Discussion Paper

Connecting correction methods for linkage error in capture-recapture

Peter-Paul de Wolf
Jan van der Laan
Daan Zult

May 2018

Contents

A commonly known problem in population size estimation using registers, is that registers do not necessarily cover the whole population. This may be because they intend to cover part of the population (e.g., students), due to administrative delay or because part of the target population is not registered by default (e.g., illegal persons). One of the methods to estimate the population size in the presence of undercount, is the capture-recapture method that combines the information of two or more samples. In the context of census estimation, it combines information from two (or more) registers. The method however assumes perfect linkage between the registers can be achieved. It is known that often this assumption is violated. Ding and Fienberg (1994) proposed a correction for linkage error in the setting of evaluating the population coverage of a census using a post-enumeration survey. Di Consiglio and Tuoto (2015) extended that correction by relaxing some of the conditions used by Ding and Fienberg. However, Di Consiglio and Tuoto still implicitly assumed that the two registers are of equal size. We will introduce a further generalization that includes both previously mentioned correction methods and at the same time deals with registers of different sizes. Specific parameter settings will correspond to the different correction methods. We will show that the parameters of each method can be chosen such that the resulting estimates all equal the traditional Peterson estimate that would be obtained under truly perfect linkage. [1)2)]

1 Introduction

The capture-recapture methodology goes back to the ecological setting of estimating the size of fish and wildlife populations. The basic idea is to take a first sample (capture), tag or mark the captured animals, return them to their population and take a second sample (recapture). Among the recaptures, some of the animals will be marked, others not. The relation between the tagged and non-tagged animals in the second sample is used to construct an estimate of the total population size. See e.g., Peterson (1896) and Lincoln (1930). Since then it was not only used to estimate animal population sizes, but also to estimate undercount in traditional censuses (for an overview see e.g., Fienberg (1992)). More recently, it was used to estimate the under coverage of registers used for the Dutch Census (Gerritse et al., 2016a).

In the original setting, one of the assumptions is that the units can be classified without error to belong to the first sample only, the second sample only or the overlap of the two samples. This assumption was likely to be met, when the marking of the units in the first sample would stick to the animals during the second sampling (no tag-loss). In the setting of estimating the under count of a register, this assumption is translated to the assumption that units in the two registers can be linked without error, i.e., all links are found and no erroneous links are established. With linking two records, we mean deciding that those records represent the same population unit. Whenever the registers both contain the same reliable unique identifiers like a Social Security Number, it is likely that this assumption holds. However, not all registers contain such a uniform unique identifier. Actually, when considering under coverage of registers, one can not rely on the existence of such unique identifiers only. Indeed, in order to find units that are not properly registered, one should also use sources that do not have such a unique identifier for all units.

1) The authors like to thank Jeroen Pannekoek for reviewing an earlier version of the paper.
2) The views expressed in this paper are those of the authors and do not necessarily reflect the policy of Statistics Netherlands.

In case a unique identifier is not available, one often relies on probabilistic record linkage techniques like the one developed in Fellegi and Sunter (1969). In this setting the assumption of perfect linkage is not likely to be met in practice. Contrary to the situation of tag-loss, there will be two different errors possible: the error of linking two records that should not be linked and the error of not linking records that should be linked. Only the latter is similar to tag-loss.

In the presence of linkage errors, the standard capture-recapture estimate of the unknown population size can be biased, see e.g., Gerritse et al. (2016b). In Ding and Fienberg (1994) the standard capture-recapture estimator is adjusted to correct for linkage errors. In that paper, they considered the situation where a post-enumeration survey (PES) is used to estimate the under coverage of the population census. See e.g., Wolter (1986) for an explanation of using a PES. Ding and Fienberg assume that the false match that affects the population size estimator most, occurs when a record from the subset of the PES that should not be matched is actually linked to a record from the subset of the census that should not be matched. In other words, they assume a one-way linkage error, linking PES records to census records. Moreover, they assume that all records in the PES will be linked to a record in the census.

Di Consiglio and Tuoto (Di Consiglio and Tuoto, 2015) argue that in the setting of administrative data sources, a one-way linkage direction is not guaranteed. That is, they allow for the possibility that a population unit residing in one administrative data source, but not in the other, to be (incorrectly) linked to a unit in the other administrative data source, irrespective of which data source is called `the one' and which is called `the other'. Hence they propose a two-way correction for linkage error. In their paper, they assume that the probability of a false match is equal in both linkage directions. We will call this the symmetric two-way correction for linkage error. Using the same error probability in both directions, is approriate in case the two administrative data sources are (approximately) of equal size.

A further extension would be to allow for different error probabilities in the two linkage directions. Indeed, when the two registers differ considerably in size, this is more likely to be the case, especially in case of (forced) one-to-one linkage.[3)] Obviously, the largest source contains units that are not present in the smaller source. In case of one-to-one linkage a subset of those units cannot be linked: there are just not enough `target'-records in the smaller source. Hence, those records will also not be linked incorrectly. In other words, a unit in the largest administrative data source has a smaller chance of being falsely linked with a unit in the smaller administrative data source, compared to the other way around. In the current paper we will thus introduce an asymmetric two-way correction for linkage error. The formulation of this asymmetric two-way corrections has three parameters. Choosing specific values for those parameters, the formula can cover the one-way correction and the symmetric two-way correction as well.

The outline of the paper is as follows. We start with explaining the general setting of capture-recapture and probabilistic linkage. In section 3 we briefly state the non-corrected estimator, the one-way corrected estimator, the symmetric two-way corrected estimator and an asymmetric two-way corrected estimator. The formula of the asymmetric two-way correction can be viewed as a general estimator in the sense that all introduced estimators can be expressed with this formula. Finally, we unify all estimators by choosing specific `optimal'

[3)] One-to-one linkage here means that exactly one record from PES is allowed to be linked to exactly one record from the census. Some linkage procedures do not ensure this by default.

parameters. The following section, Section 4, shows some simulation results using publicly available fictitious data on the UK population census. In Section 5 we draw conclusions and the appendices contain some technical details.

2 General setting

Let us first introduce the notation that will be used throughout the remainder of this paper. We try to stay close to the notation used in Ding and Fienberg (1994) and Di Consiglio and Tuoto (2015). We also state the general assumptions underlying the capture-recapture methodology when applied with two registers. Note that we will only discuss the situation of two registers that are linked using probabilistic record linkage methods.

