	n_{01}
Included	n_{10}	n_{11}

*Treated as structurally zero cell

Assume that for each sample, each individual has the same inclusion probability. Then Π_1, Π_2 and N can be estimated by

$$\hat{\Pi}_1 = \frac{n_{11}}{n_{11} + n_{01}} \quad , \quad \hat{\Pi}_2 = \frac{n_{11}}{n_{11} + n_{10}} \tag{5.1}$$

$$\hat{N} = \frac{n_{11}}{\hat{\Pi}_1 \hat{\Pi}_2}$$

and this result forms the basis of our development.

5.3.2 Two lists and two strata

Assume now that we have two strata, for example, two years (or time periods). Year is a stratifying variable with two categories indexed by k, where $k = 1$ denotes the first year and $k = 2$ the second year. Let $\Pi_{1|1}$ and $\Pi_{2|1}$ be the probabilities to be in list 1 and 2 in the first year and $\Pi_{1|2}$ and $\Pi_{2|2}$ be the probabilities to be in list 1 and 2 in the second year. Let the joint probabilities for the first year be $\pi_{ij|1}$ and the joint probabilities for the second year be $\pi_{ij|2}$. Let the unknown population size for year 1 and 2 be N_1 and N_2 respectively. By analyzing the data from each year (or time period) separately we can estimate $\Pi_{1|1}$, $\Pi_{2|1}$, and N_1 by

$$\hat{\Pi}_{1|1} = \frac{n_{11|1}}{n_{11|1} + n_{01|1}} \quad , \quad \hat{\Pi}_{2|1} = \frac{n_{11|1}}{n_{11|1} + n_{10|1}} \tag{5.2}$$

$$\hat{N}_1 = \frac{n_{11|1}}{\hat{\Pi}_{1|1} \hat{\Pi}_{2|1}}$$

and $\Pi_{1|2}$, $\Pi_{2|2}$, and N_2 by

$$\hat{\Pi}_{1|2} = \frac{n_{11|2}}{n_{11|2} + n_{01|2}} \quad , \quad \hat{\Pi}_{2|2} = \frac{n_{11|2}}{n_{11|2} + n_{10|2}} \tag{5.3}$$

$$\hat{N_2} = \frac{n_{11|2}}{\hat{\Pi}_{1|2}\hat{\Pi}_{2|2}}$$

Now let list 2 be observed only in the first year such that the observed table can be set out as in Table 5.3.

Table 5.3: Dual-record system problem with 2 years

		List 2		
Year	List 1	Not included	Included	Total
1	Not included	0 *	$n_{01\|1}$	
	Included	$n_{10\|1}$	$n_{11\|1}$	
2	Not included	0 *	0 *	
	Included	?	?	$n_{10\|2} + n_{11\|2}$ †

*Structurally zero cells
†Only the margin is observed

Assume that we ignore the fact that the registrations refer to different populations, by ignoring (the variable) year. Let us denote the elements in the table where year is ignored by $n_{ij|+}$. These elements are related to the elements in table 5.3 by $n_{11|+} = n_{11|1}$, $n_{01|+} = n_{01|1}$, and $n_{10|+} = n_{10|1} + n_{10|2} + n_{11|2}$. The question is: can equation (5.1) be used to estimate N? And if so, under what assumptions? In other words, when would ignoring the fact that list 2 is observed only in one year lead to an unbiased estimate of the population size? The observations to be estimated are $n_{00|+} = n_{00|1} + n_{00|2} + n_{01|2}$. Using this we find that,

$$\hat{\Pi}_{1|+} = \frac{n_{11|1}}{n_{11|1} + n_{01|1}} \quad , \quad \hat{\Pi}_{2|+} = \frac{n_{11|1}}{n_{11|1} + n_{10|1} + n_{11|2} + n_{10|2}},$$

$$\hat{N}_+ = \left(\frac{n_{11|1}}{\hat{\Pi}_{1|1}}\right)\left(\frac{n_{11|1}+n_{10|1}+n_{11|2}+n_{10|2}}{n_{11|1}}\right) \tag{5.4a}$$

$$= \left(\frac{n_{11|1}}{\hat{\Pi}_{1|1}}\right)\left(\frac{1}{\hat{\Pi}_{2|1}}+\frac{n_{11|2}+n_{10|2}}{n_{11|1}}\right) \tag{5.4b}$$

$$= \hat{N}_1 + \left(\frac{n_{11|2}+n_{10|2}}{\hat{\Pi}_{1|1}}\right) \tag{5.4c}$$

$$= \hat{N}_1 + \left(\frac{n_{11|2}}{\hat{\Pi}_{1|1}\hat{\Pi}_{2|2}}\right) \tag{5.4d}$$

Equation (5.4) shows that if $\hat{\Pi}_{1|1} = \hat{\Pi}_{1|2}$, then $\hat{N}_+ = \hat{N}_1 + \hat{N}_2$. Thus for two lists and two years (or strata) even if the joint inclusion probability of some individuals in the combined population is zero the dual record-systems estimator can still be used, as long as $\hat{\Pi}_{1|1} = \hat{\Pi}_{1|2}$, that is, the list observed in both years (or strata) has to have homogeneous inclusion probabilities.

5.3.3 Two lists and three strata

Instead of two years we now assume that we have two lists and three years. We denote the third year by $k = 3$. Assume that list 1 operates in the years 1 and 2, and list 2 operates in years 2 and 3, such that the years where list 1 and 2 are operational partly overlap but the years for list 2 are not necessarily a subset of the years where list 1 is active. The cells actually observed are; $n_{1+|1}$ in the first year, $n_{10|2}$, $n_{01|2}$, and $n_{11|2}$ in the second year, and $n_{+1|3}$ in the third year (see Table 5.4).

Only the observations in the year 2 have non-zero joint *inclusion* probabilities. Ignoring year, the elements of the resulting table are related to those in table 5.4 by $n_{11|+} = n_{11|2}$, $n_{10|+} = n_{10|1} + n_{11|1} + n_{10|2}$, and $n_{01|+} = n_{01|2} + n_{01|3} + n_{11|3}$. The estimates of $\Pi_{1|+}$, $\Pi_{2|+}$, N_+ from this table are

$$\hat{\Pi}_{1|+} = \frac{n_{11|2}}{n_{01|2}+n_{11|2}+n_{01|3}+n_{11|3}} \quad , \quad \hat{\Pi}_{2|+} = \frac{n_{11|2}}{n_{10|1}+n_{11|1}+n_{10|2}+n_{11|2}};$$

$$\hat{N}_+ = n_{11|2}\left(\frac{n_{01|2}+n_{11|2}+n_{01|3}+n_{11|3}}{n_{11|2}}\right)\left(\frac{n_{10|1}+n_{11|1}+n_{10|2}+n_{11|2}}{n_{11|2}}\right); \tag{5.5a}$$

$$= n_{11|2}\left(\frac{1}{\hat{\Pi}_{1|2}}+\frac{n_{01|3}+n_{11|3}}{n_{11|2}}\right)\left(\frac{n_{10|1}+n_{11|1}}{n_{11|2}}+\frac{1}{\hat{\Pi}_{2|2}}\right); \tag{5.5b}$$

$$= \frac{n_{11|1}}{\hat{\Pi}_{1|2}\hat{\Pi}_{2|1}} + \hat{N}_2 + \frac{n_{11|3}}{\hat{\Pi}_{1|3}\hat{\Pi}_{2|2}} + \frac{(n_{10|1}+n_{11|1})(n_{01|3}+n_{11|3})}{n_{11|2}}. \tag{5.5c}$$

Table 5.4: Dual-record system problem with 3 years

		List 2		
Year	List 1	Not included	Included	Total
1	Not included	0 *	0 *	
	Included	?	?	$n_{10\|1} + n_{11\|1}$ †
2	Not included	0 *	$n_{01\|2}$	
	Included	$n_{10\|2}$	$n_{11\|2}$	
3	Not included	0 *	?	
	Included	0 *	?	
	Total		$n_{01\|3} + n_{11\|3}$ †	

*Structurally zero cells
†Only the margin is observed

This shows that, even if $\hat{\Pi}_{1|1} = \hat{\Pi}_{1|2}$ and $\hat{\Pi}_{2|2} = \hat{\Pi}_{2|3}$, collapsing the table over years results in a positively biased estimate of population size, the bias being $(n_{10|1}+n_{11|1})(n_{01|3}+n_{11|3})/n_{11|2}$. However, as this quantity has observable values it can be subtracted from $\hat{N}_+$ to get an unbiased estimate for the population size.

In conclusion, we note that in certain cases ignoring stratification (strata can be years or time periods) is not a problem in the estimation of the population size but in some cases it is. Furthermore, it is not possible to estimate the population sizes for each of the strata (or years) separately. This shows that there is a need to develop a general approach which would work for the cases where stratification has to be incorporated in the models.

5.4 EM Algorithm

A widely used method for analysis of partially classified counts is the EM algorithm (Little and Rubin, 1987). This technique was developed for data that are "missing at random" (MAR). In a survey context, missing values are said to be MAR if the occurrence of the missing value is conditionally independent of the actual response that would have been observed given the observed responses to the other questions; that is, the occurrence of the missing value can depend on observed responses to other questions, but given these, it does not depend on the missing value itself. Loosely speaking, for our case this implies that, observations from years where all registrations are active and observations from years with non-operating registrations that have the same characteristics do not differ systematically by year. If the data are MAR, the missingness is called "ignorable"

because ignoring the missing data mechanism does not affect likelihood-based inferences, such as the maximum likelihood (ML) estimates. ML produces estimates that are asymptotically unbiased if the model is true (see Rice, 1995, p. 264).

Under the ignorability assumption the EM algorithm can be used to obtain the ML estimates (Little and Rubin, 1987). The EM algorithm is an iterative procedure with two steps in each iteration: the E-step and the M-step. The E-step of the EM algorithm computes the expected complete data sufficient statistics given the current parameter estimates and the observed data; in this case this entails distributing partially classified counts using information in other years. The M-step computes new ML estimates of the parameters based on the current values of the expected complete-data sufficient statistics.

Applying the EM algorithm to capture-recapture data with partially overlapping populations is valid, if the non-operating lists are missing by design (such that the missingness is ignorable). For epidemiological capture-recapture data populations might partially overlap, for example:

- by year due to development of registrations which are hoped to be better than active ones or the closing of obsolete existing registrations,
- and by region as some regions might have registrations that are not yet implemented in other regions.

These examples are all design based, implying that the use of the EM algorithm is valid.

5.4.1 General procedure

In this Section we illustrate the general procedure of the EM algorithm. In our procedure the EM algorithm is used to distribute the observations from years where some registrations are not operational, and as such it is similar to the standard EM algorithm, whilst Baker (1990) uses a non-standard application of the the EM algorithm in that it is used to estimate observations missed by all lists. The notation used here is similar to that in (Little and Rubin, 1987, p. 182).

Let n_{ik} denote the frequencies of the hypothetical complete data, where $i = (1, 2, \cdots, I)$ is an index denoting a cross-classification of S lists such that $I = 2^S - 1$, and $k = (1, 2, \cdots, K)$ is the index for *Year* (which is fully observed). Note that for each list, $i_s = (0, 1)$, where $s = (1, 2, \cdots, S)$ is the index for a list. The hypothetical complete data consist of $I \times K$ cells (with cells denoted by c_{ik}), such that n_{ik} denotes the observed frequency of individuals classified into cell c_{ik}, with corresponding probabilities π_{ik}, $\sum_i \sum_k \pi_{ik} = 1$.

The observed data consist of two sets of years: 1) a set of years, denoted by S_1, where all lists are operating (completely classified observations), and 2) a set of years, denoted by S_2, where not all the lists are operating (partially classified observations). Note that it is a possible that S_1 is empty; if S_2 is empty, then there

is no missing data problem and the data can be analyzed in a standard way. We partition the partially classified observations into J groups, so that within each group, all units have the same set of possible cells (during partitioning we ignore year). Suppose r_{jk} denotes the count for the partially classified observations in the k^{th} year which fall in the j^{th} group; let S_{jk} denote the set of cells to which the observations might belong. Define indicator functions $\delta(c_{ik} \in S_{jk})$, $i = (1, 2, \cdots, I)$ and $j = (1, 2, \cdots, J)$ where $\delta(c_{ik} \in S_{jk}) = 1$ if cell c_{ik} belongs to S_{jk} and 0 otherwise.

Let $\hat{\pi}_{ik}^{(t)}$ be the current estimate of the probability for cell c_{ik} after the t^{th} iteration of the M-step. The $(t+1)^{th}$ E-step of the EM algorithm calculates the expectation of the cell frequencies (n_{ik}) for the partially classified frequencies (r_{jk}) using

$$\hat{n}_{ik}^{(t+1)} = \frac{\sum_{p=1}^{K} \hat{\pi}_{ip}^{(t)} \delta(c_{ip} \in S_{jp})}{\sum_{p=1}^{K} \sum_{l=1}^{I} \hat{\pi}_{lp}^{(t)} \delta(c_{lp} \in S_{jp})} \times r_{jk}. \tag{5.6}$$

The above expression distributes the partially classified counts (r_{jk}) using the current estimates of the conditional probabilities of falling in cell c_{ik} given that an observation falls in the set of categories S_{jk}.

The $\hat{n}_{ik}^{(t+1)}$ in (5.6) denote the completed data at the $(t+1)^{th}$ iteration. In the M-step a log-linear model is fitted to the completed data, with the cells missing by design denoted as structurally zero. Thus we maximize the so-called complete data log-likelihood. Let S^* denote the set of cells corresponding to years where all lists are operating, that is, where all relevant counts are observed. The complete data likelihood at iteration $t+1$ denoted by $\ell^{(t+1)}$ is given by,

$$\ell^{(t+1)} = \sum_{c_{ik} \in S^*} n_{ik} \ln \pi_{ik} + \sum_{c_{ik} \in S_{jk}} \hat{n}_{ik}^{(t+1)} \ln \pi_{ik}. \tag{5.7}$$

The fitted probabilities, $\hat{\pi}_{ik}$, from the log-linear model fitted in the M-Step are then used in the E-step of the $(t+1)$ iteration, where they are denoted by $\hat{\pi}_{ik}^{(t+1)}$, to derive updates for the completed data. This procedure is repeated until the complete data log-likelihood converges. After convergence the parameter estimates are used to find point estimates for the structurally zero cells, and an estimate of the population size.

