Bayesian latent class models for estimating population size in multiple record systems in presence of missing data

Di Cecco D.[1], Di Zio M.[1], Liseo B. [2]

[1]Istituto Nazionale di Statistica
[2]Sapienza University, Roma

Firenze 7 Jun 2019

Admin data for capture-recapture

In recent years, the main challenge of National Statistics Institutes has become to produce statistics based solely on Admin Data

Our goal: Estimate the number of usual residents by municipality

From:

Census + PES

Main concern:
Undercoverage

➡

To:

Administrative Data (+Survey)

Main concern:
Overcoverage
(Erroneous captures)

The common approach in Official Statistics

- All admin sources are integrated in a single population register
- The register is coupled with a (overcoverage-free) survey to exploit a Dual System Estimator DSE
- The overcoverage rate is estimated and used to "adjust" the DSE in some way

Zhang and Dunne (2017) Trimmed DSE

Multiple Record Systems

We retain the sources as separate lists.

Other works on erroneous captures in Multiple Record Systems:

- da Silva (2009), Wright et al (2009), Link et al (2014)
  focus on Record Linkage duplicates only
- Overstall (2014) and Fegatelli et al (2017)
  just one list is assumed to have false captures

Our proposal

We define the erroneous captures as random classification errors.
We propose a latent class model with two subpopulations:
our target population and one consisting of out-of-scope units.

$$X = \begin{cases} 1 & \text{for in–scope units;} \\ 0 & \text{for out–of–scope units.} \end{cases}$$

Specific issues of Admin Data:

NON INDEPENDENCE OF CAPTURES

- Captures of a same unit in different sources are not independent
  (e.g. compulsory registration on a list preliminary to others)

→ Capture histories are modeled through Log–Linear models which explicitly include dependencies via interaction parameters.
Covariates (if any) are easily included.

Specific issues of Admin Data:

INCOMPLETE LISTS

- Some lists target specific subpopulations of our population of interest (e.g. different social security organizations for specific categories of workers).
- This leads to units having null probability of being captured.
- Treating them as sampling zeros would cause biased estimates.

→ We treat the capture histories of the units not covered by the incomplete lists as if they were partially observed. So the information about the capturing status is Missing At Random in some cases.

A Bayesian Approach

Why a Bayesian approach?

1. Exploit prior information
2. Model uncertainty is handled more easily
   Estimates are sensitive to model specification.
   - Interval estimates of the population size are automatically obtained
   - More thorough methodologies are possible (Model Averaging)
3. The approach is computationally simple as a Gibbs sampler is adopted (extremely simple for Decomposable models)

Priors for capture probabilities

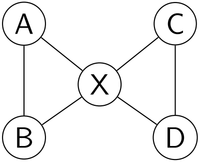
Decomposable models

$\Downarrow$

product of Dirichlet priors:

The (augmented) likelihood is a product of Multinomial distributions

$$l(X, Y \mid \Theta) \propto \prod_{x,a,b,c,d} p_x^{n_x} \, p_{ab|x}^{n_{abx}} \, p_{cd|x}^{n_{cdx}}$$



On each one we set a conjugate prior Dirichlet distribution

$$P_X \sim Beta(\alpha_0, \alpha_1) \quad P_{AB|X} \sim Dir(\alpha_{ab|x}) \quad P_{CD|X} \sim Dir(\alpha_{cd|x})$$

Priors for capture probabilities

General log-linear models

$\Downarrow$

"Constrained Dirichlet" prior (normalized over a loglinear subspace)

Bayesian Iterative Proportional Fitting (Shafer 97) to generate from a Constrained Dirichlet

We can still use a Gibbs sampler + BIPF
(MultiNormal priors requires Metropolis-Hastings)

e.g. for [AB][AX][BX][CDX] we consider [ABX][CDX]
where prior on [ABX] is restricted to no-triple

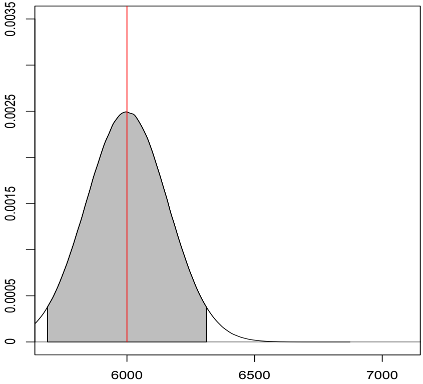
Priors on N

We test various priors for the total population size N (False+True captures including unobserved units)

We use an MCMC to sample from the posterior distribution of N conditionally on $X = 1$

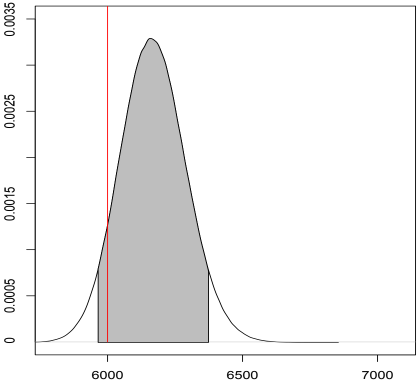
- if $\pi(N) \propto 1/N$ (improper non–informative)
  $\rightarrow$ simple Gibbs sampling
- otherwise (Poisson, Negative Binomial, Rissanen,...)
  $\rightarrow$ we add a Metropolis-Hastings step within Gibbs

Example: 5 lists, N=10000, 3 Strata, 40% False Captures, 30% Uncaptured



Estimating Model =

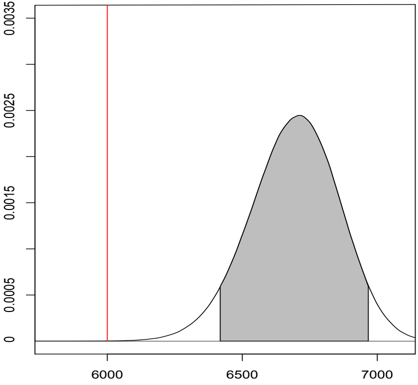
Generating Model =[ABCX][DX][EX]



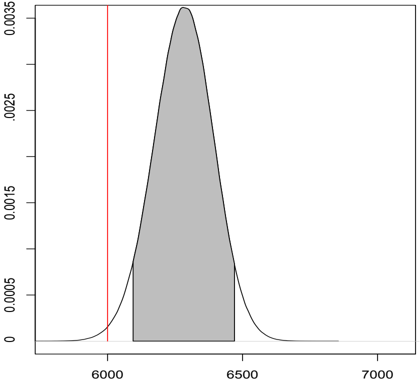
Estimating Model =

All second order interactions

Example: 5 lists, N=10000, 3 Strata, 40% False Captures, 30% Uncaptured



C.I. Model [AX][BX][CX][DX][EX]



C.I. Model + informative priors