2.1 Capture-recapture with two registers

Let R_1 and R_2 denote two registers containing units from a common population $\mathcal{X}$ of unknown size $N_{\mathcal{X}}$. Assuming we can identify population units to belong to either one or both of the registers, we get Table 2.1 and Table 2.2. In Table 2.1 the numbers correspond to the unobservable true population counts, whereas the numbers in Table 2.2 are the observed counts *after* the linkage process has taken place and thus depend on the used linkage procedure.

In the tables, the first subscript denotes whether or not a population unit resides in R_1 and the second subscript whether or not a population unit resides in R_2. So, e.g., N_{10} denote the (unobserved) number of population units that does reside in R_1 but not in R_2. Note that, assuming no duplicates in each R_i, $n_{1+} = N_1$ is the size of R_1, $n_{+1} = N_2$ the size of R_2. Moreover, $N_{\neg i}$ denotes the number of units in the population that do not reside in R_i, i.e., $N_{\neg 1} = N_{\mathcal{X}} - N_1$ and $N_{\neg 2} = N_{\mathcal{X}} - N_2$. Even after the linkage process has taken place, we still cannot observe population units that are included in neither register (i.e., N_{00}). That means that $N_{\neg 1} \geq n_{01}$ and $N_{\neg 2} \geq n_{10}$.

		unit in R_2		
		yes	no	
unit in R_1	yes	N_{11}	N_{10}	N_1
	no	N_{01}	N_{00}	$N_{\neg 1}$
		N_2	$N_{\neg 2}$	$N_{\mathcal{X}}$

Table 2.1 Counts based on population

		unit in R_2		
		yes	no	
unit in R_1	yes	n_{11}	n_{10}	n_{1+}
	no	n_{01}	0	n_{01}
		n_{+1}	n_{10}	n_{++}

Table 2.2 Counts based on linkage process

Using similar notation, we can write the probability that a population unit resides in register R_i as p_i and decompose those probabilities as follows: $p_1 = p_{11} + p_{10}$ and $p_2 = p_{11} + p_{01}$.

The general assumptions in capture-recapture estimation are:

- The population $\mathcal{X}$ is closed, i.e., units can enter nor leave the population during the capture-recapture experiment.
- There are no erroneous captures, i.e., only units from $\mathcal{X}$ can be captured.
- The event that a unit resides in R_1 is independent of the event that a unit resides in R_2.

– The probability that a unit resides in R_i is the same for all units in $\mathcal{X}$.
– There is no error in allocating the units to R_1, R_2 or both.

These assumptions imply that $N_{11}/N_1 = N_2/N_{\mathcal{X}}$ or equivalently, $N_{\mathcal{X}} = (N_1 N_2)/N_{11}$. Hence, under perfect conditions a natural estimator would be the one introduced in Peterson (1896): $\hat{N}_{\mathcal{X}} = (n_{1+}n_{+1})/n_{11}$. See Subsection 3.1 as well.

2.2 Probabilistic record linkage

The probabilistic record linkage technique we will assume in this paper, is the one described in Fellegi and Sunter (1969). In their approach, they consider the set of all possible pairs (a, b) of records from R_1 and R_2: $\{(a, b) \mid a \in R_1 \text{ and } b \in R_2\}$. They decompose that set into two disjunct sets: set $\mathcal{M}$ consisting of all pairs of records of matches and set $\mathcal{U}$ of all pairs of records of non-matches. Hence, e.g., a pair (a, b) in the set $\mathcal{U}$ of non-matches should consist of a record a from register R_1 and a record b from R_2 where a and b refer to two different population units. See Figure A.1 in Appendix A for a graphical representation of the sets $\mathcal{M}$ and $\mathcal{U}$.

Fellegi and Sunter then describe a model to decide whether an observed pair of records should be allocated to $\mathcal{M}$ or to $\mathcal{U}$. To that end they use so called comparison functions that assign a value to a pair indicating the amount of similarity between the two records. For example, in case of personal data, a comparison function could assign a value zero whenever the name of the person of record a is not exactly equal to the name of the person of record b, and a value of one whenever the names are exactly equal. Obviously, this can be more elaborate, assigning a value between zero and one in case of small spelling mistakes. Different comparison functions can be applied to different variables within a record, which would result in a comparison *vector*.

Selecting a pair of records at random from all possible pairs, the comparison function applied to that selected pair is a random variable. They define the so-called m-probability as the probability that a certain value of the comparison function is found among a pair of records that should belong to the set $\mathcal{M}$ of matches and the u-probability as the probability that a certain value of the comparison function is found among a pair of records that should belong to the set $\mathcal{U}$ of non-matches. Using those probabilities, they assign weights to each possible pair and say that a pair of records is linked whenever the weight is above a some threshold and not-linked whenever that weight is below that threshold. Since this is defined at the level of *pairs* of records, it is possible that several records from register R_1 are said to be linked to the same record in register R_2: whenever a pair has a weight above the threshold, it will be said to be linked. In practice, often a one-to-one linkage is then enforced: one of those pairs is selected and designated to be a link, while the other pairs are considered to be non-links despite their weight above the threshold.

In their paper, Fellegi and Sunter consider two error probabilities: the probability of a false link (assigning a pair of records to $\mathcal{M}$ where it should be assigned to $\mathcal{U}$) and the probability of a false non-link (assigning a pair of records to $\mathcal{U}$ where it should be assigned to $\mathcal{M}$). Note that these probabilities are thus defined at the level of *pairs* of records and not on the level of *individual* records. In the description of the correction methods (see Section 3) error probabilities are defined at the level of individual records. To be able to discuss the correction methods for linkage error, it is convenient to decompose our registers R_i each into two disjunct sets M_i and U_i. Now M_i consists of all unique records from register R_i that should appear in a pair of matches and U_i

of all other unique records from register R_i. Figure A.1 in Appendix A graphically shows the differences between the sets $\mathcal{M}$, $\mathcal{U}$, M_i and U_i.

3 Estimation of the population size

In this section we will first briefly state the existing estimators for the population size under no linkage error, one-way error correction and symmetric two-way error correction. At the end of this section we will introduce our new asymmetric two-way error correction estimator.

