Estimation of hidden populations using `singleRcapture` package

Chlebicki Piotr[1] **& Beręsewicz Maciej**[2,3]

(1) Adam Mickiewicz University Poznań, Poland
(2) Department of Statistics, Poznań University of Economics and Business, Poland
(3) Centre for Small Area Estimation, Statistical Office in Poznań, Poland

SCAN ME

Introduction

Population size estimation is an important issue in official statistics, social and natural sciences. One way to tackle this problem is by applying capture-recapture methods, which can be classified depending on the number of sources used, i.e. one or two and more sources.

In this `R` package, we focus on the first group of methods, i.e. single-source capture-recapture (SSCR). SSCR models assume that observed counts follow some positive count distribution (e.g. zero-truncated Poisson, one-inflated zero-truncated geometric). This assumption is used to estimate missing (hidden) zero counts.

We present a new `R` package called `singleRcapture`, which implements state-of-the-art SSCR models with user-friendly functions.

The package is currently in development and has not yet been submitted to `CRAN` but is already usable and presents functionalities not currently present in any other `R` packages.

Single source capture - recapture

Most of the details pertaining to the statistical models used in and developed for the package are available on the package website which can be accessed from the repository. In short, let Y_k represent the number of times k-th unit was observed in source data. Clearly, we don not know how often $Y_k = 0$ and to find the total population size N we need to estimate it.

In general, we assume that conditional distribution of Y_k given a vector of covariates $\mathbf{x}_k$ follows some version of truncated distribution such as zero truncated Poisson/geometric/negative binomial or any of their modifications

$$Y_k|\mathbf{x}_k \sim \text{ZTP}(\lambda_k) \quad \text{or} \quad Y_k|\mathbf{x}_k \sim \text{ZTOIP}(\lambda_k, \omega_k) \quad \text{or any other modification,}$$

knowing the values of λ and ω we may estimate the population size using Horwitz-Thompson type estimator:

$$\hat{N} = \sum_{k=1}^{N} \frac{I_k}{\mathbb{P}(Y_k > 0|\mathbf{x}_k, \lambda_k, \omega_k)},$$

and maximum likelihood estimate of N is obtained after substituting regression estimates for λ_k, ω_k into the equation above. Our package allows for several types of standard error estimates and their full description is available in the help page for the main function `singleRcapture::estimatePopsize`.

R implementation

Example code for the model from publication: van der Heijden, Bustami, Cruyff, Engbersen & van Houwelingen (2003).

The syntax is very similar to `stats::glm` which is supposed to ensure ease of writing code using `singleRcapture`.

```
library(singleRcapture)
model <- estimatePopsize(
  formula = capture ~ gender + age + nation + reason,
  data = netherlandsimmigrant, #dataset included in package
  popVar = "analytic", #specify variance
  model = "ztpoisson", #distribution
  method = "IRLS" #fitting method one of two currently supported
)
summary(model)
```

We most important part of the output is containing the population size estimation results we present below (the full `glm` like the output of `summary` is available using QR code)

```
#> Population size estimation results:
#> Point estimate 12691.45
#> Observed proportion: 14.8% (N obs = 1880)
#> Std. Error 2809.508
#> 95% CI for the population size:
#>           lowerBound upperBound
#> normal      7184.917   18197.99
#> logNormal   8430.749   19723.38
#> 95% CI for the share of observed population:
#>           lowerBound upperBound
#> normal     10.330814   26.16592
#> logNormal   9.531836   22.29932
```

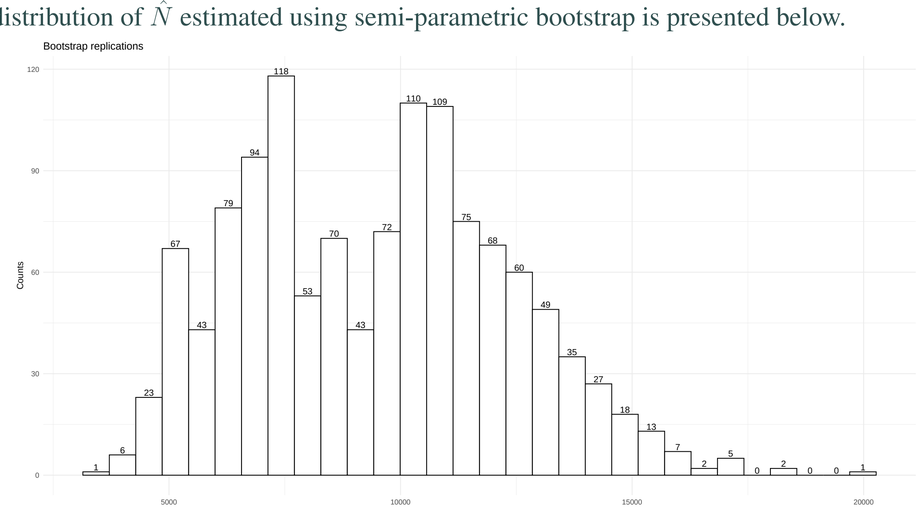
Example of a more complex model here we use one-inflated zero-truncated (OIZT) model with both parameters λ and ω depending on covariate information (cf. Godwin & Böhning, 2017). Symbolically model can be written as:

