Statistical challenges in observational cohort studies

Hof, M.H.P.

Publication date
2015
Document Version
Final published version

3 Methods for analyzing data from probabilistic linkage strategies based on partially identifying variables

In record linkage studies unique identifiers are often not available and therefore the linkage procedure depends on combinations of partially identifying variables with low discriminating power. As a consequence wrongly linked covariate and outcome pairs will be created and bias further analysis of the linked data. In this chapter, two estimators that correct for linkage error in regression analysis were investigated. We extended the estimators developed by Lahiri and Larsen and also suggested a weighted least squares approach to deal with linkage error. Both linear and logistic regression problems were considered and the performance of both methods was evaluated with simulations. Our results show that all wrong covariate and outcome pairs need to be removed from the analysis in order to calculate unbiased regression coefficients in both approaches. This removal requires strong assumptions on the structure of the data. In addition, the bias significantly increases when the assumptions do not hold and wrongly linked records influence the coefficient estimation. Our simulations showed that both methods had similar performance in linear regression problems. With logistic regression problems the weighted least squares method showed less bias. Because the specific structure of the data in record linkage problems often leads to different assumptions it is necessary that the analyst has prior knowledge on the nature of the data. These assumptions are more easily introduced in the weighted least squares approach than in the Lahiri and Larsen estimator.

Introduction

Computerized record linkage is an effective manner to aggregate information from different data sources when a unique identifier is not available. Partially identifying variables that are registered in both data sources, referred to as linking variables, are used to find records that belong to the same individual (match) or do not belong to the same individual (non-match). Because record linkage is based on existing data, it has the potential to answer research questions without the need for new data collection [65, 66, 67, 68].

To determine whether a set of records belongs to the same individual Fellegi and Sunter [29] have developed a theory for record linkage. Basically, comparison rules are defined for each of the linking variables and for each combination of rows in both datasets comparison vectors are made. The frequency of all comparison vectors along with the number of matches is used to determine the posterior probability of a match given a certain comparison vector.

A problem that arises with record linkage is the presence of wrongly matched records. Neter et al. [69] investigated the impact of these records and concluded that these records will substantially bias the estimation of the relationship between the response variable and covariates of interest. Therefore Scheuren and Winkler (SW) have developed a method to correct for linkage error in linear models [70, 71]. The idea behind their method is that each observed pair of outcome and covariate (x, z) can be described in terms of the true values (x, y) and the bias (x, b), assuming that only the outcome y is biased by the linking process.

From the observed data, the outcome z with the highest posterior probability of a match is assigned to the true value y. In addition, the outcome z with the second highest posterior probability is assumed to be the wrong value b. Although simulations by SW showed that this method is able to correct regression estimates for linkage error bias, this method does not produce unbiased results. However, a modification of the SW estimator developed by Lahiri and Larsen (LL) does produce unbiased estimators [30]. Their proposed method uses the posterior probabilities as a weighting scheme to derive the expected value y from z for all rows of x.

To derive the unbiased estimator, LL used constraints on the data structure; the outcome variable and covariates are always located in separate datsets, both datasets have similar number of records, and all records from both datasets refer

to the same population. However, these constraints are not met in most record linkage studies.

Chambers et al. [72] generalized the LL method. An estimation equation was proposed and they derived a best linear unbiased estimator that performed somewhat better than the LL estimator. Although simulations showed that Chamber's estimator is useful for dealing with bias, the required constraints on the data structure are similar to LL.

In addition to the work of Chambers et al., Kim and Chambers [73] tried to relax the constraints on the data with an extra weighting procedure in the analysis, that was based on the assumption that the dataset with the covariates x is a subsample of the population of interest and the dataset with outcomes z covers the entire population. However, their extension was based on the strong assumption that for all matches the sign of the regression errors are available. This requires information on the process that governs the assignment of matches and non-matches, which is almost never available. Kim and Chambers also considered linkage situations with non-linear relations between outcome and covariates.

In this chapter we extend the LL method and also consider alternative estimators based on a weighted least squares method to deal with bias introduced by linkage error. We investigate situations with multiple datasets and where the datasets describe different populations but contain a number of true matches. We investigate both linear and logistic regression problems. Performance of all methods was evaluated with simulations.

Record linkage

Consider there are two datasets **A** and **B** containing respectively n and m records and both datasets contain records from the same individuals. Since there is no unique identifying variable per person, other less discriminative variables, must be used in the record linkage procedure such as date of birth, surname, place of residence, or gender. These variables are registered in both datasets and are referred to as linkage variables. Each linkage variable has its own discriminative power, determined by the number of variable values and distribution of variable values [74].

Since there is no prior knowledge on likely matches in both datasets the strategy begins by comparing each record i ($i = 1, 2, \ldots, n$) from dataset **A** with

each record j ($j = 1, 2, \ldots, m$) from $\mathbf{B}$, leading to nm comparisons $\mathbf{g}_{ij}$ ($ij = 1, 2, \ldots, nm$). This vector contains measures of agreement for all k linking variables $\mathbf{r}_l$ ($l = 1, 2, \ldots, k$) and therefore we write $\mathbf{g}_{ij} = (g_{ij1}, \ldots, g_{ijl}, \ldots, g_{ijk})$.

Comparisons g_{ijl} can be defined in different ways but for the sake of simplicity only dichotomous agreement/disagreement outcomes are used, in this chapter

$$g_{ijl} = \begin{cases} 1 & \text{if } r_{il} = r_{jl}, \\ 0 & \text{if } r_{il} \neq r_{jl}. \end{cases}$$

Notice that the number of unique patterns $\mathbf{g}_{ij}$ is 2^k and for $k = 2$ the unique patterns are $(0, 1)$, $(0, 0)$, $(1, 0)$, and $(1, 1)$. All comparisons need to be divided into two groups; matches and non-matches. Therefore, we can specify $q_{ij} = p(\text{match}|\mathbf{g}_{ij})$, i.e. the probability of a match given $\mathbf{g}_{ij}$. Basically, there are two types of strategies to determine q_{ij} [26]. In the deterministic approach, all linking variables or a subset of linking variables in both datasets have to agree to consider a record pair as a link. If the linking variables are highly discriminative, the q_{ij} values will be close to zero or one, depending on the comparison vector g_{ij} [75]. With the probabilistic approach, the probability of a comparison vector, $\mathbf{g}_{ij}$ can be expressed in the mixture model [76]

$$p(g_{ij}) = \pi p(\mathbf{g}_{ij}|\text{match}) + (1 - \pi)p(\mathbf{g}_{ij}|\text{non match}), \tag{3.1}$$

where π is the relative frequency of matches among the nm records-pairs, e.g. the probability that two random records are from the same person. The parameters from this model can be estimated using, for instance, an expectation-maximization algorithm [77].