Using the notation from Subsection 2.1, we assume that the number of individuals that fall in the four interior cells of Table 2.2 have a multinomial distribution:

$$(n_{11}, n_{10}, n_{01}, N_{\mathcal{X}} - n_{++}) \sim \text{Mult}(N_{\mathcal{X}}, p_{11}, p_{10}, p_{01}, p_{00})$$

where $n_{++} = n_{11} + n_{10} + n_{01}$. Like in Ding and Fienberg (1994), we will derive the estimators using the approach of maximizing the conditional likelihood as described in Sanathanan (1972). In that approach the likelihood is written as a product of two likelihoods $L_1(\cdot)$ and $L_2(\cdot)$, where $L_1(\cdot)$ is the likelihood of (n_{11}, n_{10}, n_{01}) for fixed n_{++} and $L_2(\cdot)$ the likelihood of n_{++}, given the cell-probabilities p_{11}, p_{10} and p_{01}. In the conditional approach, first $L_1(\cdot)$ is maximized to derive the maximum likelihood estimates of the cell probabilities, after which the $N_{\mathcal{X}}$ is found that maximizes $L_2(\cdot)$, given the values of p_{11}, p_{10} and p_{01}. Sanathanan (1972) has shown that under suitable regularity conditions both the conditional and the unconditional likelihood estimates of $N_{\mathcal{X}}$ are consistent and have the same asymptotic distribution.

Using that $\mathbb{E}(n_{++}) = \mathbb{E}(n_{1+}) + \mathbb{E}(n_{+1}) - \mathbb{E}(n_{11}) = (p_1 + p_2 - p_{11})N_{\mathcal{X}}$, we derive the following generic formulation of an estimator of the population total:

$$\hat{N}_{\mathcal{X}} = \frac{n_{++}}{\hat{p}_1 + \hat{p}_2 - \hat{p}_{11}} \tag{1}$$

In the following subsections we will derive conditional ML estimators of the cell probabilities under different linkage error scenarios.

3.1 No linkage error

Under independence and perfect linkage, we would have the following equations for the probabilities of recording population units in the different observed counts n_{ij}:

$$p_{11} = p_1 p_2 \tag{2}$$

$$p_{10} = p_1 - p_{11} = p_1(1 - p_2) \tag{3}$$

$$p_{01} = p_2 - p_{11} = p_2(1 - p_1) \tag{4}$$

Using the conditional ML approach we would get the estimators

$$\hat{p}_1 = \frac{n_{11}}{n_{+1}} \quad \text{and} \quad \hat{p}_2 = \frac{n_{11}}{n_{1+}}$$

Plugging those estimators into (2) and (1), the estimator of the population total then becomes after some straightforward calculations

$$\hat{N}_{\mathcal{X}}^{P} = \frac{n_{1+} n_{+1}}{n_{11}}$$

and this is essentially the estimator as described in e.g., Peterson (1896).

3.2 One way correction (OC)

In Ding and Fienberg (1994) the situation of linkage error is considered under the assumptions that (using the notation as in Subsection 2.2) :

(a) A matching pair between records from M_1 and M_2 remains a match with probability $0 < \alpha \leq 1$.
(b) A record from M_1 is matched incorrectly with a record in M_2 with negligible probability.
(c) A false match between records from M_1 and U_2 occurs with negligible probability.
(d) A false match between records from U_1 and M_2 occurs with negligible probability.
(e) Each record from U_1 will be linked with a record in U_2 with common probability $0 \leq \beta < 1$.

the reason for assuming negligible probabilities for (b), (c) and (d) is that in those cases two errors are made: both the correct match is not made and an incorrect match is made. In cases (a) and (e) only one error is made. Note that the probabilities α and β are now defined at *record* level, i.e., different from the probabilities in the Fellegi and Sunter setting (see Subsection 2.2).[4)]

Under those assumptions we get the following relations

$$p_{11} = \alpha p_1 p_2 + \beta p_1 (1 - p_2) \tag{5}$$

$$p_{10} = p_1 - p_{11} = p_1 - \alpha p_1 p_2 - \beta p_1 (1 - p_2) \tag{6}$$

$$p_{01} = p_2 - p_{11} = p_2 - \alpha p_1 p_2 - \beta p_1 (1 - p_2) \tag{7}$$

Note that the `one-way' correction is reflected in (5): the second term on the right hand side only shows the probability of falsely linking (β) a unit that resides in R_1 (p_1) but not in R_2 ($1 - p_2$). The probability of falsely linking a unit that resides in R_2 but not in R_1 is not considered. I.e., only one linkage direction is considered.

The conditional ML estimators are then given by (Ding and Fienberg, 1994)

$$\hat{p}_1 = \frac{n_{11} - \beta n_{1+}}{(\alpha - \beta) n_{+1}} \quad \text{and} \quad \hat{p}_2 = \frac{n_{11} - \beta n_{1+}}{(\alpha - \beta) n_{1+}}$$

Plugging this into (5) and (1), the population total then can be estimated by

$$\hat{N}_X^{OC} = \frac{(\alpha - \beta) n_{11}}{n_{11} - \beta n_{1+}} \frac{n_{1+} n_{+1}}{n_{11}} = \frac{(\alpha - \beta) n_{11}}{n_{11} - \beta n_{1+}} \hat{N}_X^P$$

Note that this estimator depends on the parameters α and β which are unknown in practice and should therefore be estimated. This will be discussed in Subsection 3.5.

3.3 Symmetric two-way correction (SC)

In Di Consiglio and Tuoto (2015) it is proposed to relax the assumption of the one-way correction and to allow a two way correction. This means that assumption (e) as given in the description of the one-way correction, is relaxed to allow for a unit in U_1 that is not in U_2 still to be (incorrectly) linked to a unit in U_2 as well as to allow for a unit in U_2 that is not present in U_1 still to be (incorrectly) linked to a unit in U_1. Both events occur with the same probability $0 \leq \beta < 1$.

[4)] Note that the probabilities in the Fellegi and Sunter setting are sometimes also denoted by α and β. These α and β are thus fundamentally different from the ones used in the current paper.

This results in the following equations:

$$p_{11} = \alpha p_1 p_2 + \beta p_1(1-p_2) + \beta p_2(1-p_1) \quad (8)$$

$$p_{10} = p_1 - p_{11} = p_1 - \alpha p_1 p_2 - \beta p_1(1-p_2) - \beta p_2(1-p_1) \quad (9)$$

$$p_{01} = p_2 - p_{11} = p_2 - \alpha p_1 p_2 - \beta p_1(1-p_2) - \beta p_2(1-p_1) \quad (10)$$

Again, under certain regularity conditions and using the conditional likelihood approach, they derive that the ML estimators are then given by

$$\hat{p}_1 = \frac{n_{11} - \beta(n_{1+} + n_{+1})}{(\alpha - 2\beta)n_{+1}} \quad \text{and} \quad \hat{p}_2 = \frac{n_{11} - \beta(n_{1+} + n_{+1})}{(\alpha - 2\beta)n_{1+}}$$

Plugging this into (8) and (1), the population total then can be estimated by

$$\hat{N}_X^{SC} = \frac{(\alpha - 2\beta)n_{11}}{n_{11} - \beta(n_{1+} + n_{+1})} \frac{n_{1+}n_{+1}}{n_{11}} = \frac{(\alpha - 2\beta)n_{11}}{n_{11} - \beta(n_{1+} + n_{+1})} \hat{N}_X^P$$

3.4 Asymmetric two-way correction (AC)

As a further relaxation of the assumptions, we propose to allow for different probabilities of false links. I.e., a unit present in U_1 but not present in U_2 is still linked to a unit in U_2 with probability $0 \leq \beta_1 < 1$ and a unit present in U_2 but not present in U_1 is still linked to a unit in U_1 but with probability $0 \leq \beta_2 < 1$.

