International Statistical Review

International Statistical Review (2018), 0, 0, 1–19 doi:10.1111/insr.12295

Statistical Analysis with Linked Data

Ying Han[1] and Partha Lahiri[2]

[1]*Department of Mathematics, University of Maryland, College Park, MD, USA*
[2]*Department of Mathematics and Joint Program of Survey Methodology, University of Maryland, College Park, MD, USA*
E-mail: hanlucky@umd.edu and plahiri@umd.edu

Summary

Computerised Record Linkage methods help us combine multiple data sets from different sources when a single data set with all necessary information is unavailable or when data collection on additional variables is time consuming and extremely costly. Linkage errors are inevitable in the linked data set because of the unavailability of error-free unique identifiers. A small amount of linkage errors can lead to substantial bias and increased variability in estimating parameters of a statistical model. In this paper, we propose a unified theory for statistical analysis with linked data. Our proposed method, unlike the ones available for secondary data analysis of linked data, exploits record linkage process data as an alternative to taking a costly sample to evaluate error rates from the record linkage procedure. A jackknife method is introduced to estimate bias, covariance matrix and mean squared error of our proposed estimators. Simulation results are presented to evaluate the performance of the proposed estimators that account for linkage errors.

Key words: Record linkage; linkage errors; estimating equations; jackknife; Monte Carlo simulation.

1 Introduction

In *record linkage*, or exact matching, one compares two or more files on a single population in the absence of a unique error-free identifier for the purpose of unduplication or production of an enhanced, merged database (e.g. Newcombe *et al.*, 1959; Fellegi & Sunter, 1969; Herzog *et al.*, 2007). Record linkage differs from *statistical matching* in terms of the types of units to be linked or matched. The primary goal of record linkage is to link an entity (e.g. person and household) from one file to the same entity in other file(s). In contrast, the primary goal of statistical matching is to link similar units (e.g. matching the same demographic group from different files). In this paper, our focus is on the statistical estimation related to record linkage and not statistical matching. Readers interested in statistical matching are referred to Rässler (2002), D'Orazio *et al.* (2006), and others.

A merged or linked data set, created by record linkage, is of great interest to analysts interested in certain specialised multivariate analysis, which would be otherwise either impossible or difficult as variables are stored in different files. Record linkage is used in many applications, including population size estimation at the US Census Bureau (Winkler, 1994, 1995; Jaro, 1989, 1995), epidemiology and medical studies (Newcombe, 1988; Gill, 1997), sociological studies, survey frame improvement and, more recently, counterterrorism (Gomatam & Larsen, 2004). The National Death Index is matched to existing insurance, medical, and other databases for

studies (e.g. Livingston & Ko, 2005; Rauscher and Sandler, 2005; Thompson *et al*., 2005). For more applications, see Alvey and Jamerson (1997) and references therein.

Record linkage can be broadly classified into deterministic and probabilistic. They both use common variables available in files to be linked that are indicative of a true match status of an entity. Such variables used for comparison are called 'matching fields'. Some of these matching fields contain lots of information for identifying population units, such as name and date of birth, while others contain very little information, such as race and gender. In deterministic record linkage, a record pair is deemed a link if the two records agree on all or some available matching fields according to a pre-specified rule, and hence, there is no stochastic element in the deterministic record linkage process. On the other hand, the linkage is called probabilistic if a record pair is deemed a link with certain probability. This paper concerns probabilistic record linkage.

Whether deterministic or probabilistic, any record linkage procedure is subject to linkage or mismatch errors. There are two kinds of linkage errors. The linkage error is called *false positive* if a mismatch is deemed a link by the record linkage procedure. On the other hand, the linkage error is called *false negative* if a true match is deemed a non-link by the record linkage procedure. In the context of finite population sampling, Neter *et al.* (1965) observed that a relatively small amount of linkage errors could lead to a substantial bias in estimating the relationship between response errors and true values. If ignored, analysis of linked data could yield misleading results in a scientific study. Hence, the importance of accounting for linkage errors in statistical analysis cannot be overemphasised. To account for linkage errors in the analysis of linked data, additional data that inform about the two error rates are clearly needed.

