Discussion paper

Correcting for linkage errors in contingency tables – a cautionary tale

Sander Scholtus
Natalie Shlomo
Ton de Waal

January 2020

Content

Summary

Record linkage aims to bring records together from two or more files that belong to the same statistical entity such as an individual or a business. In this paper, we focus on incorrectly linked pairs which result from records in two or more datasets being linked incorrectly due to errors, missing values or changes over time in the variables that are used in the matching procedure. It is well known that naïvely treating a probabilistically linked file as if there are no linkage errors leads to biased inference. We present two approaches for compensating for linkage error when analysing a two-way contingency table for categorical data where one variable is coming from one file and the second variable is coming from the other file. One approach is an unbiased correction and we show that this often equates to the approach used by Chipperfield and Chambers (2015) under the exchangeable linkage error model, which is a widely used model for linkage errors and the one that we will adopt in this paper. The other approach is a biased correction, but which often leads to lower mean square error than the unbiased approach.

Under the exchangeable linkage error model, we will study the following fundamental questions: can a compensation approach for linkage error improve on the naïve approach, where linkage error is not compensated for, and, if so, under what conditions? To this end, we will compare the compensation approaches to the naïve approach. We will examine these three approaches in detail, both by means of an analytical study as well as by means of a simulation study. In particular, we examine estimation errors, bias, variance and mean square error of the three approaches. We show in this paper that the approach to take for a given situation depends on the characteristics of the table and whether the table shows dependent or independent attributes, and more specifically whether the particular cell of the table has a positive, negative or no association.

Keywords

probabilistic record linkage, exchangeable linkage error model, contingency tables, linkage error correction

The views expressed in this paper are those of the authors and do not necessarily reflect the policies of Statistics Netherlands. The authors would like to thank Arnout van Delden for his helpful comments on an earlier version of this paper.

1. Introduction

Record linkage aims to bring records together from two or more files that belong to the same statistical entity such as an individual or a business. The seminal paper by Fellegi and Sunter (1969) provides a framework for probabilistic record linkage and resulted in decades of research on linking datasets. In the official statistics framework, the first large-scale record linkage exercises were between census and coverage surveys. At Statistics Netherlands, probabilistic record linkage has, for instance, been applied to link health insurance data to the Netherlands Twin Register, community pharmacy records to a birth cohort study, and the Dutch Population Register and the Employment Register (see Van Grootheest et al., 2015).

Given the increasing use of administrative and new forms of data being used in statistical systems there is an increasing need to continue to develop and research record linkage approaches and in particular, there is a growing emphasis on accounting for linkage errors in statistical analysis. In a linked dataset there are two types of errors: incorrectly linked pairs which may potentially incur bias in statistical analyses and missed links which impacts on coverage and potential bias if those missed links differ in their characteristics from the found links. For instance, linkage error in linked health insurance data and data from the Netherlands Twin Register may lead to incorrect conclusions with respect to the statistical effects of health care.

In this paper, we focus on incorrectly linked pairs which result from records in two or more datasets being linked incorrectly due to errors, missing values or changes over time in the variables that are used in the matching procedure. Throughout the paper, we will assume that both datasets contain the same units and the aim is to link all of them (one-to-one linkage).

It is well known that naïvely treating a probabilistically linked file as if there are no linkage errors leads to biased inference. There has been early work on compensating for linkage errors in regression models, see Scheuren and Winkler (1993 and 1997) and Lahiri and Larsen (2005). Chambers (2009) proposed bias-corrected linear regression coefficients and Kim and Chambers (2012a, 2012b) also looked at more general models using estimating equations. Chipperfield et al. (2011) and Chipperfield and Chambers (2015) focused on categorical data and proposed an unbiased method for compensating for linkage errors.

Our focus in this paper is also on categorical data. We follow the Chambers (2009) assumption of the exchangeable linkage error model and this is described in Section 2. The error rates for the exchangeable linkage error model can be estimated by sub-sampling linked pairs and manually checking for errors in these linked pairs. In addition, bootstrapping approaches have been proposed, see Winglee, Valliant and Scheuren (2005) and Chambers and Chipperfield (2015). We assume that the linkage error rate is known under the exchangeable linkage error

model and we address the problem of how to compensate for the linkage error when the variables of interest are categorical.

We present two approaches for compensating for linkage error when analysing a two-way contingency table where one variable is coming from one file and the second variable is coming from the other file that are linked through a probabilistic record linkage process. One approach is an unbiased correction and we show that this often equates to the approach used by Chipperfield and Chambers (2015) under the exchangeable linkage error model. The other approach is a biased correction, but which often leads to lower mean square error than the unbiased approach. Under the exchangeable linkage error model, we will study the following fundamental questions: can a compensation approach for linkage error improve on the naïve approach, where linkage error is not compensated for, and, if so, under what conditions? To this end, we will compare the compensation approaches to the naïve approach. We will examine the three approaches in detail, both by means of an analytical study as well as by means of a simulation study. In particular, we examine estimation errors, bias, variance and mean square error of the three approaches.

Results show in this paper that the compensation approach to take for a given situation depends on the characteristics of the table and whether the table shows dependent or independent attributes, and more specifically whether the particular cell of the table has a positive, negative or no association.

Section 2 provides the motivation for correcting for linkage errors in contingency tables with a small toy example and defines the two compensation approaches based on the biased and unbiased correction techniques. In Section 3, we present theoretical results on which approach outperforms the naïve approach in terms of the estimation error in individual cell values, depending on the type of association of the cell value. Section 4 shows theoretical results of the bias and variance of the cell values under the naïve and compensation approaches. All theory is tested in Section 5 in a simulation study based on a set of tables with dependent and independent attributes and varying cell associations and, in particular, we show the impact of the naïve and compensation approaches on statistical tests of independence. We conclude in Section 6 with a discussion.

2. Methods

In this section, we first describe the record linkage problem and the effect of linkage errors on contingency tables. We also introduce the exchangeable linkage error model (see Chambers, 2009), and the three approaches that we will study in this paper: the naïve approach and two correction approaches. These three approaches will be examined under the exchangeable linkage error model.
We briefly describe a fourth approach proposed by Chipperfield and Chambers (2015). As already mentioned in the Introduction we will show that this approach is basically the same as our unbiased correction approach under the assumptions made here.

2.1 Notation and terminology

In this paper we focus on a basic record linkage problem. We suppose that there are two data files A and B of n records ($n \geq 2$), containing variables (x, y) and (x, z), respectively. We are interested in estimating the contingency table of the categorical variables y and z. To do this, the two datasets first need to be linked on the common variable(s) x. It is assumed that both datasets contain the same units, i.e., linkage is one-to-one.

For the purpose of constructing the contingency table of y and z, variable y in the first dataset can be coded into dummy variables as a binary $n \times J$ matrix $\mathbf{Y}$, where J is the number of categories of variable y. The elements of this matrix are $y_{ij} = 1$ if the unit in record i of the first dataset belongs to category j of variable y, and $y_{ij} = 0$ otherwise. Similarly, variable z in the second dataset can be coded as a binary $n \times K$ matrix $\mathbf{Z}$, where K is the number of categories of variable z. After linking the two datasets, the $J \times K$ target contingency table is then given by $\mathbf{T} = \mathbf{Y}'\mathbf{Z}$ (where $'$ denotes taking the transpose of a matrix), with typical element $t_{jk} = \sum_{i=1}^{n} y_{ij} z_{ik}$.

During the record linking process, linkage errors may occur. Since both datasets contain the same entities, random linkage errors can be represented by a random permutation of the order of some units in the second dataset. As a result, instead of $\mathbf{Z}$ we observe $\mathbf{Z}^* = \mathbf{C}\mathbf{Z}$, where, due to the random linkage errors, $\mathbf{C}$ is a stochastic permutation matrix of order n. That is to say, each row and each column of $\mathbf{C}$ contains exactly one element equal to 1 and all other elements equal to 0. The target contingency table is observed as $\widehat{\mathbf{T}}^* = \mathbf{Y}'\mathbf{Z}^* = \mathbf{Y}'\mathbf{C}\mathbf{Z}$, with typical element $\hat{t}^*_{jk} = \sum_{i=1}^{n} y_{ij} (\sum_{l=1}^{n} c_{il} z_{lk})$. In reality, $\mathbf{C}$ would be unobserved, and correction for linkage error is non-trivial.

Example

To illustrate the above notation, we consider an example with $n = 10$ and variables y and z that both have two categories. Suppose that the two dummy variables are:

$$\mathbf{Y} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}'$$

and

$$\mathbf{Z} = \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 1 \end{pmatrix}'.$$

The true contingency table of y and z is therefore:

$$\mathbf{T} = \mathbf{Y}'\mathbf{Z} = \begin{pmatrix} 4 & 3 \\ 2 & 1 \end{pmatrix}.$$

For the purpose of this example, we suppose that linkage errors occur according to the following permutation matrix:

$$\mathbf{C} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Thus, the 4th and 10th unit in the second dataset are permuted and the other units are linked correctly. Hence,

$$\mathbf{Z}^* = \mathbf{C}\mathbf{Z} = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \end{pmatrix}'$$

and the observed contingency table is given by

$$\widehat{\mathbf{T}}^* = \mathbf{Y}'\mathbf{Z}^* = \begin{pmatrix} 3 & 4 \\ 3 & 0 \end{pmatrix} \neq \mathbf{Y}'\mathbf{Z}.$$

We assume that the linkage errors, and hence the permutation matrix $\mathbf{C}$, are stochastic. Since we want to focus here on the effect of linkage errors on estimators, we treat $\mathbf{Y}$ and $\mathbf{Z}$ as fixed. It is easy to see that, in the presence of random linkage errors, the naïve estimator $\widehat{\mathbf{T}}^* = \mathbf{Y}'\mathbf{Z}^*$ is biased for the true contingency table. Namely:

$$E\left(\widehat{\mathbf{T}}^*\right) = E(\mathbf{Y}'\mathbf{C}\mathbf{Z}) = \mathbf{Y}'E(\mathbf{C})\mathbf{Z} \equiv \mathbf{Y}'\mathbf{Q}\mathbf{Z}, \qquad (1)$$

which in general is not equal to $\mathbf{T} = \mathbf{Y}'\mathbf{Z}$. Here, $\mathbf{Q} = E(\mathbf{C})$ denotes the matrix of linkage error probabilities, with typical element $q_{il} = \Pr(c_{il} = 1)$, where $\Pr$ reflects the stochastic process by which $\mathbf{C}$ is generated. Note that all rows and columns of $\mathbf{Q}$ must sum to one, under the assumption that linkage is one-to-one.

As seen in (1) and below, the matrix of linkage error probabilities $\mathbf{Q}$ plays a key role in estimators that compensate for linkage errors. In what follows, we will assume that $\mathbf{Q}$ is an $n \times n$ matrix with each diagonal element for the probability of a correct link equal to, say, q and each off-diagonal element equal to $\delta = (1 - q)/(n - 1)$. That is to say,

$$\mathbf{Q} = \begin{pmatrix} q & \delta & \delta & \cdots & \delta & \delta \\ \delta & q & \delta & \cdots & \delta & \delta \\ \delta & \delta & q & \cdots & \delta & \delta \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \delta & \delta & \delta & \cdots & q & \delta \\ \delta & \delta & \delta & \cdots & \delta & q \end{pmatrix}.$$

In matrix-vector notation, a matrix of this form can be written as:

$$\mathbf{Q} = q\mathbf{I} + \delta(\mathbf{u}\mathbf{u}' - \mathbf{I}), \tag{2}$$

where $\mathbf{u}$ denotes the n vector of ones and $\mathbf{I}$ the identity matrix of order n. This corresponds to a simple, but widely used model for linkage errors that is known as the *exchangeable linkage error model* (Chambers, 2009). Despite the apparent simplicity of the model, we believe that it provides important insights into the properties of various compensation approaches for linkage error, and that these insights are also useful for more complicated linkage error models.

In practice, the exchangeable linkage error model is often applied within blocks defined by a common variable x in files A and B, with possibly different values of q for different blocks.

We will make the weak technical assumption that

$$1/n < q < 1. \tag{3}$$

The left inequality implies that the linking process is at least better than linking the records completely at random. The right inequality rules out the trivial case where the linkage process is deterministic and perfect ($q = 1$). For future reference, we note that a matrix $\mathbf{Q}$ of the form (2) has an inverse of the form

$$\mathbf{Q}^{-1} = \frac{1}{nq - 1}\{(n - 1)\mathbf{I} - (1 - q)\mathbf{u}\mathbf{u}'\} \tag{4}$$

and that (3) implies that $2 - \delta n > 1$. The left inequality in assumption (3) is needed to ensure that this inverse exists.

2.2 Correcting for linkage errors

In this paper, we consider several estimators to correct for linkage errors:
a. $\widehat{\mathbf{T}}^{Q} = (\mathbf{QY})'\mathbf{Z}^{*}$, an estimator that uses the $\mathbf{Q}$ matrix;
b. $\widehat{\mathbf{T}}^{BC} = \mathbf{Y}'\mathbf{Q}^{-1}\mathbf{Z}^{*}$, an estimator that uses the inverse of the $\mathbf{Q}$ matrix;
c. $\widehat{\mathbf{T}}^{CC}$ proposed by Chipperfield and Chambers (2015) (see below).

The idea of the first approach is that $(\mathbf{CY})'\mathbf{Z}^{*} = (\mathbf{CY})'\mathbf{CZ} = \mathbf{Y}'\mathbf{Z} = \mathbf{T}$. The second equality follows from the fact that $\mathbf{C}$ is a permutation matrix so $\mathbf{C}'\mathbf{C} = \mathbf{I}$. By replacing the unknown matrix $\mathbf{C}$ in this expression by its expectation $\mathbf{Q}$, we obtain $\widehat{\mathbf{T}}^{Q} = (\mathbf{QY})'\mathbf{Z}^{*}$.

The idea of the second approach is that $\mathbf{Y}'\mathbf{Q}^{-1}\mathbf{Z}^{*}$ is an unbiased estimator for the true contingency table of y and z. Namely:

$$E(\mathbf{Y}'\mathbf{Q}^{-1}\mathbf{Z}^{*}) = E(\mathbf{Y}'\mathbf{Q}^{-1}\mathbf{CZ}) = \mathbf{Y}'\mathbf{Q}^{-1}E(\mathbf{C})\mathbf{Z} = \mathbf{Y}'\mathbf{Q}^{-1}\mathbf{QZ} = \mathbf{Y}'\mathbf{Z} = \mathbf{T}.$$

We use the superscript BC to indicate that this is a bias-corrected estimator.

The approach of Chipperfield and Chambers (2015) applies the following additive correction to the naïve estimated contingency table:

$$\widehat{\mathbf{T}}^{CC} = \mathbf{Y}'\mathbf{Z}^{*} + \mathbf{Y}'(\mathbf{I} - \mathbf{Q})\hat{\Delta}^{CC}, \tag{5}$$

with $\hat{\Delta}^{CC} = \mathbf{Y}\widehat{\mathbf{\Pi}}^{CC}$, and $\widehat{\mathbf{\Pi}}^{CC}$ is an estimate for the $J \times K$ matrix $\mathbf{\Pi}$ of conditional probabilities $\pi_{jk} = \Pr(z = k|y = j)$. Chipperfield and Chambers (CC) proposed to estimate $\widehat{\mathbf{\Pi}}^{CC}$ by the following iterative algorithm:

1. Start with $\widehat{\mathbf{\Pi}}_0$ based on the naïve estimate $\widehat{\mathbf{T}}_0 = \widehat{\mathbf{T}}^{*}$: compute $\widehat{\mathbf{\Pi}}_0 = \mathbf{D}_0^{-1}\widehat{\mathbf{T}}_0$, where $\mathbf{D}_0$ is a diagonal matrix with the row sums of $\widehat{\mathbf{T}}_0$ on the main diagonal. Let $h = 0$.
2. Compute a new estimate according to (5): $\widehat{\mathbf{T}}_{h+1} = \mathbf{Y}'\mathbf{Z}^{*} + \mathbf{Y}'(\mathbf{I} - \mathbf{Q})\hat{\Delta}_h$, with $\hat{\Delta}_h = \mathbf{Y}\widehat{\mathbf{\Pi}}_h$.
3. Compute a new estimate for $\mathbf{\Pi}$: $\widehat{\mathbf{\Pi}}_{h+1} = \max\{\mathbf{D}_{h+1}^{-1}\widehat{\mathbf{T}}_{h+1}, 0\}$, where $\mathbf{D}_{h+1}$ is a diagonal matrix with the row sums of $\widehat{\mathbf{T}}_{h+1}$ on the main diagonal and the maximum is taken element-wise.
4. Stop when the estimates have converged. Otherwise, let $h = h + 1$ and return to step 2.

Upon convergence the algorithm yields $\widehat{\mathbf{\Pi}}^{CC}$ and $\hat{\Delta}^{CC}$. Note that the truncation in step 3 is built in to prevent the occurrence of estimated conditional probabilities outside the interval $[0,1]$.