Parameters need to be estimated for $p(\mathbf{g}_{ij}|\text{match})$ and $p(\mathbf{g}_{ij}|\text{non match})$ for each unique comparison value g_{ij} in (3.1). Because there are 2^k unique patterns for k linking variables, the number of parameters that need to be estimated may become impracticable. To decrease the number of parameters in the model, Fellegi and Sunter [29] suggested to assume independence between the comparison outcomes of each linking variable. Other assumptions and extensions of the mixture model that have been investigated are, among others, the introduction of approximate field estimators [78], use of approximate string comparison [79], the introduction of clerical review in the estimation [76], and the addition of interactions among comparison fields [76, 80].

After estimating the parameters of the mixture model in (3.1), the posterior probability of a match given $\mathbf{g}_{ij}$ can be introduced in the regression of $\mathbf{y}$ on $\mathbf{x} = (x_1, \ldots, x_p)$ by creating a $n \times m$ weighting matrix $\mathbf{Q} = \{q_{ij} : i = i, \ldots, n \text{ and } j = 1, \ldots, m\}$, where q_{ij} equals

$$q_{ij} = p(\text{match}|\mathbf{g}_{ij}) = \frac{\pi p(\mathbf{g}_{ij}|\text{match})}{\pi p(\mathbf{g}_{ij}|\text{match}) + (1-\pi)p(\mathbf{g}_{ij}|\text{non match})}.$$

Linear regression

Lahiri-Larsen estimator

After dataset **A** and dataset **B** are linked together, the relationship between outcome variable y and covariates $\mathbf{x}$ may be estimated. Simple regression models are not applicable in this situation since the true pairs of $(\mathbf{x}, y)$ are not observed. Only all possible combinations of values with their corresponding posterior probability derived from the record linkage procedure are available.

LL showed that their assumptions generate unbiased estimators of a linear relationship. Covariates must be located in dataset **A** and the outcome in dataset **B**, and both datasets must contain the same number of records, therefore implying $n = m$. Kim and Chambers [73] relaxed the assumption of $n = m$ and showed that the LL estimator is still unbiased if the records from dataset **A** are a subset of the population described by dataset **B**, resulting in $n \leq m$. Notice that all records from dataset A must be located in dataset B. Now consider the following linear relation

$$y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \ldots + x_{ip}\beta_p + \epsilon_i, \qquad i = 1, \ldots, n, \tag{3.2}$$

where the βs are unknown regression coefficients. In addition we assume $E(\epsilon_i) = 0$, $var(\epsilon_i) = \sigma^2$, and $cov(\epsilon_i, \epsilon_j) = 0$ for $j \neq i, j = 1, \ldots, n$.

Because it is unknown which record from dataset **A** belongs to which record in dataset **B**, the record pairs $(\mathbf{x}_i, z_i)$ are observed instead of the true $(\mathbf{x}_i, y_i)$. Therefore the relation from LL is not based on $E(y)$ but on $E(z)$ and [70]

$$z_i = \begin{cases} y_i & \text{with probability } q_{ii}, \\ y_j & \text{with probability } q_{ij} \text{ for } j \neq i, j = 1, \ldots, n. \end{cases} \tag{3.3}$$

Furthermore LL require $\sum_{j=1}^{n} q_{ij} = 1, i = 1, \ldots, n$. Given $\mathbf{z} = (z_1, \ldots, z_m)^T$ and $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$ and proposed the estimator of $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{Q}^T\mathbf{Q}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Q}^T\mathbf{z}, \tag{3.4}$$

which is unbiased because

$$E(z_i) = E[E(z_i|\mathbf{y})] = \sum_{j=1}^{n} q_{ij}\mathbf{x}_i\boldsymbol{\beta}^T, \tag{3.5}$$

where $\mathbf{y} = (y_1, \ldots y_m)$. LL propose a parametric bootstrap procedure for variance estimation of $\hat{\boldsymbol{\beta}}$, which captures both the uncertainty of $\hat{\boldsymbol{\beta}}$ and also the uncertainty of $\mathbf{Q}$ caused by the estimation in the record linkage procedure.

Because the LL estimator gives unbiased regression coefficients, we propose a method that uses this estimator in more complex linkage situations. We no longer assume that all covariates $\mathbf{X}$ from (3.2) are located in the same dataset. The only assumption that is required is that the covariate datasets contain records that are a subset of the population described by the dataset containing $\mathbf{y}$.

Suppose we have the situation in which all covariates $\mathbf{x}_k$ $(k = 1, \ldots, p)$ and the outcome are located in $p + 1$ separate datasets. This means that the direct use of (3.4) is not possible because the structure of the $\mathbf{Q}$ matrix becomes far too complex. Each separate covariate $\mathbf{x}_k$ has its own $\mathbf{Q}$ matrix with the outcome variable $\mathbf{y}$. In addition, the covariance of $\mathbf{x}_k$ with $\mathbf{x}_s$, for $s \neq i, s = 1, \ldots, p$ is also influenced by different $\mathbf{Q}$ matrices.

However, if the influence of all $\mathbf{x}_s$ is removed from $\mathbf{x}_k$, it is still possible to estimate β_k with the LL method. By regressing all $\mathbf{x}_s$ on $\mathbf{x}_k$ using the LL method, residuals $\mathbf{x}_k^*$ are derived that are orthogonal to all other covariates. The vector $\mathbf{x}_k^*$ can be used in formula (3.4) to calculate the unbiased $\hat{\beta}_k$. This method is iteratively performed for all p covariates $\mathbf{x}_k$.