5.4.2 Dual list examples

To illustrate the EM algorithm we use the examples presented in Section 5.3.2 and 5.3.3. For the example given in Section 5.3.2 there are two lists and two years (strata), with only list 1 operating in the second year. In this example there is only one partially classified frequency and it occurs in the second year, and using the notation presented in the general procedure the partially classified observation is denoted by r_{12} ($r_{12} \equiv n_{1+|2}$, see table 5.3, and for simplicity we proceed with

$n_{1+|2}$). As detailed earlier in the E-step we compute the conditional expectations for $n_{10|2}$ and $n_{11|2}$ using

$$\hat{n}_{10|2}^{(t+1)} = \frac{\hat{\pi}_{10|1}^{(t)}}{\hat{\pi}_{10|1}^{(t)} + \hat{\pi}_{11|1}^{(t)}} \times n_{1+|2};$$

$$\hat{n}_{11|2}^{(t+1)} = \frac{\hat{\pi}_{11|1}^{(t)}}{\hat{\pi}_{10|1}^{(t)} + \hat{\pi}_{11|1}^{(t)}} \times n_{1+|2}.$$

In the M-step the complete data log-likelihood is maximized with respect to the unknown parameters $\pi_{ij|k}$ with $n_{10|2}$ and $n_{11|2}$ replaced by their conditional expectations $\hat{n}_{10|2}^{(t+1)}$ and $\hat{n}_{11|2}^{(t+1)}$. Thus we maximize (compare with equation (5.7))

$$\ell^{(t+1)} = \sum_{i,j} n_{ij|1} \ln(\pi_{ij|1}) + \hat{n}_{10|2}^{(t+1)} \ln(\pi_{10|2}) + \hat{n}_{11|2}^{(t+1)} \ln(\pi_{11|2}).$$

To maximize this log-likelihood we use a log-linear model with structural zeros for the unknown counts, that is $n_{00|1}$, $n_{00|2}$ and $n_{01|2}$. Thus the complete data likelihood is maximized over 5 cells. The parameters can then be used to estimate the frequencies for the structurally zero cells. It can easily be verified that the total number of observations missed by all lists $n_{00|1} + n_{00|2} + n_{01|2}$ is equal to that obtained by collapsing the table as shown in Section 5.3.2. Thus the EM is also able to provide the solution to the problem.

Collapsing over years in the example in Section 5.3.3 resulted in a biased estimate of the population size. Here we show that the EM algorithm results in an unbiased estimate of the population size. This example has two partially classified counts and they are denoted by r_{11} and r_{23} ($r_{11} \equiv n_{1+|1}$ and $r_{23} \equiv n_{+1|3}$, see table 5.4). In the E-Step we compute the conditional expectations for $n_{10|1}$, $n_{11|1}$, $n_{01|3}$, and $n_{11|3}$ using

$$\hat{n}_{10|1}^{(t+1)} = \frac{\hat{\pi}_{10|1}^{(t)}}{\hat{\pi}_{10|1}^{(t)} + \hat{\pi}_{11|1}^{(t)}} \times n_{1+|1} \quad ; \quad \hat{n}_{11|1}^{(t+1)} = \frac{\hat{\pi}_{11|1}^{(t)}}{\hat{\pi}_{10|1}^{(t)} + \hat{\pi}_{11|1}^{(t)}} \times n_{1+|1};$$

$$\hat{n}_{01|3}^{(t+1)} = \frac{\hat{\pi}_{01|3}^{(t)}}{\hat{\pi}_{01|3}^{(t)} + \hat{\pi}_{11|3}^{(t)}} \times n_{+1|3} \quad ; \quad \hat{n}_{11|3}^{(t+1)} = \frac{\hat{\pi}_{11|3}^{(t)}}{\hat{\pi}_{01|3}^{(t)} + \hat{\pi}_{11|3}^{(t)}} \times n_{+1|3}.$$

In the M-step the complete data log-likelihood is maximized with respect to the unknown parameters $\pi_{ij|k}$ with $n_{10|1}$, $n_{11|1}$, $n_{01|3}$, and $n_{11|3}$ replaced by their conditional expectations $\hat{n}_{10|1}^{(t+1)}$, $\hat{n}_{11|1}^{(t+1)}$, $\hat{n}_{01|3}^{(t+1)}$, and $\hat{n}_{11|3}^{(t+1)}$ respectively. In this instance we maximize

$$\ell^{(t+1)} = \sum_{i,j} n_{ij|2} \ln(\pi_{ij|2}) + \hat{n}_{10|1}^{(t+1)} \ln(\pi_{10|1}) + \hat{n}_{11|1}^{(t+1)} \ln(\pi_{11|1}) + \hat{n}_{01|3}^{(t+1)} \ln(\pi_{01|3}) + \hat{n}_{11|3}^{(t+1)} \ln(\pi_{11|3}). \quad (5.8)$$

In the M-step a log-linear model with structurally zero cells is fitted to maximize the likelihood. In this instance the complete data has 7 cells and there are 5 structurally zero cells. After convergence the parameters can then be used to estimate the frequencies for the structurally zero cells. It can be easily verified that the estimate of population size obtained using the EM algorithm is unbiased, and thus it results in an estimate of population size from the collapsed table corrected for the bias found in equation (5.5).

5.4.3 Model selection

As there might be several competing models that can be entertained, it is imperative to find a parsimonious model which best fits the data. The likelihood ratio test can be used to discriminate between two competing (log-linear) models. This test compares the difference in deviance (-2 $\times$ log-likelihood) of the two models with the chi-squared distribution for a given the number of degrees of freedom (difference in number of parameters). For this the observed-data likelihood should be used. The researcher can also use the AIC or BIC statistics, which penalize the maximized likelihood for a model by number of parameters. Thus models with more parameters receive a high penalty. The model with the lowest AIC or BIC is preferred. The AIC is the usually preferred model selection method in capture-recapture studies (see Hook and Regal, 2000, p. 776).

When there are several adequate models which result in different estimates of the population size, basing inferences on one selected model alone is risky (Hoeting et al., 1999). This is because selecting one model ignores the uncertainty in model selection, leading to overconfident inferences. A striking example in the capture-recapture problem is given in Regal and Hook (1991). In this instance it is important to take model uncertainty into account when making inferences (Hoeting et al., 1999). The simplest way to incorporate the uncertainty of the model selection process is to use model averaging. If model selection uses the AIC, one can use AIC weights in the model averaging process (Stanley and Burnham, 1998; Burnham and Anderson, 2002).

The main advantages of model averaging are that it improves predictive performance and coverage (Hoeting et al., 1999). Model averaging avoids the problem of having to defend model choice and thus simplifies presentation. This is because model averaging allows users to incorporate several competing models in the estimation process. Furthermore, model averaging is more correct in that it takes into account a source of uncertainty that analysis based on model selection ignore. In general, this leads to higher estimates of variance than do estimates that ignore model uncertainty (see Hoeting et al., 1999, p. 398–399).

5.4.4 Variance estimation

We propose to use the parametric bootstrap (see Buckland and Garthwaite, 1991; Norris and Pollock, 1996) to calculate confidence intervals for the point estimates

(of the population size). The advantage of the bootstrap method over asymptotic methods is that it is simple. Also, formulae for asymptotic standard errors are available only for the usual approach of multiple-record systems estimation, but not for the situation where some of the registrations are not operating in some strata.

To illustrate the parametric bootstrap we use the example in Section 3.2, where there are two lists and two years. The initial step in the bootstrap is to use the EM algorithm to compute $\hat{p}_{00|1}$, $\hat{p}_{10|1}$, $\hat{p}_{01|1}$, $\hat{p}_{11|1}$, $\hat{p}_{00|2}$, $\hat{p}_{10|2}$, $\hat{p}_{01|2}$, $\hat{p}_{11|2}$, $\hat{N}_1$, and $\hat{N}_2$, where,

$$\hat{p}_{00|1} = \frac{\hat{n}_{00|1}}{\hat{N}_1}, \quad \hat{p}_{10|1} = \frac{n_{10|1}}{\hat{N}_1}, \quad \hat{p}_{01|1} = \frac{n_{01|1}}{\hat{N}_1}, \quad \hat{p}_{11|1} = \frac{n_{11|1}}{\hat{N}_1}; \tag{5.9a}$$

$$\hat{p}_{00|2} = \frac{\hat{n}_{00|2}}{\hat{N}_2}, \quad \hat{p}_{10|2} = \frac{\hat{n}_{10|2}}{\hat{N}_2}, \quad \hat{p}_{01|2} = \frac{\hat{n}_{01|2}}{\hat{N}_2}, \quad \hat{p}_{11|2} = \frac{\hat{n}_{11|2}}{\hat{N}_2}. \tag{5.9b}$$

To compute the confidence intervals (or variances) the following steps have to be used;

- Sample from a multinomial distribution with index $\hat{N}_1$ and probability vector $(\hat{p}_{00|1}, \hat{p}_{10|1}, \hat{p}_{01|1}, \hat{p}_{11|1})$. Do the same for the second year. If $\hat{N}_1$ and $\hat{N}_2$ are not integers it is simplest to round to the nearest integer (Buckland and Garthwaite, 1991).

- Remove cells corresponding to cells not observed in the original data table. That is, delete the observations for $n_{00|1}$, $n_{00|2}$, and $n_{01|2}$. Finally, add $n_{10|2} + n_{11|2}$ such that the form of the resulting table is identical to the form of the observed table.

- Use the EM algorithm to get the estimated population sizes for both years.

- Repeat the above steps B times, to get estimates of $\hat{N}_{1(j)}$ and $\hat{N}_{2(j)}$ $(j = 1, \cdots, B)$.

The variance of $\hat{N}_1$ and $\hat{N}_2$ is simply the variance of $\hat{N}_{1(j)}$ and $\hat{N}_{2(j)}$ (Buckland and Garthwaite, 1991). Using the parametric bootstrap results in standard errors which are not conditional on the observed sample size.

5.5 Application

To apply the EM algorithm to the data presented in Section 5.2 we note that for 1992 the observed array is $2 \times 2 \times 2 \times 2$, and in the E step it is spread out into a five dimensional array of $2 \times 2 \times 2 \times 2 \times 2$ using the five dimensional arrays for years 1993-1998. In 1993-1998 we have one structural zero cell in a year, namely the cell corresponding to observations missed by all lists. For 1992 we have two structural zeros, one corresponding to the observations missed by all lists and one

corresponding to the observations which are only contained in the registration not operating in 1992. The inclusion profiles for these observations are 00000 and 00010 (corresponds to observations that would have been observed in R4 only if the registration was active). In 1988-1991 there are four structural zero cells corresponding to the following inclusion profiles, 00000, 00100, 00110, and 00010. The last 3 inclusion profiles correspond to cells that could have been observed if R3 and R4 were operational.

This problem can be related to the general procedure described in Section 5.4 as follows. In 1993 to 1998 all lists are active, so these years belong to S_1. All observations in other years are in S_2. For this problem, S_2 can be classified into $J = [2^3 - 1] + [2^4 - 1] = 22$ cells as from 1988 to 1991 the observations for each inclusion profile have the same set of possible cells, and in 1992 the observations have their own set of possible cells. Rather than using i, j, and k, for the illustration we will use the corresponding capture profile and year. For example, the partially classified observation $r_{010|1989} = n_{01++0|1989} = 114$ has four possible cells, $S_{010|1989} = \{c_{01000|1989},\ c_{01100|1989},\ c_{01010|1989},\ c_{01110|1989}\}$, that is if all registrations were active, the frequency $n_{01000|1989}$ would have been distributed over these cells.

Table 5.5 presents a summary of the models fitted to the data. Year (Y_{cat}) is used as a stratifying variable in the table. The main effects only model has a poor fit. The approach we followed was to first explore heterogeneity followed by dependence (International Working Group for Disease Monitoring and Forecasting, 1995a). Thus we begin by adding heterogeneity terms (that is H1 and H2). First order heterogeneity (H1) results in a big improvement of the fit, but second order heterogeneity (H2) does not significantly fit better than model 2. We then allow the inclusion probabilities to vary by year, and it turns out that only the inclusion probabilities for R1 and R2 vary over time (model 7) but the other registrations do not (models 4-6). Model 11 shows that the registrations are pairwise related except for R3 and R4 (as the R4 and R5 interaction is set to zero in order to estimate H1). Models 12-14 allow the interactions to vary over time but none of these models lead to an improvement in fit.

The models with substantial support from the data, that is models with an AIC less than or equal to 2 from the AIC of the model with the lowest AIC (Burnham and Anderson, 2002, p. 70–71,170), are models 10 and 11. The yearly estimates of the population size for these two models is basically the same. If this was not the case, model uncertainty has to be incorporated in the estimates of the population size and their variances (Stanley and Burnham, 1998). This implies using one of these two models does not lead to overconfident inferences. The model with the lowest AIC is model 11 (see table 5.5) and this model will be used for the estimation of the yearly estimated population sizes and confidence intervals.