The CC derived estimator (5) is a bias-corrected estimator, under a slightly different framework that also treats y and z as random variables. They considered the bias of $\mathbf{Y}'\mathbf{Z}^{*}$ conditional on $\mathbf{Y}$ and assumed that $\mathbf{C}$ and $\mathbf{Z}$ are conditionally independent given $\mathbf{Y}$. Under this assumption, it follows that:

$$E(\mathbf{Y}'\mathbf{Z}^{*} - \mathbf{Y}'\mathbf{Z}|\mathbf{Y}) = \mathbf{Y}'\{E(\mathbf{C}|\mathbf{Y}) - \mathbf{I}\}E(\mathbf{Z}|\mathbf{Y}) = \mathbf{Y}'(\mathbf{Q} - \mathbf{I})\mathbf{Y}\mathbf{\Pi}.$$

Example (continued)
Suppose that in the above example with $n = 10$ the $\mathbf{Q}$ matrix is given by (2) with $q = 0.91$ and therefore $\delta = 0.01$. For the first alternative estimator $\widehat{\mathbf{T}}^{Q}$, we then find:

$$\widehat{\mathbf{T}}^{Q} = (\mathbf{QY})'\mathbf{Z}^{*} = \frac{1}{100}\begin{pmatrix} 312 & 388 \\ 288 & 12 \end{pmatrix} = \begin{pmatrix} 3.12 & 3.88 \\ 2.88 & 0.12 \end{pmatrix}.$$

In this example, this approach adjusts all entries in the contingency table a bit in the right direction.

For the second alternative estimator $\widehat{\mathbf{T}}^{BC}$, we find:

$$\widehat{\mathbf{T}}^{BC} = \mathbf{Y}'\mathbf{Q}^{-1}\mathbf{Z}^{*} = \frac{1}{90}\begin{pmatrix} 258 & 372 \\ 282 & -12 \end{pmatrix} \approx \begin{pmatrix} 2.8667 & 4.1333 \\ 3.1333 & -0.1333 \end{pmatrix}.$$

This is worse than the naïve estimate $\widehat{\mathbf{T}}^{*} = \mathbf{Y}'\mathbf{Z}^{*}$. In fact, all entries in the contingency table $\widehat{\mathbf{T}}^{BC}$ are further from the true entries in $\mathbf{T}$ than $\widehat{\mathbf{T}}^{*}$. Note also that the entry in the second row and column of $\widehat{\mathbf{T}}^{BC}$ is negative, which is not acceptable for a contingency table.

The procedure of CC yields

$$\widehat{\mathbf{\Pi}}^{CC} = \begin{pmatrix} 0.4109 & 0.5891 \\ 1.0000 & 0.0000 \end{pmatrix},$$

from which the following estimate is obtained:

$$\widehat{\mathbf{T}}^{CC} = \begin{pmatrix} 2.8763 & 4.1237 \\ 3.1237 & -0.1237 \end{pmatrix}.$$

This estimate is quite close to the bias-corrected estimate $\widehat{\mathbf{T}}^{BC}$. In particular, it has the same undesirable properties in this example.

The similarity of the estimates $\widehat{\mathbf{T}}^{BC}$ and $\widehat{\mathbf{T}}^{CC}$ in the above example is not a coincidence. In fact, the following property can be shown.

Theorem 1. *For random linkage errors that are described by a matrix* $\mathbf{Q}$ *of the form (2) that satisfies assumption (3), if the truncation in step 3 of the algorithm of CC is never used, then this algorithm converges to the solution* $\widehat{\mathbf{T}}^{\mathrm{CC}} = \widehat{\mathbf{T}}^{\mathrm{BC}}$.

The proof of this theorem is given in Appendix A.1. The above example illustrates that even when the truncation step is used in the algorithm of CC, the resulting estimated contingency table is still close to the bias-corrected estimate based on $\mathbf{Q}^{-1}$.

Given Theorem 1, we will focus on the first two estimators, since in practice the properties of $\widehat{\mathbf{T}}^{CC}$ are similar to those of $\widehat{\mathbf{T}}^{BC}$. This was confirmed by our simulation study where we obtained the same results for the two estimators; hence, we do not show the former estimator in Section 5.

3. Theoretical results on estimation errors

In this section we examine the errors of the single entries in estimated contingency tables under the exchangeable linkage error model. We begin by introducing some further notation, and then examine the $\mathbf{Q}$ and $\mathbf{Q}^{-1}$ correction methods in turn.

3.1 Further notation

Consider a single entry of the true contingency table $\mathbf{T}$: $t_{jk} = (\mathbf{Y}'\mathbf{Z})_{jk} = \mathbf{y}_j'\mathbf{z}_k$, where $\mathbf{y}_j$ denotes a column of $\mathbf{Y}$ and $\mathbf{z}_k$ a column of $\mathbf{Z}$. The corresponding entries of the estimated tables $\widehat{\mathbf{T}}^*$, $\widehat{\mathbf{T}}^Q$, and $\widehat{\mathbf{T}}^{BC}$ are:

$$\begin{aligned}\hat{t}_{jk}^* &= (\mathbf{Y}'\mathbf{C}\mathbf{Z})_{jk} = \mathbf{y}_j'\mathbf{C}\mathbf{z}_k,\\ \hat{t}_{jk}^Q &= (\mathbf{Y}'\mathbf{Q}'\mathbf{C}\mathbf{Z})_{jk} = \mathbf{y}_j'\mathbf{Q}'\mathbf{C}\mathbf{z}_k,\\ \hat{t}_{jk}^{BC} &= (\mathbf{Y}'\mathbf{Q}^{-1}\mathbf{C}\mathbf{Z})_{jk} = \mathbf{y}_j'\mathbf{Q}^{-1}\mathbf{C}\mathbf{z}_k.\end{aligned}$$

Denote the errors in these entries as $e_{jk}^* = \hat{t}_{jk}^* - t_{jk}$, $e_{jk}^Q = \hat{t}_{jk}^Q - t_{jk}$, and $e_{jk}^{BC} = \hat{t}_{jk}^{BC} - t_{jk}$:

$$\begin{aligned}e_{jk}^* &= \mathbf{y}_j'(\mathbf{C} - \mathbf{I})\mathbf{z}_k,\\ e_{jk}^Q &= \mathbf{y}_j'(\mathbf{Q}'\mathbf{C} - \mathbf{I})\mathbf{z}_k,\\ e_{jk}^{BC} &= \mathbf{y}_j'(\mathbf{Q}^{-1}\mathbf{C} - \mathbf{I})\mathbf{z}_k.\end{aligned} \tag{6}$$

Since the estimators $\widehat{\mathbf{T}}^Q$ and $\widehat{\mathbf{T}}^{BC}$ are intended to correct for linkage errors, it might be reasonable to expect that $|e_{jk}^Q| < |e_{jk}^*|$ and $|e_{jk}^{BC}| < |e_{jk}^*|$. However, in the running example of Section 2 with $n = 10$ it was seen that in fact $|e_{jk}^{BC}| > |e_{jk}^*|$ for all entries in the target table. The aim of this section is to derive necessary and sufficient conditions for the inequalities $|e_{jk}^Q| < |e_{jk}^*|$ and $|e_{jk}^{BC}| < |e_{jk}^*|$ to hold. In Section 4, we will compare the bias and variance of the different estimators.

As is noted in the proof of Theorem 1 (see Appendix A.1), the marginal counts of the contingency table of y and z can be obtained from the separate data files and are therefore not subject to linkage errors. (This can also be seen in the running example of Section 2.) We denote the marginal count for category j of variable y by $r_j = \mathbf{y}_j'\mathbf{u} = \mathbf{u}'\mathbf{y}_j$. Similarly, we denote the marginal count for category k of variable z by $s_k = \mathbf{z}_k'\mathbf{u} = \mathbf{u}'\mathbf{z}_k$ where $\mathbf{u}$ denotes the n vector of ones as before.

It will be useful below to distinguish between three different cases, based on the sign of the quantity $nt_{jk} - r_j s_k$. Note that, if there is no association between the dummy variables $\mathbf{y}_j$ and $\mathbf{z}_k$, it holds that

$$t_{jk} = n\frac{r_j}{n}\frac{s_k}{n} = \frac{r_j s_k}{n}.$$

With this in mind, we will call the variables $\mathbf{y}_j$ and $\mathbf{z}_k$ *positively associated* when $nt_{jk} > r_j s_k$, *negatively associated* when $nt_{jk} < r_j s_k$, and *not associated* when $nt_{jk} = r_j s_k$. It should be noted that this association type refers to an individual cell value in a table and not to a measure of association for the table as a whole.

3.2 The Q-adjusted contingency table

For the estimator $\hat{\mathbf{T}}^Q = \mathbf{Y}'\mathbf{Q}'\mathbf{C}\mathbf{Z}$, the following result is derived in Appendix A.2.

Theorem 2. *For random linkage errors that are described by a matrix* $\mathbf{Q}$ *of the form (2) that satisfies assumption (3),* $\left|e_{jk}^{Q}\right| > \left|e_{jk}^{*}\right|$ *when one of the following two conditions holds:*

a. $nt_{jk} > r_j s_k$ *(positive association) and* $\hat{t}_{jk}^{*}$ *satisfies*

$$\frac{r_j s_k}{n} < \hat{t}_{jk}^{*} < t_{jk} + \delta\frac{nt_{jk} - r_j s_k}{2 - \delta n}; \tag{7}$$

b. $nt_{jk} < r_j s_k$ *(negative association) and* $\hat{t}_{jk}^{*}$ *satisfies*

$$t_{jk} + \delta\frac{nt_{jk} - r_j s_k}{2 - \delta n} < \hat{t}_{jk}^{*} < \frac{r_j s_k}{n}. \tag{8}$$

In all other cases, $\left|e_{jk}^{Q}\right| \leq \left|e_{jk}^{*}\right|$*, with equality holding only at the endpoints of (7) or (8) and in the special case* $nt_{jk} = r_j s_k$ *(no association) for* $\hat{t}_{jk}^{*} = t_{jk}$.

Example (continued)

To illustrate the result stated in Theorem 2, we revisit the example from Section 2. This example concerns a 2×2 contingency table with fixed marginal counts (as noted above), which has just one degree of freedom. To study the effect of linkage errors on this table it therefore suffices to consider just one of the entries, because the errors e_{jk}^{*} in all entries will be equal up to a sign change (as could be seen in the numerical results from Section 2). Here, we will focus on the upper left entry $t_{11} = 4$.

In this example, it holds that $n = 10$, $r_1 = 7$, $s_1 = 6$. Since $nt_{11} = 40 < 42 = r_1 s_1$, the case of negative association applies here. The relevant condition (8) is given by

$$4 - \frac{2}{190} < \hat{t}_{11}^{*} < \frac{42}{10}.$$

The observed table has $\hat{t}_{11}^{*} = 3$, which lies outside this interval. Hence, for this particular observed table we find that $\left|e_{11}^{Q}\right| < |e_{11}^{*}|$, which was verified directly in Section 2.2. (In fact, $e_{11}^{*} = -1$ and $e_{11}^{Q} = -0.88$.)

Moreover, since $\hat{t}_{11}^{*}$ has to be an integer, the above condition is satisfied only when $\hat{t}_{11}^{*} = t_{11} = 4$. Thus, if we consider all $10! = 3628800$ possible permutation

matrices $\mathbf{C}$ for this example, the conditions from Theorem 2 are satisfied only for those permutations that yield an observed contingency matrix that happens to be equal to the true contingency matrix. (A non-trivial illustration of such a permutation for this example would be one that permutes only the 1st and 2nd units.) We conclude that, for the table in this example, adjusting by the $\mathbf{Q}$ matrix will improve the estimation error except in those instances where the observed table happens to be error-free.

Note that the true value t_{jk} is contained in the relevant interval (7) or (8) in Theorem 2. Thus, according to this theorem, adjusting the observed contingency table by the $\mathbf{Q}$ matrix increases the error in particular when the observed value $\hat{t}^*_{jk}$ happens to be close to the true value t_{jk}.

3.3 The bias-corrected contingency table

For the estimator $\widehat{\mathbf{T}}^{BC} = \mathbf{Y}'\mathbf{Q}^{-1}\mathbf{CZ}$, the following result is derived in Appendix A.3.

Theorem 3. *For random linkage errors that are described by a matrix* $\mathbf{Q}$ *of the form (2) that satisfies assumption (3),* $\left|\mathrm{e}^{\mathrm{BC}}_{\mathrm{jk}}\right| < \left|\mathrm{e}^{*}_{\mathrm{jk}}\right|$ *when one of the following two conditions holds:*

a. $nt_{jk} > r_j s_k$ *(positive association) and* $\hat{\mathrm{t}}^{*}_{\mathrm{jk}}$ *satisfies*

$$\frac{r_j s_k}{n} < \hat{t}^*_{jk} < t_{jk} - \delta\frac{nt_{jk} - r_j s_k}{2 - \delta n}; \tag{9}$$

b. $nt_{jk} < r_j s_k$ *(negative association) and* $\hat{\mathrm{t}}^{*}_{\mathrm{jk}}$ *satisfies*

$$t_{jk} - \delta\frac{nt_{jk} - r_j s_k}{2 - \delta n} < \hat{t}^*_{jk} < \frac{r_j s_k}{n}. \tag{10}$$

In all other cases, $\left|\mathrm{e}^{\mathrm{BC}}_{\mathrm{jk}}\right| \geq \left|\mathrm{e}^{*}_{\mathrm{jk}}\right|$*, with equality holding only at the endpoints of (9) or (10) and in the special case* $\mathrm{nt_{jk}} = \mathrm{r_j s_k}$ *(no association) for* $\hat{\mathrm{t}}^{*}_{\mathrm{jk}} = \mathrm{t_{jk}}$.

Example (continued)

To illustrate the result in Theorem 3, we again use the example from Section 2, with $n = 10$, $r_1 = 7$, $s_1 = 6$, and $t_{11} = 4$. The observed table has $\hat{t}^*_{11} = 3$. As noted above, we are in the case of negative association. Since $\hat{t}^*_{11} < t_{11}$ and $2 - \delta n > 1$, it follows immediately that $\left|e^{BC}_{11}\right| > |e^*_{11}|$, because the lower bound in (10) is strictly larger than t_{11}. Thus, according to the theory, adjusting the observed entry by means of the $\mathbf{Q}^{-1}$ matrix increases the error in this example. It was seen in Section 2.2 that this is indeed true ($e^*_{11} = -1$ and $e^{BC}_{11} \approx -1.1333$).

In fact, we have checked the conditions of Theorem 3 for all $10!$ possible permutation matrices $\mathbf{C}$ for this example, and found that they were never satisfied. Thus, a striking and unexpected result for this example is that the bias-corrected estimator increases the absolute size of the error in the estimated table compared to the naïve estimator in every possible instance. The estimator $\widehat{\mathbf{T}}^{BC}$ is theoretically unbiased over repeated draws from the linkage error model, with the error being sometimes negative and sometimes positive, but it increases the

absolute size of the error for every possible realisation. In other examples, the bias-corrected estimator may perform better than the naïve approach.

From the similarities in the conditions of Theorem 2 and Theorem 3, it is easy to derive the following interesting connection between the behaviour of the bias-corrected and $\mathbf{Q}$-adjusted estimators.

Corollary 1. *For random linkage errors that are described by a matrix* $\mathbf{Q}$ *of the form (2) that satisfies assumption (3), whenever* $\left|\mathrm{e}_{\mathrm{jk}}^{\mathrm{BC}}\right| < \left|\mathrm{e}_{\mathrm{jk}}^{*}\right|$ *it must also hold that* $\left|\mathrm{e}_{\mathrm{jk}}^{\mathrm{Q}}\right| > \left|\mathrm{e}_{\mathrm{jk}}^{*}\right|$.

Proof. First suppose that there is a positive association and condition (9) holds. In this case

$$\delta \frac{nt_{jk} - r_j s_k}{2 - \delta n} > 0.$$

Therefore, it follows from (9) that

$$\frac{r_j s_k}{n} < \hat{t}_{jk}^{*} < t_{jk} - \delta \frac{nt_{jk} - r_j s_k}{2 - \delta n} < t_{jk} + \delta \frac{nt_{jk} - r_j s_k}{2 - \delta n}.$$

Hence, condition (7) is satisfied and it follows from Theorem 2 that $\left|e_{jk}^{Q}\right| > \left|e_{jk}^{*}\right|$.

In the same way, it can be shown that when there is a negative association, condition (10) implies that condition (8) holds and it follows again from Theorem 2 that $\left|e_{jk}^{Q}\right| > \left|e_{jk}^{*}\right|$. Since these are the only two possible scenarios where $\left|e_{jk}^{BC}\right| < \left|e_{jk}^{*}\right|$, the result is established. ■

Note that the converse statement of Corollary 1 does not hold: it is possible to find cases where both $\left|e_{jk}^{Q}\right| > \left|e_{jk}^{*}\right|$ and $\left|e_{jk}^{BC}\right| \geq \left|e_{jk}^{*}\right|$.