Concisely, this method works as follows. Consider the situation where an outcome and two covariates are divided over three datasets $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$, where $\mathbf{A}$ contains

the outcome $\mathbf{y}$, $\mathbf{B}$ contains the first covariate vector $\mathbf{x}_1$, and $\mathbf{C}$ contains the second covariate vector $\mathbf{x}_2$. In addition, we define three $\mathbf{Q}$ matrices $\mathbf{Q}_{AB}$, $\mathbf{Q}_{AC}$, and $\mathbf{Q}_{BC}$ as the weighting matrices obtained from the record linkage procedures linking $\mathbf{A}$ to $\mathbf{B}$, $\mathbf{A}$ to $\mathbf{C}$, and $\mathbf{B}$ to $\mathbf{C}$, respectively. We use the following algorithm to estimate the weights β_1 and β_2 of the regression model, i.e. $E(y|x_{i1}, x_{i2}) = \beta_1 x_{i1} + \beta_2 x_{i2}$:

1. Estimate the regression parameter of $\mathbf{x}_1$ on $\mathbf{x}_2$

$$\boldsymbol{\beta}_{(x_1 \sim x_2)} = \left(\mathbf{x}_2^T \mathbf{Q}_{BC}^T \mathbf{Q}_{BC} \mathbf{x}_2\right)^{-1} \mathbf{x}_2^T \mathbf{Q}_{BC}^T \mathbf{x}_1.$$

2. Calculate the residuals $\mathbf{x}_1 | \mathbf{x}_2$

$$\mathbf{x}_{1res} = \mathbf{x}_1 - \mathbf{Q}_{BC} \mathbf{x}_2 \beta_{(x_1 \sim x_2)}.$$

3. Estimate the regression parameter of $\mathbf{x}_{1res}$ and $\mathbf{x}_2$ on y

$$\begin{aligned}
\beta_{(y \sim x_{1res})} &= \left(\mathbf{x}_{1res}^T \mathbf{Q}_{AB}^T \mathbf{Q}_{AB} \mathbf{x}_{1res}\right)^{-1} \mathbf{x}_{1res}^T \mathbf{Q}_{AB}^T \mathbf{y}, \\
\beta_{(y \sim x_2)} &= \left(\mathbf{x}_2^T \mathbf{Q}_{AC}^T \mathbf{Q}_{AC} \mathbf{x}_2\right)^{-1} \mathbf{x}_2^T \mathbf{Q}_{AC}^T \mathbf{y}.
\end{aligned}$$

4. Derive the true regression coefficients

$$\begin{aligned}
\beta_1 &= \beta_{(y \sim x_{1res})}, \\
\beta_2 &= \beta_{(y \sim x_2)} - \beta_{(y \sim x_{1res})} \beta_{(x_1 \sim x_2)}.
\end{aligned}$$

Another extension is the relaxation of the fact that $\sum_{j=1}^{n} q_{ij} = 1$, meaning that all records from $\mathbf{A}$ must have at least one match in $\mathbf{B}$. Now consider a record linkage situation in which records in $\mathbf{A}$ do not necessarily have a true match in $\mathbf{B}$. With perfect linkage variables the sum of posterior probabilities in this row will be zero and therefore these rows cannot be transformed to sum up to one. Since we focus on situations with non-perfect linkage variables the cumulative posterior probabilities of these rows will not be non-zero. However, if the discriminative power of the linkage variables is high these sums will be relatively small. Therefore we propose to only transform the rows in $\mathbf{Q}$ where it is likely that at least one match has been found. A likely match can be defined as combination of records

from **A** and **B** with a probability higher than an (arbitrary) threshold λ and the **Q** matrix is transformed as

$$\sum_{j=1}^{n} q_{ij} = \begin{cases} 1 & \text{if } max(q_{ij}) \geq \lambda, \\ 0 & \text{if } max(q_{ij}) < \lambda. \end{cases}$$

Because non-matches are likely to have a low q_{ij} and matches a q_{ij} close to one, the λ value directly determines the number of true and false data pairs in the analysis. Choosing a λ value close to one will guarantee that few non-matches are included in the analysis. Conversely, choosing λ close to zero will result in many non-matches. Basically, the choice of λ is a trade-off between specificity and sensitivity of the categorization of matches and non-matches [81, 82].

Weighted least squares estimator

Again consider the linear relation defined in formula (3.2) and that the real pairs of covariates and outcomes are not observed. The two datasets **A** and **B** with n and m records are linked to each other and the **Q** matrix is calculated. For now, we assume that the covariates **X** are located in **A**, and the outcome y in **B**.

To analyze the data we propose a (WLS) approach, which requires some data restructuring. We consider all nm combinations of all n records in A and all m records in **B**. Define **R** as the operator matrix which adds m multiples of **X**, and **P** as the operator matrix which extends **y** n times as follows

$$\tilde{\mathbf{X}} = \mathbf{R}\mathbf{X},$$
$$\tilde{\mathbf{y}} = \mathbf{P}\mathbf{y},$$

$$\mathbf{R} = \begin{array}{c} \\ 1 \\ \\ 2 \\ \\ \vdots \\ \\ m \\ \\ \end{array} \begin{pmatrix} & & & & n \\ 1 & \begin{pmatrix} 1 & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 \end{pmatrix} & & & \\ & \vdots & \vdots & \vdots & \vdots \\ n & \begin{pmatrix} 1 & 0 & \dots & 0 \end{pmatrix} & & & \\ & & & & n \\ 1 & \begin{pmatrix} 0 & 1 & \dots & 0 \\ 0 & 1 & \dots & 0 \end{pmatrix} & & & \\ & \vdots & \vdots & \vdots & \vdots \\ n & \begin{pmatrix} 0 & 1 & \dots & 0 \end{pmatrix} & & & \\ & & \vdots & & \\ & & & & n \\ 1 & \begin{pmatrix} 0 & 0 & \dots & 1 \\ 0 & 0 & \dots & 1 \end{pmatrix} & & & \\ & \vdots & \vdots & \vdots & \vdots \\ n & \begin{pmatrix} 0 & 0 & \dots & 1 \end{pmatrix} & & & \end{pmatrix}, \quad \mathbf{P} = \begin{array}{c} \\ 1 \\ \\ 2 \\ \\ \vdots \\ \\ n \\ \\ \end{array} \begin{pmatrix} & & & & m \\ 1 & \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \end{pmatrix} & & & \\ & \vdots & \vdots & \ddots & \vdots \\ m & \begin{pmatrix} 0 & 0 & \dots & 1 \end{pmatrix} & & & \\ & & & & m \\ 1 & \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \end{pmatrix} & & & \\ & \vdots & \vdots & \ddots & \vdots \\ m & \begin{pmatrix} 0 & 0 & \dots & 1 \end{pmatrix} & & & \\ & & \vdots & & \\ & & & & m \\ 1 & \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \end{pmatrix} & & & \\ & \vdots & \vdots & \ddots & \vdots \\ m & \begin{pmatrix} 0 & 0 & \dots & 1 \end{pmatrix} & & & \end{pmatrix},$$

