Estimating the population size of drunk-drivers in Poland

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

06.09.2023

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*.
- The goal of this work is to present new methods for modelling one-inflation in single-source capture-recpature (SSCR) studies and with their applications to estimating the size of drunk-driving in Poland.
- All relevant calculations were done using R package `singleRcapture` version 0.2.0.1 (0.2.0 on CRAN newer version on GitHub) and codes are publicly available on GitHub repository ncn-foreigners/paper-one-inflation

Estimating population size with only one register

Let Y_k represent the number of times k-th unit was observed in source data. 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

$$Y_k|\mathbf{x}_k \sim \text{ZTPoisson}(\lambda_k) \quad \text{or} \quad Y_k|\mathbf{x}_k \sim \text{ZTNegBinomial}(\lambda_k, \alpha_k) \quad \text{or any other distribution,}$$

Estimate the population size using Horwitz-Thompson type estimator:

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

after substituting regression estimates for λ_k, ω_k into the equation above.

Assumptions

Basic SSCR models assume that:

1. Population is closed in relevant time period
2. Observations of distinct units are independent
3. Unobserved part of the population may be modelled using observed counts
4. There is no dependence between subsequent counts

We focus on the fourth one, since in practice it is most often violated, for such reason such as e.g.:

- An observation of a unit may prevent this unit from being observed another time (e.g. police apprehension resulting in jail time) or
- A change in unit behaviour may occur as a result of observation (e.g. driver's licence revocation)

Let f_Y denote PMF of 'base' distribution the (corresponding) one-inflated distributions can be described as:

Zero truncated one inflated distribution. (ZTOI)
Full distribution ($\mathbb{P}(Y = y)$):

$$\begin{cases} f_y(\cdot) & y = 0, \\ \omega \{1 - f_0(\cdot)\} + (1-\omega) f_y(\cdot) & y = 1, \\ (1-\omega) f_y(\cdot) & y > 1 \end{cases}$$

Truncated distribution ($\mathbb{P}(Y = y | Y > 0)$):

$$\begin{cases} \omega + (1-\omega) f_{y|y>0}(\cdot) & y = 1, \\ (1-\omega) f_{y|y>0}(\cdot) & y > 1 \end{cases}$$

Population size estimate

$$\hat{N}^{(ztoi)} = \sum_{k=1}^{N_{obs}} \frac{1}{1 - f_y(0)}$$

One inflated zero truncated distribution. (OIZT)
Full distribution ($\mathbb{P}(Y = y)$):

$$\begin{cases} (1-\omega) f_y(\cdot) & y = 0, \\ \omega + (1-\omega) f_y(\cdot) & y = 1, \\ (1-\omega) f_y(\cdot) & y > 1 \end{cases}$$

Truncated distribution ($\mathbb{P}(Y = y | Y > 0)$):

$$\begin{cases} \frac{\omega + (1-\omega) f_y(\cdot)}{1 - (1-\omega) f_y(0)} & y = 1, \\ \frac{1}{1 - (1-\omega) f_y(0)} (1-\omega) f_y(\cdot) & y > 1 \end{cases}$$

Population size estimate

$$\hat{N}^{(oizt)} = \sum_{k=1}^{N_{obs}} \frac{1}{1 - (1-\omega) f_y(0)}$$

Zero truncated hurdle distribution. (ZTHURDLE)
Full distribution ($\mathbb{P}(Y = y)$):

$$\begin{cases} \frac{f_y(\cdot)}{1-f_y(1)} & y = 0, \\ \pi(1 - f_y(1)) & y = 1, \\ (1-\pi)\frac{f_y(\cdot)}{1-f_y(1)} & y > 1 \end{cases}$$

Truncated distribution ($\mathbb{P}(Y = y|Y > 0)$):

$$\begin{cases} \pi & y = 1, \\ (1-\pi)\frac{f_y(\cdot)}{1-f_y(0)-f_y(1)} & y > 1 \end{cases}$$