To compute the confidence intervals for the yearly estimates of population size the parametric bootstrap with 500 replications is used (see table 5.5). The confidence intervals show that, most often, the distribution of the estimates by

Table 5.5: Selected models with deviance and AIC

Model	Design matrix	Number of parameters	Degrees of freedom †	Deviance	AIC	$\hat{N}$ §
1	R1+R2+R3+R4+R5+Y_{cat}	16	213	409 ‡	441	2229
2	1 + H1	17	212	359 ‡	393	3009
3	2 + H2	18	211	359 ‡	395	2822
4	2 + (R1+ R2+R3+R4+R5) * Y_{cat}	67	162	191	325	2702
5	2 + (R1+ R2+R3+R5) * Y_{cat}	57	172	193	307	2708
6	2 + (R1+ R2+R3) * Y_{cat}	47	182	203	297	2697
7	2 + (R1+ R2) * Y_{cat}	37	192	213	287	2697
8	2 + R1 * Y_{cat}	27	202	280 ‡	334	3212
9	2 + R2 * Y_{cat}	27	202	256 ‡	310	2683
10	7 + R1*(R2+R3+R4+R5)+ R2* (R3+R4+R5)+ R3*(R4+R5)	46	183	156	248 ¶	2777
11	7 + R1*(R2+R3+R4+R5)+ R2* (R3+R4+R5)+ R3*R5	45	184	156	246 ¶	2778
12	11 + (R1*(R2+R3+R4+R5)+ R2* (R3+R4+R5)+ R3*R5)*Y_{cat}	135	94	102	372	3034
13	11 + (R1*(R2+R3+R5))*Y_{cat}	85	144	118	288	2988
14	11 + R1*R2*Y_{cat}	55	174	140	250	2990

†NB: There are 229 observed cells (see Table 1).
§There are 1783 cases observed at least once (see Table 1).
‡Significant at the 5 % level of significance.
¶Model has substantial support from the data (Burnham and Anderson, 2002, p. 70–71,170).

year is skewed. Furthermore, years with a higher number of structurally zero cells (1988-1992) have somewhat wider confidence intervals.

The table also shows estimates from the standard capture-recapture methods, that is log-linear models (see Bishop et al., 1975; Fienberg, 1972; Cormack, 1989) and the sampling coverage approach (Chao et al., 2001; Tsay and Chao, 2001). These models were fitted to each year separately using the program CARE-1 (which is downloadable from `http://chao.stat.nthu.edu.tw/`). Most of the yearly estimates from the sampling coverage approach are consistent with the estimates from the EM algorithm, except for a couple of years. This is not true for the log-linear models. As the estimates from the traditional approaches do not use information from the other years they tend to be more variable.

In conclusion, we stress that although in our example it is possible to use traditional approaches within each year, this is not possible where in one or more strata only one list is operating (compare Section 5.3). If a stratum has only two active lists, traditional approaches assume independence between the lists, whereas the EM algorithm utilizes the dependence between the lists in other strata. A joint model also decreases the possibility of chance capitalization. Fitting a joint model

Table 5.6: Estimates of population size and 95% confidence intervals by year

		EM algorithm		Log-linear			Sample Coverage	
Year	Observed	$\hat{N}$	95 % C.I.	Model [§]	$\hat{N}$	95 % C.I. [¶]	$\hat{N}$	95 % C.I.
1988	145	238	[161 , 290]	[12,5]	311	[200 , 648]	231 [†]	[188 , 318]
1989	163	204	[163 , 243]	[1,25]	174	[161 , 192]	181	[164 , 606]
1990	170	231	[185 , 269]	[1,25]	177	[168 , 189]	190	[171 , 581]
1991	150	187	[152 , 227]	[12,15]	191	[149 , 282]	185	[156 , 360]
1992	172	286	[211 , 319]	[12,23,5,H1]	782	[326 , 2687]	303 [†]	[249 , 395]
1993	160	220	[193 , 264]	[12,15,24,34,45]	320	[207 , 957]	233	[187 , 356]
1994	162	275	[235 , 355]	[14,15,24,34,35]	232	[197 , 293]	265	[197 , 464]
1995	174	307	[263 , 396]	[12,13,23,34,35,45]	206	[188 , 231]	200	[182 , 257]
1996	153	269	[233 , 345]	[12,13,24,25,34,45]	317	[220 , 583]	255	[181 , 527]
1997	180	306	[268 , 380]	[12,14,15,24,34,35,45]	351	[259 , 595]	340	[213 , 952]
1998	154	254	[220 , 319]	[14,23,24,25,34,45]	212	[179 , 266]	248	[200 , 346]

[§] 1,2,3,4, and 5 refer to R1, R2, R3, R4 and R5 respectively. H1 refers to first order heterogeneity (International Working Group for Disease Monitoring and Forecasting, 1995a).

[¶] C.I. computed using parametric bootstrap with 500 replications.

[†] One step estimator used. In 1988 the coverage is low, *i.e.* less than 55% (Chao et al., 2001, p. 3137), and in 1992 the *s.e.* is very large rendering the estimate useless.

is also more efficient and the resulting estimates are more stable.

5.6 Conclusions and Discussion

We have shown how the population size can be estimated using the multiple system estimator when the registrations emanate from partially overlapping populations. In epidemiology, there is a tendency for different institutions to collect data on the same diseased population, and in most cases not all cases are ascertained. Furthermore, some registrations might concentrate on special subgroups of the population, for example children or the elderly, such that the usual multiple system estimator cannot be used. The approach we presented can be useful in such situations.

This method will also be attractive to ecologists in cases where due to the nature of their surveys, certain groups of animals are excluded a-priori from the surveys. For instance in the two sample case: both large and small animals are captured and attached tags in the first sample, whereas only large animals are permitted to be caught in the second sample, the Peterson-Lincoln estimator is still valid because sample one is a random sample under the assumption that the samples observed in both strata have homogeneous capture probabilities. Similar statements can be said for cases where large, medium and small animals denote a strata.

In the two list case the traditional multiple system estimator does not account for any possible dependence between the lists. If there is dependence, this

approach also does not provide a correct estimate of the population size. This problem though can be minimized by the inclusion of (categorical) covariates, such that independence is assumed at each level of the covariates.

References

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

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

Bishop, Y., Fienberg, S., and Holland, P. (1975). *Discrete Multivariate Analysis: Theory and Practice*. McGraw-Hill, New York, USA.

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

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

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

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

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

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*, 152:771–778.

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*, 142:1047–1058.

International Working Group for Disease Monitoring and Forecasting (1995b). Capture-recapture and multiple record systems estimation 2: applications. *American Journal of Epidemiology*, 142:1059–1068.

Little, R. and Rubin, D. (1987). *Statistical analysis with missing data*. J. Wiley & Sons, New York.

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

Regal, R. and Hook, E. (1991). The effects of model selection on confidence intervals for the size of a closed population. *Statistics in Medicine*, 10:717–721.

Rice, J. (1995). *Mathematical Statistics and Data Analysis*. Duxbury Press, Carlifornia.

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

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

Tsay, P. and Chao, A. (2001). Population size estimation for capture-recapture models with application to epidemiological data. *Journal of Applied Statistics*, 28:25–36.

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. *Eur. J. Obset. Gynecol. Reprod. Biology*, 108:33–39.

Chapter 6

Structurally missing covariates in capture-recapture models

Registrations in epidemiological studies suffer from incompleteness, thus a general consensus is to use capture-recapture models. Lately there has been a thrust to incorporate covariates which relate to the capture probabilities in order to improve the estimate of population size. Current approaches utilize covariates observed in all the pertinent registrations. In epidemiological applications, one often encounters capture-recapture data where some covariates are not measured by one or more of the registrations under scrutiny. In this article, the usefulness of these types of covariates is evaluated and furthermore a methodology for analyzing such capture-recapture data is presented. The methodology is illustrated by applying it to data on neural tube defects in the Netherlands.

6.1 Introduction

Capture-recapture methods are frequently used to correct for underascertainment in epidemiological populations (see Chao et al., 2001, for a recent review). These methods utilize two or more existing incomplete overlapping lists (or registrations). A typical approach to analyse such data uses log-linear models, where the data are regarded as a 2^S contingency table (S being the total number of lists) for which the cell corresponding to those individuals absent in all the lists is missing (see Fienberg, 1972, p. 592). Several log-linear models are fitted to the incomplete contingency table, and the simplest plausible model (assuming that there is no S-factor interaction) that fits the observed cells is projected to cover the missing cell, thus yielding an estimate of the total population size (Fienberg, 1972).

Log-linear models are particularly useful for modelling both capture dependencies between samples and heterogeneity (Schwarz and Seber, 1999). In epidemiology list dependence is the norm, and it is easily handled using log-linear models (Schwarz and Seber, 1999, p. 439). "Observable" heterogeneity is usually handled using stratification based on covariate information. The current approach is to use covariates which are measured/observed in all the lists (hereafter *fully* observed covariates). These covariates are meant to account for any differences in "ascertainment" by stratum, where the strata are defined by the cross classification of the categorical covariates.

In epidemiology, unlike ecology, the lists used in the capture-recapture problem are usually started for different purposes. For example, for the same diseased population the researcher may use a combination of hospital records or patient group surveys, among others. This implies that in most cases each list will have a set of covariates that are not necessarily identical to the set of covariates in another list. The method presented above can only be used with covariates that are observed in all lists (that is, the fully observed covariates). A related problem is when the lists do not measure the same population (see Zwane et al., 2004).

In this paper we evaluate the usefulness of covariates that might affect the probability of inclusion to some of the lists, but are not observed in all of them.

The standard approach is to ignore covariates not available in every list. (Therefore we never see examples that involve these covariates in the literature.) We will show that ignoring these covariates may lead to biased results. Other naïve approaches like replacing (imputing) each of the missing values with a single reasonable proxy (alternatives are the mean, random hot-deck, and model-based) or regarding the missing values as an additional category can also be sources of bias or result in overestimation of gained precision (Vach, 1997).

Therefore we take the maximum likelihood approach. We will assume that the covariates are missing at random (MAR) in the sense of Little and Rubin (1987); that is, the probability of missingness depends only on the observed data (including the response). This is a reasonable assumption here, because the missingness is due to unasked questions. As a result, the missingness provides no information about the underlying process, implying that the missing data mechanism is ignorable (see Little and Rubin, 1987). For this presentation we concentrate on problems where all the covariates are categorical and thus use the EM algorithm. Continuous covariates can be categorized to meet this requirement.

The structure of the paper is as follows. In section 6.2, we introduce an example on neural tube defects to motivate the ideas of the paper. In Section 6.3, we show under what assumption(s) ignoring the partially overlapping covariates result in unbiased estimates of the populations size for the two list case. In other words, we show what assumptions are implicitly made when available covariates are ignored; these assumptions will often not hold in practice. In Section 6.4, we show for three lists that utilizing these covariates offers the researcher a rich choice of models from which to select those which describe the data best. In Section 6.5, we return to the neural tube defects example, where we will compare results using all covariates with results utilizing only fully overlapping covariates. Section 6.6 provides a concluding discussion.

6.2 Neural tube defects data

This work is motivated by data gathered routinely on children born with neural tube defects (NTD's) in the Netherlands. The data are for children registered between 2000 and 2002 by one of three registrations. The first (LVR_1) is a registry for low risk pregnancies and births in primary care, the second (LVR_2) registers births in secondary care, and the third (LNR) registers admission and re-admissions of newborns to paediatric department within the first 28 days of life.

In each of the three registries duration of pregnancy (0 if < 24 weeks, which is the legal limit for pregnancy terminations in the Netherlands, and 1 if ≥ 24 weeks), and the birth (or delivery) weight of the child (0 if $< 2,500$ grams, that is children with a low birth weight, and 1 if $\geq 2,500$ grams) are recorded (fully overlapping covariates). LVR_1 and LVR_2 also have information on age of mother (0 if < 25 years, 1 if ≥ 25 & < 30 years, 2 if ≥ 30 & < 35 years) which is not

measured in LNR. For this analysis we consider only the children who are not aborted, that is, children with pregnancy duration above 24 weeks. A summary of the data is shown in table 6.1. Birth weight had two missing values in observations

Table 6.1: Neural tube defects data : 2000-2002

Birth		Ascertainment history *							
weight	Age in years	100	010	001 †	110	101	011	111	Total
< 2.5 kg	< 25	0	8	?	4	1	4	0	31
	≥ 25 & < 30	3	10	?	8	0	1	2	24
	≥ 30 & < 35	10	20	?	8	1	2	2	43
	≥ 35	6	4	?	7	0	1	1	19
	Missing	0	0	14	0	0	0	0	32
≥ 2.5 kg	< 25	7	10	?	1	7	5	3	65
	≥ 25 & < 30	37	25	?	8	3	11	3	87
	≥ 30 & < 35	31	22	?	11	12	6	8	90
	≥ 35	16	7	?	1	2	2	4	32
	Missing	0	0	32	0	0	0	0	32
Total		110	106	46	48	26	32	23	391

*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).

†Observations listed in LNR only have a value for birth weight only.

listed in LVR_1 only, and another observation listed by LVR_2 only had a missing value for gender. We replaced these missing values by their predicted values given the capture profile and the observed covariates. For other details about the registrations, see Van der Pal et al. (2003).

In the capture-recapture problem, the age of the mother was used by York et al. (1995) to estimate the number of children born with birth defects in New York, and they found that the model including maternal age substantially improved predictive ability (see York et al., 1995, p. 237). Note that this variable is also known to be a risk factor of neural tube defects (Olney and Mulinare, 1998; Davidoff et al., 2002).

6.3 Dual list problem with partially overlapping covariates

6.3.1 General case

Consider a study with 2 lists (L_1 and L_2) and $a + b + c$ categorical variables, where a is the number of variables observed in L_1 only, b is the number of variables

observed in L_2 only, and c is the number of variables observed in both L_1 and L_2. The goal is to estimate the numbers missed by both lists resulting in an estimate of the population size N (the estimate is denoted by $\hat{N}$). Denote the set of variables observed by L_1 only, L_2 only, and in both L_1 and L_2 by $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ respectively. Let $\boldsymbol{\upsilon} = (\upsilon_1, \cdots, \upsilon_a)$, $\boldsymbol{\nu} = (\nu_1, \cdots, \nu_b)$, and $\boldsymbol{\omega} = (\omega_1, \cdots, \omega_c)$ index the specific combinations of levels of the variables in $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ respectively. Let U, V, and W denote the number of levels in $\boldsymbol{\upsilon}$, $\boldsymbol{\nu}$, and $\boldsymbol{\omega}$, respectively. Simplified cases, that is, when one of $\mathbf{A}$, $\mathbf{B}$, or $\mathbf{C}$ is empty will be discussed at the end of this section.