The information on the quality of record linkage procedure may come from a new sample of the linked data. The main purpose of this sample is to determine the true matching status of the linked records in the sample thorough clerical review. Such a sample may be available as a part of evaluation of record linkage methodology. If the linked data as well as the evaluation sample are available to researchers, there is a scope for correcting bias in the secondary statistical analysis. Neter *et al.* (1965) discussed this secondary analysis using an audit sample. In the context of understanding the effects of low-level radiation exposure on cancer death rate, Lahiri (1996) and Krewski *et al.* (2001) suggested analysis of the Cox proportional hazard model using information contained in a sample to correct for linkage error biases. More recently, following Neter *et al.* (1965), Chambers (2009) put forward a variety of methods for different secondary data analyses that uses a sample to correct for linkage error biases. Following the work of Chambers (2009), researchers advanced the secondary data analysis of linked data in several different directions; see Chambers *et al.* (2009), Chipperfield *et al.* (2011), Kim and Chambers (2012a, 2012b, 2013), Samart and Chambers (2014), Dasylva (2014), Chipperfield and Chambers (2015), and Chambers and Kim (2016). Kandari and Lahiri (2016), following up on Lahiri (1996), suggested a theory for predicting a function of misclassified binary variables using information from a sample.

The information about linkage errors can be obtained as a part of the record linkage process. This is generally not available to the secondary data users but is accessible to primary data users. The information is contained in the common matching fields that are used to perform the record linkage operation. Some matching fields have more discriminatory power than the others in distinguishing matches from mismatches. A comparison vector, derived from the matching fields, is a vector that records the pattern of agreement and disagreement on different matching fields for a record pair. The components of a comparison vector could be binary (1 if a record pair agrees on the matching field and 0 if otherwise) or continuous (e.g. a score produced by a string comparator; see Winkler, 1990b). If an evaluation sample is not available as a part of record linkage process, it might be difficult to convince funding for a costly evaluation sample.

On the other hand, the primary users, if desired, can think of collecting appropriate summary information that measures the quality of the record linkage operation, which can be potentially used to correct for the linkage biases in statistical procedures.

Scheuren and Winkler (1991, 1993, 1997) showed how to use record linkage process information in correcting linkage bias of the ordinary least square (OLS) estimator of the regression coefficient in a standard multiple linear regression model. Their approach involves first estimating an analytical expression of linkage bias of the OLS estimator using the record linkage process information and then applying the estimated bias correction to the OLS estimator. Lahiri and Larsen (2005) obtained an exact unbiased estimator of the regression coefficient by deriving the expected value of the linked response variable when linkage errors are uncorrelated with the true response given the comparison vectors. The unbiased estimator, however, depends on parameters of the assumed linkage model, which are then estimated using a two-class mixture model (see, e.g. McCutcheon, 1987; McLachlan & Peel, 2000) on comparison vectors. The mixture model uses the so-called conditional independence assumption, introduced in the well-celebrated paper by Fellegi and Sunter (1969), to reduce the number of parameters. Although one-to-one matching was assumed under the linkage error model proposed by Lahiri and Larsen (2005), it was not enforced in the actual implementation of record linkage. Holf and Zwinderman (2012) followed up on the Lahiri–Larsen approach and showed how to extend it to link multiple files or when one-to-one matching is not desired. Estimation under the conditional independence assumption could still provide accurate decision rules even though the assumption is violated; see Thibaudeau (1993) and Winkler (1989aa, 1992, 1994). The conditional independence assumption can, however, be relaxed; see Armstrong and Mayda (1993), Thibaudeau (1993) and Larsen and Rubin (2001). Mixture models have been used to model the data arising from comparing records in two files; see Winkler (1988, 1994, 1995a), Jaro (1989, 1995) and Larsen and Rubin (2001).

Bayesian approaches to record linkage have been suggested by Larsen (1999a, 2002), Fortini *et al*. (2000, 2002), McGlinchy (2004), Tancredi and Liseo (2011, 2015), Steorts *et al*. (2017) and Tancredi *et al.* (2017). A new procedure that explicitly uses the one-to-one matching assumption and allows parameter values to vary by blocks, subsets of the data being linked, is likely to benefit from a Bayesian approach. This is because of the relatively small sample sizes within blocks and the difficulty of calculating expectations under complex restrictions on unobserved data. A Bayesian or an empirical best prediction approach can take advantage of borrowing strength across sites and deal with expectations via simulation. Bayesian record linkage alternatives can be developed to make use of prior information and experience and to formally deal with one-to-one matching.

In Section 2, we extend Lahiri and Larsen (2005) to a general model that accounts for linkage errors in the analysis of a wide range of variables – discrete and continuous. We provide a general class of estimating equations that can generate different estimators corrected for linkage bias using the record linkage process information rather than information from a new sample. In Section 3, we give a brief review of the two-class mixture model built on comparison vectors, which is used to estimate the probability of a record pair being a true match and designate the record pairs into links and non-links. In Section 4, we propose a jackknife method to estimate bias, variance and mean squared error, which incorporates variability due to linkage errors. In Section 5, we devise a Monte Carlo simulation study to compare the performance of different estimators. In Section 6, we offer a few concluding remarks and scope for future work in this area.