where $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{y}}$ contain all nm combinations $(\mathbf{x}_i, y_j)$ from the original $\mathbf{X}$ matrix and $\mathbf{y}$ vector. In addition we need to restructure the $\mathbf{Q}$ matrix to a $nm \times nm$ diagonal weighting matrix $\mathbf{W}$ as follows

$$\mathbf{W} = \begin{pmatrix} q_1 \otimes \mathbf{I} & 0 & \dots & \dots & \dots & 0 \\ 0 & q_2 \otimes \mathbf{I} & \vdots & \vdots & \vdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & q_i \otimes \mathbf{I} & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \dots & \dots & q_n \otimes \mathbf{I} \end{pmatrix},$$

where all rows from the $\mathbf{Q}$ matrix form the diagonal values for the weighting matrix $\mathbf{W}$, where $\otimes$ is the elementwise multiplication function and $\mathbf{I}$ the $m \times m$ identity matrix. $\mathbf{W}$ can be introduced in the analysis and $\boldsymbol{\beta}$ can be estimated by minimizing the weighted sum of squares of the transformed data pairs

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \left(\tilde{\mathbf{X}}^T \mathbf{W} \tilde{\mathbf{X}}\right)^{-1} \tilde{\mathbf{X}}^T \mathbf{W} \tilde{\mathbf{y}}, \\ &= \left(\mathbf{X}^T \mathbf{R}^T \mathbf{W} \mathbf{R} \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{R}^T \mathbf{W} \mathbf{P} \mathbf{y}. \end{aligned}$$

This is, however, a biased estimator under imperfect linkage and the bias is

$$\text{bias}(\hat{\boldsymbol{\beta}}) = \left[\left(\mathbf{X}^T \mathbf{R}^T \mathbf{W} \mathbf{R} \mathbf{X}\right)^{-1} \left(\mathbf{X}^T \mathbf{R}^T \mathbf{W} \mathbf{P} \mathbf{X}\right) - \mathbf{I}\right] \boldsymbol{\beta},$$

where the bias depends on both $\boldsymbol{\beta}$ and the weighting matrix $\mathbf{W}$ and the estimate is unbiased if $\mathbf{R}^T \mathbf{W} \mathbf{R} = \mathbf{R}^T \mathbf{W} \mathbf{P}$ or when all βs are zero. Notice that $\mathbf{R}^T \mathbf{W} \mathbf{R} = \mathbf{R}^T \mathbf{W} \mathbf{P}$ will only occur if all true pairs $(\mathbf{x}_i, y_i)$ have weight $w_{ii} = 1$ and all wrong pairs $(\mathbf{x}_i, y_j)$ have weight $w_{ij} = 0$ for $j \neq ij = 1, \dots, m$. In good linkage situations all w_{ij} will be small resulting in relatively low bias.

Although this estimator is biased in most linkage situations, its simplicity and the potential to improve its performance with some simple operations makes it an interesting alternative to the LL estimator. By manipulating the $\mathbf{q}_i$ vectors the weights in the $\mathbf{W}$ matrix are directly affected. Similarly to the LL method, we assume that only *one* of the non-zero values in the vector $\mathbf{q}_i$ describes a true covariate and outcome pair. These properties allow us to specify the following operations for each q_i that contains *more* than one non-zero value

1. Assign zero to all values of the vector $\left(\sum_{j=1}^{m} q_{ij} = 0\right)$

   This operation removes all the records from $\mathbf{A}$ with more than one match

in dataset **B** from the analysis. This operation ensures that all the records that accidentally have two or more matches are not included in the analysis. All wrong data pairs are removed from the analysis with this operation if the LL assumptions hold.

2. Weigh the vector with its cumulative probability of a match $\left(\sum_{j=1}^{m} q_{ij} = 1\right)$

   Create one weighted match for each record in dataset **A**. Because it is not known which one of the non-zero weights in $\mathbf{q}_i$ resembles the true match, all q_{ij} values are weighted to have a cumulative probability of one. Although this method does not remove wrongly linked data pairs from the analysis, it reduces the weight that is assigned to potential wrong data pairs.

3. Randomly select one non-zero value and assign zero values to the others

   In this operation, one non-zero weight is randomly chosen as the true match and its weight will be the original posterior probability q_{ij}. Zero is attributed to all other non-zero weights. Notice that this method can be performed an arbitrary number of times to reflect the uncertainty that is present in the $\mathbf{q}_i$ vector.

A disadvantage of using the WLS approach might seem that the required **P**, **R**, and **Q** matrices can become extremely large in certain linkage studies. For instance, the Rochester Epidemiology Project has linked 1,145,856 medical records to 486,564 individuals [68]. Analyzing these datasets with the WLS method will require matrices with unrealistic dimensions of nm (5.58×10^{11}) rows.

However, most of the comparisons between datasets **A** and **B** will have a weight q_{ij} close to zero and $q_{ij} < \lambda$. We remove these records from **R** and **P** and the associated rows and columns from **W**. The resulting **R**, **P**, and **W** matrices will be considerably smaller. In the case of the Rochester Epidemiology Project the number of rows of **P** and **R** will be close to 1145856 instead of 5.58×10^{11}. This is of the same order as the LL method.

Similarly to the variance estimation of LL, the variance of $\hat{\boldsymbol{\beta}}$ can be derived from a bootstrap procedure. Furthermore the WLS method can also be extended to fit situations with more than two datasets. To capture all the unique combinations of records from the available datasets, we need to multiply all datasets with permutation matrices. These matrices will have similar structures as the **P** and **R** matrices, and the greatest difference is that their inner matrices are multiplied in more dimensions instead of one. In addition the **W** matrix can be calculated by

using a generalized Fellegi-Sunter framework to calculate the posterior probabilities of a data pair combination for more than two datasets.