4. Theoretical results on bias and variance

In this section we derive analytical results on bias and variance for our three estimators for a contingency table $\mathbf{T} = \mathbf{Y}'\mathbf{Z}$. Obviously, this automatically also leads to analytical results for the mean square error of these estimators.

4.1 Bias

We already noted in Section 2.2 that the estimator $\widehat{\mathbf{T}}^{BC}$ is unbiased for the true contingency table $\mathbf{T} = \mathbf{Y}'\mathbf{Z}$, whereas the observed contingency table $\widehat{\mathbf{T}}^{*} = \mathbf{Y}'\mathbf{Z}^{*}$ is biased in general. In fact, the bias of $\widehat{\mathbf{T}}^{*}$ as an estimator for $\mathbf{T}$ is seen to be:

$$B\left(\widehat{\mathbf{T}}^{*}\right) = E(\mathbf{Y}'\mathbf{C}\mathbf{Z}) - \mathbf{Y}'\mathbf{Z} = \mathbf{Y}'E(\mathbf{C})\mathbf{Z} - \mathbf{Y}'\mathbf{Z} = \mathbf{Y}'(\mathbf{Q} - \mathbf{I})\mathbf{Z}.$$

Using the fact that

$$\mathbf{Q} - \mathbf{I} = (q-1)\mathbf{I} + \frac{1-q}{n-1}(\mathbf{u}\mathbf{u}' - \mathbf{I}) = \frac{1-q}{n-1}(\mathbf{u}\mathbf{u}' - n\mathbf{I}),$$

we find for an individual entry $\hat{t}^{*}_{jk}$ in the observed table that

$$B\left(\hat{t}^{*}_{jk}\right) = \mathbf{y}_j'(\mathbf{Q} - \mathbf{I})\mathbf{z}_k = \frac{1-q}{n-1}\mathbf{y}_j'(\mathbf{u}\mathbf{u}' - n\mathbf{I})\mathbf{z}_k = -\delta\left(nt_{jk} - r_j s_k\right). \qquad (11)$$

The corrected estimator based on the $\mathbf{Q}$ matrix also yields a biased estimator, since

$$E\left(\widehat{\mathbf{T}}^{Q}\right) = E\{(\mathbf{Q}\mathbf{Y})'\mathbf{C}\mathbf{Z}\} = \mathbf{Y}'\mathbf{Q}'E(\mathbf{C})\mathbf{Z} = \mathbf{Y}'\mathbf{Q}'\mathbf{Q}\mathbf{Z}.$$

Therefore:

$$B\left(\widehat{\mathbf{T}}^{Q}\right) = \mathbf{Y}'\mathbf{Q}'\mathbf{Q}\mathbf{Z} - \mathbf{Y}'\mathbf{Z} = \mathbf{Y}'(\mathbf{Q}'\mathbf{Q} - \mathbf{I})\mathbf{Z},$$

which is non-zero in general since $\mathbf{Q}'\mathbf{Q} \neq \mathbf{I}$ except in the trivial case where $\mathbf{Q} = \mathbf{I}$.

For an easy comparison of the bias of the individual entries $\hat{t}^{Q}_{jk}$ and $\hat{t}^{*}_{jk}$, it is useful to consider the associated error $e^{Q}_{jk} = \mathbf{y}_j'(\mathbf{Q}'\mathbf{C} - \mathbf{I})\mathbf{z}_k$ defined in (6). We note that

$$\begin{aligned}\mathbf{Q}'\mathbf{C} - \mathbf{I} &= q\mathbf{C} + \frac{1-q}{n-1}(\mathbf{u}\mathbf{u}'\mathbf{C} - \mathbf{C}) - \mathbf{I} \\ &= \frac{1}{n-1}\{(nq-1)\mathbf{C} + (1-q)\mathbf{u}\mathbf{u}'\mathbf{C} - (n-1)\mathbf{I}\} \\ &= \frac{1}{n-1}\{(nq-1)\mathbf{C} + (1-q)\mathbf{u}\mathbf{u}' - (n-1)\mathbf{I}\}\end{aligned}$$

$$= \frac{nq-1}{n-1}(\mathbf{C}-\mathbf{I}) + \frac{1-q}{n-1}(\mathbf{uu}'-n\mathbf{I}).$$

In the third line, we used property (16) from Appendix A.2, which says that $\mathbf{uu}'\mathbf{C} = \mathbf{uu}'$. It follows that

$$\begin{aligned} B\left(\hat{t}_{jk}^{Q}\right) &= E\left(e_{jk}^{Q}\right) \\ &= \frac{nq-1}{n-1}\mathbf{y}_j'(\mathbf{Q}-\mathbf{I})\mathbf{z}_k + \frac{1-q}{n-1}\mathbf{y}_j'(\mathbf{uu}'-n\mathbf{I})\mathbf{z}_k \\ &= \frac{nq-1}{n-1}B\left(\hat{t}_{jk}^{*}\right) + B\left(\hat{t}_{jk}^{*}\right) \\ &= -\left(1+\frac{nq-1}{n-1}\right)\delta\left(nt_{jk}-r_j s_k\right), \end{aligned}$$

where we used (11) in the third line.

In summary:

$$\begin{aligned} B\left(\hat{t}_{jk}^{*}\right) &= -\delta\left(nt_{jk}-r_j s_k\right), \\ B\left(\hat{t}_{jk}^{Q}\right) &= -\left(1+\frac{nq-1}{n-1}\right)\delta\left(nt_{jk}-r_j s_k\right), \\ B\left(\hat{t}_{jk}^{BC}\right) &= 0. \end{aligned} \tag{12}$$

From (12) we can draw several general conclusions about the bias of these estimators:

- Since $0 < nq-1 < n-1$ by assumption (3), it always holds that

  $$\left|B\left(\hat{t}_{jk}^{Q}\right)\right| \geq \left|B\left(\hat{t}_{jk}^{*}\right)\right| \geq \left|B\left(\hat{t}_{jk}^{BC}\right)\right| = 0.$$

  That is to say, the **Q**-adjusted estimator is at least as biased as the naïve estimator, and the naïve estimator is (obviously) at least as biased as the bias-corrected estimator.

- $B\left(\hat{t}_{jk}^{Q}\right)$ and $B\left(\hat{t}_{jk}^{*}\right)$ have the same sign, which is completely determined by the type of association: the bias of these estimators is positive for entries with a negative association ($nt_{jk} < r_j s_k$) and negative for entries with a positive association ($nt_{jk} > r_j s_k$). In the special case that $nt_{jk} = r_j s_k$, the bias of these estimators is zero.

- For given values of n and q, and assuming that $B\left(\hat{t}_{jk}^{*}\right) \neq 0$, the bias ratio

  $$\frac{B\left(\hat{t}_{jk}^{Q}\right)}{B\left(\hat{t}_{jk}^{*}\right)} = 1 + \frac{nq-1}{n-1}$$

  is invariant for all entries in any contingency table. Furthermore, under assumption (3), this ratio always lies between 1 (for $q \downarrow \frac{1}{n}$) and 2 (for $q \uparrow 1$). Thus, although in absolute terms the bias of $B\left(\hat{t}_{jk}^{Q}\right)$ and $B\left(\hat{t}_{jk}^{*}\right)$ is maximal when $q \downarrow \frac{1}{n}$, the relative bias of $B\left(\hat{t}_{jk}^{Q}\right)$ compared to $B\left(\hat{t}_{jk}^{*}\right)$ is maximal when $q \uparrow 1$. Furthermore, for large n this bias ratio tends to $1+q$.

Example (continued)
For the example from Section 2 with $n = 10$, the bias of the observed contingency table $\widehat{\mathbf{T}}^*$ is:

$$B(\widehat{\mathbf{T}}^*) = \mathbf{Y}'(\mathbf{Q} - \mathbf{I})\mathbf{Z} = \begin{pmatrix} 4.02 & 2.98 \\ 1.98 & 1.02 \end{pmatrix} - \begin{pmatrix} 4 & 3 \\ 2 & 1 \end{pmatrix} = \begin{pmatrix} 0.02 & -0.02 \\ -0.02 & 0.02 \end{pmatrix}.$$

The bias for the alternative estimator $\widehat{\mathbf{T}}^Q$ is:

$$B(\widehat{\mathbf{T}}^Q) = \mathbf{Y}'(\mathbf{Q}'\mathbf{Q} - \mathbf{I})\mathbf{Z} = \begin{pmatrix} 4.038 & 2.962 \\ 1.962 & 1.038 \end{pmatrix} - \begin{pmatrix} 4 & 3 \\ 2 & 1 \end{pmatrix} = \begin{pmatrix} 0.038 & -0.038 \\ -0.038 & 0.038 \end{pmatrix},$$

which is indeed larger than the bias of the uncorrected estimator. More precisely, the bias increases for all entries by the same factor

$$1 + \frac{nq - 1}{n - 1} = 1 + \frac{10 \times 0.91 - 1}{9} = 1.9.$$

4.2 Variance

To analyse the variance of the estimators $\widehat{\mathbf{T}}^*$, $\widehat{\mathbf{T}}^Q$, and $\widehat{\mathbf{T}}^{BC}$, it is useful to re-examine the errors e_{jk}^*, e_{jk}^Q and e_{jk}^{BC} defined in (6). It holds that:

$$\begin{aligned} e_{jk}^* &= \mathbf{y}_j'(\mathbf{C} - \mathbf{I})\mathbf{z}_k, \\ e_{jk}^Q &= \frac{nq - 1}{n - 1}\mathbf{y}_j'(\mathbf{C} - \mathbf{I})\mathbf{z}_k + \frac{1 - q}{n - 1}\mathbf{y}_j'(\mathbf{u}\mathbf{u}' - n\mathbf{I})\mathbf{z}_k, \\ e_{jk}^{BC} &= \frac{n - 1}{nq - 1}\mathbf{y}_j'(\mathbf{C} - \mathbf{I})\mathbf{z}_k - \frac{1 - q}{nq - 1}\mathbf{y}_j'(\mathbf{u}\mathbf{u}' - n\mathbf{I})\mathbf{z}_k. \end{aligned} \tag{13}$$

The expression for e_{jk}^Q was derived in Section 4.1. For e_{jk}^*, we used expression (6). For e_{jk}^{BC}, we used expression (20) from Appendix A.3.

Now using that $\mathrm{Var}(\hat{t}_{jk}^*) = \mathrm{Var}(e_{jk}^*) = \mathrm{Var}(\mathbf{y}_j'\mathbf{C}\mathbf{z}_k)$ (and similarly for the other estimators), we can conclude immediately from (11) and (13) that

$$\begin{aligned} \mathrm{Var}(\hat{t}_{jk}^*) &= \mathrm{Var}(\mathbf{y}_j'\mathbf{C}\mathbf{z}_k), \\ \mathrm{Var}(\hat{t}_{jk}^Q) &= \frac{(nq - 1)^2}{(n - 1)^2}\mathrm{Var}(\mathbf{y}_j'\mathbf{C}\mathbf{z}_k), \\ \mathrm{Var}(\hat{t}_{jk}^{BC}) &= \frac{(n - 1)^2}{(nq - 1)^2}\mathrm{Var}(\mathbf{y}_j'\mathbf{C}\mathbf{z}_k). \end{aligned} \tag{14}$$

Thus, in order to compute the variance of all three estimators, it suffices to compute the variance of the quadratic form $\mathbf{y}_j'\mathbf{C}\mathbf{z}_k$. Using a result from Chambers (2009), the following approximation for large n is derived in Appendix A.4:

$$\mathrm{Var}\left(\mathbf{y}_j'\mathbf{C}\mathbf{z}_k\right) \approx q(1-q)\left(1-2\frac{s_k}{n}\right)t_{jk} + (1-q)\frac{r_j s_k}{n}\left[1-(1-q)\frac{s_k}{n}\right]. \qquad (15)$$

By substituting this expression into (14), large-sample approximations are obtained to $\mathrm{Var}\left(\hat{t}_{jk}^{*}\right)$, $\mathrm{Var}\left(\hat{t}_{jk}^{Q}\right)$, and $\mathrm{Var}\left(\hat{t}_{jk}^{BC}\right)$. We demonstrate the approximation of the variance in (15) in the simulation study described in Section 5 and compare the variance for the case of the naïve approach to the empirical variance obtained from the study in Appendix B.

From (14) we can draw two general conclusions about the variance of these estimators:

- Since $0 < nq - 1 < n - 1$ by assumption (3), it always holds that

  $$\mathrm{Var}\left(\hat{t}_{jk}^{BC}\right) > \mathrm{Var}\left(\hat{t}_{jk}^{*}\right) > \mathrm{Var}\left(\hat{t}_{jk}^{Q}\right).$$

  That is to say, of the three estimators considered here, the bias-corrected estimator and the $\mathbf{Q}$-adjusted estimator always have the largest and smallest variance, respectively. Note that the reverse order was found above for the bias.

- For given values of n and q, the variance ratios

  $$\frac{\mathrm{Var}\left(\hat{t}_{jk}^{Q}\right)}{\mathrm{Var}\left(\hat{t}_{jk}^{*}\right)} = \frac{(nq-1)^2}{(n-1)^2},$$
  $$\frac{\mathrm{Var}\left(\hat{t}_{jk}^{BC}\right)}{\mathrm{Var}\left(\hat{t}_{jk}^{*}\right)} = \frac{(n-1)^2}{(nq-1)^2}$$

  are invariant for all entries in any contingency table. For $q \downarrow \frac{1}{n}$, the variance of $\hat{t}_{jk}^{Q}$ tends to zero, whereas the variance of $\hat{t}_{jk}^{BC}$ explodes. For $q \uparrow 1$, the variances of the three estimators all converge to the same value. Furthermore, for large n these variance ratios tend to q^2 and $1/q^2$, respectively.

4.3 Mean squared error (MSE)

We can combine the bias formulas in (12) with the variance formulas in (14) to obtain expressions for the MSEs of the three estimators under linkage errors:

$$\mathrm{MSE}\left(\hat{t}_{jk}^{*}\right) = \delta^2\left(nt_{jk} - r_j s_k\right)^2 + \mathrm{Var}\left(\mathbf{y}_j'\mathbf{C}\mathbf{z}_k\right),$$
$$\mathrm{MSE}\left(\hat{t}_{jk}^{Q}\right) = \left(1+\frac{nq-1}{n-1}\right)^2\delta^2\left(nt_{jk} - r_j s_k\right)^2 + \frac{(nq-1)^2}{(n-1)^2}\mathrm{Var}\left(\mathbf{y}_j'\mathbf{C}\mathbf{z}_k\right),$$
$$\mathrm{MSE}\left(\hat{t}_{jk}^{BC}\right) = \frac{(n-1)^2}{(nq-1)^2}\mathrm{Var}\left(\mathbf{y}_j'\mathbf{C}\mathbf{z}_k\right).$$

Which of these estimators has the smallest MSE depends on the value of $\mathrm{Var}\left(\mathbf{y}_j'\mathbf{C}\mathbf{z}_k\right)$ and may therefore differ between applications.

5. Simulation study

In this section, we present a simulation study to demonstrate the theoretical findings shown in Sections 3 and 4.

The simulation is based on a dataset of 300 records and two error matrices $\mathbf{Q}$ of size 300×300, the first with 0.9 on the diagonal and all off-diagonal equal to 0.1/299, and the second with 0.8 on the diagonal and all off-diagonal equal to 0.2/299. We generated an auxiliary variable $x \sim N(20, 10^2)$ and two target variables $z \sim N(20, 15^2)$ and $y = 20 + 2x + 1z + \epsilon$, where $\varepsilon \sim N(0, 4^2)$. We also generated: $w \sim N(40, 30^2)$. We then discretized y, z and w to 5 categories each and converted them to vectors of dummy variables denoted $\mathbf{Y}_g$, $\mathbf{Z}_g$ and $\mathbf{W}_g$ respectively. The 5×5 table defined by $\mathbf{Y}_g'\mathbf{Z}_g$ has dependent attributes ($\chi^2 = 96.57, p < 0.0001$) and the 5×5 table defined by $\mathbf{Y}_g'\mathbf{W}_g$ has independent attributes ($\chi^2 = 7.32, p = 0.9666$). Note that these tables have high power to the statistical test of independence. The dataset was split into two files: File A contains the vector $\mathbf{Y}_g$ and matching variables and File B contains the vectors $\mathbf{Z}_g$ and $\mathbf{W}_g$ and matching variables.

For each of the error matrices, we generated 1000 permutation matrices using the unbiased zero-restricted controlled rounding for 2-dimensional tables according to the procedure of Cox (1987) and presented in Willenborg and De Waal (2001). The procedure control-rounds the $\mathbf{Q}$ matrix to base 1 according to the probabilities provided in this matrix and ensures that there is a one in each column/row and the rest of the entries are zero. The expectation of these generated permutation matrices therefore equals the corresponding $\mathbf{Q}$ matrix of the exchangeable linkage error model (2).

