An Analysis on Generalized Models for Capture–Recapture Studies with Individual Covariates

Rajani Kotyal[1] Dr. Kamleshwar Dutta[2]

[1]Research Scholar, CMJ University, Shillong, Meghalaya

[2]Professor of Statistics, CMJ University, Shillong, Meghalaya

Abstract – Capture-recapture methods are used to estimate the incidence of a disease, using a multiple-source registry. Usually, log-linear methods are used to estimate population size, assuming that not all sources of notification are dependent. Where there are categorical covariates, a stratified analysis can be performed. The multinomial logit model has occasionally been used. In this paper, the authors compare log-linear and logit models with and without covariates, and use simulated data to compare estimates from different models. The crude estimate of population size is biased when the sources are not independent. Analyses adjusting for covariates produce less biased estimates. In the absence of covariates, or where all covariates are categorical, the log-linear model and the logit model are equivalent.

A 4exible method for modelling capture–recapture data with continuous covariates that describe heterogeneous catch ability 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.

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.

---♦-------------------------------------

INTRODUCTION

Accurate information on the prevalence or incidence of a disease may be required by epidemiologists, health services researchers, or health care planners and providers. One way to obtain such data is to set up a specialist registry, with the aim of registering all cases in a population. However, even when the methodology and resources for such a registry are well-established (such as with regional cancer registries in the United Kingdom and around the developed world), it is unlikely that every case will be registered. Thus, some estimate of the coverage of the registry, and hence of the true number of affected individuals, is needed.

The capture-recapture method was initially developed to estimate the size of wildlife populations. Animals are trapped, marked, and released on a number of occasions, and the individual trapping histories are then used to estimate die size of die population. The method has been increasingly used to estimate the size of human populations (such as the number of people with a given disease), using overlapping multiple sources of notification as die "captures." A major epidemiologic application has been to the estimation of die completeness of cancer registries, but die method has been used to estimate me sizes of many populations, including homeless people and children born with genetic disorders.

Two recent reviews of capture-recapture method s in epidemiology have emphasized the use of log-linear models (using Poisson regression) to investigate die relations between die sources used and die number of people "captured". This method makes two major and related assumptions about die probability of capture. Firstly, it is assumed that the capture probabilities for different notification sources are not all dependent. Thus, if there are only two sources, they are assumed to be independent. Applying standard capture-recapture method s to two dependent sources has been shown algebraically to underestimate the population size if die two sources are positively dependent, and to overestimate the population size in the case of negative dependence. Secondly, die probability of capture by a given source is assumed to be the same for each individual in die population.

These assumptions are violated when die probabilities of capture depend on covariates, such as severity of disease or age. One proposed solution using log linear models is to stratify die data according to the covariate(s), estimate the total number of cases for each stratum, and then combine these estimates.

Where covariate information is not available and there are more man two sources, multiple different two source estimates can be made, considering each source against all of die others pooled. 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, pp. 645–646) recently noted that when there are unmeasured sources of heterogeneity, accounting only for the measured ones will not eliminate all sources of bias. In support, Pollock (2002, p. 88) comments that “although using individual covariates has the purpose of accounting for heterogeneity, some inherent heterogeneity may still remain due to other unobserved variables”. This remaining heterogeneity may result in some registrations to be dependent even after controlling for the observed covariates.

An important feature of capture-recapture modeling is the ability to include covariates, including individual-specific ones (Lebreton et al. 1992, Schwarz et al. 1993, Bonner and Schwarz 2006, King et al. 2008, Catchpole et al. 2008, Bonner et al. 2010). The development of

methods for including individual covariates has focused on models that condition on the first capture of each individual. A consequence is that likelihood based inference is restricted to statements about survival or recapture probabilities and related quantities.