Now the equations for the probabilities of recording population units in the different observed counts become

$$p_{11} = \alpha p_1 p_2 + \beta_1 p_1(1-p_2) + \beta_2 p_2(1-p_1) \quad (11)$$

$$p_{10} = p_1 - p_{11} = p_1 - \alpha p_1 p_2 - \beta_1 p_1(1-p_2) - \beta_2 p_2(1-p_1) \quad (12)$$

$$p_{01} = p_2 - p_{11} = p_2 - \alpha p_1 p_2 - \beta_1 p_1(1-p_2) - \beta_2 p_2(1-p_1) \quad (13)$$

Under certain regularity conditions, we then get the following ML estimators:

$$\hat{p}_1 = \frac{n_{11} - \beta_1 n_{1+} - \beta_2 n_{+1}}{(\alpha - (\beta_1 + \beta_2))n_{+1}} \quad \text{and} \quad \hat{p}_2 = \frac{n_{11} - \beta_1 n_{1+} - \beta_2 n_{+1}}{(\alpha - (\beta_1 + \beta_2))n_{1+}} \quad (14)$$

See Appendix B for a discussion on admissibility to obtain proper values for the probabilities $\hat{p}_1$ and $\hat{p}_2$ in the interval $[0, 1]$.

Plugging (14) into (11) and (1), the population total then can be estimated by

$$\hat{N}_X^{AC} = \frac{(\alpha - (\beta_1 + \beta_2))n_{11}}{n_{11} - \beta_1 n_{1+} - \beta_2 n_{+1}} \frac{n_{1+}n_{+1}}{n_{11}} = \frac{(\alpha - (\beta_1 + \beta_2))n_{11}}{n_{11} - \beta_1 n_{1+} - \beta_2 n_{+1}} \hat{N}_X^P$$

Note that this formulation covers all previous situations by choosing appropriate α, β_1 and β_2:

- Peterson estimator: $\alpha = 1$ and $\beta_1 = \beta_2 = 0$
- One way correction: $\alpha = \alpha$, $\beta_1 = \beta$ and $\beta_2 = 0$
- Symmetric two way correction: $\alpha = \alpha$, $\beta_1 = \beta_2 = \beta$

3.5 Linking the correction methods

We consider the Peterson estimator in case of perfect linkage, i.e., knowing the true N_1, N_2 and N_{11}, the `optimal' estimator and call it the `true Peterson estimator' (TP):

$$N_X^{TP} = \frac{N_1 N_2}{N_{11}} = \frac{n_{1+} n_{+1}}{N_{11}}$$

Equating the AC estimator to the true Peterson estimator, i.e., setting $\hat{N}_X^{AC} = N_X^{TP}$, we get the following relationship between the parameters:

$$\alpha N_{11} + \beta_1 (N_1 - N_{11}) + \beta_2 (N_2 - N_{11}) = \alpha N_{11} + \beta_1 N_{10} + \beta_2 N_{01} = n_{11} \quad (15)$$

Note that the left hand side equals the expected number of links under the model for linkage error.

Let us first explore this relationship under the unrealistic assumption that we know the true N_{11}. A natural choice for the parameter α would then be the fraction of true population matches among the links from the linkage process. We will denote this natural choice by $\check{\alpha}$. Substituting that natural choice in (15) and setting $\beta_1 = \beta^{OC}$ and $\beta_2 = 0$, we get

$$\alpha^{OC} = \check{\alpha} = \frac{m_{11}}{N_{11}} \quad \text{and} \quad \beta^{OC} = \frac{n_{11} - m_{11}}{N_1 - N_{11}}$$

where m_{11} is the number of true population matches among the links from the linkage process. We will call this choice of parameters the `optimal OC-parameters'.

In case of the symmetric two-way correction, using the natural choice for α and setting $\beta_1 = \beta_2 = \beta^{SC}$ leads to

$$\alpha^{SC} = \check{\alpha} = \frac{m_{11}}{N_{11}} \quad \text{and} \quad \beta^{SC} = \frac{n_{11} - m_{11}}{N_1 + N_2 - 2N_{11}}$$

We will call this choice of parameters the `optimal SC-parameters'.

In case of the asymmetric two-way correction, we need an additional constraint to uniquely define `optimal AC-parameters'. In practice, it is convenient to enforce one-to-one linkage in the process. Under that assumption, we can derive the following relationship between the parameters of the asymmetric two-way estimator (see the Appendix C for a derivation):

$$\beta_1 = \frac{(\alpha n_{+1} - n_{11})\beta_2}{(\alpha n_{1+} - n_{11}) - 2\beta_2 (n_{1+} - n_{+1})} \quad (16)$$

In case we want to satisfy both (16) and (15), using the natural $\check{\alpha}$ parameter, we get either

$$\alpha^{AC} = \check{\alpha} = \frac{m_{11}}{N_{11}}, \quad \beta_1^{AC} = \frac{n_{11} - m_{11}}{2(N_1 - N_{11})} \quad \text{and} \quad \beta_2^{AC} = \frac{n_{11} - m_{11}}{2(N_2 - N_{11})}$$

or

$$\tilde{\alpha}^{AC} = \check{\alpha} = \frac{m_{11}}{N_{11}}, \quad \tilde{\beta}_1^{AC} = \frac{m_{11} N_2 - n_{11} N_{11}}{m_{11}(N_2 - N_1)} \quad \text{and} \quad \tilde{\beta}_2^{AC} = \frac{m_{11} N_1 - n_{11} N_{11}}{m_{11}(N_1 - N_2)}$$

where m_{11} again is the number of true population matches among the links from the linkage process. For the second set of parameters ($\tilde{\alpha}^{AC}$, $\tilde{\beta}_1^{AC}$ and $\tilde{\beta}_2^{AC}$) it holds that the $\tilde{\beta}$'s will be undefined in case $N_1 = N_2$. Moreover, when $N_1 \neq N_2$, one of them will be negative, what contradicts the fact that the $\tilde{\beta}$'s should be probabilities. We will hence call the first set of parameters the `optimal AC-parameters'. Note that, in case register R_1 is the largest and hence under one-to-one linkage $N_1 - N_{11} > N_2 - N_{11}$, we get $\beta_1^{AC} < \beta_2^{AC}$ as expected (see discussion in introduction).