For each of the 1000 permutation matrices, we linked a row of file A with the row of file B according to the indicator 1 of the permutation matrix. Note that for the 0.9 error matrix, this resulted in 90% of the links being correct and 10% of the links being incorrect in expectation and similarly for the 0.8 error matrix, 80% of the links are correct and 20% of the links are incorrect in expectation. On each of the 1000 linked files, we then reproduced the 5×5 tables of $\mathbf{Y}_g'\mathbf{Z}_g$ and $\mathbf{Y}_g'\mathbf{W}_g$ according to the following:

a. naïve approach: $\mathbf{Y}_g'\mathbf{Z}_g$ and $\mathbf{Y}_g'\mathbf{W}_g$ without any correction
b. $\mathbf{Q}$ approach: $(\mathbf{Q}\mathbf{Y}_g)'\mathbf{Z}_g$ and $(\mathbf{Q}\mathbf{Y}_g)'\mathbf{W}_g$
c. $\mathbf{Q}^{-1}$ approach (bias-corrected): $\mathbf{Y}_g'\mathbf{Q}^{-1}\mathbf{Z}_g$ and $\mathbf{Y}_g'\mathbf{Q}^{-1}\mathbf{W}_g$
d. Chipperfield and Chambers (2015) approach described in Section 2.2.

In accordance with Theorem 1, the Chipperfield and Chambers (2015) approach gives similar results as the $\mathbf{Q}^{-1}$ (bias-corrected) approach throughout, and hence we do not show these results below.

Table 1 contains the percentage of cases out of the tables produced from the 1000 linked files where the naïve approach is expected to perform better than the $\mathbf{Q}$ and $\mathbf{Q}^{-1}$ (bias corrected) approaches for each of the 25 individual cells of the

table. In other words, Table 1 contains the percentage of times that $\left|e_{jk}^{Q}\right| > \left|e_{jk}^{*}\right|$ or $\left|e_{jk}^{BC}\right| > \left|e_{jk}^{*}\right|$ respectively in the individual cell. These percentages correspond to the bounds defined in Theorem 2 for the Q approach and the bounds defined in Theorem 3 for the $\mathbf{Q}^{-1}$ (bias corrected) approach. Note that the bounds described in the theorems depend on whether there is a positive association, negative association or no association for the individual cell value and these are denoted in the column labelled 'Ass. Type'. The percentages are provided separately for the error matrices with 0.9 and 0.8 on the diagonal and for the dependent attribute table (variables y and z) and independent attribute table (variables y and w). Each of the individual cells show different percentages for when the naïve approach outperforms the alternative approaches depending on the type of association of the cell. For example, under the dependent table, cell $R_1 \times C_3$ is on the boundary of the bounds described in Theorem 2 and Theorem 3 and therefore, the naïve approach is never better than the $\mathbf{Q}$ approach but always better than the $\mathbf{Q}^{-1}$ (bias-corrected) approach.

Focusing on the tables derived from the dependent attribute table, the smaller error level defined by 0.9 on the diagonal of the error matrix has generally more of a chance that the naïve approach will outperform the other approaches compared to the larger error level defined by 0.8 on the diagonal in almost all the cells. In other words, when the error is small, the naïve approach is likely just as good as an alternative approach when the table has a dependent attribute. Furthermore, for the 0.9 diagonal error matrix and the dependent attribute table, the average percentage where the naïve approach outperforms the alternative approaches across all cells is about 75% under the $\mathbf{Q}$ approach and 66% under the $\mathbf{Q}^{-1}$ (bias-corrected) approach. For the 0.8 diagonal error matrix and the dependent attribute table, the average percentage where the naïve approach outperforms the alternative approaches across all cells is about 70% under the $\mathbf{Q}$ approach and about 55% under the $\mathbf{Q}^{-1}$ (bias-corrected) approach. These findings in Table 1 provide some evidence that when the original table has dependent attributes (small p-value for the Chi-square test of independence) and the linking process has a relatively high error rate, the best approach for correcting for linkage error is to use the $\mathbf{Q}^{-1}$ (bias-corrected) approach. Nevertheless, we will show in the next evaluation based on the statistical properties of the approaches that even the $\mathbf{Q}^{-1}$ approach is (slightly) outperformed by the naïve approach.

Focusing on the tables derived from the independent attribute table, the smaller error level defined by 0.9 on the diagonal of the error matrix has generally more of a chance that the naïve approach will outperform the other approaches compared to the higher error level defined by 0.8 on the diagonal of the error matrix, similar to the case for the dependent attribute table. Furthermore, for the 0.9 diagonal error matrix and the independent attribute table, the average percentage where the naïve approach outperforms the alternative approaches across all cells is about 58% under the $\mathbf{Q}$ approach and 88% under the $\mathbf{Q}^{-1}$ (bias-corrected) approach. For the 0.8 diagonal error matrix and the independent attribute table, the average percentage where the naïve approach outperforms the alternative approaches across all cells is about 42% under the $\mathbf{Q}$ approach and about 86% under the $\mathbf{Q}^{-1}$ (bias-corrected) approach. These findings in Table 1 provide some evidence that when the original table has independent attributes (large p-value for

the Chi-square test of independence) and the linking process has a relatively high error rate, the best approach for correcting for linkage error is to use the **Q** approach.

Table 1: Percentage in each cell entry of the 1000 tables produced from the linked files where the naïve approach outperforms the respective alternative approaches based on Q and $\mathbf{Q}^{-1}$ (bias-corrected) and the type of association (labelled 'Ass. Type') of the cell entry (denoted by $R_j \times C_k$ where j is the row, $j = 1,..,5$ and k is the column $k = 1,\ldots,5$)

Cell position	Dependent Attributes					Independent Attributes				
	Ass. Type	0.9		0.8		Ass. Type	0.9		0.8	
		Q	$\mathbf{Q}^{-1}$	Q	$\mathbf{Q}^{-1}$		Q	$\mathbf{Q}^{-1}$	Q	$\mathbf{Q}^{-1}$
$R_1 \times C_1$	pos	92.8	58.7	91.1	34.8	pos	80.1	80.0	64.8	72.8
$R_1 \times C_2$	pos	95.5	44.1	99.6	42.4	neg	66.7	100.0	44.7	100.0
$R_1 \times C_3$	equal	0.0	100.0	0.0	100.0	neg	79.9	78.5	60.9	75.2
$R_1 \times C_4$	neg	97.4	61.3	97.1	32.8	neg	71.5	78.8	55.3	74.0
$R_1 \times C_5$	neg	98.4	45.1	100.0	43.8	pos	77.5	70.2	67.1	56.6
$R_2 \times C_1$	pos	87.5	58.3	86.2	38.4	neg	55.1	100.0	33.7	100.0
$R_2 \times C_2$	pos	78.4	66.0	67.6	58.3	pos	52.3	100.0	32.4	100.0
$R_2 \times C_3$	pos	77.7	66.1	67.5	57.3	neg	45.4	100.0	29.0	100.0
$R_2 \times C_4$	neg	43.8	100.0	28.0	100.0	neg	37.5	100.0	26.0	100.0
$R_2 \times C_5$	neg	90.7	43.3	98.5	43.5	pos	60.6	78.3	43.4	77.8
$R_3 \times C_1$	pos	47.1	100.0	27.2	100.0	neg	78.7	67.6	69.2	58.5
$R_3 \times C_2$	pos	41.5	100.0	24.5	100.0	pos	45.1	100.0	28.8	100.0
$R_3 \times C_3$	pos	79.4	61.5	75.7	44.7	pos	57.8	77.8	43.5	80.4
$R_3 \times C_4$	pos	37.1	100.0	23.5	100.0	neg	35.8	100.0	20.0	100.0
$R_3 \times C_5$	neg	77.5	53.2	74.8	39.0	pos	36.4	100.0	19.6	100.0
$R_4 \times C_1$	neg	85.3	59.1	82.4	42.1	neg	46.1	100.0	25.2	100.0
$R_4 \times C_2$	neg	87.3	52.0	87.4	31.5	neg	48.2	100.0	28.8	100.0
$R_4 \times C_3$	neg	61.5	79.6	45.1	77.8	pos	57.8	78.8	41.6	78.8
$R_4 \times C_4$	pos	78.8	55.8	79.4	37.2	pos	52.8	79.3	33.5	83.0
$R_4 \times C_5$	pos	79.1	51.5	77.2	38.5	neg	54.4	79.3	36.8	82.7
$R_5 \times C_1$	neg	100.0	58.6	100.0	26.1	pos	85.5	64.2	82.2	45.6
$R_5 \times C_2$	neg	97.4	57.3	96.0	31.0	neg	54.5	100.0	35.6	100.0
$R_5 \times C_3$	neg	94.7	57.3	93.2	31.5	neg	49.2	100.0	28.3	100.0
$R_5 \times C_4$	neg	46.3	100.0	25.8	100.0	pos	44.2	100.0	23.3	100.0
$R_5 \times C_5$	pos	96.9	25.5	100.0	29.3	neg	79.2	63.1	70.0	54.3

We now turn to the statistical properties for the 0.9 and 0.8 error matrices and according to whether the table has dependent or independent attributes. Table 2 presents the empirical MSE for the naïve approach, and the relative difference from the naïve approach for the alternative approaches, for example for the Q approach: $(MSE - MSE^{naive})/MSE^{naive}$. Negative values mean that the alternative approach performs better than the naïve approach in the individual cell. Similarly, Table B1 of Appendix B presents the standard error component of the empirical MSE and Table B2 of Appendix B presents the bias component of the empirical MSE.

Table 2: Empirical MSE of the cell entries across tables from 1000 linked files according to the 0.9 or 0.8 error matrices and dependent or independent attributes on the original tables for the naïve approach; and the relative difference from the naïve empirical MSE for the $\mathbf{Q}$ and $\mathbf{Q}^{-1}$ approaches. We denote row by R_j where $j = 1,..,5$ and column by C_k where $k = 1,\dots,5$

Cell Row		Dependent Attributes									
		0.9					0.8				
		C_1	C_2	C_3	C_4	C_5	C_1	C_2	C_3	C_4	C_5
Naïve:	R_1	1.05	1.57	0.85	0.96	1.57	1.66	2.38	1.15	1.53	2.30
	R_2	1.19	1.11	1.16	1.14	1.73	1.72	1.59	1.57	1.54	2.56
	R_3	1.08	1.21	1.34	1.37	1.76	1.56	1.65	1.96	1.88	2.58
	R_4	1.19	1.42	1.25	1.57	1.91	1.81	2.13	1.75	2.34	2.63
	R_5	1.02	1.07	1.13	1.11	2.58	1.63	1.66	1.66	1.60	4.26
Q:	R_1	0.22	0.36	-0.10	0.28	0.36	0.29	0.51	-0.20	0.38	0.54
	R_2	0.12	-0.02	-0.02	-0.10	0.28	0.20	-0.04	-0.03	-0.19	0.43
	R_3	-0.10	-0.10	0.03	-0.10	0.06	-0.19	-0.19	0.05	-0.19	0.09
	R_4	0.10	0.19	-0.09	0.08	0.08	0.14	0.29	-0.17	0.15	0.12
	R_5	0.32	0.29	0.24	-0.10	0.48	0.43	0.40	0.36	-0.19	0.60
Q^{-1}:	R_1	-0.02	-0.12	0.11	-0.05	-0.13	-0.04	-0.22	0.25	-0.12	-0.25
	R_2	0.02	0.08	0.08	0.11	-0.08	0.02	0.17	0.18	0.24	-0.15
	R_3	0.11	0.11	0.06	0.11	0.04	0.25	0.25	0.14	0.24	0.10
	R_4	0.03	-0.01	0.11	0.06	0.03	0.08	-0.03	0.24	0.07	0.10
	R_5	-0.10	-0.08	-0.05	0.11	-0.22	-0.19	-0.11	-0.08	0.24	-0.39
Cell Row		**Independent Attributes**									
		0.9					**0.8**				
		C_1	C_2	C_3	C_4	C_5	C_1	C_2	C_3	C_4	C_5
Naïve:	R_1	0.80	0.68	0.84	1.02	1.13	1.11	0.96	1.22	1.43	1.55
	R_2	0.88	0.94	1.09	1.29	1.30	1.24	1.30	1.48	1.78	1.84
	R_3	1.11	1.08	1.33	1.51	1.50	1.58	1.51	1.86	2.03	2.18
	R_4	1.12	1.16	1.35	1.53	1.60	1.54	1.46	1.88	2.27	2.22
	R_5	1.07	0.90	1.05	1.25	1.25	1.53	1.16	1.44	1.73	1.82
Q:	R_1	-0.06	-0.10	-0.03	-0.06	-0.03	-0.11	-0.19	-0.08	-0.11	-0.05
	R_2	-0.08	-0.10	-0.09	-0.10	-0.07	-0.15	-0.20	-0.19	-0.20	-0.13
	R_3	0.01	-0.09	-0.06	-0.10	-0.10	0.00	-0.19	-0.14	-0.20	-0.20
	R_4	-0.09	-0.10	-0.09	-0.08	-0.08	-0.18	-0.20	-0.16	-0.17	-0.15
	R_5	0.05	-0.10	-0.09	-0.10	-0.01	0.10	-0.20	-0.18	-0.19	-0.03
Q^{-1}:	R_1	0.10	0.11	0.08	0.10	0.09	0.22	0.25	0.21	0.21	0.18
	R_2	0.10	0.11	0.11	0.11	0.10	0.23	0.25	0.25	0.25	0.22
	R_3	0.07	0.11	0.09	0.11	0.11	0.16	0.24	0.24	0.25	0.25
	R_4	0.11	0.11	0.11	0.11	0.11	0.24	0.25	0.24	0.24	0.23
	R_5	0.05	0.11	0.11	0.11	0.08	0.09	0.25	0.25	0.25	0.17

From Table 2, we can see from the negative values for which of the cells of the table the MSE of the alternative approach is smaller than the MSE of the naïve approach and therefore outperforms the naïve approach. For the dependent attribute table on the top half of Table 2, both the $\mathbf{Q}$ and $\mathbf{Q}^{-1}$ approaches have approximately the same number of cells with negative values although they are appearing in different cells. Adding up the relative differences across the 25 cells of the table for the dependent table, under the 0.9 error matrix, the $\mathbf{Q}$ approach has a total of 2.78 and $\mathbf{Q}^{-1}$ approach has a total of 0.30 whilst for the 0.8 error matrix, the $\mathbf{Q}$ approach has a total of 3.56 and $\mathbf{Q}^{-1}$ a total of 0.983. Clearly, the more error in the linked data, the higher the overall total difference in favour of the naïve approach. Based on the original table with the dependent attribute, the

$\mathbf{Q}$ approach has larger total difference compared to the naïve approach and the $\mathbf{Q}^{-1}$ approach smaller total difference compared to the naïve approach, but all are positive and there is no evidence that the alternative approaches outperform the naïve approach when the table has a dependent attribute.

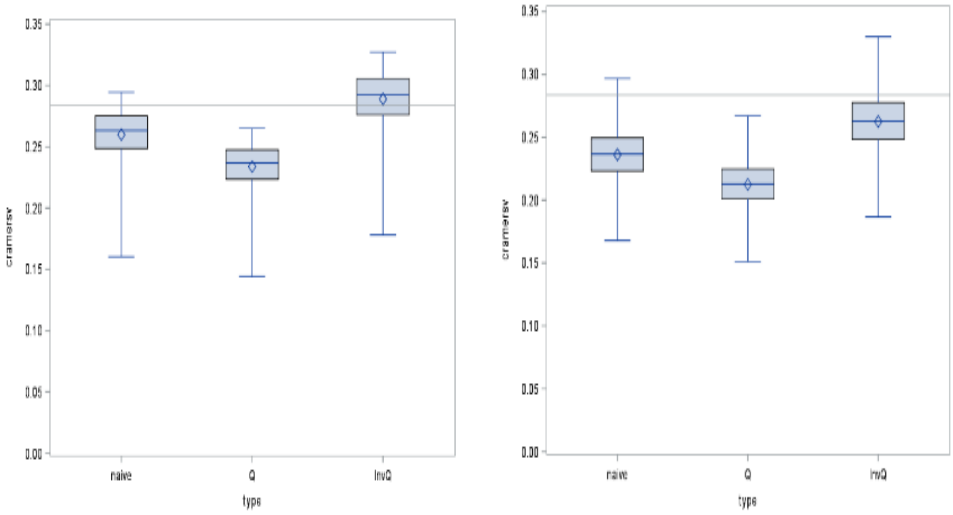
For the independent attribute table on the bottom half of Table 2, the $\mathbf{Q}$ approach is clearly outperforming the $\mathbf{Q}^{-1}$ and the naïve approaches with all cells having negative values meaning that the MSE is always smaller under the $\mathbf{Q}$ approach compared to the naïve approach. Adding up the relative differences across the 25 cells of the table for the independent attribute table, under the 0.9 error matrix, the $\mathbf{Q}$ approach has a total of -1.75 and the $\mathbf{Q}^{-1}$ approach has a total of 2.51 whilst for the 0.8 error matrix, the $\mathbf{Q}$ approach has a total of -3.49 and the $\mathbf{Q}^{-1}$ approach a total of 5.60. The larger the errors in the Q matrices, the higher the overall total difference in favour of the $\mathbf{Q}$ approach. Clearly, the $\mathbf{Q}$ compensation approach, which has a large negative total difference, is the best approach.