2 Analysis of General Model with Linked Data

Suppose we are interested in estimating a parameter or a vector of parameters associated with the conditional distribution of a scalar random variable y given a $p \times 1$ vector of auxiliary

variables x. File F_x of size N contains observations on x. Let x_i denote the value of x for unit i in F_x. We will treat x_i as the value x for the correct unit i. In other words, x-observations are correctly aligned to the units of F_x. File F_y of size $n \leq N$ contains observations on y for a subset of F_x. Let y_i be the true value of y for correct unit i, which corresponds to x_i. But y_i is not observed in any of the two files. Instead, y observations for a subset of F_x of size n are available in F_y. Let z_i be the value of y for the unit i in F_y. Because the units in F_y are not in the same order as the units in F_x, z_i may or may not be the value of y for unit i of F_x, but z_i takes one of the values from the set $\{y_1, \ldots, y_N\}$. Define $y = (y_1, \ldots, y_N)'$, $X' = (x_1, \ldots, x_N)$ and $z = (z_1, \ldots, z_n)'$, then z is a permutation of y. Let l be a binary variable specifying the true matching status of a record pair; that is, $l_{ij} = 1$ if the record pair (i, j) is a match, otherwise $l_{ij} = 0$, for any $i \in F_y$ and $j \in F_x$. Then, the randomness of the linkage process can be characterised by the following linkage error model: $z = Ly$, where $L = (l_{ij})$ is an $n \times N$ random permutation matrix.

Han (2018) classified the linkage error mechanisms into three categories: linkage completely at random (LCAR), linkage at random (LAR), and linkage not at random (LNAR). In this paper, we assume that the linkage mechanism is at random (LAR). To be specific, the conditional probability of L given y comparison vectors γ (to be defined in Section 3) and X depends on γ and X but not on y. Thus,

$$P(l_{ij} = 1|y, \gamma, X) = q_{ij}(\psi), \;\; i = 1, \ldots, n, \;\; j = 1, \ldots, N. \tag{1}$$

Here, $q_{ij}(\psi) \equiv q_{ij}(\psi; \gamma, X)$ represents the matching probability of record pair (i, j), which could depend on γ and/or X. In this paper, we will view ψ as parameters of a mixture model assumed on comparison vectors, which are derived from a set of matching fields that contain information about the matching status. We postpone the discussions on matching fields, comparison vectors, mixture model on the comparison vector and estimation of ψ until Section 3.

To illustrate our methodology, we assume

$$E(y|X) = \mu(\beta; X) \equiv \mu(\beta). \tag{2}$$

Here, β is a $p \times 1$ vector of unknown parameters, and $\mu(\beta) = (\mu(\beta; x_1), \ldots, \mu(\beta; x_N))'$ is an $N \times 1$ vector, where the functional form of $\mu(\cdot)$ is known.

Note that, under LAR, $E(z|y, \gamma, X) = Q(\psi)y$, where $Q(\psi) = (q_{ij}(\psi))$ is an $n \times N$ matrix of matching probabilities so that

$$E(z|\gamma, X) = Q(\psi)\mu(\beta). \tag{3}$$

When ψ is known, we can focus on the following class of system of p estimating equations:

$$H(\beta, \psi)\,[z - Q(\psi)\mu(\beta)] = 0, \tag{4}$$

where $H(\beta, \psi) \equiv H(\beta, \psi; X, \gamma)$ is a given $p \times n$ matrix. In order to simplify the methodology, one can replace the matrix Q in (4) by its simplified versions $Q^{\boldsymbol{M}}$ or $Q^{\boldsymbol{M}2}$. Here, $Q^{\boldsymbol{M}}$ ($Q^{\boldsymbol{M}2}$) is obtained by setting all entries in each row of Q to zeros except for the largest (two largest). In this way, $E(z_i|\gamma, X)$ is truncated to utilise the most likely one or two links. Suppose that q_{ij} and $q_{ij'}$ are the largest and the second largest among $\{q_{it} : t = 1, \ldots, N\}$, respectively, $i = 1, \ldots, n$. Then, $E(z_i|\gamma, X) = q_{ij}\mu_j$ when $Q^{\boldsymbol{M}}$ is used, and $E(z_i|\gamma, X) = q_{ij}\mu_j + q_{ij'}\mu_{j'}$ when $Q^{\boldsymbol{M}2}$ is used. The same idea can be found in Scheuren and Winkler (1993) and Lahiri and Larsen (2005). It can reduce the computational burden without losing much accuracy and

is especially useful to simplify the expression of $Var(z|\gamma, X)$ when the variance–covariance matrix is considered; see Han (2018).