Denote the indices of the general contingency table by $(i, j, \upsilon, \nu, \omega)$, where i $(i = 0, 1)$ and j $(j = 0, 1)$ denote the inclusion to L_1 and L_2 respectively, with 1 denoting ascertainment. The observations actually observed are $n_{10(\upsilon+\omega)}$, $n_{01(+\nu\omega)}$, and $n_{11(\upsilon\nu\omega)}$. Cells corresponding to (i, j)=$(0, 0)$ are structurally zero. The (conditional) probabilities are given by $\pi_{ij(\upsilon\nu\omega)}$, where $\sum_{\substack{i,j,\upsilon,\nu,\omega \\ i,j\neq 0,0}} \pi_{ij(\upsilon\nu\omega)} = 1$.

The approach we take to analyze such data (data with partially overlapping covariates) is to identify the problem as a missing information problem, and then use the EM algorithm to obtain maximum likelihood estimates. The EM algorithm is an iterative procedure with two steps, namely, the Expectation and Maximization step. The EM algorithm starts with initial values for the parameters, which we denote by $\hat{\pi}^{(0)}_{ij(\upsilon\nu\omega)}$. In the E-step we compute the expected log-likelihood of the complete data conditional on the available data under the values of the parameters in that iteration. If we let $\hat{\pi}^{(t)}_{ij(\upsilon\nu\omega)}$ denote the current parameter values at the t^{th} iteration of the EM algorithm, the expected log-likelihood (at the t^{th} iteration) is given by,

$$\ell^{(t)} = \sum_{\substack{i,j,\upsilon,\nu,\omega \\ i,j\neq 0,0}} \hat{n}^{(t)}_{ij(\upsilon\nu\omega)} \ln \hat{\pi}^{(t)}_{ij(\upsilon\nu\omega)}$$

where,

$$\hat{n}^{(t)}_{10(\upsilon\nu\omega)} = \frac{\hat{\pi}^{(t)}_{10(\upsilon\nu\omega)}}{\hat{\pi}^{(t)}_{10(\upsilon+\omega)}} n_{10(\upsilon+\omega)}, \qquad \hat{n}^{(t)}_{01(\upsilon\nu\omega)} = \frac{\hat{\pi}^{(t)}_{01(\upsilon\nu\omega)}}{\hat{\pi}^{(t)}_{01(+\nu\omega)}} n_{01(+\nu\omega)}. \tag{6.1}$$

In the M-step a *log-linear model* is fitted to the completed data, $\hat{\pi}^{(t)}_{ij(\upsilon\nu\omega)}$, with cells corresponding to (i, j)=$(0, 0)$ denoted as structurally zero. For this simple problem, these cells correspond to the numbers missed by all lists. The fitted probabilities, $\hat{\pi}^{(t+1)}_{ij(\upsilon\nu\omega)}$, from the log-linear model fitted in the M-Step are then used in the E-step of the $(t + 1)$ iteration, to derive updates for the completed data.

The algorithm is iterated until it converges. After convergence the parameter estimates are used to find point estimates for the structurally zero cells, and an estimate of the population size. We note that, rather than use $\hat{\pi}^{(t)}_{10(\upsilon\nu\omega)}$ and $\hat{\pi}^{(t)}_{01(\upsilon\nu\omega)}$ in (6.1), we can use $\hat{\pi}^{(t)}_{11(\upsilon\nu\omega)}$. A proof is given in the Appendix.

Instead of using the parameter estimates to complete the table, that is estimate the frequencies corresponding to the cells (i, j)=$(0, 0)$, it is possible to use the fitted values, $\hat{n}_{ij(\upsilon\nu\omega)} = n \times \hat{\pi}_{ij(\upsilon\nu\omega)}$, where n is the total number of people listed by at least one of the registrations. In this instance the estimate of the numbers missed for each cross classification of all the covariates is given by,

$$\hat{n}_{00(\upsilon\nu\omega)} = \frac{\hat{n}_{10(\upsilon\nu\omega)}\hat{n}_{01(\upsilon\nu\omega)}}{\hat{n}_{11(\upsilon\nu\omega)}}. \tag{6.2}$$

This yields UVW estimates of the numbers missed that have to be added to arrive at the estimate of the number of individuals missed in the population. Notice that in (6.2) we assume that the lists are independent at each cross classification of the covariates.

6.3.2 Maximal model

The most complex log-linear model (*maximal* model) that can be fitted to this data is the log-linear model given by $[L_1\text{BC}, L_2\text{AC}, \text{ABC}]$. The maximal model does not include the interactions between L_1 and A, and L_2 and B, due to that A exist only when $L_1 = 1$ and B exists only when $L_2 = 1$, resulting in them being inestimable. Under this model we can use the observed frequencies such that, $\hat{\pi}_{11(\upsilon\nu\omega)}/\hat{\pi}_{11(\upsilon+\omega)} = n_{11(\upsilon\nu\omega)}/n_{11(\upsilon+\omega)}$ and $\hat{\pi}_{11(\upsilon\nu\omega)}/\hat{\pi}_{11(+\nu\omega)} = n_{11(\upsilon\nu\omega)}/n_{11(+\nu\omega)}$. Using this we can rewrite (6.2) using (6.1) as,

$$\begin{aligned}
\hat{n}_{00(\upsilon\nu\omega)} &= \frac{\hat{n}_{10(\upsilon\nu\omega)}\hat{n}_{01(\upsilon\nu\omega)}}{\hat{n}_{11(\upsilon\nu\omega)}}; && \text{(6.3a)}\\
&= \frac{\left[n_{10(\upsilon+\omega)}\frac{n_{11(\upsilon\nu\omega)}}{n_{11(\upsilon+\omega)}}\right]\left[n_{01(+\nu\omega)}\frac{n_{11(\upsilon\nu\omega)}}{n_{11(+\nu\omega)}}\right]}{n_{11(\upsilon\nu\omega)}}; && \text{(6.3b)}\\
&= n_{10(\upsilon+\omega)}n_{01(+\nu\omega)}\left[\frac{n_{11(\upsilon\nu\omega)}}{n_{11(\upsilon+\omega)}n_{11(+\nu\omega)}}\right]. && \text{(6.3c)}
\end{aligned}$$

This result is essential in showing the implicit assumptions that researchers make by ignoring the variables in **A** and **B**, as we will elaborate below.

6.3.3 Simplified situations

Under certain conditions (6.3c) can be simplified, for instance, when the covariates in **A** are conditionally independent of those in **B** given **C**, and when there are no covariates in either **A**, **B** or **C**. In all cases we will assume that whenever a set of covariates is present, then those covariates influence the inclusion probability or probabilities.

6.3.3.1 Conditional independence of A and B given C

Here we show that if the partially overlapping covariates in **A** and **B** are conditionally independent given the fully observed covariates in **C**, then **A** and **B** can be ignored. Under conditional independence of **A** and **B** given **C** the following condition holds:

$$n_{11(\upsilon\nu\omega)} = \frac{n_{11(\upsilon+\omega)} n_{11(+\nu\omega)}}{n_{11(++\omega)}}.$$

Thus under conditional independence, the term $n_{11(\upsilon\nu\omega)} / \left[n_{11(\upsilon+\omega)} n_{11(+\nu\omega)}\right]$ in (6.3c) can be replaced by $1/n_{11(++\omega)}$ so that

$$\hat{n}_{00(\upsilon\nu\omega)} = \frac{n_{10(\upsilon+\omega)} n_{01(+\nu\omega)}}{n_{11(++\omega)}}. \tag{6.4}$$

If we sum (6.4) over strata, the estimate of the numbers missed in the population reduces to the estimate attained from using only the variables observed in both lists **C** as

$$\begin{aligned}
\sum_{\upsilon,\nu,\omega} \hat{n}_{00(\upsilon\nu\omega)} &= \sum_{\upsilon,\nu,\omega} \frac{n_{10(\upsilon+\omega)} n_{01(+\nu\omega)}}{n_{11(++\omega)}}; \\
&= \sum_{\omega} \frac{\sum_{\upsilon} n_{10(\upsilon+\omega)} \sum_{\nu} n_{01(+\nu\omega)}}{n_{11(++\omega)}}; \\
&= \sum_{\omega} \frac{n_{10(++\omega)} n_{01(++\omega)}}{n_{11(++\omega)}}.
\end{aligned}$$

This results shows that by ignoring the variables in **A** and **B** in the analysis the researcher implicitly assumes that the two sets of covariates are conditionally independent. In most cases this condition will not be fulfilled, resulting in a biased estimate of the population size.

6.3.3.2 No variables in A or in B

Here we show that if covariates are measured by only one list, these covariates can be ignored. For example, if there are no covariates in **A** but some covariates in **B** and **C**, then summing (6.3) over ν and ω results in

$$\begin{aligned}
\sum_{\nu,\omega} \hat{n}_{00(\nu\omega)} &= \sum_{\nu,\omega} n_{10(+\omega)} n_{01(\nu\omega)} \left[\frac{n_{11(\nu\omega)}}{n_{11(+\omega)} n_{11(\nu\omega)}}\right]; \\
&= \sum_{\omega} \frac{n_{10(+\omega)} \sum_{\nu} n_{01(\nu\omega)}}{n_{11(+\omega)}}; \\
&= \sum_{\omega} \frac{n_{10(+\omega)} n_{01(+\omega)}}{n_{11(+\omega)}}.
\end{aligned}$$

This shows that only the variables in **C** have to be used and those in **B** can be ignored.

6.3.3.3 No variables in C

A scenario similar to Section 6.3.3.1 is when there are no covariates in **C**, but some variables present **A** and **B**. In this case adding (6.3) over υ and ν results in,

$$\sum_{\upsilon,\nu} \hat{n}_{00(\upsilon\nu)} = \sum_{\upsilon,\nu} n_{10(\upsilon+)} n_{01(+\nu)} \left[\frac{n_{11(\upsilon\nu)}}{n_{11(\upsilon+)} n_{11(+\nu)}} \right]. \tag{6.5}$$

This shows that the covariates in **A** and **B** cannot be ignored. Notice that under independence of **A** and **B**, $n_{11(\upsilon\nu)} / \left[n_{11(\upsilon+)} n_{11(+\nu)} \right] = 1/n_{11(++)}$ such that (6.5) is

$$\begin{aligned} \sum_{\upsilon,\nu} \hat{n}_{00(\upsilon\nu)} &= \sum_{\upsilon,\nu} \frac{n_{10(\upsilon+)} n_{01(+\nu)}}{n_{11(++)}}; \\ &= \frac{n_{10(++)} n_{01(++)}}{n_{11(++)}}. \end{aligned}$$

This implies that **A** and **B** can be ignored only if they are independent.

6.3.3.4 No variables in A and C or B and C

If there are no covariates in **A** and **C** adding (6.3) over ν results in,

$$\begin{aligned} \sum_{\nu} \hat{n}_{00(\nu)} &= \sum_{\nu} n_{10(+)} n_{01(\nu)} \left[\frac{n_{11(\nu)}}{n_{11(+)} n_{11(\nu)}} \right]; \\ &= \frac{n_{10(+)} \sum_{\nu} n_{01(\nu)}}{n_{11(+)}}; \\ &= \frac{n_{10(+)} \sum_{\nu} n_{01(\nu)}}{n_{11(+)}}, \end{aligned}$$

showing that the variables in **B** can be ignored.

6.3.4 Concluding remarks

The results in Section 6.3.3 imply that, only the log-linear models between the model assuming conditional independence of the partially observed covariates given by $[L_1\mathrm{BC}, L_2\mathrm{AC}]$, and the *maximal* log-linear model (see Section 6.3.2) written as $[L_1\mathrm{BC}, L_2\mathrm{AC}, \mathrm{ABC}]$ may result in an estimate of the population size different from the independence model. A summary of the results is shown in Table 6.2. Note that there are models of lower complexity with interactions between the partially observed covariates, for example $[L_1\mathrm{C}, L_2\mathrm{C}, \mathrm{ABC}]$. These models assume that there are no stratum effects, or alternatively that the sets of partially overlapping covariates do not affect being observed in any list, as a result these covariates can be ignored.

Table 6.2: Summary of results

Situation	Model	Decision
$\mathbf{A} \otimes {}^{\dagger}\mathbf{B}\vert\mathbf{C}$	$[L_1\mathrm{BC}, L_2\mathrm{AC}]$	Ignore **A** and **B**
No **A**	$[L_1\mathrm{BC}, L_2\mathrm{C}]$	Ignore **B**
No **B**	$[L_1\mathrm{C}, L_2\mathrm{AC}]$	Ignore **A**
No **C**	$[L_1\mathrm{B}, L_2\mathrm{A}, \mathrm{AB}]$	Ignore **A** and **B** if $\mathbf{A} \otimes {}^{\dagger}\mathbf{B}$
No **A** and **C**	$[L_1\mathrm{B}, L_2]$	Ignore **B**
No **B** and **C**	$[L_1, L_2\mathrm{A}]$	Ignore **A**
No **A**, **B** and **C**	$[L_1, L_2]$	Basic CRC model

$^{\dagger}\otimes$ denotes independence.

In summary when the researcher is confronted with a problem where there are covariates in **A**, **B** and **C** it is advisable to use the EM algorithm. In some cases it may turn out that there is no need to include the covariates in **A** and **B** or even those in **C**.

6.4 Triple list problem with partially overlapping covariates

6.4.1 General case

In the three list case (L_1, L_2 and L_3), there are seven combinations of the sets of covariates that can occur. Let $a+b+c+d+e+f+g$ denote the number of categorical covariates, where a is the number of covariates observed by L_1 only, b is the number of covariates observed by L_2 only, c is the number of covariates observed by L_3 only, d is the number of covariates observed by L_1 and L_2 only, e is the number of covariates observed by L_1 and L_3 only, f is the number of covariates observed by L_2 and L_3 only, and g is the number of covariates observed in all lists.