We can examine the consistency of the results between Table 1 and Table 2 by comparing the individual cell values in Table 1 to corresponding cell values in Table 2. As an example, we consider cell $R_2 \times C_4$. This cell in Table 1 shows that under the 0.9 error matrix for the $\mathbf{Q}$ compensation approach and dependent attributes table, 43.8% of the tables generated show that the naïve approach performs better than the $\mathbf{Q}$ approach. In Table 2 we find that the MSE is indeed lower for the $\mathbf{Q}$ approach with a relative difference of -0.10% from the naïve MSE. On the other hand, we can see in Table 1 under the cell $R_2 \times C_4$that 100% of the tables generated show that the naïve approach performs better compared to the $\mathbf{Q}^{-1}$ approach. In Table 2 we find that the MSE is higher for the $\mathbf{Q}^{-1}$ approach with a relative difference of 0.11 from the naïve MSE. Other consistencies can be found and hence we find that the bounds in Theorems 2 and 3 and the type of association of the cell as well as whether the table has dependent or independent attributes can provide useful a priori information regarding which approach is the best approach to compensate for the linkage error.

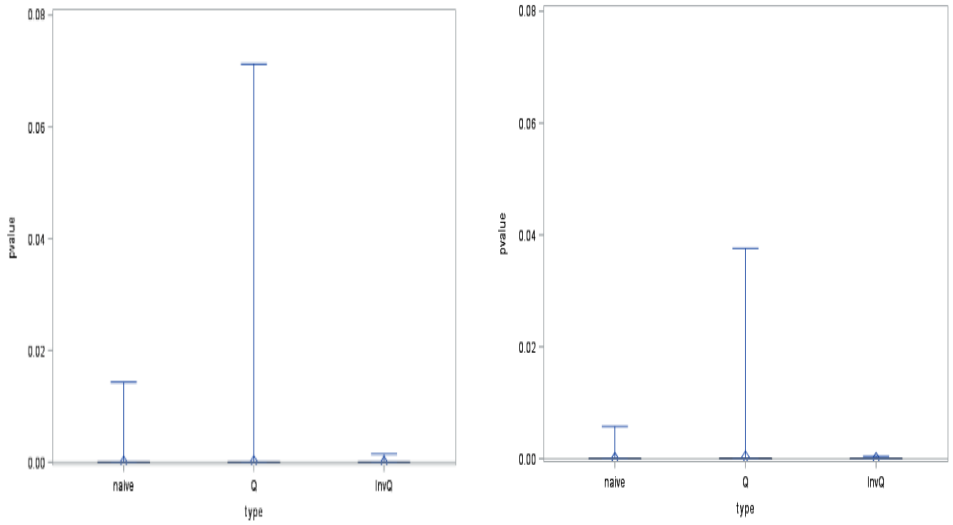
In Figure 1 we show results of the Chi-square test for independence on each of the tables derived from the 1000 linked files for the dependent attribute table with box plots of Cramer's V and the p-values from the statistical test, respectively, for the naïve, $\mathbf{Q}$ and $\mathbf{Q}^{-1}$ approaches. The horizontal lines in the box plots show the true values. The figure on the left is for the error matrix with 0.9 on the diagonal and the figure on the right is for the error matrix with 0.8 on the diagonal. The $\mathbf{Q}^{-1}$ approach is outperforming the other approaches demonstrating that there would be little error in rejecting the null hypothesis of independence. The same holds for the naïve approach. Under the $\mathbf{Q}$ approach, however, there are many instances where we would fail to reject the null hypothesis at a 0.05 significance level, particularly where the error matrix has 0.9 on the diagonal.

Figure 1: Cramer's V and p-values for a Chi-square test of independence for the 1000 linked files under the dependent attribute tables. True Cramer's V =0.2835, True p-value <0.0001

Cramer's V **Diagonal 0.9** **Diagonal 0.8**



p-values **Diagonal 0.9** **Diagonal 0.8**

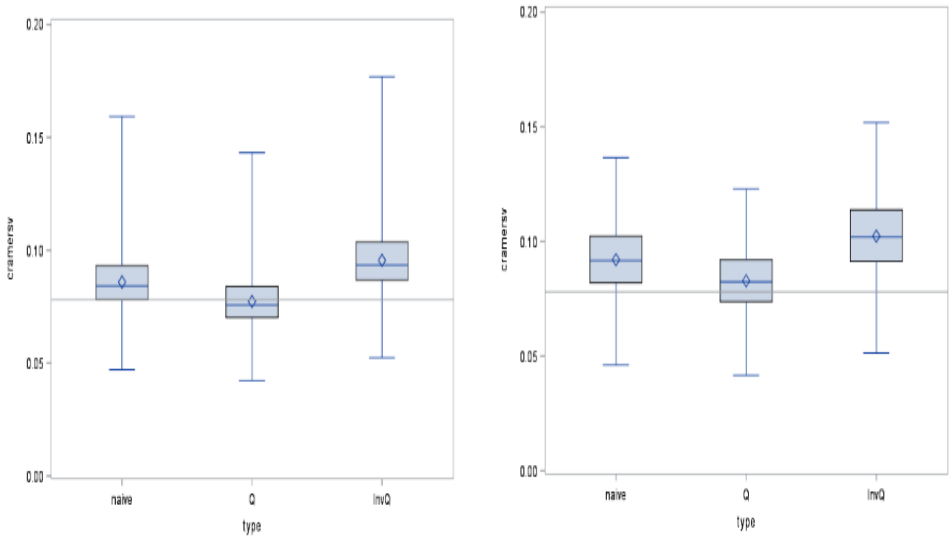


In Figure 2 we show results of the Chi-square test for independence on each of the tables derived from the 1000 linked files for the independent attribute table with box plots of Cramer's V and the p-values from the statistical test, respectively, for the naïve, $\mathbf{Q}$ and $\mathbf{Q}^{-1}$ approaches. The horizontal lines in the box plots show the true values. The figure on the left is for the error matrix with 0.9 on the diagonal and the figure on the right is for the error matrix with 0.8 on the diagonal. When the original table has an independent attribute, we see clearly that the $\mathbf{Q}$ approach is outperforming the other approaches demonstrating that we would have no error in failing to rejecting the null hypothesis of independence under this approach. This is not the case for the $\mathbf{Q}^{-1}$ approach where there may be many

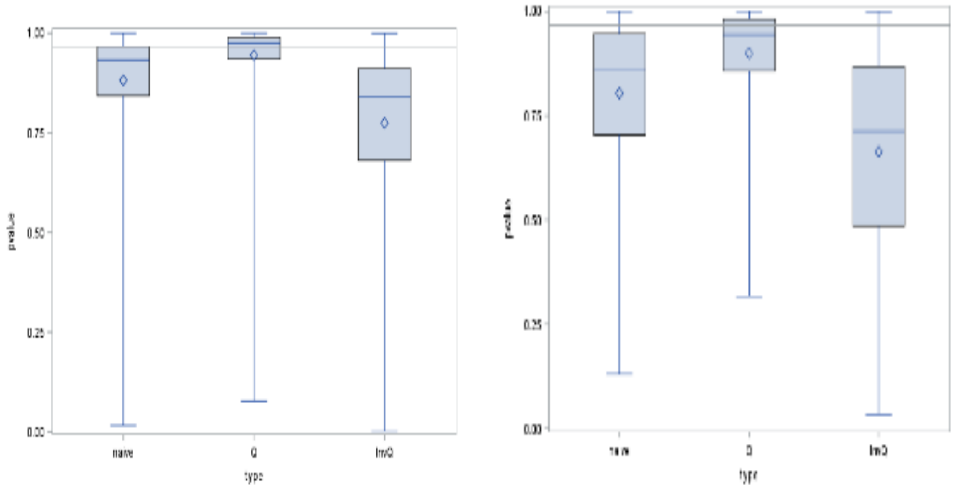
instances where we would reject the null hypothesis at a 0.05 significance level under both error matrices.

Figure 2: Cramer's V and p-values for a Chi-square test of independence for the 1000 linked files under the independent attribute tables. True Cramer's V =0.0781, True p-value= 0.9666

Cramer's V **Diagonal 0.9** **Diagonal 0.8**



p-values **Diagonal 0.9** **Diagonal 0.8**



6. Discussion

The main finding of this paper is that it strongly depends on the kind of contingency table one is dealing with to determine what is the best estimator, the naïve, $\mathbf{Q}$ or $\mathbf{Q}^{-1}$ (bias-corrected) approach, for its cell values. For dependent tables, i.e. tables where the target variables in the two data files are dependent on each other, the naïve approach performs just as well or even better than the two compensation approaches. For strongly dependent tables, we therefore recommend using the naïve approach, without trying to compensate for linkage error by either the $\mathbf{Q}$ or $\mathbf{Q}^{-1}$ approach when the error matrix has more than 0.8 on the diagonal.

For independent tables, where the target variables in the two data files are independent of each other, our results show a strikingly different picture. For such tables, the $\mathbf{Q}$ approach outperforms both the naïve approach as well as the $\mathbf{Q}^{-1}$ approach. Especially, the $\mathbf{Q}^{-1}$ performs rather badly for such tables. Based on the theoretical findings in Section 4, these good results of the $\mathbf{Q}$ approach for independent tables may be explained by the fact that this approach reduces the variance due to linkage errors, whereas the resulting bias is small when the target variables are truly independent. For independent tables, we recommend using the $\mathbf{Q}$ approach when the error matrix has more than 0.8 on the diagonal.

What remains to be examined is at what level of dependency the naïve approach performs better than the $\mathbf{Q}$ approach. This question may be very hard to answer in general as the answer may be dependent on the precise relation between the two target variables and the relations between the target variables and the common variables.

Overall, the $\mathbf{Q}^{-1}$ compensation approach performs rather poor. For dependent tables it is outperformed by the naïve approach, for independent tables it is outperformed by the $\mathbf{Q}$ approach. This is a surprising result, since it is the only unbiased approach of the three approaches – the naïve, $\mathbf{Q}$ and $\mathbf{Q}^{-1}$ – that we have examined. Using an unbiased approach would intuitively be the most obvious choice for many statisticians. However, in many instances the variance of the $\mathbf{Q}^{-1}$ estimator is much higher than the variances of the naïve and the $\mathbf{Q}$ estimator.

In theory, in between strongly dependent tables, where the naïve approach performs the best, and independent tables, where the $\mathbf{Q}$ estimator performs the best, there may be some situations where the $\mathbf{Q}^{-1}$ estimator performs the best. In our simulation study, we have not encountered these situations. More research is required to examine if there are indeed practical cases where the $\mathbf{Q}^{-1}$ estimator outperforms the other two estimators, and if so, under which level or kind of dependency.

In this paper, we have not derived analytical results for the three estimators with respect to validity of the results of hypothesis tests for independence on data that have been obtained by means of probabilistic record linkage. This is a topic for

potential future research. However, we did examine the validity of the results of hypothesis tests for independence in our simulation study. Those results suggest that the validity of the results of hypothesis tests are largely in line with the results with respect to the preservation of contingency tables. That is, for dependent tables the naïve approach leads to little error in rejecting the null hypothesis of independence. For the $\mathbf{Q}$ approach, which performs the worst with respect to preservation of contingency tables for dependent tables, we would fail to reject the null hypothesis for such tables. In contrast to the results with respect to preservation of contingency tables, the $\mathbf{Q}^{-1}$ approach performs best with respect to validity of the results of hypothesis tests for dependent tables. For independent tables, where the $\mathbf{Q}$ approach performs best with respect to the preservation of contingency tables, it also performs best with respect to the validity of the results of hypothesis tests. Similarly, for independent tables, the $\mathbf{Q}^{-1}$ approach performs worst with respect to the preservation of contingency tables and it also performs worst with respect to the validity of the results of hypothesis tests.

As we already noted in Section 2, we feel that the exchangeable linkage error models allow important insights into the properties of various compensation approaches for linkage error that are also useful for more complicated linkage error models. Whether this is indeed the case, remains to be confirmed by analytical or simulation studies. We leave this as a topic for future research.

References

Chambers, R. (2009), Regression Analysis of Probability-Linked Data. *Official Statistics Research Series*, Vol. 4, Wellington, New Zealand.

Chipperfield, J.O., Bishop, G.R. and P. Campbell (2011), Maximum Likelihood Estimation for Contingency Tables and Logistic Regression with Incorrectly Linked Data. *Survey Methodology* **37** (1), pp. 13-24.

Chipperfield, J.O. and R.L. Chambers (2015), Using the Bootstrap to Account for Linkage Errors when Analysing Probabilistically Linked Categorical Data. *Journal of Official Statistics* **31** (3), pp. 397–414, http://dx.doi.org/10.1515/JOS-2015-0024

Cox, L.H. (1987), A Constructive Procedure for Unbiased Controlled Rounding. *Journal of the American Statistical Association* **82**, pp. 520-524.

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

Hager, W.A. (1989), Updating the Inverse of a Matrix. *SIAM Review* **31** (2), pp. 221-239.

Kim, G. and R. Chambers, R. (2012a), Regression Analysis under Incomplete Linkage. *Comput. Stat. Data Anal.* **56**, pp. 2756–2770.

Kim, G. and R. Chambers (2012b), Regression Analysis under Probabilistic Multi-Linkage. *Statistica Neerlandica* **66**, pp. 64–79.

Lahiri, P. and M.D. Larsen (2005), Regression Analysis with Linked Data. *Journal of the American Statistical Association* **100**, pp. 222-230.

Scheuren, F., and W.E. Winkler (1993), Regression Analysis of Data Files that are Computer Matched. *Survey Methodology* **19**, pp. 39-58.

Scheuren, F., and W. E. Winkler (1997), Regression Analysis of Data Files that are Computer Matched II. *Survey Methodology* **23**, pp. 157-165.

Van Grootheest G., M.C.H. de Groot, D.J. van der Laan, J.H. Smit and B.F.M. Bakker (eds.) (2015), *Record Linkage for Health Studies: Three Demonstration Projects*. Statistics Netherlands, The Hague.

Willenborg, L. and De Waal, T. (2001), *Elements of Statistical Disclosure Control in Practice*. Lecture Notes in Statistics, 155. New York: Springer-Verlag.

Winglee, M., R. Valliant and F. Scheuren (2005), A Case Study in Record Linkage. *Survey Methodology* **31** (1), pp. 3-11.

Appendix A. Proofs and derivations

A.1 Proof of Theorem 1

Theorem 1. *For random linkage errors that are described by a matrix* $\mathbf{Q}$ *of the form (2) that satisfies assumption (3), if the truncation in step 3 of the algorithm of CC is never used, then this algorithm converges to the solution* $\widehat{\mathbf{T}}^{\mathrm{CC}} = \widehat{\mathbf{T}}^{\mathrm{BC}}$.

Proof. First of all, we observe that the row sums of $\widehat{\mathbf{T}}_0 = \widehat{\mathbf{T}}^*$ in the diagonal matrix $\mathbf{D}_0$ are equal to the marginal counts of the variable y, which are known from the first dataset alone (without record linkage). Crucially, provided that the truncation in step 3 of the algorithm of CC is never used, the row sums of $\widehat{\mathbf{T}}_h$ in later iterations will remain equal to these known marginal counts, and therefore $\mathbf{D}_h = \mathbf{D}$ will remain fixed throughout the algorithm. In fact, it holds that $\mathbf{D} = \mathbf{Y}'\mathbf{Y}$.

Hence, if the truncation if never used, then we may write the algorithm concisely as:

$$\begin{aligned} \hat{\Delta}_h &= \mathbf{Y}\mathbf{D}^{-1}\widehat{\mathbf{T}}_h, \\ \widehat{\mathbf{T}}_{h+1} &= \mathbf{Y}'\mathbf{Z}^* + \mathbf{Y}'(\mathbf{I}-\mathbf{Q})\hat{\Delta}_h. \end{aligned}$$

Starting with $\hat{\Delta}_0 = \mathbf{Y}\mathbf{D}^{-1}\widehat{\mathbf{T}}_0 = \mathbf{Y}\mathbf{D}^{-1}\mathbf{Y}'\mathbf{Z}^*$ and carrying out the first few iterations of this algorithm, it is seen that, for $h \in \{0,1,2\}$,

$$\hat{\Delta}_h = \left\{\sum_{g=0}^{h}[\mathbf{Y}\mathbf{D}^{-1}\mathbf{Y}'(\mathbf{I}-\mathbf{Q})]^g\right\}\mathbf{Y}\mathbf{D}^{-1}\mathbf{Y}'\mathbf{Z}^*.$$

Suppose that this expression holds for all iterations up to h. One further iteration then yields:

$$\begin{aligned} \hat{\Delta}_{h+1} &= \mathbf{Y}\mathbf{D}^{-1}\{\mathbf{Y}'\mathbf{Z}^* + \mathbf{Y}'(\mathbf{I}-\mathbf{Q})\hat{\Delta}_h\} \\ &= \mathbf{Y}\mathbf{D}^{-1}\mathbf{Y}'\mathbf{Z}^* + \left\{\sum_{g=1}^{h+1}[\mathbf{Y}\mathbf{D}^{-1}\mathbf{Y}'(\mathbf{I}-\mathbf{Q})]^g\right\}\mathbf{Y}\mathbf{D}^{-1}\mathbf{Y}'\mathbf{Z}^* \\ &= \left\{\sum_{g=0}^{h+1}[\mathbf{Y}\mathbf{D}^{-1}\mathbf{Y}'(\mathbf{I}-\mathbf{Q})]^g\right\}\mathbf{Y}\mathbf{D}^{-1}\mathbf{Y}'\mathbf{Z}^*. \end{aligned}$$

So, by induction, each $\hat{\Delta}_h$ encountered during the algorithm can be written in this form.

