Indicator for the representativeness of linked sources

Dingeman Jan van der Laan (dj.vanderlaan@cbs.nl), Bart F. M. Bakker

Keywords: representativeness, linkage quality, LR-indicator, R-indicator

1. INTRODUCTION

Administrative data are increasingly used in official statistics. These data have many advantages: e.g. a much smaller response burden, the possibility of large sample sizes for the production of small domain statistics, opportunities for following units through time to create longitudinal data and comparatively low collection costs. However, the wider use of administrative data has also revealed more and more quality issues [1,2]. One of the limitations of administrative data is that they usually have a small number of variables. It is not possible to produce the desired crosstables, if the two or more required variables are not in the same source. Therefore, data linkage techniques are used to combine data from different administrative sources. However, missed links could lead to biased estimates if the missed links are selective. This is more or less similar to selective non-response in surveys [3,4].

The linkage effectiveness is the most used indicator of the quality of the linkage. However, that is not always a good indicator of the quality of the linkage process of two sources. Estimates based on a linked data set with a high linkage percentage can still be biased if the missed links are selective. Other often used measures are based on the numbers of false positives (false links) and false negatives (false non-links). However, this can only be computed if the true links are already known [5] which in practice is difficult. Furthermore, these measures still suffer from the fact that the effect of missed links on estimates depends largely on the selectivity of the links.

We propose an indicator for the similarity of the linked records to the target population under investigation. We make use of the notion that missed links lead to similar errors as selective non-response in surveys. To measure representativeness of the response of a survey, the R-indicator has been developed [6,7]. The R-indicator is based on the idea that the response of a survey is representative of a target population if the response probabilities are the same for all units in the population. Note that this corresponds to the idea of Missing Completely At Random (MCAR) with respect to all variables. Because these response probabilities are unknown, a weaker version of this idea is used: the response of a survey is representative of a target population if the average response probabilities over variable X (with H categories) is constant. This weaker definition corresponds to a missing data mechanism that is MCAR with respect to X. For X also a vector of variables could be used. The indicator is called Linkage Representativeness Indicator (LR-indicator).

2. METHODS

The LR-indicator is based on the concept of linkage probabilities, i.e. the probability to be linked. If the records in two sources are linked, the resulting links are representative of a target population if all units have the same linkage probability. In that hypothetical situation, it is very easy to evaluate the linkage result by measuring the amount of variation in linkage probabilities. The more variation, the less representative. However, the linkage probability is a theoretical concept that cannot be observed. What can be
observed is the value of $R_i$, which has the value 1 if element $i$ links (with probability $\rho_i$) and otherwise has the value 0 (with probability $1-\rho_i$). The idea is to estimate the linkage probabilities using auxiliary variables, chosen in such a way that the linkage probabilities are optimally explained. If a set of explanatory variables $X$ can be found and their values $X_i$ are observed in the linked sources, the linkage probabilities $\rho_i$ can be replaced by the linkage propensity.

$$\rho_i (X) = \Pr (R_i = 1 \mid X = X_i).$$

To estimate the linkage propensities, one could use a logistic regression model like

$$\text{logit} \rho_i (X) = \log \left( \frac{\rho_i (X)}{1-\rho_i (X)} \right) = \sum_j X_{ij} \beta_j$$

Following Schouten et al. (2009), the LR-indicator for the representativeness of the linked records of two sources can be defined by

$$LR = 1 - 2S_\rho,$$

Where $S_\rho$ is the standard deviation of the estimated linkage probabilities. LR equals one if all linkage probabilities are equal and then there is complete representativeness. The smaller the value of LR (the minimum value is zero), the larger the lack of representativeness.