Importantly, these models do not include abundance parameters (or parameters related to abundance, such as population growth rates or stopover time) in the likelihood. Instead, inference about abundance has relied on ad-hoc Horvitz-Thompson-type approaches (Huggins 1989, McDonald and Amstrup 2001). The estimation of the population size based on multiple incomplete lists has a long history (Chao et al., 2001, Schwarz and Seber, 1999). The advantages of using these methods as a substitute for direct counting in epidemiology has been strongly emphasized (International Working Group for Disease Monitoring and Forecasting, 1995). The basic assumptions are that the population being estimated is closed, i.e., births, deaths and migrations are negligible, the individuals can be matched without error, and for the traditional approach an additional assumption is that all individuals have the same probability of being ascertained by a registration. In recent times this additional assumption is relaxed by allowing the capture probabilities to depend on covariate information or by allowing some of the registrations to be dependent.

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 catch ability. Item missing values are usually handled by imputation with a reasonable proxy (Zwane and Van der Heijden, 2004) or by excluding those observations (Hwang and Huang, 2003; Wang and Yip, 2003). The missing data problem is more acute when some of the registrations do not contain some of the covariates which define heterogeneous catch ability. 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 (Wang and Yip, 2003). A related problem is when the lists do not measure the same population (Zwane et al., 2004).

The Capture-Recapture Method is one of the most common method to estimate the size of an unknown population. This methodology was initially developed in ecology to estimate the size of wildlife populations. Animals were trapped, marked, and released on a number of occasions, and the individual trapping histories were then used to estimate the size of the whole population.

The first application to human populations data occurred in 1949 by Sekar and Deming. In this case, “being captured by the sample i” is replaced by “being included in the list i ”. In epidemiology the Capture-Recapture method is attempt to estimate or adjust for the extent of incomplete ascertainment using information from overlapping lists of cases from distinct sources.

This technique has been widely used to estimate the prevalence of drug users (see for example Frischer, 2001; Gemmell, 2004; Hope, 2005) and the number of people infected with the Human Immunodeficiency Virus (Abeni, 1994; Davies et al., 1999; Bartolucci and Forcina, 2006). Other areas of application include the estimation of deaths due to traffic accidents (Razzak and Luby, 1998), prostitution (Roberts and Brewer, 2006) and the prevalence of other diseases.

In a closed Capture-Recapture Model, we assume that there are no births, deaths or migrations, so that the population size is constant over trapping times. The demographic closure assumption is usually valid for data collected in a relatively short time. Traditionally, discrete-time capture-recapture models assume that the samples are independent, but in epidemiology lists dependence and heterogeneity (the behaviour component) are the norm and Log-Linear Models are particularly useful in modeling these phenomena (Schwarz and Seber, 1999).

MODEL FRAMEWORK

Capture-recapture models involve complex missing data mechanisms. Traditional approaches to inference focus on deriving the likelihood for the observed data (the ODL) by integrating over all missing data. Instead, we use the modeling framework of Schofield (2007), Schofield and Barker (2008, 2009) that uses data augmentation (Tanner and Wong 1987) to allow us to model in terms of the complete data likelihood (CDL). Similar ideas have also been proposed by Royle and Dorazio (2008). The likelihood we use for inference is in terms of the complete data, which for a capture-recapture study with individual-specific time-varying covariate data are the (i) times of birth, (ii) times of death, and (iii) complete covariate values for each individual ever available for capture. The main advantage of using this likelihood over the ODL is that we are able to focus on modeling the processes of interest rather than having to account for the complexities caused by missing data that result from sampling methods. Importantly, in adopting the CDL approach to inference, we do not need to make any additional assumptions to those made when using the ODL. It is simply a reformulation of the model in terms of the easier-to-understand CDL where we use computational algorithms, such as Markov chain Monte Carlo (MCMC) or the expectation-maximization (EM) algorithm, to integrate over all missing data.

THE MULTINOMIAL LOGIT, MODEL

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 $w_i = [i_1 i_2 \cdots i_S]$, which is a series of binary variables with 1 denoting ascertained and 0 otherwise. The ascertainment profile ivi can be redefined as a nominal categorical variable Y{ 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 x_i and z_i of length p and q respectively, where x_i are the covariates observed in all the registrations and z_i are the covariates not observed in all the registrations. Denoting the multinomial logit for individual i as $\eta_i' = [\eta_1(x_i, z_i), \eta_2(x_i, z_i), ..., \eta_K(x_i, z_i)]$, the category probabilities are then given by,

$$\mathbb{P}(Y_i = k | x_i, z_i) = \exp[\eta_k(x_i, z_i)] / \sum_{r=1}^{K} \exp[\eta_r(x_i, z_i)].$$

(1)

This model has to be constrained in some way for it to be used in the capture- recapture problem (Zwane and Van der Heijden, 2003, 2004). 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 leastonce. 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 (Huggins, 1989). Rather than use (1), the current standard is to use only the covariates observed in all lists, that is

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

(2)

Equation (2) will result in a biased estimate of the population size if the covariates in z_i are related to the inclusion probabilities. In this article we will complete the data set using the multiple imputation approach such that all covariates and lists are utilized.