Denote the sets of covariates observed in L_1 only, L_2 only, L_3 only, L_1 and L_2 only, L_1 and L_3 only, L_2 and L_3 only, and all lists by **A**, **B**, **C**, **D**, **E**, **F** and **G** respectively. Let $\boldsymbol{\kappa} = (\kappa_1, \cdots, \kappa_a)$, $\boldsymbol{\upsilon} = (\upsilon_1, \cdots, \upsilon_b)$, $\boldsymbol{\nu} = (\nu_1, \cdots, \nu_c)$, $\boldsymbol{\omega} = (\omega_1, \cdots, \omega_d)$, $\boldsymbol{\tau} = (\tau_1, \cdots, \tau_e)$, $\boldsymbol{\xi} = (\xi_1, \cdots, \xi_f)$, and $\boldsymbol{\zeta} = (\zeta_1, \cdots, \zeta_g)$ denote the number of levels in **A**, **B**, **C**, **D**, **E**, **F** and **G**. Let T, U, V, W, X, Y, and Z denote the number of levels in $\boldsymbol{\kappa}$, $\boldsymbol{\upsilon}$, $\boldsymbol{\nu}$, $\boldsymbol{\omega}$ $\boldsymbol{\tau}$, $\boldsymbol{\xi}$, and $\boldsymbol{\zeta}$ respectively.

The observations actually observed are $n_{100(\kappa++\omega\tau+\zeta)}$, $n_{101(+\upsilon+\omega+\xi\zeta)}$, $n_{001(++\nu+\tau\xi\zeta)}$, $n_{110(\kappa\upsilon+\omega\tau\xi\zeta)}$, $n_{101(\kappa+\nu\omega\tau\xi\zeta)}$, $n_{011(+\upsilon\nu\omega\tau\xi\zeta)}$ and $n_{111(\kappa\upsilon\nu\omega\tau\xi\zeta)}$. This problem can be identified as a missing information problem and the EM algorithm used to obtain unbiased estimates of the population size. In the E-step, the expected values of the complete data are imputed, conditional on the observed data

and the current estimated probabilities, $\pi^{(t)}_{ijk(\kappa\upsilon\nu\omega\tau\xi\zeta)}$. In the M-step a parsimonious log-linear model is fitted to the potentially observable cells. This process is iterated until convergence. At convergence the model is projected to the cells corresponding to $(i, j, k) = (0, 0, 0)$. As the problem with 3 lists is "very" complex, we consider a simplified case in the next Section. Simplified cases are those where one or more of the sets of covariates are empty.

6.4.2 Simplified problem

Consider a case where there are $a + f + g$ covariates, where a is the number of covariates observed in L_1 and L_2 only, f is the number of covariates observed in L_2 and L_3 only, and g is the number of covariates observed in all lists. Denote the sets of covariates observed in L_1 only, L_2 and L_3 only, and in all lists by $\mathbf{A}$, $\mathbf{F}$, and $\mathbf{G}$ respectively. Let $\boldsymbol{\kappa} = (\kappa_1, \cdots, \kappa_a)$, $\boldsymbol{\xi} = (\xi_1, \cdots, \xi_f)$, and $\boldsymbol{\zeta} = (\zeta_1, \cdots, \zeta_g)$ index the specific combinations of levels of the variables in $\mathbf{A}$, $\mathbf{F}$, and $\mathbf{G}$ respectively. Furthermore, let T, Y, and Z denote the number of levels in $\boldsymbol{\kappa}$, $\boldsymbol{\xi}$, and $\boldsymbol{\zeta}$, respectively.

The "hypothetical" complete contingency table has elements $n_{ijk(\kappa\xi\zeta)}$, where $i, j, k = 0, 1$ denote ascertainment by L_1, L_2 and L_3 respectively, with 1 denoting ascertained. The observations actually observed are, $n_{100(\kappa+\zeta)}$, $n_{010(+\xi\zeta)}$, $n_{001(+\xi\zeta)}$, $n_{011(+\xi\zeta)}$, and $n_{ijk(\kappa\xi\zeta)}$ for the other possible inclusion profiles. In the E-Step of the EM algorithm we calculate the conditional expected log-likelihood functions given the current estimates, denoted by $\hat{\pi}^{(t)}_{ijk(\kappa\xi\zeta)}$, and the log-likelihood,

$$\ell^{(t)} = \sum_{i,j,k,\kappa,\xi,\zeta} \hat{n}^{(t)}_{ijk(\kappa\xi\zeta)} \ln \hat{\pi}^{(t)}_{ijk(\kappa\xi\zeta)},$$

where,

$$\hat{n}^{(t)}_{100(\kappa\xi\zeta)} = \frac{\sum_{i,j,k} \hat{\pi}^{(t)}_{ijk(\kappa\xi\zeta)}}{\sum_{i,j,k}\sum_{\xi} \hat{\pi}^{(t)}_{ijk(\kappa\xi\zeta)}} n_{100(\kappa+\zeta)}, \quad \hat{n}^{(t)}_{010(\kappa\xi\zeta)} = \frac{\sum_{i,j,k} \hat{\pi}^{(t)}_{ijk(\kappa\xi\zeta)}}{\sum_{i,j,k}\sum_{\kappa} \hat{\pi}^{(t)}_{ijk(\kappa\xi\zeta)}} n_{010(+\xi\zeta)}$$

$$\hat{n}^{(t)}_{001(\kappa\xi\zeta)} = \frac{\sum_{i,j,k} \hat{\pi}^{(t)}_{ijk(\kappa\xi\zeta)}}{\sum_{i,j,k}\sum_{\kappa} \hat{\pi}^{(t)}_{ijk(\kappa\xi\zeta)}} n_{001(+\xi\zeta)}, \quad \hat{n}^{(t)}_{011(\kappa\xi\zeta)} = \frac{\sum_{i,j,k} \hat{\pi}^{(t)}_{ijk(\kappa\xi\zeta)}}{\sum_{i,j,k}\sum_{\kappa} \hat{\pi}^{(t)}_{ijk(\kappa\xi\zeta)}} n_{011(+\xi\zeta)}$$

and (t) denotes the t^{th} iteration. In the M-Step a log-linear model is fitted to the completed data with cells corresponding to $(i, j, k) = (0, 0, 0)$ denoted as structurally zero, yielding an update of the parameter estimates which is then used in the next E-Step. The procedure is repeated until convergence. At convergence the parameter estimates are used to estimate the number of individuals in the structurally zero cells.

As in the dual list problem, it is also possible to use the fitted values, $\hat{n}_{ijk(\kappa\xi\zeta)} = n \times \hat{\pi}_{ijk(\kappa\xi\zeta)}$ to estimate the numbers missed. Here an estimate of the numbers

missed for each cross-classification of the covariates is given by,

$$\hat{n}_{000(\kappa\xi\zeta)} = \frac{\hat{n}_{100(\kappa\xi\zeta)}\hat{n}_{010(\kappa\xi\zeta)}\hat{n}_{001(\kappa\xi\zeta)}\hat{n}_{111(\kappa\xi\zeta)}}{\hat{n}_{110(\kappa\xi\zeta)}\hat{n}_{101(\kappa\xi\zeta)}\hat{n}_{011(\kappa\xi\zeta)}} \tag{6.6}$$

This yields UVW estimates of the populations missed by all lists that have to be added to arrive at a single estimate of the numbers missed in the population. Using (6.6) assumes that in each level of the cross classified covariates the cross product ratio of the complete table, that is $\frac{[\hat{n}_{001(\kappa\xi\zeta)}\hat{n}_{111(\kappa\xi\zeta)}]}{[\hat{n}_{101(\kappa\xi\zeta)}\hat{n}_{011(\kappa\xi\zeta)}]}$ is the same as in the table involving the missing cell, that is $\frac{[\hat{n}_{100(\kappa\xi\zeta)}\hat{n}_{010(\kappa\xi\zeta)}]}{[\hat{n}_{000(\kappa\xi\zeta)}\hat{n}_{110(\kappa\xi\zeta)}]}$ (see Darroch et al., 1993, p.1139).

6.4.3 Finding the maximal model

To find the maximal model we have to consider the following (list effects);

- Heterogeneity due to **A** cannot be estimated for L_1 as this set of variables is observed by this list only;
- Heterogeneity due to **F** cannot be estimated for the interaction of L_2 and L_3 as this set of variables is observed by L_2 and L_3 only;

Using this information, we find that the maximal model for these data is given by the log-linear model, $[L_1L_2\text{FG}, L_1L_3\text{FG}, L_2L_3\text{AG}]$. This model has no interaction between the variables in **A** and L_1, and also no interaction of the variables in **F** and the interaction of L_2 and L_3. Under the maximal model (6.6) can be written as,

$$\hat{n}_{000(\kappa\xi\zeta)} = \frac{\left[\frac{\sum_{i,j,k} n_{ijk(\kappa\xi\zeta)}}{\sum_{i,j,k} n_{ijk(\kappa+\zeta)}} n_{100(\kappa+\zeta)}\right] n_{010(+\xi\zeta)} \left[\frac{\sum_{i,j,k} n_{ijk(\kappa\xi\zeta)}}{\sum_{i,j,k} n_{ijk(+\xi\zeta)}} n_{001(+\xi\zeta)}\right] n_{111(\kappa\xi\zeta)}}{n_{110(\kappa\xi\zeta)} n_{101(\kappa\xi\zeta)} n_{011(+\xi\zeta)}}. \tag{6.7}$$

Here the fitted probabilities are replaced by the observed frequencies.

6.4.4 Concluding remarks

In typical capture-recapture models only the covariates in **G** are used, which might result in a loss of information on heterogeneity if the other sets of covariates are related to the capture probabilities. A surprising result for the three list case is that even variables observed in one list only cannot be ignored if they are related to the inclusion probabilities. Contrast this with the two list case where these variables can be ignored (see Section 6.3.3.2).

6.5 Application

In this Section we illustrate the proposed methods using the neural tube defects data set presented in Section 6.2. The AIC is used for model selection as from simulations it tends to pick the data generating model more frequently (see Stanley and Burnham, 1998, p. 492). Uncertainty in model selection is incorporated using model averaging, see Stanley and Burnham (1998).

In the first set of analyses we consider models utilizing birth weight only, as this is the usual way to analyze these data (see Table 6.3). The assumption of

Table 6.3: Models utilizing birth weight only

		4-way table ‡		5-way table ‡			Point	
	Model §	*d.f.*	Deviance	*d.f.*	Deviance	AIC	Estimate	95% C.I. ¶
00*a*	$[R_1, R_2, R_3]$	10	107.9	46	226.1	234.1	569	[519 , 627]
00*b*	$[R_1, R_2, R_3, B]$	9	43.0	45	161.2	171.2	569	[519 , 625]
01	$[R_1B, R_2B, R_3B]$	6	21.9	42	140.1	156.1	562	[515 , 620]
02	$[R_1R_2B, R_3B]$	4	9.4 *	40	127.6	147.6	558	[498 , 659]
03	$[R_1R_3B, R_2B]$	4	18.7	40	136.8	156.8	577	[518 , 654]
04	$[R_1B, R_2R_3B]$	4	11.5	40	129.7	149.7	600	[538 , 685]
05	$[R_1R_2B, R_1R_3B]$	2	9.3	38	127.4	151.4	550	[483 , 706]
06	$[R_1R_2B, R_2R_3B]$	2	4.2 *	38	122.4	146.4	645	[519 ,– –]
07	$[R_1R_2B, R_1R_3B, R_2R_3B]$	0	0.0	36	118.2	146.2	798	[564 ,– –]
08	$[R_1B, R_2B, R_3B,$ H1.$B]$	4	16.8	40	135.0	155.0	792	[552, 1357]

*Model fits the data, i.e., deviance is not significant.

§B, R_1, R_2, and R_3 denote birth weight, LVR_1, LVR_2, and LNR respectively. H1 stands for first order heterogeneity (see Darroch et al., 1993).

‡The five-way table includes age of the mother and its used to compare with models using all variables. The usual test of fit for models utilizing fully overlapping covariates uses the four-way table deviance.

¶Quantile confidence interval based on the parametric bootstrap with 1000 replications.

independence between the lists without covariates (i.e., model 00*a* and 00*b*) is challenged by the data. Assuming birth weight influences the inclusion probabilities greatly improves the fit in terms of the AIC. Dependencies between the lists at each level of birth weight are incorporated in models 02 to 07, and the inclusion of $LVR_1 : LVR_2$ and $LVR_2 : LNR$ interactions greatly improves the fit. Model 08, which assumes that the lists are of the same kind at each level of the covariates (see Darroch et al., 1993) fits poorly.

We then considered models with substantial support from the data for model

averaging. These are models with an AIC less than 4 from the model with the lowest AIC (see Burnham and Anderson, 2002, p. 70-71,171). These are models 02, 04, 06 and 07 (see table 6.3). A model averaged estimate of the population size is 685 children with a model averaged confidence interval of $[528, - -]$.

To compute the model averaged confidence interval we used a bootstrap-like procedure. This procedure requires the model averaged population size and the model averaged fitted probabilities on the completed data, that is, including the cells that cannot be observed by design. A multinomial sample is drawn given these parameters. Subsequently, the frequencies that are structurally zero in the original data are dropped, and the frequencies for "LNR only" are added over age of the mother, such that the resulting data structure is identical to the observed data. The models used in the model averaging are then fitted to the resulting data, resulting in several (equal to the number of models used in the model averaging) estimates of the population size, which are then averaged using the model probabilities used in computing the model averaged estimate of the population size (*i.e.*, model probabilities are not computed for each bootstrap sample).

We then considered models utilizing all covariates and a summary of the fitted models is shown in Table 6.4. Including an interaction between birth weight and age of the mother to the independence model leads to a significant drop in the AIC, see model 00*c*. In the second step we considered models where the covariates affect the inclusion probabilities to all lists, see models 09 to 12. The model with the lowest AIC among these models, that is model 11, assumes that the interaction between the covariates is not related to the inclusion probabilities.