Define $\mathbf{H} = \mathbf{Y}\mathbf{D}^{-1}\mathbf{Y}'$. Clearly, $\mathbf{H}$ is symmetric. Using the above fact that $\mathbf{D} = \mathbf{Y}'\mathbf{Y}$, it is easily shown that $\mathbf{H}$ is an idempotent matrix (i.e., $\mathbf{H}^2 = \mathbf{H}$) and that $\mathbf{Y}'\mathbf{H} = \mathbf{Y}'$.

Moreover, for the exchangeable linkage error model (2), it turns out that $\mathbf{H}$ and $\mathbf{Q}$ commute:

$$\mathbf{HQ} = q\mathbf{H} + \delta(\mathbf{Huu}' - \mathbf{H}) = q\mathbf{H} + \delta(\mathbf{uu}'\mathbf{H} - \mathbf{H}) = \mathbf{QH},$$

since $\mathbf{uu}'$ is a matrix of ones which commutes with every symmetric matrix. By extension, it also follows that $\mathbf{H}(\mathbf{I} - \mathbf{Q}) = (\mathbf{I} - \mathbf{Q})\mathbf{H}$ and [using expression (4)] that $\mathbf{HQ}^{-1} = \mathbf{Q}^{-1}\mathbf{H}$.

By re-arranging the factors within each term, the above expression for $\hat{\Delta}_{h+1}$ can be rewritten as:

$$\hat{\Delta}_{h+1} = \left\{\sum_{g=0}^{h+1}[\mathbf{H}(\mathbf{I} - \mathbf{Q})]^g\right\}\mathbf{HZ}^* = \left\{\sum_{g=0}^{h+1}(\mathbf{I} - \mathbf{Q})^g\mathbf{H}^g\right\}\mathbf{HZ}^* = \left\{\sum_{g=0}^{h+1}(\mathbf{I} - \mathbf{Q})^g\right\}\mathbf{HZ}^*.$$

It follows that the algorithm converges in this case to a solution of the form

$$\hat{\mathbf{T}}^{CC} = \lim_{h\to\infty}\hat{\mathbf{T}}_h = \mathbf{Y}'\mathbf{Z}^* + \mathbf{Y}'(\mathbf{I} - \mathbf{Q})\left\{\sum_{g=0}^{\infty}(\mathbf{I} - \mathbf{Q})^g\right\}\mathbf{HZ}^*.$$

Since assumption (3) implies that $\mathbf{Q}$ has an inverse, it follows from the Woodbury matrix identity (see, e.g., Hager, 1989) that

$$\sum_{g=0}^{\infty}(\mathbf{I} - \mathbf{Q})^g = \{\mathbf{I} - (\mathbf{I} - \mathbf{Q})\}^{-1} = \mathbf{Q}^{-1}.$$

Therefore:

$$\begin{aligned}\hat{\mathbf{T}}^{CC} &= \mathbf{Y}'\mathbf{Z}^* + \mathbf{Y}'(\mathbf{I} - \mathbf{Q})\mathbf{Q}^{-1}\mathbf{HZ}^* \\ &= \mathbf{Y}'\mathbf{Z}^* + \mathbf{Y}'\mathbf{Q}^{-1}\mathbf{HZ}^* - \mathbf{Y}'\mathbf{HZ}^* \\ &= \mathbf{Y}'\mathbf{Z}^* + \mathbf{Y}'\mathbf{HQ}^{-1}\mathbf{Z}^* - \mathbf{Y}'\mathbf{Z}^* \\ &= \mathbf{Y}'\mathbf{Q}^{-1}\mathbf{Z}^*.\end{aligned}$$

In the third and fourth lines, it was used that $\mathbf{Y}'\mathbf{H} = \mathbf{Y}'$ as noted above. In the third line it was also used that $\mathbf{H}$ commutes with $\mathbf{Q}^{-1}$. We conclude that, under the assumptions made here, $\hat{\mathbf{T}}^{CC} = \hat{\mathbf{T}}^{BC}$. ■

A.2 Proof of Theorem 2

As a starting point for the analysis of the estimation error in a $\mathbf{Q}$-adjusted contingency table $\hat{\mathbf{T}}^{Q}$, it is convenient to first derive the following lemma.

Lemma 1. *For random linkage errors that are described by a matrix* $\mathbf{Q}$ *of the form (2), the errors* $\mathrm{e}^{*}_{\mathrm{jk}}$ *and* $\mathrm{e}^{\mathrm{Q}}_{\mathrm{jk}}$ *satisfy the following identity:*

$$\left(e_{jk}^{Q}\right)^2 - \left(e_{jk}^{*}\right)^2 = -\delta\left[2\left(n\hat{t}_{jk}^{*} - r_j s_k\right)\left(\hat{t}_{jk}^{*} - t_{jk}\right) - \delta\left(n\hat{t}_{jk}^{*} - r_j s_k\right)^2\right].$$

Proof. Since $e_{jk}^{*} = \mathbf{y}_j'(\mathbf{C}-\mathbf{I})\mathbf{z}_k$ and $e_{jk}^{Q} = \mathbf{y}_j'(\mathbf{Q}'\mathbf{C}-\mathbf{I})\mathbf{z}_k$ from (6), it follows immediately that

$$\begin{aligned}(e_{jk}^{*})^2 &= [\mathbf{y}_j'(\mathbf{C}-\mathbf{I})\mathbf{z}_k]'[\mathbf{y}_j'(\mathbf{C}-\mathbf{I})\mathbf{z}_k] = \mathbf{z}_k'(\mathbf{C}-\mathbf{I})'\mathbf{y}_j\mathbf{y}_j'(\mathbf{C}-\mathbf{I})\mathbf{z}_k,\\ \left(e_{jk}^{Q}\right)^2 &= \mathbf{z}_k'(\mathbf{Q}'\mathbf{C}-\mathbf{I})'\mathbf{y}_j\mathbf{y}_j'(\mathbf{Q}'\mathbf{C}-\mathbf{I})\mathbf{z}_k.\end{aligned}$$

Using the fact that $(n-1)\delta = (1-q)$, we find that

$$\begin{aligned}\mathbf{Q}'\mathbf{C}-\mathbf{I} &= q\mathbf{C} + \delta(\mathbf{u}\mathbf{u}'\mathbf{C}-\mathbf{C}) - \mathbf{I}\\ &= (\mathbf{C}-\mathbf{I}) + \delta[\mathbf{u}\mathbf{u}' - \mathbf{C} - (n-1)\mathbf{C}]\\ &= (\mathbf{C}-\mathbf{I}) + \delta(\mathbf{u}\mathbf{u}' - n\mathbf{C}).\end{aligned}$$

In the second line, we used the fact that $\mathbf{C}$ is a permutation matrix and therefore

$$\mathbf{u}\mathbf{u}'\mathbf{C} = \mathbf{u}\mathbf{u}'. \tag{16}$$

Substituting the above expression for $\mathbf{Q}'\mathbf{C}-\mathbf{I}$, we find that

$$\begin{aligned}\left(e_{jk}^{Q}\right)^2 = (e_{jk}^{*})^2 &+ \delta^2\mathbf{z}_k'(\mathbf{u}\mathbf{u}'-n\mathbf{C})'\mathbf{y}_j\mathbf{y}_j'(\mathbf{u}\mathbf{u}'-n\mathbf{C})\mathbf{z}_k\\ &+ 2\delta\mathbf{z}_k'(\mathbf{C}-\mathbf{I})'\mathbf{y}_j\mathbf{y}_j'(\mathbf{u}\mathbf{u}'-n\mathbf{C})\mathbf{z}_k.\end{aligned}$$

Note that $\mathbf{z}_k'(\mathbf{C}-\mathbf{I})'\mathbf{y}_j\mathbf{y}_j'(\mathbf{u}\mathbf{u}'-n\mathbf{C})\mathbf{z}_k = \mathbf{z}_k'(\mathbf{u}\mathbf{u}'-n\mathbf{C})'\mathbf{y}_j\mathbf{y}_j'(\mathbf{C}-\mathbf{I})\mathbf{z}_k$, since we are dealing with scalar quantities. Hence,

$$\begin{aligned}\left(e_{jk}^{Q}\right)^2 - \left(e_{jk}^{*}\right)^2 = \delta\big[&2\mathbf{z}_k'(\mathbf{C}-\mathbf{I})'\mathbf{y}_j\mathbf{y}_j'(\mathbf{u}\mathbf{u}'-n\mathbf{C})\mathbf{z}_k\\ &+ \delta\mathbf{z}_k'(\mathbf{u}\mathbf{u}'-n\mathbf{C})'\mathbf{y}_j\mathbf{y}_j'(\mathbf{u}\mathbf{u}'-n\mathbf{C})\mathbf{z}_k\big].\end{aligned} \tag{17}$$

Using the notation that was introduced in Section 3 ($r_j = \mathbf{y}_j'\mathbf{u}$, $s_k = \mathbf{z}_k'\mathbf{u}$, $t_{jk} = \mathbf{y}_j'\mathbf{z}_k$ and $\hat{t}_{jk}^{*} = \mathbf{y}_j'\mathbf{C}\mathbf{z}_k$), we find that

$$\begin{aligned}\mathbf{z}_k'(\mathbf{C}-\mathbf{I})'\mathbf{y}_j\mathbf{y}_j'(\mathbf{u}\mathbf{u}'-n\mathbf{C})\mathbf{z}_k &= \mathbf{z}_k'\big[\mathbf{C}'\mathbf{y}_j\mathbf{y}_j'\mathbf{u}\mathbf{u}' - n\mathbf{C}'\mathbf{y}_j\mathbf{y}_j'\mathbf{C} - \mathbf{y}_j\mathbf{y}_j'\mathbf{u}\mathbf{u}' + n\mathbf{y}_j\mathbf{y}_j'\mathbf{C}\big]\mathbf{z}_k\\ &= r_j s_k\hat{t}_{jk}^{*} - n\left(\hat{t}_{jk}^{*}\right)^2 - r_j s_k t_{jk} + n t_{jk}\hat{t}_{jk}^{*}\\ &= (r_j s_k - n\hat{t}_{jk}^{*})(\hat{t}_{jk}^{*} - t_{jk})\end{aligned}$$

and

$$\begin{aligned}&\mathbf{z}_k'(\mathbf{u}\mathbf{u}'-n\mathbf{C})'\mathbf{y}_j\mathbf{y}_j'(\mathbf{u}\mathbf{u}'-n\mathbf{C})\mathbf{z}_k\\ &= \mathbf{z}_k'\big[\mathbf{u}\mathbf{u}'\mathbf{y}_j\mathbf{y}_j'\mathbf{u}\mathbf{u}' - n\mathbf{u}\mathbf{u}'\mathbf{y}_j\mathbf{y}_j'\mathbf{C} - n\mathbf{C}'\mathbf{y}_j\mathbf{y}_j'\mathbf{u}\mathbf{u}' + n^2\mathbf{C}'\mathbf{y}_j\mathbf{y}_j'\mathbf{C}\big]\mathbf{z}_k\\ &= r_j^2 s_k^2 - 2n r_j s_k\hat{t}_{jk}^{*} + n^2\left(\hat{t}_{jk}^{*}\right)^2\\ &= \left(n\hat{t}_{jk}^{*} - r_j s_k\right)^2.\end{aligned}$$

Substituting these expressions into (17), we find:

$$\left(e_{jk}^{Q}\right)^2 - \left(e_{jk}^{*}\right)^2 = \delta\left[2(r_j s_k - n\hat{t}_{jk}^{*})(\hat{t}_{jk}^{*} - t_{jk}) + \delta\left(n\hat{t}_{jk}^{*} - r_j s_k\right)^2\right].$$

The proof of Lemma 1 is completed by multiplying both factors on the right-hand-side by -1. ■

Theorem 2. *For random linkage errors that are described by a matrix* $\mathbf{Q}$ *of the form (2) that satisfies assumption (3),* $\left|\mathrm{e}_{\mathrm{jk}}^{\mathrm{Q}}\right| > \left|\mathrm{e}_{\mathrm{jk}}^{*}\right|$ *when one of the following two conditions holds:*

a. $\mathrm{nt_{jk}} > \mathrm{r_j s_k}$ *(positive association) and* $\hat{\mathrm{t}}_{\mathrm{jk}}^{*}$ *satisfies*

$$\frac{r_j s_k}{n} < \hat{t}_{jk}^{*} < t_{jk} + \delta\frac{nt_{jk} - r_j s_k}{2 - \delta n}; \tag{7}$$

b. $nt_{jk} < r_j s_k$ *(negative association) and* $\hat{\mathrm{t}}_{\mathrm{jk}}^{*}$ *satisfies*

$$t_{jk} + \delta\frac{nt_{jk} - r_j s_k}{2 - \delta n} < \hat{t}_{jk}^{*} < \frac{r_j s_k}{n}. \tag{8}$$

In all other cases, $\left|\mathrm{e}_{\mathrm{jk}}^{\mathrm{Q}}\right| \leq \left|\mathrm{e}_{\mathrm{jk}}^{*}\right|$*, with equality holding only at the endpoints of (7) or (8) and in the special case* $\mathrm{nt_{jk}} = \mathrm{r_j s_k}$ *(no association) for* $\hat{\mathrm{t}}_{\mathrm{jk}}^{*} = \mathrm{t_{jk}}$.

Proof. We start from the expression stated in Lemma 1. Applying the identity

$$n\hat{t}_{jk}^{*} - r_j s_k = n\left(\hat{t}_{jk}^{*} - t_{jk}\right) + \left(nt_{jk} - r_j s_k\right) \tag{18}$$

to this expression, we obtain the following alternative form:

$$\begin{aligned}\left(e_{jk}^{Q}\right)^2 - \left(e_{jk}^{*}\right)^2 &= -\delta\left[2n\left(\hat{t}_{jk}^{*} - t_{jk}\right)^2 + 2\left(nt_{jk} - r_j s_k\right)\left(\hat{t}_{jk}^{*} - t_{jk}\right)\right.\\ &\quad - \delta n^2\left(\hat{t}_{jk}^{*} - t_{jk}\right)^2 - 2\delta n\left(nt_{jk} - r_j s_k\right)\left(\hat{t}_{jk}^{*} - t_{jk}\right)\\ &\quad \left. - \delta\left(nt_{jk} - r_j s_k\right)^2\right]\\ &= -\delta\left[n(2 - \delta n)\left(\hat{t}_{jk}^{*} - t_{jk}\right)^2 + 2(1 - \delta n)\left(nt_{jk} - r_j s_k\right)\left(\hat{t}_{jk}^{*} - t_{jk}\right)\right.\\ &\quad \left. - \delta\left(nt_{jk} - r_j s_k\right)^2\right].\end{aligned}$$

Hence, introducing the function

$$f(u) = n(2 - \delta n)u^2 + 2(1 - \delta n)\left(nt_{jk} - r_j s_k\right)u - \delta\left(nt_{jk} - r_j s_k\right)^2,$$

it holds that

$$\left(e_{jk}^{Q}\right)^2 - \left(e_{jk}^{*}\right)^2 = -\delta f(\hat{t}_{jk}^{*} - t_{jk}). \tag{19}$$

Clearly, $\left|e_{jk}^{Q}\right| > \left|e_{jk}^{*}\right|$ if, and only if, $f\left(\hat{t}_{jk}^{*} - t_{jk}\right) < 0$.

It is not difficult to check that $f(u)$ can be factorised as follows:

$$f(u) = n(2 - \delta n)\left(u + \frac{nt_{jk} - r_j s_k}{n}\right)\left(u - \delta\frac{nt_{jk} - r_j s_k}{2 - \delta n}\right).$$

Thus, the quadratic polynomial $f(u)$ has zeroes at the following points:

$$u_1 = -\frac{nt_{jk} - r_j s_k}{n},$$
$$u_2 = \delta \frac{nt_{jk} - r_j s_k}{2 - \delta n}.$$

Together with $2 - \delta n > 1 > 0$, this implies that $f(u) < 0$ for all u in between u_1 and u_2, and that $f(u) > 0$ for all u to the left of the left-most zero point and all u to the right of the right-most zero point.