MULTIPLE IMPUTATION IN THE CAPTURE-RECAPTURE PROBLEM

In this paper, we will briefly describe the idea of multiple imputation methods. Multiple imputation is now standard in statistical literature and thus we will highlight only the most important points (Rubin, 1996). 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.

Multiple imputation aims at imputing the missing values in **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 (Horton and Lipsiz, 2001). Below we highlight the features of each of these approaches and situations where they can be used for creating the multiple imputations.

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 variables forming the inclusion profiles are binary, but because they have no missings they can enter the model as continuous covariates (Schafer, 1997). 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.

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 to make the missing at random (MAR) assumption more plausible (Rubin, 1996). In some instances, especially in the general location model use of a large number of categorical covariates results in an unestimable model. Belin et al. (1999) 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.

METHODS

Capture-Recapture Method in a closed population:

The simplest capture-recapture model consists of two catches and can be set out in a 2x2 table. The goal is to estimate the number of subjects not caught in both the occasions (n_{00}). This number can be estimated using the information on

subjects captured in both samples and on subjects captured only in one sample, thus providing the total population size N.

The capture-recapture model requires that three assumptions are satisfied:

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

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

c) The two samples are independent. This assumption actually follows from assumption B since the latter implies that marked and unmarked individuals have the same probability of being caught in the second sample, so that the capture in the first sample does not affect the capture in the second sample.

If the three assumptions hold, then the estimated number of subjects not caught in both the occasions (n_{00}) is given by the well known Petersen-Estimator (or Dual-System Estimator):

$$\hat{n}_{00} = \frac{n_{10} n_{01}}{n_{11}} \tag{3}$$

and die resulting estimate of the total population size will be

$$\hat{N} = \hat{n}_{00} + n_{10} + n_{01} + n_{11}$$

Usually, the first assumption may be controlled by the researcher as it is sufficient to carry the two captures at a relatively short time. In contrast, the second and third assumption may not always be controlled because they are related to intrinsic characteristics of individuals belonging to the population.

In this case the estimate obtained by (3) will be distorted. For example, consider the situation where two groups of the same species have different sizes and hence the larger has a higher probability of being captured than the smaller.

Ignoring the size of the animal we violate the assumption B and hence the C, as we induce dependence between the two catches. When there are only two sources of capture, the information regarding covariates is available and the covariates may somehow affect the capture probability, a commonly used approach is to stratify the population by the covariates, to estimate the missing number in each strata by using the estimator (1) and then to pool these estimates to obtain the total population size.

Moreover, when two or more sources of capture are available, instead of stratifying according to the observed covariates, it is possible to handle the direct dependence between sources and to model the observed heterogeneity induced by covariates by using the Generalized Linear Model (GLM).

This class of models is certainly one of the most common in Epidemiology to solve problems in the capture-recapture field because it allows to treat in an easy way simultaneously both the dependence among sources and the heterogeneity.

Multi-Model approach:

To overcome difficulties regarding the selection of the best model a methodology known as multi-model estimation has been proposed in literature (Burn- ham and Anderson 2004) in order to mitigate the error we make assuming the existence of a single optimal model. It's based on a weighted average of those models having a maximum distance of 10, in terms of AIC or BIC values, from the model with minimum index.

Once the best model is selected according to AIC or BIC, the following difference $\Delta_i = AIC_i - AIC_{\min}$ is calculated for each i-th model, and all models with $\Delta_i > 10$ are excluded from the analysis.

