Journal of Statistical Planning and Inference 176 (2016) 52–63

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Accounting for contamination and outliers in covariates for open population capture–recapture models

Jakub Stoklosa[a,*], Wen-Han Hwang[b], Paul S.F. Yip[c], Richard M. Huggins[d]

[a] School of Mathematics and Statistics and Evolution & Ecology Research Centre, The University of New South Wales, New South Wales 2052, Australia
[b] Department of Applied Mathematics, National Chung Hsing University, Taiwan
[c] Department of Social Work and Social Administration, The University of Hong Kong, Hong Kong
[d] Department of Mathematics and Statistics, The University of Melbourne, Victoria 3010, Australia

ARTICLE INFO

Article history:
Received 14 March 2016
Received in revised form 15 March 2016
Accepted 20 March 2016
Available online 1 April 2016

Keywords:
Cormack–Jolly–Seber model
Heterogeneity
Horvitz–Thompson estimator
Measurement error
Outliers
Robust statistics

ABSTRACT

In many capture–recapture experiments, covariates are collected on individuals and their inclusion in the study enhances the analysis. Typical examples of individual covariates include: gender, body weight, age, whether an individual was radio tagged, location strata, etc. To estimate open population sizes, McDonald and Amstrup (2001) used the ratio of the number of captured individuals divided by the estimated capture probabilities obtained from fitting the well-known Cormack–Jolly–Seber model (which permits modelling of both survival and capture probabilities as functions of individual and/or environmental covariates). However, this population estimator can result in bias or give unrealistically large values when contaminated covariates are used, *e.g.*, when outliers or false recordings are present, or due to measurement error in covariates. In this short note a new robust open population size estimator is proposed to account for outliers via a lower bound approach. These estimators are used on real data in the context of social sciences and ecology where capture–recapture experiments are commonly applied. The first case study concerns the population size of drug addicts in Hong Kong in 2004 collected by the Central Registry of Drug Abuse, and the second case study examines the population size estimation of the Yellow-bellied Prinia bird *Prinia flaviventris* also collected in Hong Kong in the Mai Po Bird Sanctuary in 1991. A simple simulation study is conducted to examine bias, robustness and efficiency.

1. Introduction

The modelling of population demographics such as population size or survival probabilities can be greatly enhanced through the use of covariates. In capture–recapture studies, such covariates are usually collected on individuals from the population, *e.g.*, gender, body weight, age, whether an individual was radio tagged or location strata. As discussed in McCrea and Morgan (2014), the use of covariates in such analysis helps for various reasons, they can: simplify models; control for heterogeneity, remove redundancy when present; absorb lack of fit; validate models; and explain variation and generate hypothesis for further study. Ideally, these covariates are measured error free with little (or ideally no) contaminated observations. Contaminated observations can occur in several ways. There can be contamination in the covariates or an

[*] Corresponding author. Tel.: +61 2 93854723.
E-mail address: j.stoklosa@unsw.edu.au (J. Stoklosa).

http://dx.doi.org/10.1016/j.jspi.2016.03.004
0378-3758/© 2016 Elsevier B.V. All rights reserved.

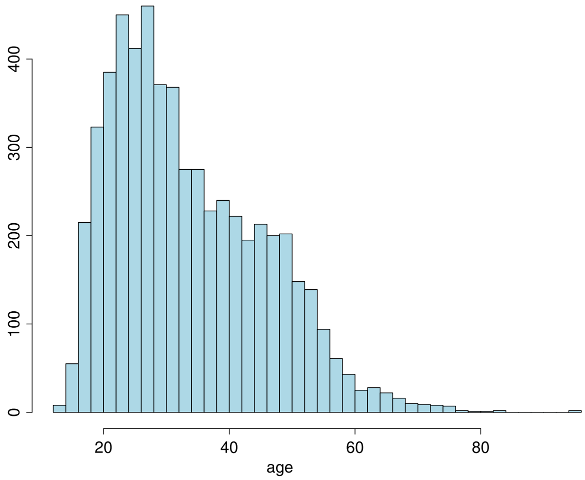

Fig. 1. Histogram of the *ages* of drug users detected in Hong Kong in 2004. Notice the skewness to the right due to two observations aged 95 and 96.

individual's capture status can be miss-recorded, say, through a simple mistake such as misreading a tag or even tag loss itself.

