singleRcapture – an R package for single-source capture-recapture models

Piotr Chlebicki
Adam Mickiewicz University
Maciej Beręsewicz
Poznań University of Economics and Business
Statistical Office in Poznań

20.07.2023

Outline

1. Introduction
2. The main functionalities
3. Methods for `singleR` class and useful functions
4. Summary and future development

Introduction

- This work is supported by the National Science Center, OPUS 22 grant no. 2020/39/B/HS4/00941 *Towards census-like statistics for foreign-born populations – quality, data integration and estimation*.
- In this study I would like to present the results of about one and a half years of development on `singleRcapture` package for single-source capture-recapture modeling and show how to use the package.
- The package is available on CRAN: `CRAN.R-project.org/package=singleRcapture`
- Current development version is always on our GitHub: `github.com/ncn-foreigners/singleRcapture`

Estimating population size with only one register

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

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

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

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

and maximum likelihood estimate of N is obtained after substituting regression estimates for λ_k, ω_k into the equation above.

Why we made `singleRcapture`

- There are plenty of R packages that allow to fit zero-truncated models (i.e. `extraDistr`, `countreg`, `VGAM`), but the choices for modified zero truncated distributions are sparse.
- There are plenty of R packages to fit capture-recapure models based on two and more sources (e.g. `Rcapture`, `RMark`).
- However, there are no packages that allow to easily deal with single-source capture-recapture (SSCR) models.
- The goal of this package is to make the models proposed in the literature available for wider audience.
- The package should provide general function that allows to fit various of SSCR models and enable assessing their quality.

Available models

- zero-truncated Poisson, geometric, NB type II regression (including modelling dispersion parameter),
- zero-one truncated Poisson, geometric, NB type II regression (including modelling dispersion parameter),
- zero-truncated one-inflated (ztoi) and one-inflated zero-truncated (oizt) Poisson, geometric, NB type II regression (including modelling dispersion parameter and inflation parameter).
- alternative approach to model one-inflation that mimic hurdle models, paper describing and comparing them to other methods is in the making,
- Chao's and Zelterman's regression.

Main function

The main function that our package is built around is `estimatePopsize`. The leading design principle was to make using `estimatePopsize` as close to standard `stats::glm` as possible.
Some of the parameters that `estimatePopsize` allows for are:

- `formula` – The main formula (i.e for λ parameter).
- `data, weights, subset, modelFrame, x, y` – Analogous to `glm`.
- `model` - either a function a string or a family class object specifying which model should be used possible values are listed in documentation.
- `method` – Numerical method used to fit regression `IRLS` or `optim`.
- `popVar` – A method for estimating variance of $\hat{N}$ and interval estimation.
- `controlMethod, controlModel, controlPopVar` - control parameters for numerical fitting, specifying additional formulas (inflation, dispersion) and population size estimation respectively.

Standard call for estimatePopsize

Simple model from P. G. v. d. Heijden et al. 2003:

```
model <- estimatePopsize(
  formula = capture ~ gender + age + nation,
  data = netherlandsimmigrant,
  popVar = "analytic",
  model = "ztpoisson",
  # optional
  method = "IRLS"
)
```

the netherlandsimmigrant data frame is exported by the package and contains data on unregistered immigration in four cities in Netherlands ($N_{obs} = 1880$), collected in 1995.

An example of a much more complicated call:

```
modelInflated <- estimatePopsize(
    formula = capture ~ age,
    data = netherlandsimmigrant,
    popVar = "bootstrap",
    model = oiztgeom(omegaLink = "cloglog"),
    method = "IRLS",
    controlPopVar = controlPopVar(
        B = 500, alpha = .01,
        bootType = "semiparametric",
        bootstrapFitcontrol = controlMethod(
          epsilon = 2.220446e-16,
          silent = TRUE,
          stepsize = 2
        )
    ),
    controlModel = controlModel(omegaFormula = ~ gender)
)
```

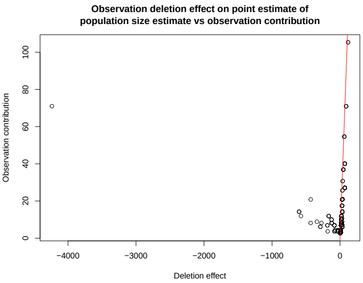
Standard output via summary

```
> summary(model)
#> 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
```

