Bayesian Dual Systems Population Estimation Adjusting for Linkage Error and Misclassification.

Patrick Graham, Rebecca Sarah, Matt Hendtlass, Abby Morgan, Andrew Richens, Lucianne Varn
Statistical methods, Statistics New Zealand
patrick.graham@stats.govt.nz; patrick.graham.br@gmail.com

1. Introduction

- Dual systems estimation (DSE) is a form of closed population capture-recapture analysis.
- Different emphasis from ecological applications:
  - rarely interested primarily in total population size;
  - distributions of the population over combinations of covariates, such as age, sex, ethnic group and location is important - this can be very fine-grained stratification, easily running to hundreds of thousands of combinations.
- There are complications due to:
  - Size of lists,
  - population unit record file that is adjusted for under-coverage in both census and administrative data?

2. Motivation for this paper

- Response to the 2018 New Zealand census was lower than expected.
- Could DSE be used to combine information from census data and an administrative list to build a population unit record file that is adjusted for under-coverage in both census and administrative data?

3. Table 1. Probability model for the population distribution at covariate setting in inclusion indicators are as shown in Table 1.

4. Gibbs sampler for dual systems estimation - simple case without linkage error or misclassification

Suppose the covariate distribution depends on \( \theta \), let \( n_{\text{mix}} \) denote the number of observed records and let \( D^{\text{obs}} \) denote observed data (cell locations and covariate values for the \( n_{\text{mix}} \) records counted in the data). The Gibbs sampler alternates between:
- \( f(1) \) drawing \( (X_{obs}, N) \) from \( p(X_{obs}|D^{\text{obs}}; \phi, \theta) \).
- \( f(2) \) updating \( \phi \) using \( p(\phi|X_{obs}, N)=\prod_{i=1}^{n_{\text{mix}}}(1-c_{i}(X_{obs})(1-c_{i}(X_{obs}))p(X_{i}(\theta)) \) (1)
- \( f(3) \) updating \( \theta \) using \( p(\theta|N, \phi)=p(\phi|X_{obs}, N)N!/(N-n_{\text{mix}})!(N-n_{\text{mix}}) \) (2)

Steps (1) and (3) condition on the completed population data and are thus standard.

5. Adjusting for linkage error

- We can concentrate on mixed matches and assume no erroneous links. This can be approximated by requiring strict criteria for declaring a link.
- A missed match results in an individual being represented in both the \( (1,0) \) and \( (0,1) \) cells, when they should be in the \( (1,1) \) cell.
- Solution
  - Conduct a linkage validation study. Sample records from List 1; manually check true match status.
  - Let \( Y \) denote the true cell that would be observed in the absence of linkage error, and let \( Y_{L} \) the observed cell.
  - The link validation study informs the linkage sensitivity model:
    \( \alpha(X,Y)=Pr(Y_{L}=(1,1)|X=(1,1))p(1,1) \).
  - The key addition to the Gibbs sampler is a step to adjust for missing cells by sampling missed matches from the observed \( (1,0) \) cell with probability:
    \( Pr(Y=(1,1)|X=(1,1), Y_{L}=(0,1), X=(1,1)) \).
  - An updating step for \( \phi \) is also added to the Gibbs sampler. This conditions on the corrected cell locations \( Y_{L} \).

6. Adjusting for misclassification

- Need to extend the model to accommodate covariates recorded differently on the two lists (e.g. location).
- Expand the covariate vector to \( (X_{1}, X_{2}, Z) \) where \( Z \) denotes covariates recorded accurately on both lists and \( X_{1}, X_{2} \) denote, respectively, the List 1 and List 2 versions of the remaining variables.
- Both \( X_{1} \) and \( X_{2} \) are defined for all individuals in the population, but we only see \( X_{1} \) for people on List 1 and only see \( X_{2} \) for people on List 2.
- \( X_{1} \) and \( X_{2} \) are jointly observed only for linked group; i.e. the observed \( (1,1) \) cell. Our method adjusts for this inherent bias.
- For the application it is reasonable to equate \( X_{1} \) with the true values; hence we model the covariates using:
  \( p(X_{1}|X_{2}, Z)=\gamma(Y_{L}=(1,1)|Y=(0,1), Z, \gamma)p(X_{1}|Z) \).

7. Simulated data example

- We constructed a simulated population of size 1,000,000 with covariate distribution based on 2013 census data.
- We specified models for coverage, misclassification of location and linkage error, such that each of these issues varied by age, sex, ethnic group and small area geography (2,000 areas).
- Area effects were modelled hierarchically; age effects were model using cubic splines.
- Applying the models to the synthetic population produces two lists with under-coverage that are linked with error, with misclassification of location (small area geography) on List 2.
- We applied our Gibbs sampler to the generated data; a weak test of our method that nevertheless tests our implementation. However, it does not speak to robustness to model misspecification.
- We are able to recover the underlying population structure. We have run many examples of this type.

8. Discussion

- Implementation is currently in R and has been tested on simulated data.
- Applications of DSE involving large administrative datasets, will often have to deal with linkage error and measurement error. Dealing with these issues jointly is not common in the DSE literature.
- Gibbs sampling provides a natural framework for Bayesian DSE but, unsurprisingly, is slow with large lists.

Key References


Figure 1: Data generating model for dual systems estimation with linkage error and misclassification

Figure 2: Traceplots for total population size, from application of our Gibbs sampler to simulated data. The true population size is 1,000,000.