We are motivated by capture–recapture data collected on drug users by the Hong Kong Police Force in Hong Kong (Policy 21 Limited, 2013; Huggins et al., in press). These data are only a subset from a full data set containing daily observations of approximately 150,000 individuals from 1977 to 2011 collected by the Central Registry of Drug Abuse in Hong Kong. As in Huggins et al. (in press), we only considered observations from 2004, and rather than using daily occasions we aggregated the data to monthly occasions. This yields 5715 uniquely marked individuals for the twelve month period. The *age* of an individual was also recorded by the police department on capture. For this period, we only take the first observed age in years of an individual, and give the distribution of the observed ages using a histogram in Fig. 1. These data are very skewed to the right, notably due to two observations aged 95 and 96. These two recordings are unusually large and may be an artifact of the data, particularly since the third largest observed age was 84. If age were used as a covariate in the model, then such observations may impact the overall inference (see below). Rather than completely removing these observations or, say, log-transforming the covariate, we retain all values on the original scale in the overall analysis.

The Cormack–Jolly–Seber model (CJS, Lebreton et al., 1992; McCrea and Morgan, 2014) is a well-known open population capture–recapture model which permits for survival and capture probabilities to be modelled as functions of environmental covariates (Gimenez et al., 2006a; Stoklosa and Huggins, 2012a), individual/trait covariates (Gimenez et al., 2006b; Bonner and Schwarz, 2006) or both (Van Duesen, 2001; McDonald and Amstrup, 2001; Amstrup et al., 2005). Model parameters are usually estimated by maximizing a product multinomial likelihood which is derived from modelling the first recaptures of individuals that were captured and released on each occasion. A number of different software packages are available that have the CJS model (and its variants) implemented in their routines, we found that the `marked` R-package (Laake et al., 2013) is efficient and easy to use, particularly when dealing with covariates. The `marked` R-package uses an *Automatic Differentiation Model Builder* through a C++ language extension, making it very fast even for big data sets (*i.e.*, many capture occasions and observed individuals), see also the interactive website http://www.mbr-pwrc.usgs.gov/software/capture.html. We therefore used `marked` throughout this study although other R-packages, such as `mra` (McDonald, 2012) can alternatively be used. Generally, the CJS model is used for estimating and conducting inferences on survival probabilities, in this study we focus on the estimation of open population sizes across capture occasions. We follow McDonald and Amstrup (2001) and Section 9.5 of Amstrup et al. (2005) and use a Horvitz–Thompson type estimator (Horvitz and Thompson, 1952; Huggins, 1989); throughout we denote this as the MA (open population) estimator. The MA estimator is a function of capture probability estimates that are directly obtained from the fitted CJS model. It is these capture probabilities that are modelled via covariates.

For any analysis that involves modelling with covariates, statistical inference crucially relies on accurate readings/measurements. There are two effects of outlying covariates on MA estimation. First, if contaminated observations or measurement error are present and ignored then the analysis can result in biased parameter estimates and misleading inferences (Huber, 1981; Carroll et al., 2006). Estimation in the CJS model is based on a product multinomial distribution and there are several approaches to robust inference in the multinomial setting (Mebane and Sekhon, 2004; Tabatabai et al., 2014) that potentially may be adapted but we do not consider this further here. Secondly and more importantly, no matter how the parameters are estimated, if a covariate is involved in the model for the capture probabilities then an outlying value of this covariate will produce an outlying predicted capture probability. This is crucial for the MA estimator because it is based on the sum of the reciprocals of fitted capture probabilities. That is, if the predicted capture probabilities are too small then the MA estimator and its standard error becomes unrealistically large. In this paper a robust MA-type open population size estimator is proposed. This can also be seen as an extension to the work of Stoklosa and Huggins (2012b) but in the