Logistic regression

Both the LL and WLS methods can be generalized to deal with logistic regression problems

$$logit(p(y_i = 1|x_i) = x_{i1}\beta_1 + x_{i2}\beta_2 + \ldots + x_{ip}\beta_p + \epsilon_i, \qquad i = 1, \ldots, n \tag{3.6}$$

The regression estimates $\boldsymbol{\beta}$ can be calculated with a iterative re-weighted least squares procedure [83]), which maximizes the log-likelihood function with Newton Rapshon updates. For the LL estimator one single update is

$$\begin{aligned} \text{LL}(\boldsymbol{\beta}^{new}) &= \boldsymbol{\beta}^{old} + \left(\mathbf{X}^T\mathbf{Q}^T\boldsymbol{\Omega}\mathbf{Q}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Q}^T\left[\mathbf{z} - \hat{p}\left(\mathbf{z}|\mathbf{X},\mathbf{Q},\boldsymbol{\beta}^{old}\right)\right], \\ \hat{p}(\mathbf{z}|\mathbf{X},\mathbf{Q},\boldsymbol{\beta}^{old}) &= e^{\mathbf{Q}\mathbf{X}\boldsymbol{\beta}^{old}}/(1 + e^{\mathbf{Q}\mathbf{X}\boldsymbol{\beta}^{old}}), \\ \boldsymbol{\Omega} &= \text{diag}\left\{\hat{p}(\mathbf{z}|\mathbf{X},\mathbf{Q},\boldsymbol{\beta}^{old})\left[1 - \hat{p}(\mathbf{P}\mathbf{y}|\mathbf{z},\mathbf{X},\mathbf{Q},\boldsymbol{\beta}^{old})\right]\right\}, \end{aligned} \tag{3.7}$$

and for the WLS method

$$\begin{aligned} \text{WLS}(\boldsymbol{\beta}^{new}) &= \boldsymbol{\beta}^{old} + \\ &\left(\mathbf{X}^T\mathbf{R}^T\boldsymbol{\Omega}\mathbf{R}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{R}^T\left[\mathbf{W}\mathbf{y} - \hat{p}(\mathbf{P}\mathbf{y}|\mathbf{R}\mathbf{X},\boldsymbol{\beta}^{old})\right], \\ \hat{p}(\mathbf{P}\mathbf{y}|\mathbf{R}\mathbf{X},\boldsymbol{\beta}^{old}) &= e^{\mathbf{R}\mathbf{X}\boldsymbol{\beta}^{old}}/(1 + e^{\mathbf{R}\mathbf{X}\boldsymbol{\beta}^{old}}), \\ \Omega &= \text{diag}\left\{\hat{p}(\mathbf{P}\mathbf{y}|\mathbf{R}\mathbf{X},\boldsymbol{\beta}^{old})\left[1 - \hat{p}(\mathbf{P}\mathbf{y}|\mathbf{R}\mathbf{X},\boldsymbol{\beta}^{old})\right]\right\}, \end{aligned} \tag{3.8}$$

Both estimators are unbiased when the covariates are located in the same dataset. However, it is not possible to use formula 3.7 in the LL method to calculate $\hat{\boldsymbol{\beta}}$ if the covariates are located in more than one dataset. This is because the method proposed at page 41 to calculate each $\hat{\beta}_k$ separately does not give unbiased estimates, as the covariates are linked to the outcome variable with a non-linear logit

function. Therefore the $\hat{\beta}_k$ that is derived from $p(y_i = 1|x^*_{ik})$ will not be similar to β_k, since our proposed method implicitly assumes linear relations between outcome and all covariates.

Simulation

Scenarios

To measure the performance of the LL estimator and the WLS estimator we performed simulations of different record linkage scenarios (table 3.1). Situations with different locations of the covariates, and the number of records in dataset **A** without a true match in dataset **B** were investigated. Scenario 1_a represents the situation that fits all the assumptions made by Lahiri and Larsen [30] and Kim and Chambers [73], required for unbiased estimators in the LL model. Covariates are located in **A**, the outcome in **B**, and all records from **A** have one true match in **B**. Therefore the linkage procedure only introduced error in the outcome variable. In scenario 1_b not all records from **A** are located in **B**. Because a number of matches were falsely identified in the record linkage procedure, wrong covariate and outcome pairs were introduced in the analysis.

In scenarios 2_a and 2_b covariate $\mathbf{x}_1$ was located in **A** and both the covariate $\mathbf{x}_2$ and the outcome $\mathbf{y}$ in **B**. Similarly to scenario 1, in 2_a all records from dataset **A** are located in dataset **B**, and some records from **A** were not located in **B** in scenario 2_b.

Scenario	Dataset **A**		Real matches	Dataset **B**	
	Regression Variables	Number of Records		Regression Variables	Number of Records
1_a	$\mathbf{x}_1$, $\mathbf{x}_2$	200	200 ($\mathbf{A} \subset \mathbf{B}$)	$\mathbf{y}$	800
1_b	$\mathbf{x}_1$, $\mathbf{x}_2$	300	200 ($\mathbf{A} \subsetneq \mathbf{B}$)	$\mathbf{y}$	800
2_a	$\mathbf{x}_1$	200	200 ($\mathbf{A} \subset \mathbf{B}$)	$\mathbf{x}_2$, $\mathbf{y}$	800
2_b	$\mathbf{x}_1$	300	200 ($\mathbf{A} \subsetneq \mathbf{B}$)	$\mathbf{x}_2$, $\mathbf{y}$	800

Table 3.1: Characteristics of all simulated scenarios.

In the scenarios either good or relatively bad linking variables were available, drawn from discrete uniform distributions. All linkage variables were free from error and therefore all true matches were present in the analysis. In the good situation five linkage variables were present in both datasets with respectively 30, 8, 7, 4, and 2 unique values (13440 unique patterns $\mathbf{g}$). In the bad situation only four variables were available with 30, 8, 7, and 2 unique values (3360 unique patterns $\mathbf{g}$). In the scenarios with 200 records in $\mathbf{A}$ and 800 in $\mathbf{B}$ we would expect 12 wrong matches (6% of all matches) with the good linking variables and 48 wrong matches (19% of all matches) with the bad linkage variables. In addition, the maximum posterior probability of a match q_{ij} was approximately 0.9 with good linking variables and approximately 0.8 for the scenarios with bad linkage variables.