According to the error correction model, a false match between a record from U_1 with a record from U_2 occurs with probability β_1 and, independently, a false match between a record from U_2 with a record from U_1 occurs with probability β_2. Considering these events independently, we would count such a link twice. However, enforcing one-to-one linkage, these two events can only happen at the same time. This is reflected in the factor $1/2$ in the `optimal AC-parameters' β_i^{AC}.

Given the true N_{11} and choosing the parameters such that they satisfy equation (15) would thus lead to the optimal estimator. Indeed, using the `optimal OC-parameters', the `optimal SC-parameters' or the `optimal AC-parameters' will all yield the same estimator, i.e., the true Peterson estimator TP (with perfect linkage).

Unfortunately, in practice we do not know the true N_{11}. Hence, we need to estimate the α and β_i parameters. As long as the estimated parameters satisfy relation (15), the resulting estimates will be exactly the same for all estimators. This would for example be the case when we would estimate the optimal parameters by plugging in some estimate for N_{11}, since N_1, N_2, n_{11} and m_{11} are the same in all settings. Indeed, the resulting estimators would then be given by the simple formula

$$\hat{N}_{\mathcal{X}} = \frac{N_1 N_2}{\hat{N}_{11}} \tag{17}$$

where $\hat{N}_{11}$ is a (consistent) estimator of the `true' overlap between the two registers.

Another possibility would be to use a sample of one of the registers and determine the true matches for that sample. Dividing that number by the sampling fraction would yield a direct estimate of N_{11}. Similarly, we could obtain direct estimates of n_{11} and m_{11}. Note that a direct estimate of n_{11} is needed instead of the original n_{11} to prevent the estimated m_{11} getting larger than the original n_{11}.

Yet another approach would be to use expert knowledge on the linkage errors, e.g., asking experts to give estimates of the parameters. In that case, these expert guesses would not necessarily satisfy relation (15) and the estimators could thus yield different values.

4 Simulations

Simulated data has the advantage that the population as well as the registers are completely known: we know all entries of Table 2.1 as well as Table 2.2. We can thus easily determine how well the estimators approximate the true population size. An additional advantage is, that we can derive the `optimal' Peterson estimator: the Peterson estimator with truly no linkage error. Since this is the maximum likelihood estimator using population information, the resulting estimate is the best one could get. We will call this estimator the True Peterson estimator and use it as benchmark for our other estimators in our simulations.

Since ensuring that the parameter estimates satisfy relation (15) will result in the same estimates of the population size for estimators introduced in section 3, we will concentrate on different ways to estimate the parameters. We will use different methods to estimate N_{11} and m_{11} and plug those estimates into the formulas of our `optimal' parameters, to show empirically that these estimates indeed lead to the same estimate of the population total.

4.1 Setup

For the simulation we will make use of the fictitious data on the UK population census as created for the ESSnet DI (McLeod et al., 2011). The ESSnet DI was a European project on data integration (Record Linkage, Statistical Matching, Micro Integration Processing), running from 2009 to 2011. We used three files from that fictitious dataset: the files `Person` (a fictional list of persons, acting as the population), `CIS` (fictional observations from a Customer Information System, being a combination of tax and benefit data) and `PRD` (fictional observations from the Patient Register Data of the National Health Service). The `Person` dataset comprised of 26,625 individuals, the `CIS` has a coverage probability of that population of $\tau_1 = 0.930$ and the `PRD` of $\tau_2 = 0.924$.

To reduce computation time and to be able to apply the linkage process without blocking, we repeatedly constructed a smaller population and corresponding registers from those files, using the following steps:

1. Draw a sample of size 10,000 from `Person`. This will be our population $\mathcal{X}$ with size $N_{\mathcal{X}}$.
2. Select the records from `CIS` that are present in population $\mathcal{X}$ to get register R_1.
3. Select the records from `PRD` that are present in population $\mathcal{X}$. Randomly select a fraction f of those records to get register R_2.

This way we obtained multiple instances of a population and the corresponding registers where one of them covers the population for about 93% and the other for about f times 92.4 %.

In Di Consiglio and Tuoto (2015) several linkage scenarios were mentioned: a bronze, a silver and a gold scenario. In the current paper we will only use their silver scenario, i.e., we only use the full date of birth (day (`DB_D`), month (`DB_M`) and year (`DB_Y`)) as key variables in the linkage process. For the comparison function of the probabilistic record linkage process (see Subsection 2.2), we simply used `equality' on all key variables separately. That is, whenever two records a and b are compared, the comparison function for key variable V_i is 1 when $V_i(a) = V_i(b)$ or 0 when $V_i(a) \neq V_i(b)$. Whenever V_i is missing in at least one of the two records, the comparison function is defined to be 0 as well. To perform the probabilistic record linkage as described in Subsection 2.2, we used our own R-code. In that code we also forced one-to-one linkage. See `https://github.com/djvanderlaan/reclin` for the R-package `reclin` that we used.

We implemented four methods to obtain values for the N_{11}, m_{11} and n_{11} needed in the formulas for the `optimal' parameters:

A Since we use simulated data, we know the true m_{11} and N_{11} by design. The n_{11} follows from the linkage process.

B Using the EM-algorithm on the complete registers to estimate the posterior m-probabilities. Those posterior probabilities were used to estimate the m_{11} and N_{11}. The n_{11} follows from the linkage process.

C Using a sample of size 200 from the smallest register of which we determine the true match-status. Using that sample we fitted a logistic model (see the Appendix D for more information on the used model) and used that to predict the m-probabilities for the complete registers. Those posterior probabilities were used to estimate the m_{11} and N_{11}. The n_{11} follows from the original linkage process.

D Using a sample of size 200 from the smallest register of which we determine the true match-status. Using that sample we calculated the direct estimates of n_{11}, N_{11} and m_{11} for the complete registers.

In methods B and C, summing the posterior m-probabilities over all linked pairs yields an estimate of m_{11}, whereas summing those probabilities over all possible pairs yields an estimate of N_{11}. For a definition of posterior m-probabilities and why summing them is appropriate, we refer to Fellegi and Sunter (1969).

Using those estimated sizes, we then used the formulas for the `optimal' parameters as derived in section 3.5 to get estimates of the population size. As discussed in that section, we expect to obtain exactly the same estimates for all approaches (OC, SC and AC).

