Available at
www.ElsevierMathematics.com
POWERED BY SCIENCE DIRECT®

Statistics & Probability Letters 65 (2003) 121–125

STATISTICS & PROBABILITY LETTERS

www.elsevier.com/locate/stapro

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

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

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

Received February 2003; received in revised form March 2003; accepted July 2003

Abstract

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.

Keywords: Bootstrap; Population-size estimation; Capture–recapture; Multinomial logit model; Log-linear model

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 Garthwire, 1991). They presented both the non-parametric and parametric bootstrap for the case where there are no covariates.

* Corresponding author.
E-mail address: e.zwane@fss.uu.nl (E.N. Zwane).

0167-7152/$ - see front matter
doi:10.1016/j.spl.2003.07.010

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 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 3 presents a simulation to evaluate the properties of these methods.

2. Variance estimation methods

Assume that the true population size is N and the individuals are indexed by i $(i = 1, 2, \ldots, N)$ 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 $\mathbf{x_i}$. The inclusion profile for individual i (denoted by $\mathbf{w_i}$) 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 p_i (the estimated probability is denoted by $\hat{p}_i$). 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,

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

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.

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.

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

This 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 Garthwire (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.

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 $1(\Pi_1)$ and list $2(\Pi_2)$ 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 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" bootstrap. A comparison of the

Table 1
Coverage of the bootstrap methods

						Coverage	
N	A	Mean	Median	Minimum	Maximum	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

means and the medians in Table 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., Garthwire, 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., O'Brien, T.E., 1996. An application of profile-likelihood confidence interval to capture–recapture estimators. J. Agric. Biol. Envir. Statist. 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. Am. J. Epidemiol. 142, 1047–1058.

Norris, J., Pollock, K., 1996. Including model uncertainty in estimating variances in multiple capture studies. Environ. Eco. Statist. 3, 235–244.

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

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

Zwane, E., Van der Heijden, P., 2002. The multiple-system estimator in the presence of covariates. In: Stasinopoulos, M., Touloumi, G. (Eds.), 18th International Workshop on Statistical Modelling. Chania University, Chania, Greece, pp. 697–701.