We maximized the likelihood to estimate the parameters from the Fellegi-Sunter model. In the simulation both linear and logistic relations between one outcome measure and two covariates were simulated

$$y = f^{-1}(a + x_1\beta_1 + x_2\beta_2 + \epsilon) \tag{3.9}$$

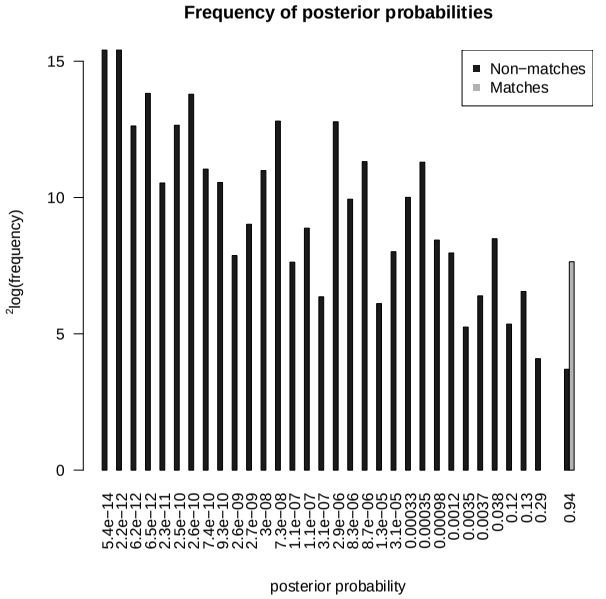
where f^{-1} is a link function, which is the identity function with continuous outcome y and the logit function with binary outcome y. We simulated an intercept $a = 0$, $\beta_1 = 1.5$, and $\beta_2 = -2$. The residuals ϵ were normally distributed with mean 0 and standard deviation 2. Two covariates x_1 and x_2 were drawn from a multivariate normal distribution with means 0, variances 1, and covariance 0.2.

Regression models

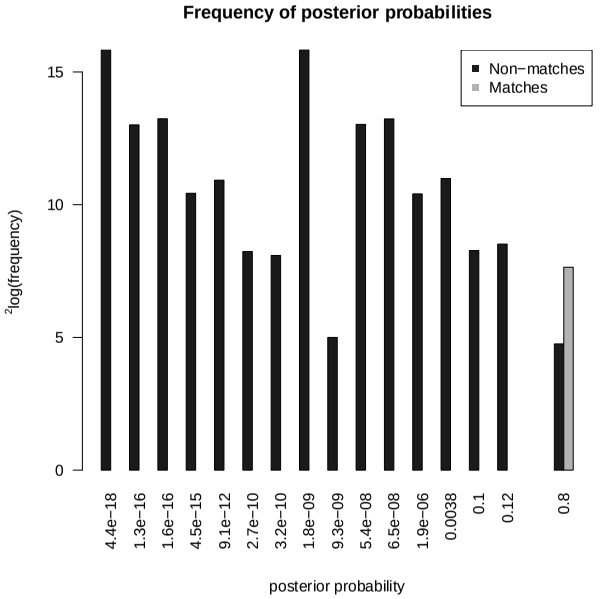
In linkage problems $\mathbf{Q}$ often contains a large number of comparisons with a very low probability. To reduce bias a simple pre-analysis procedure is to assign the value zero to all these highly unlikely data pairs. Because we simulated a fairly simple linkage problem this phenomenon was clearly seen in all $\mathbf{Q}$ matrices (figures 3.1 and 3.2). Notice that we assigned the value zero to all comparisons that did not have the highest probability of a match, resulting in the matrix $\mathbf{Q}'$.

For all of the following analysis approaches the bias, mean squared error, and coverage of the 95% confidence interval were calculated

LL: extended LL estimator.

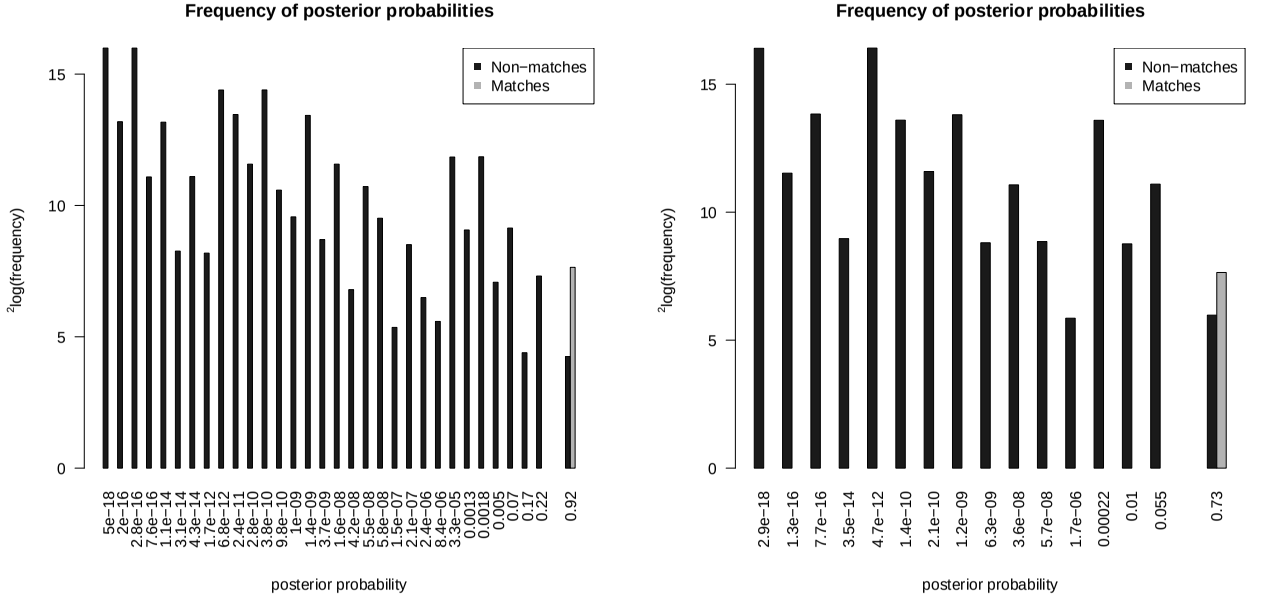


(a) Scenario a with good linkage variables



(b) Scenario a with bad linkage variables

Figure 3.1: Frequencies of posterior probabilities in the $\mathbf{Q}$ matrix taken from one simulated run for scenario a with either good or bad linkage variables.



(a) Scenario b with good linkage variables

(b) Scenario b with bad linkage variables

Figure 3.2: Frequencies of posterior probabilities in the $\mathbf{Q}$ matrix taken from one simulated run for scenario b with either good or bad linkage variables.

WLS_1: use the $\mathbf{Q}'$ matrix to generate the weighting matrix W.

WLS_2: all rows from $\mathbf{Q}'$ with one or more non-zero values are weighted so that $\sum_{i=1}^{n} q_{ij} = 1$, following the assumption that all records from **A** have one true match in **B**.