For known ψ, let $\hat{\beta}_F(\psi)$ denote an estimator of β obtained as a solution to (4) for a given choice of $H(\beta, \psi)$. The corresponding estimator of β when Q is replaced by Q^M (or Q^{M2}) in (4) is denoted by $\hat{\beta}_M(\psi)$ (or $\hat{\beta}_{M2}(\psi)$). When ψ is unknown, one can estimate β by $\hat{\beta}_F(\hat{\psi})$, where $\hat{\psi}$ is a consistent estimator of ψ. In this paper, we view ψ as parameters of a mixture model and estimate it using the expectation maximisation (EM) algorithm; see Section 3 for details. The corresponding estimator of β when Q is replaced by Q^M (or Q^{M2}) in (4) is denoted by $\hat{\beta}_M(\hat{\psi})$ (or $\hat{\beta}_{M2}(\hat{\psi})$).

In case of blocking, if the linkages are assumed to occur only within blocks, the estimating Equation (4) can be simplified. Suppose that there are B blocks. For the bth block, define $z_b = (z_{b1}, \ldots, z_{bn_b})'$ as an $n_b \times 1$ vector of observed response, $\mu_b(\beta) = \left(\mu_{b1}(\beta), \ldots, \mu_{bN_b}(\beta)\right)'$ as an $N_b \times 1$ vector of conditional means, and $Q_b(\psi) = \left(q_{ij}^b(\psi)\right)$ as an $n_b \times N_b$ matrix of matching probabilities, where $b = 1, \ldots, B$. Thus, $z' = (z_1', \ldots, z_B')$, $\mu' = (\mu_1', \ldots, \mu_B')$, and $Q(\psi) = \text{block-diag}\,(Q_1(\psi), \ldots, Q_B(\psi))$. Then (4) can be written as

$$\sum_{b=1}^{B} H_b(\beta, \psi)\,[z_b - Q_b(\psi)\mu_b(\beta)] = 0, \tag{5}$$

where $H_b(\beta, \psi)$ is a given $p \times n_b$ matrix.

Remark 1. *As a special case, consider the linear model: $E(y|X) = X\beta$. In the absence of linkage errors, the OLS estimator is obtained from (4) when $H(\beta, \psi) \equiv H(\psi) = X'Q'$ because $Q = [I_n \;\; 0]$, where I_n is the identity matrix of order n and 0 is an $n \times N$ matrix of zeros. The choices $H(\beta, \psi) \equiv H(\psi) = X'Q'(\psi)$ in (4) and $H_b(\beta, \psi) \equiv H_b(\psi) = X_b'Q_b'(\psi)$ in (5) yield the estimator proposed by Lahiri and Larsen (2005).*

Remark 2. *Our proposed method is different from the one proposed by Chambers (2009) in three different aspects:*

(1) Assumptions on data availability: The method proposed by Chambers (2009) focused on secondary data analysis. The researchers can only have access to the linked data, which contain nothing but designated links. Chambers (2009) assumed that the linked data were generated by linking two files (say, F_y and F_x) of the same size, which respectively contain the observations on y and x for the same population, without duplicate. It was also assumed that the resulting linkage was complete and one-to-one. However, these assumptions are likely to be violated in practice. For example, the record linkage techniques may be used to merge survey and administrative data sets in order to increase the number of variables for the sampled units. Also, the problem of one-to-many or many-to-one linkage commonly exists in most record linkage processes. For example, any record pair with its likelihood ratio score above the upper threshold would be designated as a link for the well-known method proposed by Fellegi and Sunter (1969). It is possible that one record in one file is linked to two different records in the other file. In contrast, our proposed method is for primary data analysis that starts with the original data from the files to be linked. The observed data include the observations on y and w in F_y and on x and w in F_x, where w represents a vector of matching fields. Also, the sizes of F_y and F_x are allowed to be different, and one-to-one linkage is not required.

(2) Assumptions on the linkage error model: Chambers (2009) proposed an exchangeable linkage error model, which is built on the assumption that the probability of a designated link (or non-link) being a true match is the same within each block. These probabilities depend only on the block-specific parameters $\{\lambda_b, b = 1, \ldots, B\}$. To be specific, $q_{ij}^b = \lambda_b$ if the record pair (i, j) is designated as a link, and $q_{ij}^b = (1 - \lambda_b)/(N_b - 1)$, otherwise. It implies that the linkage mechanism is assumed to be LCAR. However, our linkage error model is built on the assumption that the probability of a record pair being a true match depends on its corresponding comparison vector; that is, $q_{ij}^b = P(l_{ij} = 1|\gamma_{ij}^b; \psi)$, which varies across record pairs. Hence, our proposed model is built on the LAR assumption, rather than LCAR.
(3) Dependence on training data: In practice, λ_b and ψ are unknown and need to be estimated. Estimation of λ_b requires a clerically reviewed training sample of the linkage data for each block. It can be simply estimated by the sample proportion of the correctly linked record pairs among all the designated links in block b, $b = 1, \ldots, B$. The estimate of the block-specific parameter λ_b can be unreliable if there are not enough samples in block b, which occurs often because smaller blocks are usually preferred in order to reduce the computational burden. In contrast, training data is not required by our proposed method. The maximum likelihood estimate of the global parameter ψ is approximated using the EM algorithm. In this way, ψ is estimated by using values of the comparison vector for all record pairs.

