LR3_1 Capture-recapture method and log-linear models to estimate register undercoverage

<table>
<thead>
<tr>
<th>General setting</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Configuration</strong></td>
<td>Configuration 3.</td>
</tr>
<tr>
<td><strong>Type of data sources</strong></td>
<td>Two or three registers</td>
</tr>
<tr>
<td><strong>Statistic of interest</strong></td>
<td>Size of the population and its domains</td>
</tr>
<tr>
<td><strong>Type of errors</strong></td>
<td>Register under-coverage of the population, uncertainty of the population size estimates</td>
</tr>
<tr>
<td><strong>Quality measure</strong></td>
<td>Parametric bootstrap estimate of the confidence interval for the population size and its domain size. Case study: deviations of the population size estimates for comparison.</td>
</tr>
</tbody>
</table>

**Method (use one block per method)**

<table>
<thead>
<tr>
<th>Description</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>The beginning of using capture-recapture method for population size estimation is found in the nature sciences to estimate, for example, the number of fish in a lake. Some amount of fish is caught, counted, marked and released back to the lake. It is captured. The fish is being caught repeatedly, counted and the proportion of the recaptured marked fish is used to estimate the number of fish in the lake. For estimation of a human population size two lists of the same population are used instead of capture and recapture. One of the first applications of the method to the human population size estimation is found in India in 1947, in the paper by Sekar and Deming (1949). Nowadays the method is used often to estimate census undercount and its importance is especially increasing for register based censuses and register based statistics in general.</td>
<td></td>
</tr>
</tbody>
</table>

Let us suppose two registers I and II of the same population are linked. Some elements are included into both of them, denote their number by \( m_{11} \), some elements are included into the first of them, but not to the second \( m_{10} \), some elements are included into the second register but not into the first one \( m_{01} \). Based on that, the number of elements of the population which are not included into any of the registers, \( m_{00} \), can be estimated by

\[
\hat{m}_{00} = \frac{m_{10} m_{01}}{m_{11}}.
\]

Assumptions made are:

1) Inclusion of the element into the register I is independent of its inclusion in the register II.
2) Inclusion probabilities of the element in at least one of the registers are homogeneous.
3) The population is closed.
4) It is possible to link the elements of registers I and II perfectly.
The first and the second assumptions are usually violated in human populations. This violation should influence the accuracy of the population size estimates obtained. There are several approaches to make the impact of these violations less severe:

1) To use covariates, levels of which have heterogeneous inclusion probabilities for both registers. The loglinear models can be fitted to the contingency table of the inclusion indicators to registers I and II and the auxiliary variable – covariate. The independency assumption of the elements to be included in registers I and II is then replaced by the weaker assumption of the conditional independence of the elements to be included in register I and register II conditionally on the covariates available.

2) To use a third register and include the two-factor interactions into the loglinear model (to relax the assumption of the independence), but assume that the three-factor interaction is absent. The assumption of independence is mitigated here.

3) To use a latent variable taking heterogeneity of the inclusion probabilities into account.

The mixture of the approaches recited is also possible to use. The authors of the papers Heijden et al. (2012) and Gerritse et al. (2015) are dealing mostly with the first approach showing that it is possible to join it with the second approach as well.

It is explained in the paper by Heijden et al. (2012), that the covariates in the loglinear model allow:

a) to take into account heterogeneity of the inclusion probabilities with respect to the levels of the covariate;

b) to obtain a subdivision of the estimated population size by the levels of the covariates.

Theoretical properties of the loglinear models in the context of the population size estimation are discussed in the paper, and a graphical representation of loglinear models is given. Several cases are studied: the case when covariates are available in both of the registers; the case when a covariate is available in one of the registers and not in other register. The case of three registers available is discussed. The concept of active and passive covariates is introduced and their role in the loglinear model is studied.

In the case when there are no covariates, the loglinear independence model is fitted and missing counts in one cell are estimated. If covariates are available in both of the registers, the saturated log-linear model is fitted to the observed counts, and missing counts of the contingency table are estimated afterwards.