WLS_{3a}: assign the value zero to all entries in the rows from $\mathbf{Q}'$ with more than one non-zero value. This strategy is based on the assumption that there is a maximum of one true match for all rows in **A**. If it is unclear which record from **B** is the true match for record from **A** this record from **A** is discarded from the analysis.

WLS_{3b}: first perform the procedure suggested in WLS_{3a}. Then, treat the removal of rows from **Q** as a missing data problem by estimating the covariate or outcome from **B** that should belong to the regression variables from **A** by multiple imputation.

WLS_4: randomly select one non-zero value in the rows from $\mathbf{Q}'$ with more than one non-zero value. To all other values zero is assigned. This procedure is repeated an arbitrary number of times, which was 50 in our simulations.

For the variance estimation of the regression coefficients the bootstrap procedure from LL was used. All simulations were performed in R [84] and all scenarios were repeated 5000 times.

Results

The results of the simulations with continuous outcome and continuous covariates are summarized in tables 3.2 and 3.3. In all scenarios the LL, WLS_{3a}, and WLS_{3b} approaches gave the lowest bias with both good and bad linkage variables. The methods WLS_2, WLS_4, and especially WLS_1 gave more biased estimates.

In scenarios 1_a all methods were comparable in performance with good linking variables. In the regression models both βs were accurately estimated. The highest bias was present in the WLS_1 method and the lowest bias in the LL method (-0.083 compared to -0.001 in β_1 and 0.112 compared to 0.003 in β_2). In addition, the MSE ranged from approximately 0.02 to 0.03 for all methods in both βs. The coverage of the 95% confidence interval of β_1 was accurately estimated in all methods. This was also observed for β_2, with exception of WLS_2 that slightly underestimated the coverage.

With bad linking variables the differences between methods in scenario 1_a increased. Unbiased estimators were still achieved by using the LL, WLS_{3a}, or WLS_{3b} method. However, the other methods were biased ranging from -0.165 (for WLS_2, WLS_4) to -0.283 (for WLS_1) for β_1. For β_2 the bias ranged from 0.223 (for WLS_2, WLS_4) to 0.378 (for WLS_1). In addition, this over- and underestimation decreased the coverage of the 95% confidence interval.

In scenarios 1_b all methods gave biased results. The bias was approximately -0.04 for LL, WLS_{3a}, and WLS_{3b}, -0.11 for WLS_1, and both the WLS_2 and WLS_4 method underestimated β_1 by -0.08. Similar differences between methods were found in the estimation of β_2, in which all methods overestimated the regression coefficient.

With bad linking variables in scenario 1_b all estimations of the regression coefficients were highly biased. Similar to the other simulations LL, WLS_{3a}, and WLS_{3b} gave the best estimations, followed by WLS_2, and WLS_4 and the worst estimations were obtained with WLS_1. In addition, the coverage of the 95% confidence interval was low and ranged from 0.49 to 0.87 for β_1 and from 0.27 to 0.80 for β_2.

The bias in scenarios 2 was comparable to scenarios 1 for all the correction methods. In scenario 2_a the LL, WLS_{3a}, and WLS_{3b} methods gave unbiased results regardless of the discriminative power of the linking variables. The WLS_2 and WLS_4 methods were biased and the WLS_1 method gave the highest biased estimates.

In scenario 2 the LL method structurally underestimated the coverage of the 95% confidence interval of β_1, because we ignored the uncertainty that arose with the stepwise procedure which removed the correlation between covariates.

With binary outcome data in scenario 1 similar differences were found for the WLS and the LL methods (therefore the data is not shown). All methods were unbiased in scenario 1_a and in scenario 1_b the LL, WLS_{3a}, and WLS_{3b} approaches gave the lowest bias, and the other WLS methods performed worse.

With binary outcomes in scenario 2 the performance of all WLS approaches was similar to the scenarios with continuous outcome. The LL method was, however, biased in scenario 2 because of the non-linear transformation problem recognized previously at page 47. In scenario 2_a, where the LL estimator gave unbiased results for the linear model, the logistic version showed large bias.

			β_1			β_2		
			$E(\hat{\beta}_1 - \beta_1)$	MSE	Coverage of the 95% C.I.	$E(\hat{\beta}_2 - \beta_2)$	MSE	Coverage of the 95% C.I.
Scenario 1_a	Good	LL	-0.001 (0.145)	0.021	0.951	0.003 (0.148)	0.021	0.946
		WLS_1	-0.083 (0.151)	0.021	0.951	0.112 (0.158)	0.030	0.933
		WLS_2	-0.045 (0.144)	0.020	0.943	0.061 (0.148)	0.023	0.926
		WLS_{3a}	-0.002 (0.147)	0.021	0.953	0.002 (0.148)	0.021	0.949
		WLS_{3b}	-0.003 (0.148)	0.022	0.949	0.004 (0.15)	0.022	0.945
		WLS_4	-0.045 (0.021)	0.020	0.951	0.061 (0.022)	0.023	0.933
	Bad	LL	-0.001 (0.158)	0.023	0.953	0.002 (0.159)	0.025	0.952
		WLS_1	-0.283 (0.163)	0.023	0.845	0.378 (0.169)	0.107	0.742
		WLS_2	-0.165 (0.150)	0.020	0.789	0.222 (0.154)	0.05	0.666
		WLS_{3a}	-0.001 (0.163)	0.025	0.948	0.004 (0.166)	0.027	0.945
		WLS_{3b}	-0.006 (0.170)	0.027	0.932	0.011 (0.173)	0.029	0.925
		WLS_4	-0.166 (0.023)	0.020	0.845	0.223 (0.024)	0.050	0.742
Scenario 1_b	Good	LL	-0.040 (0.154)	0.021	0.940	0.056 (0.154)	0.025	0.935
		WLS_1	-0.119 (0.157)	0.022	0.920	0.161 (0.159)	0.039	0.890
		WLS_2	-0.083 (0.152)	0.021	0.910	0.113 (0.152)	0.03	0.878
		WLS_{3a}	-0.041 (0.156)	0.022	0.936	0.058 (0.155)	0.026	0.932
		WLS_{3b}	-0.043 (0.157)	0.022	0.934	0.06 (0.157)	0.026	0.923
		WLS_4	-0.083 (0.023)	0.021	0.920	0.113 (0.023)	0.030	0.890
	Bad	LL	-0.150 (0.172)	0.025	0.866	0.201 (0.173)	0.052	0.808
		WLS_1	-0.395 (0.163)	0.025	0.570	0.526 (0.171)	0.183	0.340
		WLS_2	-0.295 (0.159)	0.023	0.493	0.393 (0.163)	0.112	0.274
		WLS_{3a}	-0.159 (0.180)	0.030	0.841	0.212 (0.181)	0.058	0.761
		WLS_{3b}	-0.165 (0.185)	0.031	0.824	0.219 (0.188)	0.061	0.732
		WLS_4	-0.296 (0.026)	0.024	0.570	0.394 (0.027)	0.113	0.340

Table 3.2: Simulation results for regression estimates in scenario 1 with continuous outcome and continuous covariates. Covariates $\mathbf{x}_1$ and $\mathbf{x}_2$ are located in dataset **A** and the outcome variable **y** in dataset **B**.