Generally speaking, the difference between the two approaches can be attributed to the fact that Chambers (2009) focused on secondary data analysis while we focus on primary data analysis. In contrast, the ability to access the original data and the detailed information about record linkage enables us to build a more sophisticated model to capture the characteristic of linkage errors.

Remark 3. *Although our method is designed for primary data analysis, it can be used for secondary analysis under certain conditions. It is not rare that record linkage and statistical analysis are separately performed by two different groups of persons, linkers and analysts. Our proposed method requires linkers to provide the following information to the analysts: (1) the observations on y and x from F_y and F_x for each block in the same order as they were in the original file and (2) the matrix $Q_b(\hat{\psi})$ for each block. Note that the linkers have to use the mixture-model-based record linkage approach to estimate the linkage probabilities. Under this case, point estimation is quite straightforward. However, it is not possible to estimate variance using the standard jackknife method introduced in Section 4, if the linkers do not provide the replicate estimates $\hat{\psi}_{-b}$. We could simplify the jackknife method by using the full estimate $\hat{\psi}$ instead of the replicate estimate $\hat{\psi}_{-b}$ when estimating $\hat{\beta}_{-b}$ using the system of estimating equations shown in (7). In this case, the variance would be underestimated because it does not take account of the uncertainty of $\hat{\psi}$. However, the simulation result in Han (2018) shows that the difference of the variance estimates obtained from the standard and simplified jackknife methods is negligible.*

Remark 4. *Researchers (e.g. Chambers, 2009) considered estimator of β as a solution to the system of optimal estimating equations derived under an exchangeable linkage model. Evidently, such an optimal estimator will not be optimal under a more complex linkage model such as the one suggested here needed to incorporate information from the record linkage process. It is possible to derive a system of optimal $H_b(\beta, \psi)$ in (5) by assuming a covariance structure for y. However, expression for such a system of optimal estimating equations will be complex, and the resulting estimator will be optimal only if $H_b(\beta, \psi)$ is known, which is not the case in practice.*

Remark 5. *Consider the situation when F_x is the sampling frame for a finite population sampling and F_y is a probability sample drawn from F_x. In this situation, let w_i denote the survey-weight associated with z_i in File F_y. We can carry out the survey-weighted analysis in a straightforward way by simply introducing weights w_i in $H(\beta, \psi)$ in (4) or $H_b(\beta, \psi)$ in (5). For example, the survey-weighted version of Lahiri and Larsen (2005) is obtained for the choices $H(\beta, \psi) \equiv H(\psi) = X'Q'(\psi)W$ in (4) and $H_b(\beta, \psi) \equiv H_b(\psi) = X_b'Q_b'(\psi)W_b$ in (5), where $W = \text{block-diag}(W_1, \ldots, W_B)$ and W_b is a $n_b \times n_b$ matrix of survey wights $(b = 1, \ldots, B)$. For the use of estimating equations in survey statistics, readers are referred to Binder (1983), Särndal et al. (1992, pp. 494–500), Rao (2002) and others.*

3 Mixture Model for Record Linkage

Comparison vectors are used to record patterns of agreement and disagreement on matching fields between records from two files. Let K be the number of matching fields used for comparison. The comparison vector for a record pair can be simply constructed based on whether the record pair agrees on a specific matching field or not. The comparison vector is a vector of order K and can be denoted by $\gamma = (\gamma_1, \ldots, \gamma_K)$. Define $\gamma_k = 1$ if the record pair agrees on matching field k and $\gamma_k = 0$ otherwise. For example, all the possible outcomes of a comparison vector for a record pair from two data sets with $K = 3$ matching fields are (0,0,0), (0,0,1), (0,1,0), (1,0,0), (0,1,1), (1,0,1), (1,1,0) and (1,1,1).