Finally, a new estimate is calculated as weighted average of the estimates obtained by all models i with $\Delta_i \leq 10$ according to the following weights:

$$w_i = \frac{\exp(-\Delta_i / 2)}{\sum_{r=1}^{R} \exp(-\Delta_r / 2)} \tag{4}$$

where R is the total number of considered models.

Because it is impractical to evaluate all possible models, with increasing sources and/or covariates it's necessary to try different strategies for obtaining Multi- Model estimates:

(a) to evaluate all possible models without covariates effects and to select those with $\Delta_i \leq 10$ in terms of AIC or BIC values. To add and select covariates effects and, finally, to select the best model among the final models with covariates effects. Once checked again the distances Δ_i, a weighted average is calculated according to $\sum_{i=1}^{n} w_i \hat{N}_i$

(b) to evaluate all possible hierarchical models with covariates effects and to select those with $\Delta_i \leq 10$ to be used in Multi-Model estimates. In order to restrict the analysis to a smaller number of models, it can be convenient to evaluate covariates effects only on those models with a distance $\Delta_i \le 10$ when the covariates are without the model.

(c) to select directly the best model from all possible hierarchical models without covariates effects according to AIC or BIC values. Then, after adding and selecting the effects of the covariates, to use all models with a $\Delta_i \le 10$ into the weighted average.

CONCLUSION

For capture-recapture data without covariates, the usual log-linear model is equivalent to the logic model. Thus, either approach can be used in this case. The crude estimate of the population size can be seriously biased when there is dependence between the two sources. It is therefore necessary to examine the association of capture probability with measured covariates, and to include such covariates in capture-recapture analyses.

Both the log-linear model and the logit model can be extended to include the effects of categorical covariates. However, the need to include interaction terms and baseline probabilities means that this is more complicated for the log-linear model. For instance, with two covariates with four categories each, the logit model has 14 parameters, whereas the equivalent loglinear model has 30. While the logit model can include the effect of continuous covariates (10), there is no equivalent log-linear model.

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 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 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 measured by all registrations (or they contain missing data), the commonly used approaches of dropping (or ignoring) these covariates may give biased estimates of the population size. Multiple imputation is proposed to handle the missing covariate problem in the capture-recapture models.

Our results show that mean imputation also performs well with respect to the estimate of the population size but seemingly underestimates the standard error, resulting in narrow confidence intervals. The estimate of the population size from mean imputation is similar to the estimate derived from multiple imputation because the proportion of observations with missing data is very low.

Multiple imputation is applicable to missing covariate problems with arbitrary missing data patterns and arbitrary number of covariates (the categorical covariates do not necessarily have to be binary). Though our application is in epidemiology with only three lists this approach is applicable to wide ranging capture-recapture problems.

REFERENCES

- Agresti A. Simple capture-recapture models permitting unequal catchability and variable sampling effort Biometrics 1994;50:494-500.
- Alho, J., 1990. Logistic regression in capture–recapture models. Biometrics 46, 623–635.
- Brenner H. Uses and limitations of the capture-recapture method in disease monitoring with two dependent sources. Epidemiology 1995;6:42-8.
- Dupuis, J. and Schwarz, C. (2007), “A Bayesian approach to the multistate Jolly-Seber capture-recapture model,” Biometrics, 63, 1015–1022.
- Hook EB, Regal RR. Capture-recapture methods in epidemiology: methods and limitations. Epidemiol Rev 1995; 17: 243-64.
- Huggins, R. (1989), “On the Statistical Analysis of Capture Experiments,” Biometrika, 76, 133 – 140.
- International Working Group for Disease Monitoring and Forecasting, 1995. Capture–recapture and multiple record systems estimation 1: history and theoretical development. Amer. J. Epidemiology 142, 1047–1058.
- Link, W. A. and Barker, R. J. (2005), “Modeling Association among Demographic Parameters in Analysis of Open Population Capture-Recapture Data,” Biometrics, 61, 46 –54.
- Pollock, K., 2002. The use of auxiliary variables in capture–recapture modeling : an overview. J. Appl. Statist. 27, 85–102.
- Schouten LJ, Straatman H, Kiemeney LA, et al. The capturerecapture method for estimation of cancer

registry completeness: a useful tool? Int J Epidemiol 1994;23:1111-16.

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

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