4.2 Results

For three different values of f, we performed 100 replications of the procedure mentioned in the previous subsection and, as expected, we indeed found that all `optimal' parameters led to the same estimates in all four methods. In Table 4.1 the mean, median and standard deviation over the 100 replications is given for the estimates of the population size: TP (method A, the benchmark), P (Peterson, using the counts from the linkage process), EM (method B), model (method C) and sample (method D). Note that TP and P are both based on Petersons formula (Peterson, 1896), but TP is using the (in practice unobservable) true population counts, whereas P uses the observed counts.

f		TP	P	EM	model	sample
0.15	mean	10,010.9	11,033.9	8,125.8	9,986.3	10,010.9
	median	10,001.3	11,032.3	8,232.6	9,994.8	9,981.8
	st. dev.	66.0	124.7	1,211.5	283.2	212.9
0.50	mean	10,006.0	11,186.2	8,173.1	10,012.3	10,011.6
	median	10,004.5	11,186.8	8,141.0	9 969.9	10,029.4
	st. dev.	32.6	58.4	794.2	263.9	202.2
0.90	mean	10,005.6	11,398.0	8,194.2	10,019.4	10,053.8
	median	10,006.7	11,397.5	8,148.6	10 009.8	10,015.7
	st. dev.	11.8	39.3	791.1	302.4	205.8

Table 4.1 Mean, median and variance of the 100 replications of each estimator, for $N_X = 10,000$, different relative sizes f of the second register, and sample size 200.

The first thing to notice, is that the Peterson estimator using the observed counts indeed leads to a heavily biased estimate of the population size, due to the linkage errors that are present. Moreover, we see that the EM-based estimator (method B) has a very large variance compared to the other estimators and at the same time has a larger bias. This indicates that this method is not well suited to be used for correcting linkage error.

Varying the relative size of the second register (i.e., the f) does not really influence the correction for linkage error. Indeed, the bias as well as the variance of those estimators seems to be more or less the same in all situations.

In case the registers include a unique identifier for some of the records that could be used as an alternative for taking a sample, under the assumption that the absence of the identifier is not (too) selective. When such a unique identifier is not present, it could in practice be quite costly to determine the true match status of pairs. Hence, probably only a small sample would be considered by an NSI and that's why we used a relatively small sample from the second register for methods C and D.

Figure 4.1 shows a smooth estimate of the distribution of the estimators for $f = 0.5$. For the other values of f the distributions look similar. We did not plot the EM-based estimator in this figure to be able to see more clearly the differences between the other estimation methods.

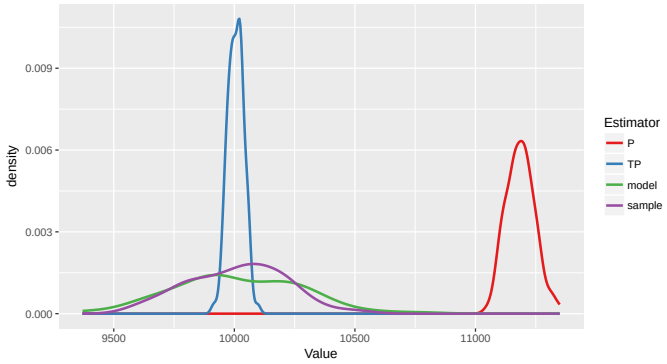

Figure 4.1 Distributions of the P, TP, model based and sample based estimators for $N_X = 10,000$, $f = 0.5$ and sample size 200.

The figure again shows clearly that the Peterson estimator using the counts from the linkage process has a large bias (due to linkage error) and that the model and sample estimators nicely correct for that. The TP estimates are obviously performing the best, since they use the true knowledge about the number of matches. However, in practice that estimator is not possible.

5 Conclusions

In estimating the population size using capture-recapture, linkage errors (false links and missed links) affect the Peterson estimator. Indeed, the Peterson estimator then becomes heavily biased. To reduce that bias, some correction methods have been proposed in the literature. These methods introduce some additional parameters that should reflect the probability of occurrence of the two possible types of linkage error. They then model how linkage errors occur and use those error-probabilities to incorporate that model into the estimator. In this paper we have introduced a general formulation for such a correction method. That general formulation incorporates all previously introduced correction methods of that type as special cases.

Looking more closely to the general correction method, it turned out that the parameters could actually be chosen in such a way that the general estimator equals the optimal estimator: the Peterson estimator with known number of true matches. These `optimal' parameters can be estimated using different methods. We have shown that for at least two methods, the results improve the traditional Peterson estimator considerably. Those two methods make use of a relatively small sample for which the true match status of the records needs to be determined.

In case it is not possible to make use of a such a sample, the general correction method could still be useful. In that situation, the model for the occurrence of the linkage errors should be assessed to estimate the error probabilities. We would like to note that the model assumes that `double errors' occur with negligible probabilities. With `double errors' we mean errors like missing a true match of a record and at the same time linking that record incorrectly to some record in the

other register. In estimating the error-probabilities this should be taken into account in some way, because in practice such double errors do occur and would influence the error-probabilities.

More elaborate methods to estimate the population size in the presence of under coverage make e.g., use of covariates or of more than two registers. In these cases, more complex loglinear or Poisson models can be used to obtain a capture-recapture estimate. Similarly, the Fellegi and Sunter based linkage procedure can also be applied more elaborately, e.g., by making use of blocking(s). This would affect the (estimates of the) posterior m-probabilities. In our view, the ideas expressed in the current paper, as well as the introduced general formulation of the linkage error correction methods, will lead to a better understanding of the implications of such extensions and will be of help in deriving new, linkage error correcting, consistent estimates of the population size.

References

Di Consiglio, L. and T. Tuoto (2015). ``Coverage Evaluation on Probabilistically Linked Data'', *Journal of Official Statistics*, 31, 3: 415--429, DOI: https://doi.org/10.1515/jos-2015-0025.

Ding, Y. and S.E. Fienberg (1994). ``Dual system estimation of Census undercount in the presence of matching error'', *Survey Methodology*, 20: 149--158.

Fienberg, S.E. (1992). ``Bibliography on capture-recapture modelling with application to census undercount adjustment'', *Survey Metodology*, 18: 143--154.

Fellegi, I.P. and A.B. Sunter (1969). ``A Theory for Record Linkage'', *Journal of the American Statistical Association*, 64: 1183--1210.

Gerritse, S.C., B.F.M. Bakker, P.P. de Wolf and P.G.M. van der Heijden (2016a). ``Under coverage of the population register in the Netherlands, 2010'', *Discussion paper 2016-02 (Centraal Bureau voor de Statistiek, Den Haag/Heerlen)*.