The set of all comparison vectors can be partitioned into two classes. One class represents the group of matches $\{(i, j) : l_{ij} = 1, i = 1, \ldots, n, j = 1, \ldots, N\}$, and the other one represents the group of mismatches $\{(i, j) : l_{ij} = 0, i = 1, \ldots, n, j = 1, \ldots, N\}$. Patterns of agreement and disagreement on matching fields have different distributions among matches and mismatches. It is assumed that comparison vector γ follows a two-class mixture model. The probability mass function of comparison vector γ can be written as $P(\gamma) = \pi P(\gamma|l = 1) + (1-\pi)P(\gamma|l = 0)$. Here, π is probability of a record pair being a match, and $P(\gamma|l = 1)$ $(P(\gamma|l = 0))$ is the probability of observing γ in group of matches (mismatches). Conditional independence of the matching fields is assumed. That is, agreement on matching fields for a pair of records is assumed to be independent within matches and mismatches. Then,

$$P(\gamma|l = 1) = \prod_{k=1}^{K} m_k^{\gamma_k}(1 - m_k)^{(1-\gamma_k)}, \quad P(\gamma|l = 0) = \prod_{k=1}^{K} u_k^{\gamma_k}(1 - u_k)^{(1-\gamma_k)},$$

where $m_k = P(\gamma_k = 1|l = 1)$ and $u_k = P(\gamma_k = 1|l = 0)$ are the probabilities of agreement on matching fields k among matches and mismatches, respectively.

The parameters in the mixture model can be denoted by $\psi = \{\pi, m_k, u_k, k = 1, \ldots, K\}$. The maximum likelihood estimator of ψ can be obtained by using the EM (Dempster *et al.*, 1977) and the expectation conditional maximisation (Meng & Rubin, 1993) algorithms.

The probability of a record pair (i, j) with $i \in F_y$ and $j \in F_x$ being a match can be estimated using Bayes' Theorem:

$$q_{ij}(\psi) = P(l_{ij} = 1|\gamma_{ij}) = \frac{\pi P(\gamma_{ij}|l_{ij} = 1)}{\pi P(\gamma_{ij}|l_{ij} = 1) + (1-\pi)P(\gamma_{ij}|l_{ij} = 0)},$$

where γ_{ij} is the value of comparison vector γ for record pair (i, j).

As described in Larsen and Rubin (2001), the estimated probabilities can be used to partition the record pairs into designated links and non-links and to estimate error rates. Jaro (1989),

Armstrong and Mayda (1993), Winkler (1993, 1994, 1995a), Thibaudeau (1993), Larsen (1996), and others used mixture models in record linkage problems.

4 Variance Estimation

As mentioned previously, when ψ is known, estimate of β can be obtained by solving the following system of estimating equations. That is,

$$\hat{\beta}(\psi) : \sum_{b=1}^{B} H_b(\beta, \psi)\,[z_b - Q_b(\psi)\mu_b(\beta)] = 0. \tag{6}$$

In order to estimate the bias, covariance matrix and mean squared error of an estimate $\hat{\beta}(\psi)$, the unified jackknife theory proposed by Jiang *et al.* (2002), henceforth referred to as JLW, can be used. The jackknife replicate b is obtained by deleting data from block b in both files F_x and F_y, and the delete-b estimates of β, $\hat{\beta}_{-b}(\psi)$, are the solutions of

$$\hat{\beta}_{-b}(\psi) : \sum_{b' \neq b}^{B} H_{b'}(\beta, \psi)\,[z_{b'} - Q_{b'}(\psi)\mu_{b'}(\beta)] = 0, \tag{7}$$

for $b = 1, \ldots, B$. When ψ is known, the JLW jackknife estimate of bias, covariance matrix and mean squared error of $\hat{\beta}(\psi)$ are given by

$$\begin{aligned}
\text{bias}_J\left(\hat{\beta}(\psi)\right) &= (B-1)\left(\bar{\hat{\beta}}(\psi) - \hat{\beta}(\psi)\right), \\
\text{var}_J\left(\hat{\beta}(\psi)\right) &= \frac{B-1}{B}\sum_{b=1}^{B}\left(\hat{\beta}_{-b}(\psi) - \bar{\hat{\beta}}(\psi)\right)\left(\hat{\beta}_{-b}(\psi) - \bar{\hat{\beta}}(\psi)\right)', \\
\text{mse}_J\left(\hat{\beta}(\psi)\right) &= \frac{B-1}{B}\sum_{b=1}^{B}\left(\hat{\beta}_{-b}(\psi) - \hat{\beta}(\psi)\right)\left(\hat{\beta}_{-b}(\psi) - \hat{\beta}(\psi)\right)',
\end{aligned}$$

where $\bar{\hat{\beta}}(\psi)$ is the average of the replicated estimates $\hat{\beta}_{-b}(\psi)$, $b = 1, \ldots, B$.

