Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. doi: 10.18637/jss.v000.i00

Single-Source Capture-Recapture Models With singleRcapture

Piotr Chlebicki
Stockholm University

Maciej Beręsewicz
Poznań University of Economics and Bisiness

Abstract

Estimating population size is an important issue in official statistics, social sciences and natural sciences. One way to approach this problem is to use capture-recapture methods, which can be classified according to the number of sources used, the main distinction being between methods based on one source and those based on two or more sources. In this presentation we will introduce the **singleRcapture** R package for fitting SSCR models. The package implements state-of-the-art models as well as some new models proposed by the authors (e.g. extensions of zero-truncated one-inflated and one-inflated zero-truncated models). The software is intended for users interested in estimating the size of populations, particularly those that are difficult to reach or for which information is available from only one source and dual/multiple system estimation cannot be used.

Keywords: population size estimation, truncated distributuons, count regression models, R.

1. Introduction

1.1. Literature review

This work is supported by the National Science Center, OPUS 20 grant no. 2020-/-39/-B/-HS4/-00941 *Towards census-like statistics for foreign-born populations – quality, data integration and estimation*

The subject of this workshop is the **singleRcapture** package and its lightweight extension that allows for integration with other R packages called **singleRcaptureExtra**.

The package is available on CRAN: `CRAN.R-project.org/package=singleRcapture` while the extension is available on: `https://github.com/ncn-foreigners/singleRcaptureExtra`.

The **singleRcapture** package is an R language package that focuses on implementing state of the art methods for frequentist point and interval estimation of size of closed populations in single-source capture-recapture (SSCR) setting (e.g. estimation of the population size of irregular migrants at set time point in a given area).

1.2. How do we estimate population size with only one register

1.3. Example with R code

Installation:

```
R> install.packages("singleRcapture")
R> remotes::install_github("https://github.com/ncn-foreigners/singleRcaptureExtra")
```

```
R> library(singleRcapture)
R>
R> head(netherlandsimmigrant)
```

```
  capture gender    age       reason       nation
1       1   male <40yrs Other reason North Africa
2       1   male <40yrs Other reason North Africa
3       1   male <40yrs Other reason North Africa
4       1   male <40yrs Other reason         Asia
5       1   male <40yrs Other reason         Asia
6       2   male <40yrs Other reason North Africa
```

```
R> basicModel <- estimatePopsize(
+   formula = capture ~ gender + age + nation,
+   model   = ztpoisson(),
+   data    = netherlandsimmigrant
+ )
R>
R> summary(basicModel)
```

```
Call:
estimatePopsize.default(formula = capture ~ gender + age + nation,
    data = netherlandsimmigrant, model = ztpoisson())

Pearson Residuals:
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-0.486442 -0.486442 -0.298080  0.002093 -0.209444 13.910844

Coefficients:
-----------------------
For linear predictors associated with: lambda
```

```
                     Estimate Std. Error z value  P(>|z|)
(Intercept)           -1.3411     0.2149  -6.241 4.35e-10 ***
gendermale             0.3972     0.1630   2.436 0.014832 *
age>40yrs             -0.9746     0.4082  -2.387 0.016972 *
nationAsia            -1.0926     0.3016  -3.622 0.000292 ***
nationNorth Africa     0.1900     0.1940   0.979 0.327398
nationRest of Africa  -0.9106     0.3008  -3.027 0.002468 **
nationSurinam         -2.3364     1.0136  -2.305 0.021159 *
nationTurkey          -1.6754     0.6028  -2.779 0.005445 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

AIC: 1712.901
BIC: 1757.213
Residual deviance: 1128.553

Log-likelihood: -848.4504 on 1872 Degrees of freedom
Number of iterations: 8
-----------------------
Population size estimation results:
Point estimate 12690.35
Observed proportion: 14.8% (N obs = 1880)
Std. Error 2808.169
95% CI for the population size:
          lowerBound upperBound
normal      7186.444   18194.26
logNormal   8431.275   19718.32
95% CI for the share of observed population:
          lowerBound upperBound
normal     10.332927   26.16037
logNormal   9.534281   22.29793
```

2. Detailed information

2.1. Fitting method

2.2. Avaiable models

The full list of implemented models in **singleRcapture** along with the expressions for probability density functions and point estimates is found in the collective help file for all family functions:

```
R> ?ztpoisson
```

Here we limit ourselves to just listing the family functions:

- Zero-truncated and zero-one-truncated Poisson, geometric, NB type II regression where the untruncated distribution is parameterized as:

$$\mathbb{P}[Y=y|\lambda,\alpha]=\frac{\Gamma\left(y+\alpha^{-1}\right)}{\Gamma\left(\alpha^{-1}\right)y!}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\lambda}\right)^{\alpha^{-1}}\left(\frac{\lambda}{\lambda+\alpha^{-1}}\right)^{y}.$$