Gerritse, S.C., B.F.M. Bakker, D. Zult and P.G.M. van der Heijden (2016b). ``The effects of imperfect linkage and erroneous captures on the population size estimator'', Chapter 3 of PhD thesis *An Application of Population Size Estimation to Official Statistics*, S.C. Gerritse, ISBN 978-94-6233-323-9.

Lincoln, F.C. (1930). ``Calculating Waterfowl Abundance on the Basis of Banding Returns'', *United States Department of Agriculture Circular*, 118: 1--4.

McLeod, P., D. Heasman and I. Forbes (2011). ``Simulated data for the on the job training'', ESSnet DI, available at https://ec.europa.eu/eurostat/cros/content/job-training_en.

Peterson, C.G.J. (1896). ``The Yearly Immigration of Young Plaice Into the Limfjord From the German Sea'', *Report of the Danish Biological Station (1895)*, 6: 5--84.

Sanathanan, L. (1972). ``Estimating the size of a multinomial population'', *The Annals of Mathematical Statistics*, 43, 1: 142--152.

Wolter, K.M. (1986). ``Some coverage error models for census data'', *Journal of the American Statistical Association*, 81: 338--346.

Appendix

A Sets defined in the setting of probabilistic record linkage

Let R_1 be a register with records numbered $\{1, 2, 3, \ldots, 10\}$ and R_2 a register with records numbered $\{1, 2, 3, \ldots, 15\}$. The total number of *pairs* (a, b) that can be constructed from the *records* of those registers is $10 \times 15 = 150$. Figure A.1 shows all possible pairs. Moreover, an example of the set $\mathcal{M}$ of pairs of matching records and the set $\mathcal{U}$ of pairs of non-matching records is shown in that figure. In the example, the number of pairs in $\mathcal{M}$ is 8 and the number of pairs in $\mathcal{U}$ is 142.

We can write each register as the union of two disjoint sets, $R_i = M_i \cup U_i$, where the disjoint sets of unique records are given by

$$M_1 = \{1, 3, 4, 5, 6, 7, 8, 9\} \qquad U_1 = \{2, 10\}$$
$$M_2 = \{2, 3, 4, 6, 8, 9, 10, 13\} \qquad U_2 = \{1, 5, 7, 11, 12, 14, 15\}$$

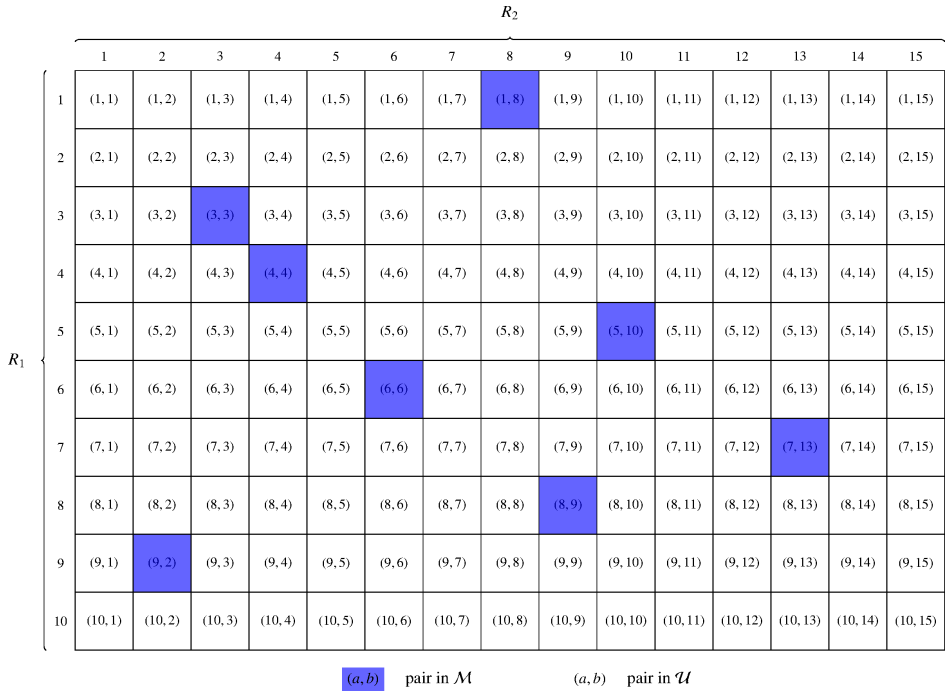

Figure A.1 Graphical representation of the sets of pairs defined in subsection 2.2

B Admissibility of asymmetric two-way correction estimators $\hat{p}_i$

The estimators for the probabilities p_i in case of the asymmetric two-way correction approach should obviously be within $[0, 1]$. This puts some restrictions on the parameters α, β_1 and β_2.

To ensure that the estimators are non-negative, straightforward calculations lead to the condition that either

$$\beta_1 n_{1+} + \beta_2 n_{+1} \leq n_{11} \qquad \text{and} \qquad \beta_1 + \beta_2 < \alpha \tag{B.1}$$

or

$$\beta_1 n_{1+} + \beta_2 n_{+1} \geq n_{11} \qquad \text{and} \qquad \beta_1 + \beta_2 > \alpha \tag{B.2}$$

Additionally, ensuring that both probabilities are not larger than one, leads under (B.1) to the condition

$$\beta_1 n_{1+} + \beta_2 n_{+1} \geq n_{11} - (\alpha - (\beta_1 + \beta_2))\,(n_{1+} \wedge n_{+1}) \tag{B.3}$$

and under (B.2) to the condition

$$\beta_1 n_{1+} + \beta_2 n_{+1} \leq n_{11} - (\alpha - (\beta_1 + \beta_2))\,(n_{1+} \vee n_{+1}) \tag{B.4}$$

where $n_{1+} \vee n_{+1}$ equals the maximum of n_{1+} and n_{+1} and $n_{1+} \wedge n_{+1}$ the minimum of n_{1+} and n_{+1}.