In practice, however, ψ is unknown. The maximum likelihood estimate of ψ, $\hat{\psi}$, can be obtained using the EM algorithm. It can also be treated as a solution of an estimating equation derived from the maximum log-likelihood function based on the distribution of comparison vector γ. In order to account for uncertainty of $\hat{\psi}$ in (6) and (7), ψ can be replaced by $\hat{\psi}$ and $\hat{\psi}_{-b}$, respectively, where $\hat{\psi}_{-b}$ is the delete-b estimate of ψ by removing values of comparison vectors in block b, $b = 1, \ldots, B$. Then the bias, covariance matrix and mean squared error of $\hat{\beta}(\hat{\psi})$ can be estimated. The properties of $\hat{\beta}(\hat{\psi})$ are expected to be similar to those of $\hat{\beta}(\psi)$ if $\hat{\psi}$ is assumed to be independent of y, that is, the distribution of the matching variables (e.g. last name, phone number) is assumed to be independent of the response variable y (e.g. income) and hence of z. This is true in many applications. The bias, variance and mean squared error of any smooth function of β can be proposed in a straightforward way. For large B, under regularity conditions, asymptotic properties of $\hat{\beta}(\hat{\psi})$ and the jackknife estimators proposed in this section can be obtained from the unified theory on jackknife given in Jiang *et al*. (2002).

Table 1. *Estimating equations used for four different estimators of regression coefficient β in simple logistic models.*

Estimators	Estimating equations
$\hat{\beta}_N$	$\sum_{b=1}^{B} X_b'\{y_b^\star - \mu_b(\beta)\} = 0$
$\hat{\beta}_F$	$\sum_{b=1}^{B} X_b' Q_b'\{z_b - Q_b \mu_b(\beta)\} = 0$
$\hat{\beta}_M$	$\sum_{b=1}^{B} X_b' Q_b^{'M}\{z_b - Q_b^M \mu_b(\beta)\} = 0$
$\hat{\beta}_{M2}$	$\sum_{b=1}^{B} X_b' Q_b^{'M2}\{z_b - Q_b^{M2} \mu_b(\beta)\} = 0$

Note: $y_b^\star$ is the vector of y values that are linked to X_b, and $\mu_b(\beta) = (u_{b1}(\beta), \ldots, u_{bn_b}(\beta))'$, where $u_{bi}(\beta) = \exp(x_{bi}\beta)/[1 + \exp(x_{bi}\beta)]$.

Remark 6. *Consider the case of finite population sampling described in Remark 5. The jackknife covariance estimator for $\hat{\beta}$ proposed previously will underestimate the true covariance matrix because it does not incorporate the sampling variability. Let* (E_L, Var_L) *and* (E_S, Var_S) *denote (expectation, variance) with respect to the randomness introduced by the linkage errors and sampling errors, respectively. Note that the total covariance matrix of $\hat{\beta}(\hat{\psi})$ is given by*

$$Var[\hat{\beta}(\hat{\psi})] = \mathrm{E}_L\left[Var_S(\hat{\beta}(\hat{\psi})\right] + Var_L\left[\mathrm{E}_S(\hat{\beta}(\hat{\psi})\right]. \tag{8}$$

Let var_S be any standard estimator of $Var_S\left(\hat{\beta}(\hat{\psi}\right)$ proposed in the survey statistics literature [e.g. Binder, 1983; Särndal et al., (1992, pp. 494–500); Rao, 2002]. Then, (8) motivates us to propose $var_S\left(\hat{\beta}(\hat{\psi})\right) + var_J\left(\hat{\beta}(\psi)\right)$ as an estimator of $Var\left(\hat{\beta}(\hat{\psi})\right)$.

5 A Monte Carlo Simulation Study

In this section, we design a simulation study to compare finite sample performances of different estimators of the regression coefficient β in a simple logistic model in the presence of linkage errors. Four different estimators are evaluated: naive estimator $\hat{\beta}_N = \hat{\beta}_N(\hat{\psi})$ that ignores linkage errors, proposed estimator $\hat{\beta}_F = \hat{\beta}_F(\hat{\psi})$ that incorporates linkage errors and two of its computational simpler versions $\hat{\beta}_M = \hat{\beta}_M(\hat{\psi})$ and $\hat{\beta}_{M2} = \hat{\beta}_{M2}(\hat{\psi})$. These estimators can be derived by solving a set of corresponding estimating equations (see Table 1), where Q_b^M and Q_b^{M2} are simplified versions of the matrix $Q_b = (q_{ij}^b)$, $b = 1, \ldots B$. All the entries except for the largest one are set to zero on each row in Q_b^M, while all the entries except the first two largest are set to zero on each row in Q_b^{M2}.