The case when a covariate is available in the first of the registers but not in the second one, is more complicated. The approach taken is to consider the problem as a missing information problem in the second register, and the assumption of a missing at random mechanism (MAR by Little and Rubin, 1987) for the data of the unknown covariate values in the second register is made. It means that the probability to be missing for the value of a covariate in the second register depends only on the observed variables in the capture-recapture model. The Expectation-Maximization algorithm (EM) is used to estimate the missing values of the partially observed covariate. It is an iterative procedure. The estimates are unbiased under the MAR assumption. After convergence, the loglinear model using the observed cells of the contingency table is fitted. The values of the missing cells are assessed. The accuracy of the estimated cell sizes is assessed by an estimate of the confidence interval (CI), obtained using the parametric bootstrap. The case study presented in the paper focuses on the estimation of the number of the people with an Afghan, Iraqi or
Iranian nationality living in the Netherlands in 2007. The registers used are an automated system of decentralized (municipal) population registers and the central Police Recognition system. The choice of the model fitting is demonstrated, estimates for population size and their CI presented, and estimates for subpopulations by the levels of covariates used are given and discussed.

The paper Gerritse et al. (2015) deals also with the independence assumption in the capture-recapture model for estimation of the population size. It is emphasized that the independence problem is unverifiable. It means that independence of the element to be included in the first register and in the second register cannot be verified from the data. Therefore estimation of the impact of the dependence of the inclusion in the both registers on the estimate of the population size is not easy. The paper deals with the study of the impact of the violation of the independence assumption on the accuracy of the population size estimates which were presented in the Heijden et al. (2012) paper. The known level of dependence between the inclusion probabilities in two registers is created, and estimates of the population size under the assumption of independence are obtained. The results are compared to the results without any additional inclusion of the dependence. The sensitivity of the population size estimates to the violation of the independence assumption is studied in the following way:

Let us assume that two registers without covariates are used. Let variables A and B respectively denote inclusion indicators in registers I and II. Let the levels of the indicator A be indexed by \( i \), where \( i = 1 \) means “included in the register”, and \( i = 0 \) means “not included in the register”. The levels of B are indexed by \( j, j=0, 1 \), correspondingly. Expected counts for the contingency table of variables A and B are denoted by \( m_{ij} \), observed values are denoted by \( n_{ij} \). The count \( n_{00} = 0 \) because the number of elements not included in any of the registers is not known. Under the assumption of independence for the inclusion indicator \( s \) in the first register A and in the second register B, the loglinear model for the counts \( n_{01}, n_{10}, \) and \( n_{11} \) can be written as

\[
\log( m_{ij} ) = \lambda + \lambda_i^A + \lambda_j^B 
\]

(1)

with restrictions \( \lambda_0^A = \lambda_0^B = 0 \). The missed count \( m_{00} \) is estimated by \( \hat{m}_{00} = \exp\left( \hat{\lambda} \right) \). On the other hand, because of the assumption of independence, the odds ratio \( \theta \) equals 1:

\[
\theta = \frac{m_{11}}{m_{00}} \cdot \frac{m_{01}}{m_{10}} = \frac{m_{00} \cdot m_{11}}{m_{10} \cdot m_{01}} = 1,
\]

and the estimate for the missed count of the contingency table of A and B is as follows:

\[
\hat{m}_{00} = \frac{\hat{m}_{10} \cdot \hat{m}_{01}}{\hat{m}_{11}} = \frac{n_{10} \cdot n_{01}}{n_{11}}.
\]

There are several ways to estimate parameters of the model (1), and the authors use one of them. They assume that \( n_{ij} \) follows a Poisson distribution, and estimate parameters of (1) through the generalised linear model. The Poisson loglinear model under independence assumes that \( \lambda_{ij}^{AB} = 0 \).
In order to introduce dependence into the model, the fixed non-zero interaction parameter $\lambda_{ij}^{AB}$ is inserted into (1):

$$
\log( m_{ij} ) = \lambda + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB}
$$

(2)

using the restrictions $\lambda_{00}^{AB} = \lambda_{10}^{AB} = \lambda_{01}^{AB} = 0$. Dependence is introduced in model (2) through the term $\lambda_{ij}^{AB} \neq 0$.