$$\begin{aligned}\lambda &= \beta_{00} + \beta_1 \cdot I(\text{age} > 40) + \beta_{21} \cdot I(\text{nation = Asia})\beta_{22} \cdot I(\text{nation = North Africa}) \\ &\quad + \beta_{23} \cdot I(\text{nation = Rest of Africa}) + \beta_{24} \cdot I(\text{nation = Suriname}) + \beta_{25} \cdot I(\text{nation = Turkey}) \\ \omega &= \beta_{01} + \beta_3 \cdot I(\text{gender = Male})\end{aligned}$$

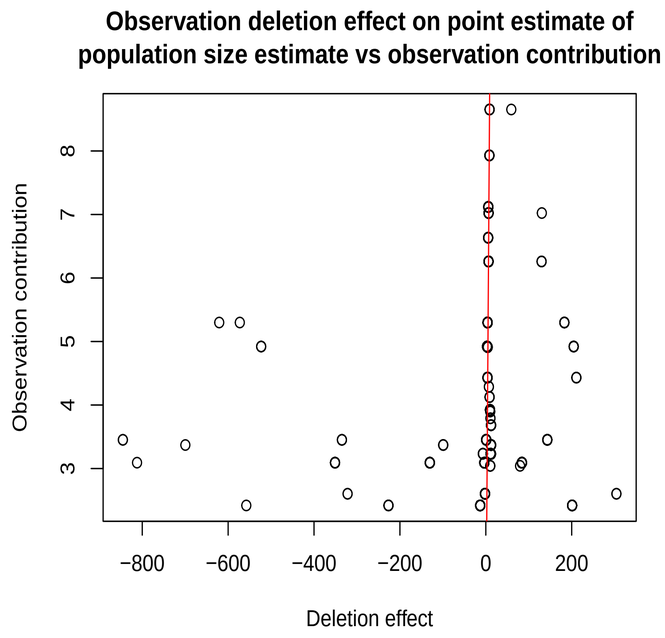
Such a model can be fitted in `singleRcapture` using the syntax:

```
set.seed(123)
modelInflated2 <- estimatePopsize(
    # Formula for lambda
    formula = capture ~ nation  + age,
    data = netherlandsimmigrant,
    # Construct confidence intervals using bootstrap
    popVar = "bootstrap",
    model = "oiztgeom",
    method = "IRLS",
    # technical information regarding the fitting procedure
    controlMethod = controlMethod(stepsize = .2),
    # information pretaining to pop size estimation
    controlPopVar = controlPopVar(
        B = 1250, # number of bootstrap replications
        alpha = .01,# confidence level
        bootType = "semiparametric",
        covType = "Fisher" # use fisher information matrix
    ),
    # put covariates on omega i.e. the inflation parameter
    controlModel = controlModel(omegaFormula = ~ gender)
)
```

The distribution of $\hat{N}$ estimated using semi-parametric bootstrap is presented below.



One type of plots available in `singleRcapture` is the contribution/deletion effect plot which is useful for determining the existence of influential observations. Using this plot we compare how the population size estimate should change if we omit one unit from the data set but regression coefficients remain the same and how it actually changes. Here by looking at this plot we clearly see that there are influential observations in the `netherlandsimmigrant` data.



Further research and development

We plan on developing SSCR methods that incorporate more theoretical advancements in glm's and vglm's such as random effects models, bias reduction in MLE estimates for regression coefficients robust regression (with sandwich type estimates already being implemented) and implementing Bayesian methods already described in the literature for example in Tuoto, Di Cecco & Tancredi (2022).

References (selected)

- Böhning, D., Bunge, J., & van der Heijden, P. G. (Eds.). (2018). Capture-recapture methods for the social and medical sciences. Boca Raton: CRC Press.
- Godwin, R. T., & Böhning, D. (2017). Estimation of the population size by using the one-inflated positive Poisson model. Journal of the Royal Statistical Society. Series C (Applied Statistics), 425-448.
- van der Heijden, P. G., Bustami, R., Cruyff, M. J., Engbersen, G., & van Houwelingen, H. C. (2003). Point and interval estimation of the population size using the truncated Poisson regression model. Statistical Modelling, 3(4), 305-322.
- Tuoto, T., Di Cecco, D., & Tancredi, A. (2022). Bayesian analysis of one-inflated models for elusive population size estimation. Biometrical Journal, 64, 912– 933. https://doi.org/10.1002/bimj.202100187

Acknowledgements

This work was supported by the National Science Center grant *Towards census-like statistics for foreign-born populations – quality, data integration and estimation* (NCN OPUS 22 2020/39/B/HS4-/00941).