Besides an overall indicator, it is also of interest to know subpopulations are under-represented and which over-represented. This information can direct further efforts in the linkage process (e.g. the search for specific linkage variables for these subpopulations) and inform on possible biases in the analyses. For this the partial LR-indicators can be used [8], of which we currently only describe the unconditional one. Let $Z$ be categorical variable with categories $k=1,2,...,K$. $Z$ is a component of $X$. Then the unconditional partial indicator for $Z$ is defined as [8]:

$$P_u (Z = k, \rho_X) = \sqrt{\frac{\sum_{k=1}^K N_k}{N} (\bar{\rho}(X) - \tilde{\rho}(X))^2} = \sqrt{\sum_{k=1}^K P_u (Z = k, \rho_X)^2},$$

with $\bar{\rho}(X)$ and $\tilde{\rho}(X, k)$ the average linkage propensity and the average linkage propensity for category $k$ of $Z$, respectively; $N$ and $N_k$ are the number of records and the number of records in category $k$, respectively. The value is bounded above by 0.5 and below by zero. The larger the value, the larger the contribution of $Z$ to the lower representativeness. $P_u (Z = k, \rho_X)$ is the unconditional partial indicator of category $k$ of $Z$. A positive value indicates an overrepresentation and a negative value an overrepresentation. The values are between -0.5 and 0.5.

3. RESULTS

The LR-indicator is applied to studies that were conducted by Statistics Netherlands in collaboration with a consortium of universities. In the first study the Population Register is linked to the Employee Register. The LR-indicator is determined using the micro-data. In the second study, the National Twin Register is linked to a register of a large Health Insurance Company, using only the aggregated data. Results from both studies will be presented. However, at this moment the results of only one study are available.

In the first example the Population Register (PR) and the Employment Register (ER) are linked. The target population was defined as the employed foreign residents in The
The registers are linked with a combination of deterministic and probabilistic methods. In the first step, the records are linked deterministically on a personal identification number that is widely available in administrative data sources in the Netherlands. The remaining records are linked probabilistically [9,5]. To reduce the number of possible pairs, the data are blocked on variables that are assumed to be of very high quality: postal code or date of birth. For the probabilistic linkage date of birth, sex, postal code, house number and extension are used. In total 84.6% of the records of the ER could be linked. The probabilistic linkage leads to an increase of 0.3% of the total number of linked records.

The covariates sex, age, nationality and region were used in the calculation of the linkage propensities. Using these propensities, we arrive at an LR-indicator of 0.50 which is significantly lower than values found in some of the surveys performed at Statistics Netherlands [6]. Therefore, one should consider correction for the linkage selectivity e.g. by weighing the data, or even end the research entirely.

It is also possible to calculate partial R-indicators, which express the contributions of the various subgroups to the overall LR-indicator [9]. The unconditional partial R-indicators are shown in figure 1. Nationality has the largest contribution to the lack of representativeness of the linked data set. Especially, persons from Poland and other new EU countries (joined EU after 2004) are underrepresented. Any analysis of this data set should, therefore, take nationality into account. Furthermore, male persons, persons aged 20-30 and persons living in the south of the Netherlands are underrepresented. However, these categories will correlate with the underrepresented nationality groups.
Figure 1. The unconditional partial LR-indicators for the linkage between the Employment Register and Population Register.

4. CONCLUSIONS

With the increasing use of linked administrative data, the need for a measure of the representativeness of the data after linkage also increases. The most frequently used indicator for the quality of the linkage process is the effectiveness. However, this is no measure for the representativeness. If all population elements have the same probability of being linked, then the resulting data are representative of the population even if the effectiveness is low.

In this paper, we present the LR-indicator to measure the representativeness of linkage of two sources. It is based on the idea that the standard deviation of the linkage probabilities estimated for a set of auxiliary variables provides sufficient indication of the representativeness: the higher the standard deviation, the lower the representativeness. It is also possible to determine partial LR-indicators to find out which variables and which categories of these variables have the largest contribution to the LR-indicator and therefore are the most misrepresented in the linked data file.

We give two examples of which one is described in this abstract. We determine the LR-indicator of the linkage of a Population Register and an Employee Register with age, sex, nationality and region. The study reveals that the representativeness is rather low. Either one should put more effort in improving the linkage, correct for the linkage selectivity or end the research.

REFERENCES