			β_1			β_2		
			$E(\hat{\beta}_1 - \beta_1)$	MSE	Coverage of the 95% C.I.	$E(\hat{\beta}_2 - \beta_2)$	MSE	Coverage of the 95% C.I.
Scenario 2_a	Good	LL	-0.003 (0.143)	0.010	0.853	0.004 (0.150)	0.021	0.951
		WLS_1	-0.046 (0.151)	0.021	0.951	0.120 (0.158)	0.025	0.932
		WLS_2	-0.025 (0.147)	0.02	0.942	0.065 (0.149)	0.022	0.924
		WLS_{3a}	0.001 (0.151)	0.022	0.945	0.004 (0.152)	0.023	0.949
		WLS_{3b}	0.001 (0.152)	0.022	0.943	0.004 (0.152)	0.023	0.947
		WLS_4	-0.024 (0.021)	0.02	0.951	0.065 (0.022)	0.022	0.932
	Bad	LL	-0.003 (0.151)	0.010	0.842	-0.001 (0.158)	0.023	0.961
		WLS_1	-0.138 (0.160)	0.022	0.936	0.396 (0.171)	0.045	0.695
		WLS_2	-0.084 (0.147)	0.020	0.901	0.231 (0.155)	0.029	0.627
		WLS_{3a}	0.001 (0.164)	0.026	0.953	-0.004 (0.163)	0.027	0.951
		WLS_{3b}	0.002 (0.169)	0.027	0.941	0.001 (0.169)	0.029	0.942
		WLS_4	-0.084 (0.022)	0.020	0.936	0.232 (0.024)	0.029	0.695
Scenario 2_b	Good	LL	-0.024 (0.143)	0.010	0.844	0.060 (0.155)	0.021	0.934
		WLS_1	-0.065 (0.153)	0.022	0.942	0.172 (0.162)	0.028	0.868
		WLS_2	-0.045 (0.148)	0.021	0.936	0.120 (0.154)	0.024	0.854
		WLS_{3a}	-0.023 (0.153)	0.022	0.946	0.061 (0.157)	0.024	0.923
		WLS_{3b}	-0.023 (0.154)	0.023	0.945	0.061 (0.158)	0.024	0.920
		WLS_4	-0.045 (0.022)	0.021	0.942	0.120 (0.024)	0.024	0.868
	Bad	LL	-0.080 (0.137)	0.010	0.790	0.214 (0.172)	0.025	0.787
		WLS_1	-0.185 (0.161)	0.023	0.862	0.548 (0.169)	0.060	0.271
		WLS_2	-0.143 (0.155)	0.022	0.811	0.412 (0.162)	0.044	0.210
		WLS_{3a}	-0.081 (0.176)	0.029	0.918	0.225 (0.183)	0.038	0.728
		WLS_{3b}	-0.083 (0.183)	0.030	0.907	0.231 (0.189)	0.040	0.717
		WLS_4	-0.143 (0.024)	0.022	0.862	0.412 (0.026)	0.044	0.271

Table 3.3: Simulation results for regression estimates in scenario 2 with continuous outcome and continuous covariates. Covariate $\mathbf{x}_1$ is located in dataset **A** and both covariate $\mathbf{x}_2$ and the outcome variable $\mathbf{y}$ in dataset **B**.

Discussion

In this chapter, we proposed a number of approaches to regression analysis of data derived from record linkage. The analysis method developed by Lahiri and Larsen [30] has been extended to relax its assumptions and we investigated the use of weighted least squares methods.

In record linkage problems bias is introduced by wrong covariates and outcome pairs. LL made it possible to remove these pairs from the analysis with a number of strong assumptions on the structure of the data. However, in most record linkage situations these assumptions are not met and the LL estimator is then also biased.

The impact of wrong data pairs could be seen in the WLS approaches. In the WLS_1 approach, in which the least assumptions were made to decrease the impact of wrong data pairs, the bias was highest. By introducing all the LL assumptions in a WLS approach (WLS_{3a} and WLS_{3b}), we showed that WLS is also able to create unbiased results regardless of the discriminative strength of the linking variables. Furthermore, the performance of the WLS_{3a}, WLS_{3b}, and LL estimators were comparable if the LL assumptions were violated.

Our results showed that the regression coefficients can only be unbiased when the assumptions are valid, thus it is necessary that the analyst has prior knowledge on the nature of the data. Note that the structure of the data is not similar in all record linkage problems and different techniques are needed to exclude wrong covariate and outcome pairs from the analysis.

The WLS method has more flexibility than the LL method and could more easily be used in different situations. The $\mathbf{W}$ matrix can be modified to fit particular assumptions, whereas more complex procedures are necessary for the LL method [73, 72]. In addition, the WLS approach can be used in more general situations where covariates and outcomes are located in multiple datasets. The WLS method was able to give unbiased estimates for the relation between a binary outcome and covariates that were located in two datasets, whereas the LL method was biased.

Another approach to the data derived from record linkage is to use error in variables models, in which records that are linked to more than one record are considered to have some measurement error. The rows with more than one match are assumed to follow either a normal or uniform distribution and for each row the corresponding sufficient statistics may be calculated. The regression coefficients can be estimated with a Bayesian approach or likelihood maximization. This

approach had similar characteristics as WLS_1 and therefore we did not show its results.

Because record linkage problems are used in a lot of different settings, more research is needed to measure the performance of the different estimators in other scenarios. Furthermore more operations could be suggested for the $\mathbf{W}$ matrix in the WLS method to fit more specific linkage situations. Another problem that requires more attention is the bootstrap procedure to estimate the variance of the regression parameters. Because the calculation of the $\mathbf{Q}$ matrix is repeated in each iteration of the bootstrap procedure, it is computationally intensive and requires cluster or grid computing.