In the third step, we considered interactions between lists in the presence of the covariates, see models 13 to 18. As we fitted the same list effect model at each level of the covariates, that is, each of the list effects (including list dependencies) had interactions with the covariates in the model, we then considered models with an AIC less than 7 from the model with the lowest AIC for refinement. (Notice that we make this rule a bit weaker than the rule used to select the final models.) The optimal model set includes model 15, 16 and 17. As model 15 is already a subset of 16 and 17, it suffices to refine only models 16 and 17 to arrive at a set of parsimonious models which best describe the data.

A general result is that incorporating the partially overlapping covariates leads to significant improvement in the fit, see Table 6.4. The best fitting models in terms of the AIC are models 16*d* and 17*b*. These models have an estimate of the population size of about 650 and 700 children delivered with an NTD in the three years, respectively. The estimates of the numbers missed from models with an AIC less that 4 from the model with the lowest AIC varies from 205 to 307. This shows some stability when compared to the models using only birth weight where the estimates ranged from 159 to 407. In addition, some models with substantial support from the data amongst models using only birth weight have large upper confidence limits, whilst the models utilizing all available covariates have reasonable confidence intervals.

Table 6.4: Models incorporating all available covariates

	Model †	*d.f.*	Deviance	AIC	Point Estimate	95% C.I. ‡
00*c*	$[R_1, R_2, R_3, BM]$	39	92.2	114.2	569	[517 , 626]
09	$[R_1B, R_2B, R_3B, BM]$	36	71.1	99.1	562	[515 , 617]
10	$[R_1M, R_2M, R_3M, BM]$	30	75.1	115.1	567	[519 , 631]
11	$[R_1B, R_1M, R_2B, R_2M, R_3B, R_3M, BM]$	27	51.8	97.8	558	[509 , 624]
12	$[R_1BM, R_2BM, R_3BM]$	18	42.2	106.2	575	[524 , 662]
13	$[R_1R_2B, R_1R_2M, R_3B, R_3M, BM]$	22	39.2	95.2	544	[484 , 652]
14	$[R_1R_3B, R_1R_3M, R_2B, R_2M, BM]$	22	41.9	97.9	576	[522 , 670]
15	$[R_1B, R_1M, R_2R_3B, R_2R_3M, BM]$	22	32.2	88.2	587	[532 , 693]
16	$[R_1R_2B, R_1R_2M, R_2R_3B, R_2R_3M, BM]$	17	24.8	90.8	626	[514 , – –]
17	$[R_1R_3B, R_1R_3M, R_2R_3B, R_2R_3M, BM]$	17	19.5	85.5	687	[577 , 1008]
18	$[R_1R_2B, R_1R_2M, R_1R_3B, R_1R_3M, R_2R_3B, R_2R_3M, BM]$	12	17.6	93.6	850	[547 , – –]
Refinements starting from model 16						
16*a*	$16 - R_1R_2M$	20	24.8	84.8	597	[481 , – –]
16*b*	$16 - R_2R_3B$	18	24.8	88.8	624	[520 , – –]
16*c*	$16 - R_2R_3B - R_1R_2M$	21	24.8	82.8	596	[484 , 835]
16*d*	$16c - R_1M$	24	29.4	81.4	644	[528 , 895]
16*e*	$16d - R_2R_3M$	28	43.7	87.7	565	[497 , 680]
16*f*	$16e - BM$	27	37.4	83.4	650	[531 , 907]
Refinements starting from model 17						
17*a*	$17 - R_1R_3M$	20	22.5	82.5	688	[582 , 950]
17*b*	$17a - R_1M$	23	27.8	81.8	698	[584 , 914]
17*c*	$17b - BM$	26	35.8	83.8	698	[580 , 930]
17*d*	$17b - R_1R_3B$	25	36.7	86.7	600	[538 , 693]
17*e*	$17b - R_2R_3M$	26	37.5	85.5	698	[570 , 927]

†B, M, R_1, R_2, and R_3 denote birth weight, the age of the mother, LVR_1, LVR_2, and LNR respectively.

‡Quantile confidence interval based on the parametric bootstrap with 1000 replications.

A model averaged estimate of the population size is 661 children born with an NTD with a confidence interval of $[567, 895]$. This estimate is different from the model averaged estimate from utilizing the fully overlapping covariate only. Furthermore, the confidence interval is reasonable.

The choice between model 16*d* and 17*b* is somewhat arbitrary. Both models indicate that the relation between LVR_2 and LNR persists in the presence of the covariates. The first model also indicates that the relation between LVR_1 and LVR_2 is present, whilst the second indicates that the relation between LVR_1 and LNR is present.

The probabilities of being missed by all registrations for the given covariates based on models 16d and 17b are shown in Table 6.5. Table 6.5 shows that,

Table 6.5: Probabilities of being missed by all registrations for given covariates

Model 16d

Delivery	Age of the mother (in years)				
weight	< 25	≥ 25 & < 30	≥ 30 & < 35	≥ 35	Total
< 2.5 kg	0.312	0.549	0.518	0.647	0.529
≥ 2.5 kg	0.148	0.344	0.295	0.392	0.307
Total	0.211	0.402	0.386	0.521	0.393

Model 17b

Delivery	Age of the mother (in years)				
weight	< 25	≥ 25 & < 30	≥ 30 & < 35	≥ 35	Total
< 2.5 kg	0.093	0.215	0.195	0.292	0.202
≥ 2.5 kg	0.284	0.544	0.489	0.595	0.503
Total	0.230	0.499	0.421	0.518	0.440

the lists miss more children born to women 35+ years and less children born to women under 25 years. This might due to that young women are more cautious compared to the older women. Between ages 25 and 35, the picture is diluted as there are high proportions of women having their first births and women already having children. This extra information is not available in models utilizing only the fully overlapping covariates. Table 6.5 also shows that for model 16d the lists miss more children with a low birth weight, while it is the contrary in model 17b.

6.6 Conclusions and Discussions

The analysis presented here incorporates partially overlapping covariates in the capture-recapture problem. This is accomplished by using the EM algorithm. For the two list case if the covariates are independent in some cases analysis can be performed using the traditional multiple systems estimator.

For mixed categorical and continuous covariates we envisage that using multiple imputation (see Schafer, 1997) would be more suitable, but this is a subject of further research.

Appendix

Under the maximal model (see Section 6.3.2), that is the log-linear models log-linear model given by $[L_1\text{BC}, L_2\text{AC}, \text{ABC}]$ the estimated probabilities are given

by,

$$\pi_{ij(\upsilon\nu\omega)} = \frac{\pi_{i+(+\nu\omega)} \times \pi_{+j(\upsilon+\omega)} \times \pi_{++(\upsilon\nu\omega)}}{\pi_{++(+\nu\omega)} \times \pi_{++(\upsilon+\omega)}}. \tag{A.1}$$

Without loss of generality we will omit the denominator in (A.1) in the rest of proof as it is fixed for a given inclusion profile or for given values of i and j. The $\pi_{ij(\upsilon\nu\omega)}$'s are then

$$\begin{aligned}
\pi_{01(\upsilon\nu\omega)} &= \pi_{0+(+\nu\omega)} \times \pi_{+1(\upsilon+\omega)} \times \pi_{++(\upsilon\nu\omega)}; && \text{(A.2a)}\\
\pi_{10(\upsilon\nu\omega)} &= \pi_{1+(+\nu\omega)} \times \pi_{+0(\upsilon+\omega)} \times \pi_{++(\upsilon\nu\omega)}; && \text{(A.2b)}\\
\pi_{11(\upsilon\nu\omega)} &= \pi_{1+(+\nu\omega)} \times \pi_{+1(\upsilon+\omega)} \times \pi_{++(\upsilon\nu\omega)}, && \text{(A.2c)}
\end{aligned}$$

such that,

$$\begin{aligned}
\pi_{01(+\nu\omega)} &= \pi_{0+(+\nu\omega)} \sum_{\upsilon} \pi_{+1(\upsilon+\omega)} \times \pi_{++(\upsilon\nu\omega)}; && \text{(A.3a)}\\
\pi_{10(\upsilon+\omega)} &= \pi_{+0(\upsilon+\omega)} \sum_{\nu} \pi_{1+(+\nu\omega)} \times \pi_{++(\upsilon\nu\omega)}; && \text{(A.3b)}\\
\pi_{11(+\nu\omega)} &= \pi_{1+(+\nu\omega)} \sum_{\upsilon} \pi_{+1(\upsilon+\omega)} \times \pi_{++(\upsilon\nu\omega)}; && \text{(A.3c)}\\
\pi_{11(\upsilon+\omega)} &= \pi_{+1(\upsilon+\omega)} \sum_{\nu} \pi_{1+(+\nu\omega)} \times \pi_{++(\upsilon\nu\omega)}. && \text{(A.3d)}
\end{aligned}$$

This shows that,

$$\frac{\text{(A.2a)}}{\text{(A.3a)}} = \frac{\pi_{0+(+\nu\omega)} \times \pi_{+1(\upsilon+\omega)} \times \pi_{++(\upsilon\nu\omega)}}{\pi_{0+(+\nu\omega)} \sum_{\upsilon} \pi_{+1(\upsilon+\omega)} \times \pi_{++(\upsilon\nu\omega)}} = \frac{\pi_{1+(+\nu\omega)} \times \pi_{+1(\upsilon+\omega)} \times \pi_{++(\upsilon\nu\omega)}}{\pi_{1+(+\nu\omega)} \sum_{\upsilon} \pi_{+1(\upsilon+\omega)} \times \pi_{++(\upsilon\nu\omega)}} = \frac{\text{(A.2c)}}{\text{(A.3c)}}$$

and,

$$\frac{\text{(A.2b)}}{\text{(A.3b)}} = \frac{\pi_{1+(+\nu\omega)} \times \pi_{+0(\upsilon+\omega)} \times \pi_{++(\upsilon\nu\omega)}}{\pi_{+0(\upsilon+\omega)} \sum_{\nu} \pi_{1+(+\nu\omega)} \times \pi_{++(\upsilon\nu\omega)}} = \frac{\pi_{1+(+\nu\omega)} \times \pi_{+1(\upsilon+\omega)} \times \pi_{++(\upsilon\nu\omega)}}{\pi_{+1(\upsilon+\omega)} \sum_{\nu} \pi_{1+(+\nu\omega)} \times \pi_{++(\upsilon\nu\omega)}} = \frac{\text{(A.2c)}}{\text{(A.3d)}}.$$

References

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

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

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*, 88:1137–1148.

Davidoff, M., Petrini, J., Damus, K., Russell, R., and Mattison, D. (2002). Neural tube defects-specific infant mortality in the United States. *Teratology*, 66:S17–S22.

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

Little, R. and Rubin, D. (1987). *Statistical analysis with missing data*. J. Wiley & Sons, New York.

Olney, R. and Mulinare, J. (1998). Epidemiology of neural tube defects. *Mental Retardation and Development Disabilities Research Reviews*, 4:241–246.

Schafer, J. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall/CRC, New York.

Schwarz, C. and Seber, G. (1999). A review of estimating animal abundance III. *Statistical Science*, 14:427–456.

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

Vach, W. (1997). Some issues in estimating the effect of prognostic factors from incomplete covariate data. *Statistics in Medicine*, 16:57–72.

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. *Eur. J. Obset. Gynecol. Reprod. Biology*, 108:33–39.

York, J., Madigan, D., Huech, I., and Lie, R. (1995). Birth defects registered by double sampling: a Bayesian approach incorporating covariates and model uncertainty. *Applied Statistics*, 44:227–242.

Zwane, E., Van der Pal, K., and Van der Heijden, P. (2004). The multiple-record systems estimator when registrations refer to different but overlapping populations. *Statistics in Medicine*, 23:2267–2281.

Chapter 7

Capture-recapture studies with incomplete mixed categorical and continuous covariates

Registrations in epidemiological studies suffer from incompleteness, thus a general consensus is to use capture-recapture models. Inclusion of covariates which relate to the capture probabilities has been shown to improve the estimate of population size. The covariates used have to be measured by all the registrations. In this article, we show how multiple imputation can be used in the capture-recapture problem when some lists do not measure some of the covariates or alternatively if some covariates are unobserved for some individuals. The approach is then applied to data on neural tube defects from the Netherlands.

7.1 Introduction

A serious problem in capture-recapture models with individual level covariates occurs when the data are missing on one or more covariates which define heterogeneous catchability. Item missing values are usually handled by imputation with a reasonable proxy (see for example, Zwane and Van der Heijden, in press) or by excluding those observations (see for example, Hwang and Huang 2003, p. 1113, and Wang and Yip 2003, p. 673). The missing-data problem is more acute when some of the registrations do not contain some of the covariates which define heterogeneous catchability. In epidemiology and public health contexts this is a common problem as the registrations used are usually compiled for different administrative purposes. The standard approach is to simply drop these covariates. On top of being a waste of data, this practice could lead to biases (invalid results) if the dropped covariates are sources of heterogeneity. As a result it is of interest to incorporate missing data techniques into capture-recapture studies (see Wang and Yip, 2003, p. 673).

This problem confronted us in the estimation of the prevalence of neural tube defects in the Netherlands using the capture-recapture (CRC) methodology. The data utilized three overlapping registrations, whether the gender of child and delivery weight are measured in all the registrations. Parity of the child and nationality (proxy for ethnicity) are measured in only two of the registrations. Under certain assumptions, two valid approaches can be used to estimate the population size, (i) use all covariates and the 2 lists where all covariates are measured, or (ii) use all lists and the covariates measured in all lists. In the first approach, one assumes that given all the covariates the lists are independent, whilst in the second approach dependencies between lists can be entertained. As a result these two approaches can result in different estimates of the population size. If given all covariates some of the lists are dependent and all the covariates influence inclusion to all lists, both these approaches will be biased.

Due to the arbitrary nature of choosing either estimating approach and that they are only valid under certain assumptions, it is preferable to formulate a model that utilizes all covariates and all lists. In the CRC problem, Zwane and Van der Heijden (2003) considered a log-linear model to describe the multinomial probabilities among discrete (and discretized) covariates and suggested the

use of the EM algorithm (see Little and Rubin, 1987, Chapter 9) for likelihood maximization when some covariates do not appear in some registrations. This approach can in-principle be used in our data but a more parsimonious model can be obtained by using the covariates in their measurement scale.