Summarizing, we need either

$$\left.\begin{array}{l} \beta_1 + \beta_2 < \alpha \\ \beta_1 n_{1+} + \beta_2 n_{+1} \leq n_{11} \\ \beta_1 n_{1+} + \beta_2 n_{+1} \geq n_{11} - (\alpha - (\beta_1 + \beta_2))\,(n_{1+} \wedge n_{+1}) \end{array}\right\} \tag{B.5}$$

or

$$\left.\begin{array}{l} \beta_1 + \beta_2 > \alpha \\ \beta_1 n_{1+} + \beta_2 n_{+1} \geq n_{11} \\ \beta_1 n_{1+} + \beta_2 n_{+1} \leq n_{11} - (\alpha - (\beta_1 + \beta_2))\,(n_{1+} \vee n_{+1}) \end{array}\right\} \tag{B.6}$$

Assuming R_1 to be the largest data set, i.e., $n_{1+} > n_{+1}$, the set of conditions (B.5) is equivalent to

$$\left.\begin{array}{l} \beta_1 \geq (n_{11} - \alpha n_{+1}) \,/\, (n_{1+} - n_{+1}) \\ \beta_1 + \beta_2 < \alpha \\ \beta_1 n_{1+} + \beta_2 n_{+1} \leq n_{11} \end{array}\right\} \tag{B.5'}$$

and the set of conditions (B.6) to

$$\left.\begin{array}{l} \beta_2 \geq (\alpha n_{1+} - n_{11}) \,/\, (n_{1+} - n_{+1}) \\ \beta_1 + \beta_2 > \alpha \\ \beta_1 n_{1+} + \beta_2 n_{+1} \geq n_{11} \end{array}\right\} \tag{B.6'}$$

Assuming the two data sets to be of equal size, i.e., $n_{1+} = n_{+1}$, the set of conditions (B.5) is equivalent to

$$\left.\begin{array}{l} \alpha \geq n_{11} \,/\, n_{1+} \\ \beta_1 + \beta_2 < \alpha \\ \beta_1 n_{1+} + \beta_2 n_{+1} \leq n_{11} \end{array}\right\} \tag{B.5''}$$

and the set of conditions (B.6) to

$$\left.\begin{array}{l} \alpha \leq n_{11} \,/\, n_{1+} \\ \beta_1 + \beta_2 > \alpha \\ \beta_1 n_{1+} + \beta_2 n_{+1} \geq n_{11} \end{array}\right\} \qquad \text{(B.6'')}$$

C Enforcing one-to-one linkage

In our asymmetric two-way correction method, we have three parameters: α, β_1 and β_2. In case we enforce one-to-one linkage, we can actually do with two, because in that situation we can write β_1 as a function of α and β_2.

In Figure C.1 the relation between (expected) counts based on the population and based on linkage are shown in the situation where we potentially would like to apply the asymmetric two-way correction with enforced one-to-one linkage. Under the assumption of one-to-one

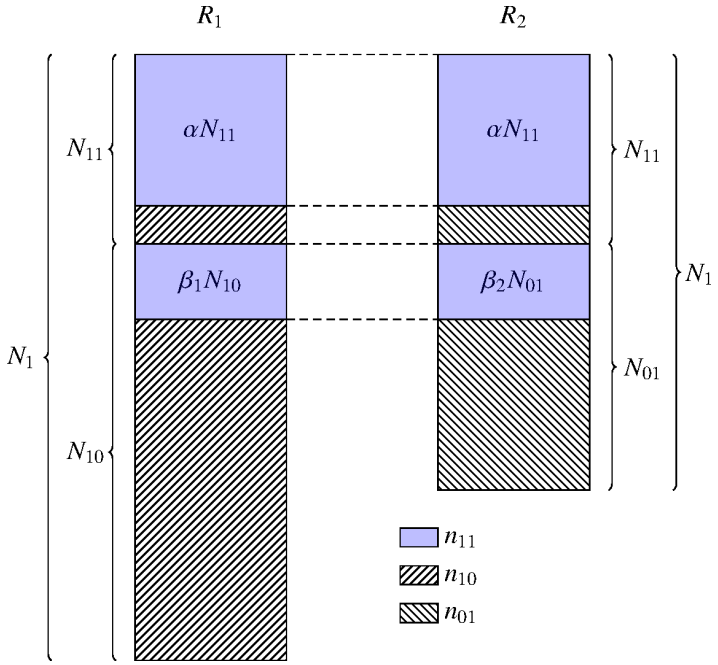

Figure C.1 Relations between counts based on population and based on one-to-one linkage

linkage, it should hold that $\beta_1 N_{10} = \beta_2 N_{01}$, as can be seen in the figure. Noting that $\mathbb{E}N_{10} = p_1(1-p_2)N_{\mathcal{X}}$ and $\mathbb{E}N_{01} = p_2(1-p_1)N_{\mathcal{X}}$ and plugging in the estimators $\hat{p}_1$ and $\hat{p}_2$ from (14), we can derive the following relation:

$$\beta_1 = \frac{(\alpha n_{+1} - n_{11})\beta_2}{(\alpha n_{1+} - n_{11}) - 2\beta_2(n_{1+} - n_{+1})} \qquad \text{(C.1)}$$

Note that, assuming equal sizes of the two registers, i.e., $n_{1+} = n_{+1}$, equation (C.1) yields $\beta_1 = \beta_2$. That is, we would obtain the situation in which the symmetric two-way correction is applicable.

Moreover, from (C.1) it follows that

$$\alpha > 2\beta_2 \text{ and } n_{1+} > n_{+1} \implies \beta_1 < \beta_2$$
$$\alpha > 2\beta_2 \text{ and } n_{1+} < n_{+1} \implies \beta_1 > \beta_2$$

as expected (see discussion in introduction).

D Estimation of the matching probabilities using logistic regression

For a sample of records from the smallest register it is assumed that the true match status can be determined. I.e., we assume that it is known whether or not the record should be linked to a record from the larger register and if so with which record it should be linked. Therefore, for a subset of all pairs generated in the linkage process, the true match status is known. The goal of the logistic regression model is to predict the probability that this pair is a true match, based on properties of the record pair.

In the regression model the following covariates are used:

- The result of the comparison of the linkage variables. In this case the linkage variables are the three elements of the date of birth: day (DB_D), month (DB_M) and year (DB_Y). These variables are binary: both records of the pair agree on the variables (true) or not (false). If in at least one of the records a variable is missing, we consider it a disagree (false).
- Whether or not the pair is selected when enforcing one-to-one linkage. This is also a binary variable which is false when there is a more likely match for one or both of the records. This variable is a strong predictor for true matches.

The target variable is the true match status (a binary variable). All variables are added as main effects. No interactions are used in the model. The model is estimated using the sampled pairs and then used to calculate predictions of the matching probability for all pairs.

Colophon

Publisher
Statistics Netherlands
Henri Faasdreef 312, 2492 JP The Hague
www.cbs.nl

Prepress
Statistics Netherlands, Grafimedia

Design
Edenspiekermann

Information
Telephone +31 88 570 70 70, fax +31 70 337 59 94
Via contact form: www.cbs.nl/information