We now examine separately the three cases of no association, positive association, and negative association as defined in Section 3.1.

Case 1: $\mathrm{nt_{jk}} = \mathrm{r_j s_k}$ *(no association)*
In this case, it is seen that $u_1 = u_2 = 0$, and hence that $f(u) > 0$ for all $u \neq 0$ (and $f(0) = 0$). Thus, it follows from (19) that $\left(e_{jk}^{Q}\right)^2 \leq \left(e_{jk}^{*}\right)^2$, with equality holding if, and only if, $\hat{t}_{jk}^{*} = t_{jk}$ (in which case $e_{jk}^{*} = e_{jk}^{Q} = 0$). We conclude that in this case adjusting the observed entry of the contingency table by the $\mathbf{Q}$ matrix can only decrease the error in the estimated entry, for any realisation of the permutation matrix $\mathbf{C}$. If the originally observed entry happened to be error-free, then the adjusted entry will also be error-free (i.e., $\hat{t}_{jk}^{*} = \hat{t}_{jk}^{Q} = t_{jk}$).

Case 2: $\mathrm{nt_{jk}} > \mathrm{r_j s_k}$ *(positive association)*
In this case, $u_1 < 0 < u_2$. Hence, $f(u) > 0$ for all $u < u_1$ and $u > u_2$, while $f(u) < 0$ for all $u_1 < u < u_2$. It follows from (19) that $\left(e_{jk}^{Q}\right)^2 > \left(e_{jk}^{*}\right)^2$ when $u_1 < \hat{t}_{jk}^{*} - t_{jk} < u_2$. Re-arranging terms, we find that $\left(e_{jk}^{Q}\right)^2 > \left(e_{jk}^{*}\right)^2$ if, and only if, condition (7) holds.

If $\hat{t}_{jk}^{*}$ is exactly equal to one of the endpoints in (7), then $e_{jk}^{Q} = e_{jk}^{*}$. For all other values of $\hat{t}_{jk}^{*}$ that do not satisfy condition (7), it holds that $\left(e_{jk}^{Q}\right)^2 < \left(e_{jk}^{*}\right)^2$. We conclude that in this case adjusting the observed entry of the contingency table by the $\mathbf{Q}$ matrix decreases the error in the estimated entry unless the permutation matrix $\mathbf{C}$ happens to be such that the original observed entry satisfies (7).

Case 3: $\mathrm{nt_{jk}} < \mathrm{r_j s_k}$ *(negative association)*
In this case, $u_2 < 0 < u_1$. Hence, $f(u) > 0$ for all $u < u_2$ and $u > u_1$, while $f(u) < 0$ for all $u_2 < u < u_1$. Analogously to Case 2, we find from (19) that $\left(e_{jk}^{Q}\right)^2 > \left(e_{jk}^{*}\right)^2$ if, and only if, condition (8) holds.

Again, $e_{jk}^{Q} = e_{jk}^{*}$ if $\hat{t}_{jk}^{*}$ is exactly equal to an endpoint in (8), and otherwise $\left(e_{jk}^{Q}\right)^2 < \left(e_{jk}^{*}\right)^2$. We conclude that in this case adjusting the observed entry of the contingency table by the $\mathbf{Q}$ matrix decreases the error in the estimated entry unless the permutation matrix $\mathbf{C}$ happens to be such that the original observed entry satisfies (8). ■

A.3 Proof of Theorem 3

As a starting point for the analysis of the estimation error in the bias-corrected contingency table $\widehat{\mathbf{T}}^{BC}$, it is convenient to first derive the following lemma.

Lemma 2. *For random linkage errors that are described by a matrix* $\mathbf{Q}$ *of the form (2) that satisfies assumption (3), the errors* $\mathrm{e}^{*}_{\mathrm{jk}}$ *and* $\mathrm{e}^{\mathrm{BC}}_{\mathrm{jk}}$ *satisfy the following identity:*

$$\begin{aligned}\left(e_{jk}^{BC}\right)^2 - \left(e_{jk}^{*}\right)^2 &= \left[\left(\frac{n-1}{nq-1}\right)^2 - 1\right]\left(\hat{t}_{jk}^{*} - t_{jk}\right)^2 \\ &+ 2\frac{(1-q)(n-1)}{(nq-1)^2}\left(nt_{jk} - r_j s_k\right)\left(\hat{t}_{jk}^{*} - t_{jk}\right) \\ &+ \left(\frac{1-q}{nq-1}\right)^2\left(nt_{jk} - r_j s_k\right)^2.\end{aligned}$$

Proof. Under the assumptions of the lemma, the inverse of $\mathbf{Q}$ is given by expression (4). Substituting this expression into e_{jk}^{BC} given by (6), we obtain:

$$\begin{aligned}e_{jk}^{BC} &= \frac{1}{nq-1}\mathbf{y}_j'\{(n-1)\mathbf{C} - (1-q)\mathbf{u}\mathbf{u}'\mathbf{C} - (nq-1)\mathbf{I}\}\mathbf{z}_k \\ &= \frac{1}{nq-1}\mathbf{y}_j'\{(n-1)\mathbf{C} - (1-q)\mathbf{u}\mathbf{u}' - [n-1-n(1-q)]\mathbf{I}\}\mathbf{z}_k \\ &= \frac{n-1}{nq-1}\mathbf{y}_j'(\mathbf{C}-\mathbf{I})\mathbf{z}_k - \frac{1-q}{nq-1}\mathbf{y}_j'(\mathbf{u}\mathbf{u}' - n\mathbf{I})\mathbf{z}_k.\end{aligned} \tag{20}$$

In the second line, we again used identity (16) from the proof of Lemma 1. Thus, we find that

$$\begin{aligned}(e_{jk}^{*})^2 &= [\mathbf{y}_j'(\mathbf{C}-\mathbf{I})\mathbf{z}_k]'[\mathbf{y}_j'(\mathbf{C}-\mathbf{I})\mathbf{z}_k] = \mathbf{z}_k'(\mathbf{C}-\mathbf{I})'\mathbf{y}_j\mathbf{y}_j'(\mathbf{C}-\mathbf{I})\mathbf{z}_k, \\ \left(e_{jk}^{BC}\right)^2 &= \left(\frac{n-1}{nq-1}\right)^2\mathbf{z}_k'(\mathbf{C}-\mathbf{I})'\mathbf{y}_j\mathbf{y}_j'(\mathbf{C}-\mathbf{I})\mathbf{z}_k \\ &\quad + \left(\frac{1-q}{nq-1}\right)^2\mathbf{z}_k'(\mathbf{u}\mathbf{u}' - n\mathbf{I})'\mathbf{y}_j\mathbf{y}_j'(\mathbf{u}\mathbf{u}' - n\mathbf{I})\mathbf{z}_k \\ &\quad - 2\frac{(1-q)(n-1)}{(nq-1)^2}\mathbf{z}_k'(\mathbf{C}-\mathbf{I})'\mathbf{y}_j\mathbf{y}_j'(\mathbf{u}\mathbf{u}' - n\mathbf{I})\mathbf{z}_k.\end{aligned} \tag{21}$$

Now again using the notation that was introduced in Section 3 ($r_j = \mathbf{y}_j'\mathbf{u}$, $s_k = \mathbf{z}_k'\mathbf{u}$, $t_{jk} = \mathbf{y}_j'\mathbf{z}_k$ and $\hat{t}_{jk}^{*} = \mathbf{y}_j'\mathbf{C}\mathbf{z}_k$), we obtain:

$$\begin{aligned}\mathbf{z}_k'(\mathbf{C}-\mathbf{I})'\mathbf{y}_j\mathbf{y}_j'(\mathbf{C}-\mathbf{I})\mathbf{z}_k &= \left(\hat{t}_{jk}^{*} - t_{jk}\right)^2, \\ \mathbf{z}_k'(\mathbf{u}\mathbf{u}' - n\mathbf{I})'\mathbf{y}_j\mathbf{y}_j'(\mathbf{u}\mathbf{u}' - n\mathbf{I})\mathbf{z}_k &= \left(nt_{jk} - r_j s_k\right)^2, \\ \mathbf{z}_k'(\mathbf{C}-\mathbf{I})'\mathbf{y}_j\mathbf{y}_j'(\mathbf{u}\mathbf{u}' - n\mathbf{I})\mathbf{z}_k &= \left(\hat{t}_{jk}^{*} - t_{jk}\right)\left(r_j s_k - nt_{jk}\right).\end{aligned}$$

Upon substituting these expressions into (21) and subtracting $(e_{jk}^{*})^2$ from $\left(e_{jk}^{BC}\right)^2$, the lemma follows immediately. ■

Theorem 3. *For random linkage errors that are described by a matrix* $\mathbf{Q}$ *of the form (2) that satisfies assumption (3),* $\left|\mathrm{e}_{\mathrm{jk}}^{\mathrm{BC}}\right| < \left|\mathrm{e}_{\mathrm{jk}}^{*}\right|$ *when one of the following two conditions holds:*

a. $nt_{jk} > r_j s_k$ *(positive association) and* $\hat{\mathrm{t}}_{\mathrm{jk}}^{*}$ *satisfies*

$$\frac{r_j s_k}{n} < \hat{t}_{jk}^{*} < t_{jk} - \delta\frac{nt_{jk} - r_j s_k}{2 - \delta n}; \tag{9}$$

b. $nt_{jk} < r_j s_k$ *(negative association) and* $\hat{\mathrm{t}}_{\mathrm{jk}}^{*}$ *satisfies*

$$t_{jk} - \delta\frac{nt_{jk} - r_j s_k}{2 - \delta n} < \hat{t}_{jk}^{*} < \frac{r_j s_k}{n}. \tag{10}$$

In all other cases, $\left|\mathrm{e}_{\mathrm{jk}}^{\mathrm{BC}}\right| \geq \left|\mathrm{e}_{\mathrm{jk}}^{*}\right|$*, with equality holding only at the endpoints of (9) or (10) and in the special case* $\mathrm{nt_{jk}} = \mathrm{r_j s_k}$ *(no association) for* $\hat{\mathrm{t}}_{\mathrm{jk}}^{*} = \mathrm{t_{jk}}$.

Proof. We start by applying Lemma 2. By observing that

$$\frac{1-q}{nq-1} = \frac{\delta(n-1)}{n[1-\delta(n-1)]-1} = \frac{\delta(n-1)}{(n-1)-\delta n(n-1)} = \frac{\delta}{1-\delta n}$$

and, similarly,

$$\frac{n-1}{nq-1} = \frac{(n-1)}{(n-1)-\delta n(n-1)} = \frac{1}{1-\delta n},$$

the expression for $\left(e_{jk}^{BC}\right)^2 - \left(e_{jk}^{*}\right)^2$ in the lemma can be re-written as follows:

$$\begin{aligned}
&\left(e_{jk}^{BC}\right)^2 - \left(e_{jk}^{*}\right)^2 \\
&= \left[\frac{1}{(1-\delta n)^2} - 1\right]\left(\hat{t}_{jk}^{*} - t_{jk}\right)^2 + \frac{2\delta}{(1-\delta n)^2}\left(nt_{jk} - r_j s_k\right)\left(\hat{t}_{jk}^{*} - t_{jk}\right) \\
&\qquad + \frac{\delta^2}{(1-\delta n)^2}\left(nt_{jk} - r_j s_k\right)^2 \\
&= \frac{2\delta n - (\delta n)^2}{(1-\delta n)^2}\left(\hat{t}_{jk}^{*} - t_{jk}\right)^2 + \frac{2\delta}{(1-\delta n)^2}\left(nt_{jk} - r_j s_k\right)\left(\hat{t}_{jk}^{*} - t_{jk}\right) \\
&\qquad + \frac{\delta^2}{(1-\delta n)^2}\left(nt_{jk} - r_j s_k\right)^2 \\
&= \frac{\delta}{(1-\delta n)^2}\Big[n(2-\delta n)\left(\hat{t}_{jk}^{*} - t_{jk}\right)^2 + 2\left(nt_{jk} - r_j s_k\right)\left(\hat{t}_{jk}^{*} - t_{jk}\right) \\
&\qquad + \delta\left(nt_{jk} - r_j s_k\right)^2\Big].
\end{aligned}$$

Hence, introducing the function

$$g(u) = n(2-\delta n)u^2 + 2\left(nt_{jk} - r_j s_k\right)u + \delta\left(nt_{jk} - r_j s_k\right)^2,$$

it holds that

$$\left(e_{jk}^{BC}\right)^2 - \left(e_{jk}^{*}\right)^2 = \frac{\delta}{(1-\delta n)^2}\, g\left(\hat{t}_{jk}^{*} - t_{jk}\right).$$

Clearly, $\left|e_{jk}^{BC}\right| < \left|e_{jk}^{*}\right|$ if, and only if, $g\left(\hat{t}_{jk}^{*} - t_{jk}\right) < 0$.

It is not difficult to check that $g(u)$ can be factorised as follows:

$$g(u) = n(2-\delta n)\left(u + \frac{nt_{jk} - r_j s_k}{n}\right)\left(u + \delta\frac{nt_{jk} - r_j s_k}{2-\delta n}\right).$$

Thus, the quadratic polynomial $g(u)$ has zeroes at the points u_1 and $u_3 = -u_2$, with u_1 and u_2 as defined in the proof of Theorem 2. Moreover, since $2 - \delta n > 1 > 0$, it must hold that $g(u) < 0$ in between the two zero points, while $g(u) > 0$ for all u to the left of the left-most zero point and to the right of the right-most zero point.

From here, we can proceed completely analogously to the proof of Theorem 2, by considering separately the three cases of no association, positive association, and negative association. The only additional observation we need is that, under assumption (3),

$$\frac{\delta}{2-\delta n} < \delta < \frac{1}{n}.$$

From this, it follows that in the case of positive association ($nt_{jk} > r_j s_k$), it holds that $u_1 < u_3 < 0$ and hence the relevant interval where $g(u) < 0$ is given by $u_1 < u < u_3$. Similarly, in the case of negative association ($nt_{jk} < r_j s_k$), it holds that $0 < u_3 < u_1$ and hence the relevant interval where $g(u) < 0$ is given by $u_3 < u < u_1$. By substituting $u = \hat{t}^*_{jk} - t_{jk}$ into these intervals, conditions (9) and (10) follow immediately. ■

A.4 Derivation of expression (15)

To derive expression (15) for $\mathrm{Var}\left(\mathbf{y}_j'\mathbf{C}\mathbf{z}_k\right)$, we begin by observing that:

$$\begin{aligned}
\mathrm{Var}\left(\mathbf{y}_j'\mathbf{C}\mathbf{z}_k\right) &= E\left(\mathbf{y}_j'\mathbf{C}\mathbf{z}_k\left(\mathbf{y}_j'\mathbf{C}\mathbf{z}_k\right)'\right) - E\left(\mathbf{y}_j'\mathbf{C}\mathbf{z}_k\right)E\left(\mathbf{y}_j'\mathbf{C}\mathbf{z}_k\right)' \\
&= E\left(\mathbf{y}_j'\mathbf{C}\mathbf{z}_k\mathbf{z}_k'\mathbf{C}'\mathbf{y}_j\right) - E\left(\mathbf{y}_j'\mathbf{C}\mathbf{z}_k\right)E\left(\mathbf{z}_k'\mathbf{C}'\mathbf{y}_j\right) \\
&= \mathbf{y}_j'E(\mathbf{C}\mathbf{z}_k\mathbf{z}_k'\mathbf{C}')\mathbf{y}_j - \mathbf{y}_j'E(\mathbf{C}\mathbf{z}_k)E(\mathbf{z}_k'\mathbf{C}')\mathbf{y}_j \\
&= \mathbf{y}_j'\{E(\mathbf{C}\mathbf{z}_k\mathbf{z}_k'\mathbf{C}') - E(\mathbf{C}\mathbf{z}_k)E(\mathbf{z}_k'\mathbf{C}')\}\mathbf{y}_j \\
&= \mathbf{y}_j'\,\mathrm{Var}(\mathbf{C}\mathbf{z}_k)\,\mathbf{y}_j.
\end{aligned}$$

To evaluate a similar variance, Chambers (2009) proposed to extend the exchangeable linkage error model by the assumption "that the linkage errors associated with two distinct records [in data file A] can be treated as approximately independent", arguing that this assumption is reasonable when the number of records n is sufficiently large. In our notation, this assumption can be summarised as follows (cf. Chambers, 2009, p. 49):

$$\begin{aligned}
&\Pr\left(c_{mp} = 1 \middle| c_{il} = 1\right) \\
&= I(i=m)I(l=p) + I(i\neq m)I(l\neq p)\{qI(m=p) + \delta I(m\neq p)\}
\end{aligned}$$

$$= \begin{cases} 1 & \text{if } i = m \text{ and } l = p \\ 0 & \text{if } (i = m \text{ and } l \neq p) \text{ or } (i \neq m \text{ and } l = p) \\ q & \text{if } i \neq m \text{ and } l \neq p \text{ and } m = p \\ \delta & \text{if } i \neq m \text{ and } l \neq p \text{ and } m \neq p \end{cases}$$

Here, i, l, m, and p are indices over the records in the two datasets, $I(\text{statement}) = 1$ if the statement holds true, and $I(\text{statement}) = 0$ otherwise.