Due to the complexity in using a maximum likelihood based procedure for capture-recapture data with continuous covariates, we prefer the use multiple imputation (MI). MI was developed by Rubin (1987, 1996) in the context of non-response in sample surveys, for handling missing data with arbitrary patterns of missing data on both continuous and categorical covariates, but has been applied successfully in a broad variety of settings. MI has been used extensively in similar problems, *i.e.*, problems utilizing different surveys or registrations, see for example Belin et al. (1993), Gelman et al. (1998), and Raghunathan and Sciskovik (1995, 1998). In all these cases MI was found to perform well.

The rest of the paper is organized into five sections. In Section 7.2 we briefly discuss the neural tube defects data set. We also discuss the multinomial logit model for analyzing CRC data with continuous covariates in Section 7.3. Section 7.4 introduces multiple imputation in the capture-recapture problem. In Section 7.5, we report an analysis of the neural tube defects data set presented in Section 7.2, and conclude with a discussion in Section 7.6.

7.2 Neural tube defects data

The problem that triggered this work is the estimation of the number of children born with a neural tube defect in the Netherlands from 2000. The data are gathered routinely on children born with a neural tube defects (NTD's) in the Netherlands by midwives, obstetricians, or paediatric units. For this analysis we utilized three incomplete but overlapping databases which we describe briefly;

1. *The Dutch Perinatal Database 1 (LVR_1)*: This is a pregnancy and birth registry of low risk pregnancies and births.

2. *The Dutch Perinatal Database 2 (LVR_2)*: Registers data concerning the birth of a child in secondary care. If a women is referred from primary care to secondary care she may be registered in LVR_1 and LVR_2.

3. *The National Neonatal Database (LNR)*: Contains information on "all" admissions and re-admissions of newborns to paediatric departments within the first 28 days of life.

In each of these registries, the gender and the birth (or delivery) weight of the child are recorded (fully overlapping covariates). In LVR_1 and LVR_2, there is also information on parity of child and ethnicity/nationality which are not measured in LNR. For other details about the registrations, see Van der Pal et al. (2003). A summary of the data is shown in table 7.1. The variables in table 7.1 are known (or believed) to be risk factors for neural tube defects (see for example, Olney and

Table 7.1: Neural tube defects data : 2000

	Ascertainment history †							
Covariates	100	010	001 ‡	110	101	011	111	Total
Gender								
Male	19	14	7	11	5	10	3	69
Female	24	23	9	13	2	7	1	79
Total	*43*	*37*	*16*	*24*	*7*	*17*	*4*	*148*
Birth weight								
Mean	3.21	2.34	2.74	2.15	3.36	2.73	3.05	2.70
Ethnicity/Nationality								
Dutch	34	27	?	21	4	16	2	104
Non-Dutch	9	10	?	3	3	1	2	28
Missing	0	0	16	0	0	0	0	16
Parity								
Mean	1.67	2.14	NA	1.54	2.29	1.59	2.00	1.81

†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).

‡Observations listed in 'LNR only' have a value for birth weight only.

Mulinare, 1998; Davidoff et al., 2002; Vieira, 2004) or have been used successfully in the capture-recapture methodology (see, for example Madigan and York, 1997, for the use of ethnicity in the capture-recapture problem). We had no reason to believe that the gender of the child and parity influence the capture probability.

7.3 The multinomial logit model in the CRC problem

Assume that the true population size is N and the individuals are indexed by i ($i = 1, 2, ..., N$) of which n are ascertained by at-least one of S registrations. The inclusion profile for individual i is the vector $\boldsymbol{w}_i = [i_1 i_2 \cdots i_S]$, which is a series of binary variables with 1 denoting ascertained and 0 otherwise. The ascertainment profile $\boldsymbol{w}_i$ can be redefined as a nominal categorical variable Y_i with $K = 2^S - 1$ levels, indexed by k ($k = 1, \cdots, K$) with individual i falling in only one of the categories.

Now assume that for individual i there are covariate vectors $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ of length p and q respectively, where $\boldsymbol{x}_i$ are the covariates observed in all the registrations and $\boldsymbol{z}_i$ are the covariates not observed in all the registrations. Denoting the multinomial logit for individual i as $\eta_i' = [\eta_1(\boldsymbol{x}_i, \boldsymbol{z}_i), \eta_2(\boldsymbol{x}_i, \boldsymbol{z}_i), ..., \eta_K(\boldsymbol{x}_i, \boldsymbol{z}_i)]$, the

category probabilities are then given by,

$$\mathbb{P}(Y_i = k|\boldsymbol{x}_i, \boldsymbol{z}_i) = \exp[\eta_k(\boldsymbol{x}_i, \boldsymbol{z}_i)]/\sum_{r=1}^{K} \exp[\eta_r(\boldsymbol{x}_i, \boldsymbol{z}_i)]. \tag{7.1}$$

This model has to be constrained in some way for it to be used in the capture-recapture problem (see Zwane and Van der Heijden, in press, for details). Alho (1990) and Huggins (1989) constrained the logits such that the lists are independent at the individual level. After fitting the model the parameters can be used to estimate the probability that an individual is registered or listed at least once. Denoting this probability by ϕ_i (the estimated probability is denoted by $\hat{\phi}_i$), the estimate of the population size is

$$\hat{N} = \sum_{i=1}^{n} \hat{N}_i = \sum_{i=1}^{n} \frac{1}{\hat{\phi}_i},$$

where $\hat{N}_i$ is the contribution of individual i to the estimate of the population size (see Huggins, 1989).

Rather than use (7.1), the current standard is to use only the covariates observed in all lists, that is

$$\mathbb{P}(Y_i = k|\boldsymbol{x}_i) = \exp[\eta_k(\boldsymbol{x}_i)]/\sum_{r=1}^{K} \exp[\eta_r(\boldsymbol{x}_i)]. \tag{7.2}$$

Equation (7.2) will result in a biased estimate of the population size if the covariates in $\boldsymbol{z}_i$ are related to the inclusion probabilities. In this article we will complete the data set using the multiple imputation approach described in section 7.4 such that all covariates and lists are utilized.

7.4 Multiple imputation in the CRC problem

In this section, we will briefly describe the idea of multiple imputation methods. Multiple imputation is now standard in statistical literature (see Rubin, 1996, p. 486), and thus we will highlight only the most important points. MI involves three steps: 1) imputing the data under an appropriate model and repeating the imputation to obtain m copies of the filled-in data set; 2) analyzing each data set separately to obtain the desired parameter estimates and standard errors; 3) combining the results from the m parameter estimates by computing the mean of the m parameter estimates and a variance estimate that includes both within-imputation and an across-imputation components. Below we describe how the multiply imputed data sets are created and how the analysis of such data can be performed to result in one estimate of the population size and its standard error.

7.4.1 Creating multiply imputed data sets

Multiple imputation aims at imputing the missing values in $\boldsymbol{z}_i$ such that they can also be used in generally available software, like the multinomial logit model. Possible multivariate models for the data that can be used to draw the m plausible values for each missing item in the data set are the multivariate normal model, the general location model, or by using "compound conditional specification". A number of software programs are available implementing these models, some which were reviewed by Horton and Lipsitz (2001). Below we highlight the features of each of these approaches and situations where they can be used for creating the multiple imputations.

7.4.1.1 Multivariate normal model

A multivariate normal model with arbitrary covariance and correlation structure can be used for the imputation. In the capture-recapture problem this approach can be used when there are no missing values in categorical variables. (The inclusion profiles are binary variables, but because they have no missings they can enter the model as continuous covariates, see Schafer 1997, p. 203.) Note that in some cases even in the presence of missing binary or ordinal variables the multivariate normal model can still be used, but as noted by Horton et al. (2003) this practice can sometimes lead to a bias.

7.4.1.2 General location model

This model was introduces by Schafer (1997, Chapter 9) to characterize the joint distribution for data containing a mix of categorical and continuous covariates. This model assumes a multinomial distribution for the categorical variables and a multivariate normal distribution for the continuous variables within each cell of the contingency table. Belin et al. (1999) gives a discussion on the performance of the general location model with an ignorable missing data assumption in a mental health services study. They also give several considerations that have to be taken into account before using the general location model.

7.4.1.3 Compound conditional specification

Also called "incompatible Gibbs sampling". In this approach a different regression model is specified for each variable, see Van Buuren and Oudshoorn (1999) and Raghunathan et al. (2001) for two implementations. For categorical variables, the model could be logistic or multinomial model and for continuous variables, the linear regression model is sufficient. Imputation is done on an equation by equation basis. The problem of dimensionality of multivariate regression is reduced.

7.4.2 Selection of covariates

The registrations used in capture-recapture problems usually contain a wealth of covariates and these can also be used for imputations. Ideally all variables have to be used in the imputation model (Rubin, 1996) to make the missing at random (MAR) assumption more plausible. In some instances, especially in the general location model use of a large number of categorical covariates results in an inestimable model. Belin et al. (1999, p.3129) illustrated an approach which is a trade-off between trying to accommodate more detail in the incomplete data model and the ability to estimate parameters of the model.

7.4.3 Analysis

Once the model has been chosen and variables selected, the model can be used to generate via a random sampling procedure m imputed values for the missing data points, thus creating m complete data sets. For each completed data set an estimate of the population size ($\hat{N}_d$, where $d = 1, \cdots, m$) and its associated variance which we denote by $\hat{\text{var}}[\hat{N}_d]$ can be computed using capture-recapture models with continuous covariates (see Alho, 1990; Zwane and Van der Heijden, 2002). These estimates can then be combined using the approach of Rubin (1987) to arrive at a single estimate of the population size ($\hat{N}$) given by,

$$\hat{N} = \frac{1}{m}\sum_{d=1}^{m}\hat{N}_d \tag{7.3}$$

and an MI variance, $\hat{\text{var}}[\hat{N}]$ given by,

$$\hat{\text{var}}[\hat{N}] = \frac{1}{m}\sum_{d=1}^{m}\hat{\text{var}}[\hat{N}_d] + \left[1+\frac{1}{m}\right]\left[\frac{1}{m-1}\right]\sum_{d=1}^{m}\left(\hat{N}_d - \hat{N}\right)^2. \tag{7.4}$$

This variance includes two parts: the average within-imputation variance, which is the first part of (7.4), and the between-imputation variance.

7.5 Application

The method presented was applied to the data from the neural tube defects data set described in Section 7.2. In our imputation model we used all available covariates, but consider only birth weight and ethnicity as covariates that can possibly have an effect on the inclusion probability. For completeness we first present the results from the traditional approaches.

7.5.1 Traditional approach

In this section we present the estimates of the population size based on the two valid traditional approaches. We first consider the models using all covariates and

the two lists (LVR_1 and LVR_2). The estimates from all possible models where the covariates enter linearly range from 226 to 242. The model averaged estimate of the population size from models with an AIC less than 4 from the model with the lowest AIC is 237 (38.94), implying the log-based confidence interval (see Chao, 1987, p. 787) is $[183, 344]$. All models excluding delivery weight do not have support from the data.

We then consider models using all lists and delivery weight. Estimates from these models range from 183 to 274. The model averaged estimate of the population size from models with an AIC less than 4 from the model with the lowest AIC is 215 (43.17), implying the log-based confidence interval is $[170, 355]$.

It is evident that the model averaged estimates from these two approaches are different, and it is impossible to choose one approach over the other. Using two lists, one can only assume independence at the individual level but it is possible that this is fulfilled given the available covariates. With three lists, dependencies can be modelled, but it is likely that the dropping of covariates of heterogeneous catchability induces a three list dependence resulting in biased results. In the following section, multiple imputation is used to complete the data set such that all features of the data are utilized.

7.5.2 Multiple imputation

In this section we use multiple imputation techniques to analyze the neural tube defects data set. The only covariate with missing values to be used in our analysis model is ethnicity, which is a binary covariate. As recommended by Horton et al. (2003, p. 232) that it is preferable to use a discrete model even when confronted by a problem with missings in only binary variables we will use a discrete model. The discrete model we use is the general location model as implemented in the R (Ihaka and Gentleman, 1996) library MIX (Schafer, 1997, Chapter 9).

We used the EM and data augmentation (DA) algorithms in MIX to generate the posterior distribution of the parameters of the assumed model. Random draws from the posterior distribution were then taken $m = 10$ times to generate $m = 10$ complete data sets for the final analysis. As recommended the parameter estimates from the EM algorithm are used as starting values for the DA algorithm (Schafer, 1997). To ensure that the successive imputations are statistically independent, the DA algorithm was run 30000 times and at every 2500 iterations one of $m + 1 = 11$ imputations was selected. The extra data set was used for model selection (see Allison, 2001, p. 41).

All possible models were fitted to the extra data set and the models with least considerably less support from the data, or alternatively, the models with an AIC less than 7 from the model with the lowest AIC (see Burnham and Anderson, 2002, p. 70-71,170) are shown in table 7.2. In most of our analysis we opt for models with substantial support from the data, but as using the *ad-hoc* approach to model selection results is biased test statistics it is preferable to use a somewhat conservative criteria (see Allison, 2001, p. 41). This for this exercise we have

Table 7.2: Estimates of population size for the covariates models

Model	Design Matrix †	Covariate Matrix ‡	Model Selection AIC	Multiple Imputation Estimated Population	*s.e.*	95 % C.I. §
1	$[12, 3]$	$1 + B$	507.0	211	23.38	[179, 275]
2	$[13, 2]$	$1 + B + E$	505.0	206	21.08	[177, 264]
3	$[1, 23]$	$1 + B$	505.1	236	27.52	[196, 308]
4	$[1, 23]$	$1 + B + E$	506.4	236	27.53	[196, 308]
5	$[12, 13]$	$1 + B$	503.3	183	15.35	[163, 227]
6	$[12, 13]$	$1 + B + E$	500.3	213	65.18	[161, 481]
7	$[12, 13]$	$1 + B \times E$	505.3	229	196.97	[153, 1383]
8	$[12, 23]$	$1 + B$	505.6	271	95.73	[180, 622]
9	$[13, 23]$	$1 + B$	503.3	226	32.65	[184, 320]
10	$[13, 23]$	$1 + B + E$	501.8	239	42.00	[186, 363]
11	$[12, 13, 23]$	$1 + B$	506.3	193	38.19	[158, 339]
12	$[12, 13, 23]$	$1 + B + E$	504.5	309	272.05	[164, 1718]

†1 is LVR_1, 2 is LVR_2, and 3 is LNR

‡B and E refer to birth weight and ethnicity/nationality respectively

§Confidence intervals computed on the log scale, see Chao (1987, p. 787). Using asymptotic confidence intervals results in some lower endpoints being less than the observed sample.

considered models with at least considerably less support from the data (see Burnham and Anderson, 2002, p. 170). The notation used for the models table 7.2 is a summarized version, for instance for model 1, the full model is is given by, $LVR_1 + LVR_2 + LNR + LVR_1 : LVR_2 + LVR_1 : B + LVR_2 : B + LNR : B + LVR_1 : LVR_2B$. In the multinomial logit model the intercept and covariate effects drop out.