Population size estimate

$$\hat{N}^{(ztHr)} = \sum_{k=1}^{N_{obs}} \frac{1 - f_y(1)}{1 - f_y(0) - f_y(1)}$$

Hurdle zero truncated distribution. (HURDLEZT)
Full distribution ($\mathbb{P}(Y = y)$):

$$\begin{cases} (1-\pi)\frac{f_y(\cdot)}{1-f_y(1)} & y = 0, \\ \pi & y = 1, \\ (1-\pi)\frac{f_y(\cdot)}{1-f_y(1)} & y > 1 \end{cases}$$

Truncated distribution ($\mathbb{P}(Y = y|Y > 0)$):

$$\begin{cases} \pi\frac{1-f_y(1)}{1-(1-\pi)f_y(0)-f_y(1)} & y = 1, \\ (1-\pi)\frac{f_y(\cdot)}{1-(1-\pi)f_y(0)-f_y(1)} & y > 1 \end{cases}$$

Population size estimate

$$\hat{N}^{(Hrzt)} = \sum_{k=1}^{N_{obs}} \frac{1 - f_y(1)}{1 - (1-\pi)f_y(0) - f_y(1)}$$

Differences

- Whilst hurdle model naturally includes one-deflation including one-deflation in inflated models (by allowing ω parameter to be negative) leads to many numerical issues
- It has been argued that failure to account for exposure may lead to both one-inflation as well as one-deflation
- Therefore hurdle models are more robust with respect to not observing exposure than one-inflated models
- It's worth noting that OIZT and HURDLEZT models often lead to very similar population size estimates
- On the other hand interpretation of model parameters in one-inflated models is more convenient

Data provided by polish police

- We excluded under 18 year olds as we are only interested in people with possibility of having a valid drivers licence
- We only operate on data up to 20th of October 2022 with covariates in data frame being:
  - Citizenship coded as (Polish, Ukrainian or other)
  - Previous offences (levels: No criminal history, Similar offences, Different offences, No knowledge)
  - Gender (with only 2 levels)
  - Age (here just 2022 - year of birth to be clearly well defined)

Captures	1	2	3	4	5	6	9	$\sum$
Count	43256	626	145	23	7	4	1	44062

Table: Counts for each number of each apprehensions

- We have chosen hurdle zero-truncated geometric with Poisson parameter being homogeneous
- logit(π) being dependent on:
  - previous offences, gender and
  - 3rd degree orthogonal polynomial created from `age` (via R function `poly`)
- population size was estimated to be $\hat{N} = 56308$
- observed proportion was estimated to be 78.3%
- For comparison according to CEPIK[a] about 22 milion people had driving licence in 2021 so about 0.255% of drivers in Poland are included in this population (according to this model)

[a] Central Register of Vehicles and Drivers

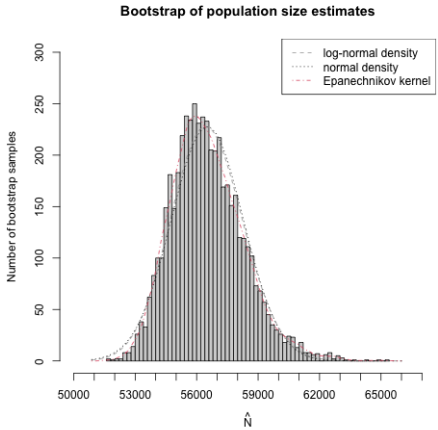

Figure: Pseudo sample from parametric bootstrap

Conf	L/H	Est	L/H	Share
95%	53534	60489	72.8%	82.3%
99%	52825	62352	70.7%	83.4%
99.9%	51999	64032	68.8%	84.7%
99.99%	51706	65115	67.7%	85.2%

Table: Bootstrap confidence intervals for population size estimates and observed proportion by confidence level