Chambers (2009) showed that, under this assumption, the variance-covariance matrix $\mathrm{Var}(\mathbf{Cf})$ for a general vector $\mathbf{f}$ can be approximated by:

$$\begin{aligned} \mathrm{Var}(\mathbf{Cf}) &\approx \mathrm{diag}\left\{(1-q)\left[q\left(f_i - \bar{f}\right)^2 + \overline{f^{(2)}} - \left(\bar{f}\right)^2\right]\right\}, \\ \bar{f} &= \frac{1}{n}\sum_{i=1}^{n} f_i\,, \\ \overline{f^{(2)}} &= \frac{1}{n}\sum_{i=1}^{n} f_i^2, \end{aligned}$$

where the neglected terms are of the order $O(1/n)$. Applying this to our situation with $\mathbf{f} = \mathbf{z}_k$ and noting that $z_{ik}^2 = z_{ik}$ and $\sum_{i=1}^{n} z_{ik} = s_k$, we find:

$$\begin{aligned} \mathrm{Var}(\mathbf{Cz}_k) &\approx \mathrm{diag}\left\{(1-q)\left[q\left(z_{ik} - \frac{s_k}{n}\right)^2 + \frac{s_k}{n} - \left(\frac{s_k}{n}\right)^2\right]\right\} \\ &= \mathrm{diag}\left\{(1-q)\left[qz_{ik}\left(1 - 2\frac{s_k}{n}\right) + \frac{s_k}{n} - (1-q)\left(\frac{s_k}{n}\right)^2\right]\right\}. \end{aligned}$$

Substituting this approximation into the above expression for $\mathrm{Var}\left(\mathbf{y}_j'\mathbf{Cz}_k\right)$ yields:

$$\begin{aligned} \mathrm{Var}\left(\mathbf{y}_j'\mathbf{Cz}_k\right) &\approx \sum_{i=1}^{n}(1-q)\left[qz_{ik}\left(1 - 2\frac{s_k}{n}\right) + \frac{s_k}{n} - (1-q)\left(\frac{s_k}{n}\right)^2\right]y_{ij}^2 \\ &= \sum_{i=1}^{n}(1-q)\left[qz_{ik}\left(1 - 2\frac{s_k}{n}\right) + \frac{s_k}{n} - (1-q)\left(\frac{s_k}{n}\right)^2\right]y_{ij} \\ &= q(1-q)\left(1 - 2\frac{s_k}{n}\right)t_{jk} + (1-q)\frac{r_j s_k}{n} - (1-q)^2\left(\frac{s_k}{n}\right)^2 r_j \\ &= q(1-q)\left(1 - 2\frac{s_k}{n}\right)t_{jk} + (1-q)\frac{r_j s_k}{n}\left[1 - (1-q)\frac{s_k}{n}\right]. \end{aligned}$$

In the second line it was used that $y_{ij}^2 = y_{ij}$ and in the fourth line it was used that $\sum_{i=1}^{n} y_{ij} = r_j$ and $\sum_{i=1}^{n} y_{ij} z_{ik} = t_{jk}$.

Appendix B. Standard error and bias components of the empirical MSE (supplement to Table 2)

Table B1: Empirical standard error of the cell entries across 1000 linked files according to the 0.9 or 0.8 error matrices and dependent or independent attributes of the original tables for the naïve approach, and the relative difference from the naïve empirical standard deviation for the Q and $\mathbf{Q}^{-1}$ approaches. We denote row by R_j where $j = 1,..,5$ and column by C_k where $k = 1,\ldots,5$

Cell Row		Dependent Attributes									
		0.9					0.8				
		C_1	C_2	C_3	C_4	C_5	C_1	C_2	C_3	C_4	C_5
Naïve:	R_1	0.93	1.25	0.84	0.82	1.23	1.28	1.48	1.15	1.08	1.38
	R_2	1.09	1.08	1.12	1.14	1.44	1.41	1.49	1.49	1.52	1.73
	R_3	1.08	1.21	1.29	1.36	1.64	1.55	1.64	1.79	1.87	2.25
	R_4	1.10	1.26	1.24	1.49	1.76	1.56	1.65	1.74	2.00	2.31
	R_5	0.83	0.89	0.96	1.11	1.81	1.05	1.18	1.21	1.59	2.07
Q:	R_1	-0.10	-0.10	-0.10	-0.10	-0.10	-0.20	-0.20	-0.20	-0.20	-0.20
	R_2	-0.10	-0.10	-0.10	-0.10	-0.10	-0.20	-0.20	-0.20	-0.20	-0.20
	R_3	-0.10	-0.10	-0.10	-0.10	-0.10	-0.20	-0.20	-0.20	-0.20	-0.20
	R_4	-0.10	-0.10	-0.10	-0.10	-0.10	-0.20	-0.20	-0.20	-0.20	-0.20
	R_5	-0.10	-0.10	-0.10	-0.10	-0.10	-0.20	-0.20	-0.20	-0.20	-0.20
Q^{-1}:	R_1	0.11	0.11	0.11	0.11	0.11	0.25	0.25	0.25	0.25	0.25
	R_2	0.11	0.11	0.11	0.11	0.11	0.25	0.25	0.25	0.25	0.25
	R_3	0.11	0.11	0.11	0.11	0.11	0.25	0.25	0.25	0.25	0.25
	R_4	0.11	0.11	0.11	0.11	0.11	0.25	0.25	0.25	0.25	0.25
	R_5	0.11	0.11	0.11	0.11	0.11	0.25	0.25	0.25	0.25	0.25
Cell Row		**Independent Attributes**									
		0.9					**0.8**				
		C_1	C_2	C_3	C_4	C_5	C_1	C_2	C_3	C_4	C_5
Naïve:	R_1	0.79	0.68	0.82	1.01	1.11	1.08	0.95	1.18	1.38	1.47
	R_2	0.88	0.94	1.08	1.29	1.28	1.22	1.30	1.48	1.78	1.79
	R_3	1.07	1.08	1.30	1.50	1.50	1.47	1.50	1.84	2.03	2.18
	R_4	1.11	1.16	1.34	1.53	1.59	1.53	1.46	1.85	2.24	2.18
	R_5	1.01	0.90	1.05	1.24	1.21	1.33	1.16	1.43	1.72	1.71
Q:	R_1	-0.10	-0.10	-0.10	-0.10	-0.10	-0.20	-0.20	-0.20	-0.20	-0.20
	R_2	-0.10	-0.10	-0.10	-0.10	-0.10	-0.20	-0.20	-0.20	-0.20	-0.20
	R_3	-0.10	-0.10	-0.10	-0.10	-0.10	-0.20	-0.20	-0.20	-0.20	-0.20
	R_4	-0.10	-0.10	-0.10	-0.10	-0.10	-0.20	-0.20	-0.20	-0.20	-0.20
	R_5	-0.10	-0.10	-0.10	-0.10	-0.10	-0.20	-0.20	-0.20	-0.20	-0.20
Q^{-1}:	R_1	0.11	0.11	0.11	0.11	0.11	0.25	0.25	0.25	0.25	0.25
	R_2	0.11	0.11	0.11	0.11	0.11	0.25	0.25	0.25	0.25	0.25
	R_3	0.11	0.11	0.11	0.11	0.11	0.25	0.25	0.25	0.25	0.25
	R_4	0.11	0.11	0.11	0.11	0.11	0.25	0.25	0.25	0.25	0.25
	R_5	0.11	0.11	0.11	0.11	0.11	0.25	0.25	0.25	0.25	0.25

From Table B1, we see that the relative difference from the empirical standard error of the naïve approach compared to the $\mathbf{Q}$ and $\mathbf{Q}^{-1}$ approaches are fixed for each cell of the table (see Section 4.2). Under the $\mathbf{Q}$ approach we have smaller standard errors compared to the naïve approach and under the $\mathbf{Q}^{-1}$ approach

standard errors are larger compared to the naïve approach. We also observe larger standard errors and larger differences in standard errors between approaches as the error matrix introduces more linkage errors from 0.9 to 0.8 on the diagonal. There is no difference in the standard error results on whether the original table has dependent or independent attributes. These results confirm the theoretical findings of Section 4.2.

Table B2: Empirical bias of the cell entries across 1000 linked files according to the 0.9 or 0.8 error matrices and dependent or independent attributes on the original tables for the naïve approach, and the relative difference from the naïve empirical bias for the $\mathbf{Q}$ and $\mathbf{Q}^{-1}$ approaches. We denote row by R_j where $j = 1,..,5$ and column by C_k where $k = 1,\dots,5$

Cell Row		**Dependent Attributes**									
		0.9					**0.8**				
		C_1	C_2	C_3	C_4	C_5	C_1	C_2	C_3	C_4	C_5
Naïve:	R_1	0.50	0.96	0.03	0.51	0.98	1.06	1.86	0.01	1.08	1.85
	R_2	0.48	0.25	0.28	0.05	0.97	0.99	0.57	0.52	0.19	1.89
	R_3	0.07	0.04	0.39	0.13	0.63	0.11	0.10	0.80	0.23	1.25
	R_4	0.45	0.66	0.12	0.48	0.75	0.92	1.36	0.20	1.22	1.26
	R_5	0.60	0.60	0.59	0.06	1.83	1.25	1.17	1.13	0.18	3.73
Q:	R_1	0.94	0.90	-0.10	0.95	0.86	0.78	0.82	-0.20	0.78	0.82
	R_2	0.88	0.89	0.85	0.97	0.87	0.75	0.69	0.83	0.34	0.79
	R_3	0.65	1.21	0.97	0.35	0.83	0.66	0.88	0.84	0.29	0.73
	R_4	0.91	0.92	0.88	1.13	0.77	0.79	0.78	0.98	0.77	0.83
	R_5	0.88	0.89	0.87	1.08	0.89	0.74	0.81	0.81	0.53	0.77
Q^{-1}:	R_1	-0.96	-1.00	0.11	-0.94	-0.95	-0.98	-0.97	0.25	-0.98	-0.97
	R_2	-0.98	-0.99	-0.94	-0.93	-0.96	-0.93	-0.86	-0.96	-0.42	-0.99
	R_3	-0.72	-0.65	-0.92	-0.39	-0.92	-0.82	-0.89	-0.94	-0.37	-0.91
	R_4	-0.99	-0.98	-0.97	-0.75	-0.85	-0.99	-0.98	-0.78	-0.96	-0.96
	R_5	-0.98	-0.99	-0.97	-0.80	-0.99	-0.92	-0.99	-0.99	-0.66	-0.96
Cell Row		**Independent Attributes**									
		0.9					**0.8**				
		C_1	C_2	C_3	C_4	C_5	C_1	C_2	C_3	C_4	C_5
Naïve:	R_1	0.13	0.03	0.20	0.13	0.22	0.25	0.06	0.33	0.37	0.51
	R_2	0.11	0.00	0.11	0.04	0.18	0.24	0.01	0.09	0.09	0.43
	R_3	0.30	0.10	0.28	0.11	0.04	0.59	0.21	0.27	0.00	0.11
	R_4	0.07	0.04	0.12	0.14	0.14	0.18	0.02	0.29	0.34	0.42
	R_5	0.34	0.03	0.08	0.07	0.30	0.76	0.11	0.14	0.12	0.62
Q:	R_1	0.95	1.44	0.83	1.21	1.06	0.92	1.08	0.91	0.73	0.79
	R_2	0.81	11.1	0.49	-0.58	0.90	0.63	3.46	1.32	0.93	0.65
	R_3	0.90	0.52	0.61	-0.04	1.21	0.81	0.37	1.25	3.66	0.67
	R_4	1.20	0.67	1.13	1.08	1.18	0.81	2.94	0.78	0.75	0.64
	R_5	0.91	0.57	1.01	0.90	0.92	0.72	0.18	1.04	0.89	0.77
Q^{-1}:	R_1	-0.95	-0.40	-0.92	-0.65	-0.83	-0.85	-0.65	-0.86	-0.91	-0.99
	R_2	-0.90	10.3	-0.54	1.58	-0.99	-0.79	6.83	-0.35	-0.84	-0.81
	R_3	-1.00	-0.58	-0.68	0.04	-0.65	-0.98	-0.46	-0.43	2.58	-0.84
	R_4	-0.67	-0.74	-0.75	-0.80	-0.69	-0.99	1.68	-0.97	-0.94	-0.80
	R_5	-0.98	-0.63	-0.88	-1.00	-0.97	-0.90	-0.22	-0.70	-0.89	-0.97

In Table B2, there is less bias under the naïve approach for the error matrix with 0.9 on the diagonal compared to 0.8 on the diagonal. Also, the bias is smaller for the table with independent attributes compared to the table with dependent attributes. Table B2 also shows that the bias is larger under the $\mathbf{Q}$ approach compared to the naïve approach and smaller under the $\mathbf{Q}^{-1}$ approach compared to the naïve approach (see Section 4.1), which confirms that the $\mathbf{Q}^{-1}$ approach is a

bias-correction approach. From the theory in Section 4.1, we expect to see a relative difference in bias of -1 for the $\mathbf{Q}^{-1}$ approach in all scenarios. For the $\mathbf{Q}$ approach the theoretical relative difference in bias is 269/299 ≈ 0.90 and 239/299 ≈ 0.80 for the error matrices with 0.9 and 0.8 on the diagonal, respectively. The simulated values in Table B2 are mostly close to these theoretical values, with some larger deviations for cells where the denominator (the bias of the naïve approach) is close to zero. There is no clear evidence that the magnitude of the bias changes as we move from smaller errors in the error matrix with 0.9 on the diagonal compared to larger errors in the error matrix with 0.8 on the diagonal.

Table B3 shows the standard errors that are obtained from (14) using the variance approximation in (15) applied to the tables with dependent and independent attributes and error matrices with 0.9 and 0.8 on the diagonal from the simulation study. Most of these values are reasonably close to the empirical values shown in Table B1, especially given that the number of records in this application ($n = 300$) is not that large.

Table B3: Analytical standard error [from (15)] of the cell entries across 1000 linked files according to the 0.9 or 0.8 error matrices and dependent or independent attributes on the original tables for the naïve approach. We denote row by R_j where $j = 1,..,5$ and column by C_k where $k = 1,\dots,5$

Cell Row		**Dependent Attributes**									
		0.9					**0.8**				
		C_1	C_2	C_3	C_4	C_5	C_1	C_2	C_3	C_4	C_5
Naïve:	R_1	0.99	1.15	0.89	0.82	1.02	1.35	1.56	1.22	1.14	1.41
	R_2	1.17	1.17	1.22	1.24	1.38	1.60	1.60	1.67	1.71	1.90
	R_3	1.33	1.43	1.56	1.63	1.85	1.83	1.97	2.14	2.24	2.55
	R_4	1.27	1.34	1.53	1.78	2.02	1.76	1.85	2.10	2.45	2.78
	R_5	0.76	0.86	0.93	1.19	1.57	1.07	1.20	1.30	1.64	2.15
Cell Row		**Independent Attributes**									
		0.9					**0.8**				
		C_1	C_2	C_3	C_4	C_5	C_1	C_2	C_3	C_4	C_5
Naïve:	R_1	0.84	0.74	0.89	1.05	1.12	1.14	1.02	1.23	1.45	1.54
	R_2	0.97	0.99	1.20	1.39	1.41	1.33	1.36	1.66	1.92	1.94
	R_3	1.21	1.28	1.60	1.80	1.79	1.67	1.75	2.20	2.48	2.47
	R_4	1.33	1.31	1.67	1.90	1.86	1.83	1.80	2.29	2.62	2.56
	R_5	1.08	0.93	1.15	1.35	1.29	1.47	1.28	1.58	1.86	1.78

Explanation of symbols

Empty cell	Figure not applicable
.	Figure is unknown, insufficiently reliable or confidential
*	Provisional figure
**	Revised provisional figure
2017–2018	2017 to 2018 inclusive
2017/2018	Average for 2017 to 2018 inclusive
2017/'18	Crop year, financial year, school year, etc., beginning in 2017 and ending in 2018
2013/'14–2017/'18	Crop year, financial year, etc., 2015/'16 to 2017/'18 inclusive

Due to rounding, some totals may not correspond to the sum of the separate figures.

Colophon

Publisher
Centraal Bureau voor de Statistiek
Henri Faasdreef 312, 2492 JP Den Haag
www.cbs.nl

Prepress
Statistics Netherlands, CCN Creation and visualisation

Design
Edenspiekermann

Information
Telephone +31 88 570 70 70, fax +31 70 337 59 94
Via contactform: www.cbsl.nl/information