Computer software is used to estimate the generalised linear model when the log link function is applied and residuals are distributed according to the Poisson distribution with $\lambda_{ij}^{AB}$ as offset. If $\lambda_{ij}^{AB} \neq 0$, the estimate of $\lambda$ for (1) differs from the estimate of $\lambda$ for (2). Using the relationship between $\lambda_{ij}^{AB}$ and the odds ratio $\theta$, the values of $\lambda_{ij}^{AB}$ can be chosen to satisfy equation:

$$
\theta = \frac{m_{00}m_{11}}{m_{10}m_{01}} = \exp \left\{ \lambda_{ij}^{AB} \right\}.
$$

Sensitivity of the population size estimates is studied for different values of the odds ratio $\theta$: 0.5, 0.67, 1.00, 1.50, 2.00, for two registers without covariates, two registers with fully observed covariates, two registers with partially observed covariates and an extension of the study to three registers. Two populations are used for a case study: people residing in the Netherlands in 2007 with Afghan, Iraqi, or Iranian nationality and people with the Polish nationality in 2009. The same two registers as in the paper of Heijden et al. (2012) are used. The results of the case study show, that population size estimate under dependence could be robust enough or not robust at all. It depends on the proportion of the population covered by the register. If the population coverage by the register is high then dependence between the inclusion in the registers introduced does not change the population size estimates much. But if the coverage of the population by the registers is low then dependence changes the population size estimates obtained under the assumption of independence dramatically. Even assumed conditional independence of fully observed covariates leads to big changes in population size estimates when the population coverage by the register is low.

The estimation methods: the EM algorithm is used to estimate the missing values for partially observed covariates and the parametric bootstrap is used to estimate standard errors for the population size estimates. The methods are the same as in the paper of Heijden et al. (2012). The bootstrap confidence intervals for population size estimates under the assumption of independence are compared to the CI under assumption of asymptotic normality of the population size estimates. Both estimates of CI are close to each other. The formula for standard errors of the population size estimate is taken from Bishop et al. 1975 and Sekar and Deming (1949). The R and SPSS codes to estimate population size under the Poisson generalized linear model and R code for estimation of standard error by parametric bootstrap are presented.

**Assumptions**

The method is based on the following assumptions:

1) Inclusion of an element into the register A is independent of inclusion of this
| Advantages | The method can be used to estimate the population size and size of its domains defined by levels of the covariates. Estimates for the confidence intervals of the population size estimates are valuable results. |
| Disadvantages | The unverifiable assumption of the independence of the element inclusion probabilities in each of the registers may leave some doubts. |

### Case study (per method if needed)

| Agency – country | Authors of the paper made a case study: Statistics Netherlands, Lancaster University, Utrecht University |
| Topic | People residing in the Netherlands in 2007 with Afghan, Iraqi or Iranian nationality and people with the Polish nationality in 2009. |
| Data sets used | Two registers of Statistics Netherlands: an automated system of decentralized population registers (with information on people that are legally allowed to reside in the Netherlands and are registered as such) and a Central Police recognition system where suspects of offences are registered. |

### Final remarks

| Gap analysis | Problem of unverifiable assumptions. Analytical expression for dependence of the population size estimate on the level of dependence for the element to be included in the first and second register would be a valuable result. Verifiable assumptions for the method used would be useful. |
| Other remarks | The papers Heijden et al. (2012) and Gerritse et al. (2015) are written very clearly, like a textbook. All details are completely explained. Case studies presented illustrate the methods presented perfectly. |

### References