Columns 4 in table 7.2 relates to the AIC of the data set used for model selection, whilst the last three columns relate to the analysis of the rest of the data sets combined using the methods discussed in section 7.4.3.

The model averaged estimate of the population size, using the AIC's in table 7.2 is 224 (68.57), implying the log-based confidence interval is $[165, 493]$. This estimate is marginally different from the estimate using all lists and birth weight, but very different from the estimate using all available covariates. What is evident from the standard error and confidence interval is that using the multiply imputation approach also adds noise to the analysis.

The model with the lowest AIC in table 7.2, that is model 6, shows that there is dependence between LVR_1 and LVR_2, and LVR_1 and LNR at the individual level. The model averaged parameter estimates from this model are given in table 7.3. To interpret the parameters in table 7.3 we focus on the

Table 7.3: Parameter estimates for model 6

	Model averaged			
Parameter	Estimate	*s.e.*	t	$\mathbb{P} > \|t\|$ †
LVR_1	-4.234	1.771	-2.391	0.008
LVR_2	-1.143	1.492	-0.766	0.222
LNR	-3.040	1.272	-2.389	0.008
$LVR_1 \times LVR_2$	3.561	1.795	1.984	0.024
$LVR_1 \times LNR$	0.131	1.989	0.066	0.474
LVR_1:B	1.005	0.481	2.091	0.018
LVR_2:B	-0.106	0.379	-0.281	0.389
LNR:B	0.321	0.301	1.066	0.143
$[LVR_1 \times LVR_2]$:B	-1.135	0.484	-2.346	0.009
$[LVR_1 \times LNR]$:B	0.384	0.563	0.682	0.248
LVR_1:E	1.913	1.380	1.386	0.083
LVR_2:E	1.839	1.262	1.457	0.073
LNR:E	1.687	1.101	1.532	0.063
$[LVR_1 \times LVR_2]$:E	-1.169	1.410	-0.829	0.204
$[LVR_1 \times LNR]$:E	-3.223	1.308	-2.464	0.007

†Degrees of freedom for t distribution computed via the increase in variance due to missing data (see Allison, 2001, p. 47–48)

significant heterogeneity parameter estimates, that is, parameters involving birth weight and ethnicity. For instance the parameter estimate for LVR_1:B indicates that the probability of being listed in LVR_1 only increases with birth weight. The parameter estimate for $[LVR_1 \times LVR_2]$:B indicates that being listed in both LVR_1 and LVR_2 decreases with birth weight. The parameter estimate for $[LVR_1 \times LNR]$:E indicates that children born from Dutch women are less likely to be listed in both LVR_1 and LNR.

7.6 Concluding Remarks

In capture-recapture models it is desirable to include individual level covariates to account for any differences in ascertainment by the registrations. When these covariates are not measure by all registrations (or they contain missing data), the commonly used approaches of dropping (or ignoring) these covariates or mean value imputation may give biased estimates of the population size. Multiple imputation is proposed to handle the missing covariate problem in the capture-recapture models.

Multiple imputation is applicable to missing covariate problems with arbitrary missing data patters and arbitrary number of covariates. Though our application is in epidemiology with only three lists this approach is applicable to wide ranging

capture-recapture problems. Based on the results presented in the previous sections we can make a strong recommendation for the use of MI in capture-recapture models with missing covariates.

In our analysis, we used an *ad hoc* approach to model selection. An alternative is to use the procedure for calculating the complete-data log likelihood ratio (and corresponding p-value) for analysis on multiply imputed data proposed by Meng and Rubin (1992). We preferred the *ad hoc* approach due to its simplicity.

References

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

Allison, P. (2001). *Missing data*. Sage University Papers Series on Quantitative Applications in Social Sciences, Thousand Oaks, CA.

Belin, T., Diffendal, G., Mack, S., Rubin, D., Schafer, J., and Zaslavsky, A. (1993). Hierarchical logistic regression model for imputation of unresolved enumeration status in undercount estimation. *Journal of the American Statistical Association*, 88:1149–1159.

Belin, T., Hu, M., Young, A., and Grusky, O. (1999). Performance of a general location model with an ignorable missing-data assumption in a multivariate mental health services study. *Statistics in Medicine*, 18:3123–3135.

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

Chao, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, 43:783–791.

Davidoff, M., Petrini, J., Damus, K., Russell, R., and Mattison, D. (2002). Neural tube defects-specific infant mortality in the United States. *Teratology*, 66:S17–S22.

Gelman, A., King, G., and Liu, C. (1998). Not asked and not answered:Multiple Imputation for multiple surveys. *Journal of the American Statistical Society*, 93:846–857.

Horton, N. and Lipsitz, S. (2001). Multiple imputation in practice: comparison of software packages for regression models with missing variables. *The American Statistician*, 55:244–254.

Horton, N., Lipsitz, S., and Parzen, M. (2003). A potential for bias when rounding in Multiple Imputation. *The American Statistician*, 57:229–232.

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

Hwang, W. and Huang, S. (2003). Estimation in capture-recapture models when covariates are subject to measurement errors. *Biometrics*, 59:1113–1122.

Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314.

Little, R. and Rubin, D. (1987). *Statistical analysis with missing data*. J. Wiley & Sons, New York.

Madigan, D. and York, J. (1997). Bayesian methods for estimation of the size of a closed population. *Biometrika*, 84:19–31.

Meng, X. and Rubin, D. (1992). Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika*, 79:103–111.

Olney, R. and Mulinare, J. (1998). Epidemiology of neural tube defects. *Mental Retardation and Development Disabilities Research Reviews*, 4:241–246.

Raghunathan, T., Lepkowski, J., Van Hoewyk, J., and Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, 27:85–97.

Raghunathan, T. and Sciskovik, D. (1995). A multiple-imputation analysis of a case-control study of the risk of primary cardiac arrest among pharmaologically treated hypertensives. *Applied Statistics*, 45:335–352.

Raghunathan, T. and Sciskovik, D. (1998). Combining exposure information from various sources in an analysis of case-control data. *The Statistician*, 47:333–347.

Rubin, D. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, New York.

Rubin, D. (1996). Multiple imputation after 18+ years. *Journal of American Statistical Society*, 91:473–489.

Schafer, J. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall/CRC, New York.

Van Buuren, S. and Oudshoorn, C. (1999). Flexible multivariate imputation by mice. *Leiden: TNO Preventie en Gezondheid, TNO/VGZ/PG 99.054*.

Vieira, A. (2004). Birth order and neural tube defects: a reappraisal. *Journal of Neurological Sciences*, 217:65–72.

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. *Eur. J. Obset. Gynecol. Reprod. Biology*, 108:33–39.

Wang, Y. and Yip, P. (2003). A semiparametric model for capture-recapture experiments. *Scandanavian Journal of Statistics*, 30:667–676.

Zwane, E. and Van der Heijden, P. (2002). The multiple-system estimator in the presence of covariates. In M. Stasinopoulos & G. Touloumi, editor, 17^{th} *International Workshop on Statistical Modelling*, pages 697–701, Chania, Greece. Chania University.

Zwane, E. and Van der Heijden, P. (2003). Partially missing covariates in capture-recapture models. In G. Verberke, G. Molenberghs, G. Aerts, & S. Fieuws, editor, 18^{th} *International Workshop on Statistical Modelling*, pages 481–485, Leuven, Belgium. Catholic University Leuven.

Zwane, E. and Van der Heijden, P. (in press). Semiparametric models for capture-recapture studies with covariates. *Computational Statistics and Data Analysis*.

Samenvatting

Dit proefschrift presenteert methoden om geobserveerde heterogeniteit in vangst-hervangstmethoden effectief te modelleren. Vangst-hervangstmethoden schatten de omvang van een populatie gebaseerd op incomplete maar overlappende registraties. Voor twee overlappende steekproeven (van dezelfde populatie) schat een vangst-hervangstmethode het niet-geobserveerde deel van de populatie (individuen die in geen van de beide steekproeven zitten). De schatting maakt gebruik van de assumptie dat de inclusiekansen onafhankelijk zijn. Een andere assumptie is de aanname van homogeniteit met betrekking tot de inclusiekansen van de individuen.

In de epidemiologie zullen sommige groepen individuen eerder in een bepaalde registratie worden opgenomen dan andere groepen. Dit betekent dat er heterogeniteit is met betrekking tot de inclusiekansen. Voorts hangen sommige registraties nauw met elkaar samen met het gevolg dat de kans op een registratie A de kans op een registratie B verhoogd of juist verlaagd. Heterogeniteit wordt vaak gemodelleerd door middel van stratificatie die is gebaseerd op waargenomen covariaten. Afhankelijkheden tussen registraties kunnen alleen worden gemodelleerd voor problemen met meer dan twee registraties. Voor categorische covariaten gebruikt men vaak loglineaire modellen die zowel heterogeniteit als afhankelijkheid modelleren. Voor continue covariaten zijn er alleen modellen die er vanuit gaan dat de registraties onafhankelijk zijn op het niveau van de individuen.

Hoofdstuk 2 is een uitbreiding van de logistische regressiemethoden van Alho (1990) en Huggins (1989) in de zin dat afhankelijkheid tussen twee registraties wordt gemodelleerd voor continue covariaten. De methoden van Alho (1990) en Huggins (1989) gaan er vanuit dat de registraties onafhankelijk zijn op het niveau van de individuen. Deze restrictie is niet houdbaar omdat er heterogeniteit zal zijn vanwege niet-geobserveerde variabelen, met het gevolg dat er afhankelijkheid tussen de registraties zal zijn. Een andere reden voor de uitbreiding is dat de opname van covariaten niet de afhankelijkheden tussen de registraties niet wegneemt. We modelleren afhankelijkheid door middel van het multinomiale logit model van Bock (1975, Hoofdstuk 8). Met simulatie laten we zien dat het negeren van aanwezige afhankelijkheid leidt tot onzuivere schattingen van de populatiegrootte en van de standaardfouten.

In hoofdstuk 3 ontwikkelen we een methode om de assumptie met betrekking

tot de lineariteit in de logit van het multinomiale logit model te omzeilen. Een veelgehoorde kritiek op de logistische functie is dat de vorm van de functie onbruikbaar is voor vangst-hervangstmethoden (zie Borchers et al., 1998). Hoofdstuk 3 is een uitbreiding van het multinomiale logit model in hoofdstuk 2 in de zin dat de covariaten niet meer lineair in de logit hoeven te zijn. De benadering kan ook worden gebruikt om de assumptie met betrekking tot de lineairiteit in de logit te checken.

In hoofdstuk 4 bespreken we de implementatie van de parametrische bootstrap wanneer er continue covariaten in het model zitten. Anderen hebben al aangetoond dat symmetrische (of asymptotische) betrouwbaarheids-intervallen niet bruikbaar zijn in vangst-hervangstmethoden, zie The International Working Group for Disease Monitoring and Forecasting (1995). Verschillende onderzoekers hebben de niet-parametrische bootstrap gebruikt wanneer er continue covariaten in het model zitten (Huggins, 1989; Tilling en Sterne, 1999), maar zoals opgemerkt door Norris en Pollock (1996): deze bootstrap resulteert in een onderschatting van de variantie. Met simulatie tonen we aan dat onze aanpak resulteert in adequate schattingen.

In hoofdstuk 5 ontwikkelen we een methode om vangst-hervangstdata te analyseren wanneer sommige registraties geen data bevatten van een deelverzameling van de populatie. Merk op dat indien sommige registraties bepaalde individuen niet meten, deze individuen kans nul hebben om in alle registraties te zitten met het gevolg dat de gebruikelijke vangst-hervangstmethoden tekort schieten. Vervolgens bespreken we het analyseren van vangst-hervangstdata wanneer sommige registraties niet alle covariaten die *heterogene vangbaarheid* meten bevatten. Het doel is om alle heterogene-vangbaarheidvariabelen te gebruiken omdat in sommige gevallen, het negeren van deze variabelen een "alle registraties"-interactie impliceert en daarom ongeldige resultaten levert.

In hoofdstuk 6 wordt een EM-algoritme gepresenteerd (Little en Rubin, Hoofdstuk 6) voor het analyseren van incomplete data als er alleen categorische covariaten zijn. In hoofdstuk 7 stellen we een multiple imputatie (Shafer, 1997) voor problemen met incomplete continue (en categorische) covariaten.

Curriculum Vitæ

Eugene N. Zwane was born in Manzini, Swaziland, on June 11, 1972. After completing secondary and high school, he received a Bachelor of Arts in Social Sciences degree on September 17, 1994, with specializations in Statistics and Demography from the University of Swaziland. In November 23, 1994 he started working as a Statistician for the Central Statistical Office of Swaziland. Whilst on study leave he obtained a Master of Science degree on October 2, 1997, with specialization in Biostatistics. In November 2000 he started as a PhD student at the Department of Methodology and Statistics (faculty of Social Sciences) at Utrecht University.