Coefficient	β	σ	z	$\mathbb{P}(>\|z\|)$
$\lambda-$Intercept	-1.23	0.07	-16.6	≈ 0
$\pi-$Intercept	1.4	0.25	5.69	$\approx 10^{-8}$
Male	-0.55	0.19	-2.88	0.004
age^1	-1.08	8.47	-0.13	0.899
age^2	31.76	8.32	3.82	$\approx 10^{-4}$
age^3	-20.91	7.93	-2.64	0.008
No criminal history	0.96	0.13	7.58	≈ 0
No knowledge	0.27	0.12	2.30	0.022
Similar offences	-0.44	0.12	-3.54	$4 \approx 10^{-4}$

Table: Regression coefficients and their summary statistics

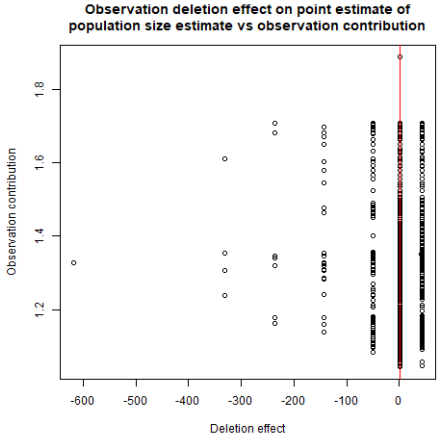

Figure: Regression deletion diagnostics

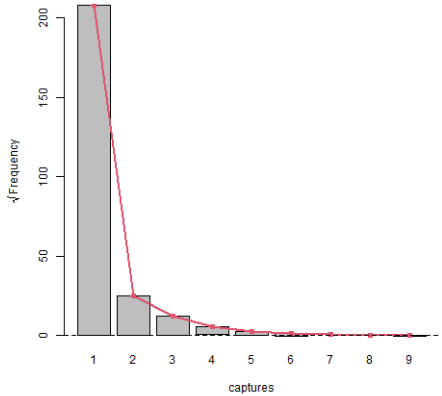

Figure: Rootogram of observed and fitted counts

Issues and limitations

Captures	1	2	3	4	5	6	9	χ^2	p value	G	p value
Count	43256	626	145	23	7	4	1	-	-	-	-
Fitted	43256	624	141	32	7	2	0	9.46	0.051	8.17	0.086

- The hurdle zero truncated geometric model fits data reasonably well (though other models, such as ztoigeometric with cloglog(ω), were 'close' in terms of performance)
- We do not have information on exposure or even on hours where each unit was driving (much less on hours spent driving under the influence)
- As previously alluded to this is most likely why hurdle model performed best on this data set. Whilst more robust that 'inflated' SSCR models in this regard it is still far from ideal
- Every unit with $Y > 3$ is an slight outlier with respect to deletion diagnostics on $\hat{N}$
- Better performance of geometric distribution when compared to poisson models indicates presence of unobserved heterogeneity

Future plans

- We have assumed a closed population and for this reason we constrained ourselves to data from 1st of January 2022 to 20th of October the same year
- This is somewhat unnatural approach and constructing a open population model for single-source data seems to be in order (it is not present in the literature right now)
- We have information on specific person's residence (up to municipality) and information on which police office 'observed' given unit on a given day whilst the later one cannot be included in this regression based approach they could be usefull in other methods
- We used the observed likelihood i.e. likelihood of $Y|Y > 0$ it is also possible to estimate N using only Y but this is not currently implemented when covariates are available and in general will not lead to radically different population size estimates

- For reasons discussed the results should be approached with some degree of scepticism
- A somewhat optimistic estimate for population size $\hat{N} = 56308$ was proposed with 99.9% confidence interval of $(51999, 64032)$
- To the best of our knowledge there are no other studies on this topic in Poland so we lack a reference for this figure

Selected literature


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.


Yee, Thomas W. (2015). *Vector Generalized Linear and Additive Models: With an Implementation in R*. 1st. Springer Publishing Company, Incorporated.


Godwin, Ryan T. (2017). "One-inflation and unobserved heterogeneity in population size estimation". In: *Biometrical Journal* 59.


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.


Chlebicki, Piotr and Maciej Beręsewicz (2023). *singleRcapture: Single-Source Capture-Recapture Models*. R package version 0.2.0.1.