- Zero-truncated one-inflated (ztoi) modifications distributions where the new probability $\mathbb{P}^*$ measure is defined in terms of count data measure $\mathbb{P}$ with support on $\mathbb{N}\cup\{0\}$ as:

$$\mathbb{P}^*[Y=y]=\begin{cases}\mathbb{P}[Y=0] & y=0,\\ \omega\left(1-\mathbb{P}[Y=0]\right)+(1-\omega)\mathbb{P}[Y=1] & y=1,\\ (1-\omega)\mathbb{P}[Y=y] & y>1,\end{cases}$$

$$\mathbb{P}^*[Y=y|Y>0]=\omega 1_{\{1\}}(y)+(1-\omega)\mathbb{P}[Y=y|Y>0].$$

- One-inflated zero-truncated (oizt) modifications distributions where the new probability $\mathbb{P}^*$ measure is defined as:

$$\mathbb{P}^*[Y=y]=\omega 1_{\{1\}}(y)+(1-\omega)\mathbb{P}[Y=y],$$

$$\mathbb{P}^*[Y=y|Y>0]=\omega\frac{1_{\{1\}}(y)}{1-(1-\omega)\mathbb{P}[Y=0]}+(1-\omega)\frac{\mathbb{P}[Y=y]}{1-(1-\omega)\mathbb{P}[Y=0]}.$$

- Generalized Chao's and Zelterman's estimators via logistic regression on variable Z defined as $Z=1$ if $Y=2$ and $Z=0$ if $Y=1$ with $Z\sim b(p)$ where $\operatorname{logit}(p)=\ln(\lambda/2)$ for poisson parameter λ,

$$\hat{N}=N_{obs}+\sum_{k=1}^{\boldsymbol{f}_1+\boldsymbol{f}_2}\left(2\exp\left(\boldsymbol{x}_k\hat{\boldsymbol{\beta}}\right)+2\exp\left(2\boldsymbol{x}_k\hat{\boldsymbol{\beta}}\right)\right)^{-1}, \qquad \text{(Chao's estimator)}$$

$$\hat{N}=\sum_{k=1}^{N_{obs}}\left(1-\exp\left(-2\exp\left(\boldsymbol{x}_k\hat{\boldsymbol{\beta}}\right)\right)\right)^{-1}. \qquad \text{(Zelterman's estimator)}$$

- Alternative approaches to modelling one-inflation that mimic hurdle models where the first type zero truncated hurdle model (ztHurdle) is defined as:

$$\mathbb{P}^*[Y=y]=\begin{cases}\frac{\mathbb{P}[Y=0]}{1-\mathbb{P}[Y=1]} & y=0,\\ \pi(1-\mathbb{P}[Y=1]) & y=1,\\ (1-\pi)\frac{\mathbb{P}[Y=y]}{1-\mathbb{P}[Y=1]} & y>1,\end{cases}$$

$$\mathbb{P}^*[Y=y|Y>0]=\pi 1_{\{1\}}(y)+(1-\pi)1_{\mathbb{N}\setminus\{1\}}(y)\frac{\mathbb{P}[Y=y]}{1-\mathbb{P}[Y=0]-\mathbb{P}[Y=1]}$$

- The Hurdle zero truncarted (Hurdlezt) is defined as:

$$\mathbb{P}^*[Y=y]=\begin{cases}\pi & y=1,\\ (1-\pi)\frac{\mathbb{P}[Y=y]}{1-\mathbb{P}[Y=1]} & y\neq 1,\end{cases}\quad \mathbb{P}^*[Y=y|Y>0]=\begin{cases}\pi\frac{1-\mathbb{P}[Y=1]}{1-\mathbb{P}[Y=0]-\mathbb{P}[Y=1]} & y=1,\\ (1-\pi)\frac{\mathbb{P}[Y=y]}{1-\mathbb{P}[Y=0]-\mathbb{P}[Y=1]} & y>1.\end{cases}$$

Structure of a family function

2.3. Marginal frequencies

A popular method of testing the model fit in single source capture-recapture studies is comparing the fitted marginal frequencies $\sum_{j=1}^{N_{obs}} \hat{\mathbb{P}}\left[Y_j = k|\boldsymbol{x}_j, Y_k > 0\right]$ with the observed marginal frequencies $\sum_{j=1}^{N} 1(Y_k = k) = \sum_{j=1}^{N_{obs}} 1(Y_k = k)$ for $k \geq 1$.
If a fitted model bears sufficient resemblance to the real data collection process these quantities should be quite close.

Affiliation:

Piotr Chlebicki
Stockholm University
Matematiska institutionen
106 91 Stockholm, Albano hus 1
E-mail: piotr.chlebicki@math.su.se
URL: https://github.com/Kertoo

Maciej Beręsewicz
Poznań University of Economics
and Bisiness
Poznań University of Economics and Business
Department of Statistics
Institute of Informatics and Quantitative Economics
Al. Niepodległosci 10
61-875 Poznań, Poland
E-mail: maciej.beresewicz@ue.poznan.pl

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
MMMMMM YYYY, Volume VV, Issue II *Submitted:* yyyy-mm-dd
doi:10.18637/jss.v000.i00 *Accepted:* yyyy-mm-dd