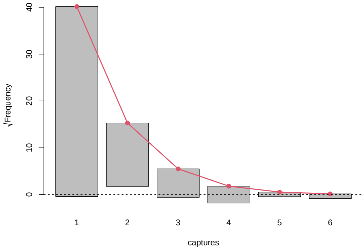
```
#> AIC: 1712.901
#> BIC: 1757.213
#> Residual deviance: 1128.553
#>
#> Log-likelihood: -848.4504
#> on 1872 Degrees of freedom
#> Number of iterations: 8
#> -----------------------
#> Population size estimation results:
#> Point estimate 12690.35
#> Observed proportion: 14.8% (N obs = 1880)
#> Std. 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
```

Diagnostic plots

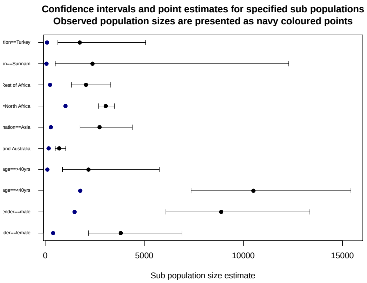
Quick diagnostic plots obtained by calling `plot(model, plotType = )`



(a) "dfpopContr"



(b) "rootogram"



(c) "strata"

Control parameters in singleRcapture

The `estimatePopsize` function accepts (up to) three different lists as control parameters allowing for great deal of customizability at call. These parameters are `controlPopVar`, `controlModel` and `controlMethod`, for each of those there is a documented function that describes what effect changing any available parameter makes. These control parameters allow in particular for:

- Specifying which of three distinct bootstrap sampling algorithms will be used.
- Allowing for linear dependence of all regression parameters on covariates.
- To control not only numerical fitting to a great degree (which is sadly often needed in SSCR since available data is usually quite sparse) but also to do the same in bootstrap (with possibly different parameters).

Auxiliary functions in the package

We built some custom functions that take `singleR` class object (returned by `estimatePopsize`) that help with assessing the fit of regression models and performing various post-hoc procedures:

- `marginalFreq` – A function that computes observed and fitted marginal counts, there is a `summary` method for objects returned by this function that performs goodness of fit tests.
- `redoPopEst` – Function for recalculating $\hat{N}$, $\text{Var}(\hat{N})$ etc.
- `dfpopsize` – Works analogously to `dfbeta` but for $\hat{N}$ instead of regression parameters.
- `stratifyPopsize` – For estimating sizes of user specified sub-populations.

Methods for `singleR` class

- Two most used methods for this class are `summary` and `plot` which were already presented.
- Standard methods in regression packages like `hatvalues`, `dfbeta`, `residuals`, `vcov`, `predict`, `cooks.distance`, `confint` etc. are implemented.
- Compatibility with `sandwich` package via methods for `estfun`, `vcovHC`, `bread` is assured.
- Most internal methods such as `nobs`, `AIC`, `BIC` so most of custom methods for `glm` and other regression objects, such as `textreg::screenreg` and `modelsummary::modelsummary`, should work reasonably well (otherwise we can make appropriate methods).

Plans for future development of `singleRcapture`

- Compatibility with `VGAM` and `countreg` packages via methods for appropriate classes.
- Methods for bias reduction for all implemented distributions (via `brglm2`).
- Implementing estimation via full (i.e. non-truncated) likelihood function.
- Robust (or resistant) regression/estimates.
- Implementation of Bayesian SSCR models.
- Performance improvements (in bootstrap).

Summary

- In this presentation a basic structure of syntax and properties of `singleRcapture` were explored.
- While developing the package we focused on making the package user-friendly.
- Basic syntax used in the package should feel familiar to active R users (similarity to `glm`).
- The package is capable of performing most of the usual workload that comes with analysing SSCR data.
- User suggestions/requests can be submitted to issues page on package GitHub page: `https://github.com/ncn-foreigners/singleRcapture/issues`.

SCAN ME

Selected literature I

- Heijden, Peter GM van der et al. (2003). "Point and interval estimation of the population size using the truncated Poisson regression model". In: *Statistical Modelling* 3.4, pp. 305–322.
- van der Heijden, Peter GM, Maarten Cruyff, and Hans C Van Houwelingen (2003). "Estimating the size of a criminal population from police records using the truncated Poisson regression model". In: *Statistica Neerlandica* 57.3, pp. 289–304.
- Cruyff, Maarten J. L. F. and Peter G. M. van der Heijden (2008). "Point and Interval Estimation of the Population Size Using a Zero-Truncated Negative Binomial Regression Model". In: *Biometrical Journal* 50.6, pp. 1035–1050.

Selected literature II

Böhning, Dankmar et al. (2013). "A Generalization of Chao's Estimator for Covariate Information". In: *Biometrics* 69.4, pp. 1033–1042.

Godwin, Ryan T. and Dankmar Böhning (2017). "Estimation of the population size by using the one-inflated positive Poisson model". In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 66.2, pp. 425–448.