In the simulation, we consider the case where F_x and F_y are of the same size within each block, that is, $N_b = n_b$. However, we allow block sizes to vary across blocks. We assume that linkage errors only exist within blocks, but not across blocks. The conditional independence assumption is also made. Two sets of simulation conditions are considered in order to compare the performances of different estimators under different levels of linkage errors. We consider $B = 100$ blocks and $R = 100$ independent simulation replications under each simulation condition.

5.1 Simulation Conditions

The number of records in each block b, n_b, across different simulation replications varies from 10 to 40 in Case 1, and from 20 to 40 in Case 2. The number of observations in each file is $N = \sum_{b=1}^{B} n_b$, and there are $\sum_{b=1}^{B} n_b^2$ potential links in total with n_b^2 potential links in block

Table 2. *Simulation conditions for Case 1 and Case 2 under the equal scenario.*

Symbol	Case 1		Case 2	
	Lower limit	Upper limit	Lower limit	Upper limit
B	100		100	
R	100		100	
X	$X \sim N(0,1)$		$X \sim N(0,1)$	
β	0	1	0	1
n_b	10	40	20	40
K	8	12	6	10
m_k	0.55	0.95	0.55	0.85
u_k	0.10	0.50	0.20	0.50

b ranging from 100 to 1600 in Case 1 and from 400 to 1600 in Case 2. The number of matching fields, K, across different simulation replications varies between 8 and 10 in Case 1, and between 6 and 10 in Case 2. Across different replications, probability of agreement on matching field k among matches, m_k, and among mismatches, u_k, takes values from interval (0.55, 0.95) and (0.10, 0.50), respectively, in Case 1, whereas they take values from interval (0.55, 0.85) and (0.20, 0.50) in Case 2. In general, linkage errors are less likely to occur under Case 1 than under Case 2, because it has smaller area sizes, more matching fields and larger probabilities of agreement among matches and smaller probabilities of agreement among mismatches. Hence, we would expect to have better estimates in Case 1 than those in Case 2. A summary of simulation conditions for Case 1 and Case 2 is shown in Table 2.

5.2 Simulation Steps

(1) *Data Generation:* N values of x in F_x and y in F_y are generated based on simulated regression coefficient β. A comparison vector γ can be generated for each record pair based on their true matching status, the number of matching fields K, probabilities of agreement on matching fields among matches $\{m_k, k = 1, \ldots, K\}$, and among mismatches $\{u_k, k = 1, \ldots, K\}$. Note that only the records within the same blocks are compared. The observed data for statistical analysis include observations of x in F_x, y in F_y (which is denoted by z), and comparison vector γ.

(2) *Record Linkage:* A two-class mixture model is fitted to observed comparison vectors γ using the EM algorithm. All the parameters $\{\pi, m_k, u_k, k = 1, \ldots, K\}$ in the mixture model are estimated. The probability of a record pair (i, j) within block b being a link, q_{ij}^b, is the same as the probability of its corresponding vector γ_{ij}^b belonging to class of matches. It can be estimated by applying Bayes' Theorem and can be used to partition the record pairs into links and non-links.

(3) *Parameter Estimation:* In order to obtain $\hat{\beta}_N$, it is essential to determine designated links. The designated link to a record i within block b in F_y is a record j within the same block in F_x whose corresponding linkage probability q_{ij}^b is the largest among $\{q_{it}^b, t = 1, \ldots, n_b\}$. In our case, it is possible that a record j in F_x is linked to two or more records in F_y because one-to-one assignment is not enforced. For our proposed estimators, the matrices Q_b, Q_b^M and Q_b^{M2} for block b need to be constructed based on the estimated probabilities $\{q_{ij}^b, i = 1, \ldots, n_b; j = 1, \ldots, n_b\}$. Then the four estimators $\hat{\beta}_N$, $\hat{\beta}_F$, $\hat{\beta}_M$ and $\hat{\beta}_{M2}$ can be estimated by solving the estimating equations shown in Table 1.

(4) *Variance Estimation:* Jackknife is used to estimate bias, variance matrix and mean squared error of each estimate of β. A jackknife replicates is generated by leaving out data of one block from the two files at a time. Hence, there are $B = 100$ jackknife replicates in total. For each jackknife replicate, Step 2 and Step 3 are performed and estimates of β are re-evaluated.

The jackknife estimates of the bias, variance and mean squared error of an estimator can be obtained by aggregating B replicate estimates of β. A 95% confidence interval can be obtained for each estimate of β.

Steps (1) to (4) are performed for $R = 100$ simulation runs.

5.3 Performance Evaluation

The performance of the four estimators can be evaluated by the average absolute deviation (AAD) and average squared deviation (ASD) over all the simulation runs. The formulas for AAD and ASD of an estimator $\hat{\beta}$ are shown in the following.

$$AAD(\hat{\beta}) = \frac{\sum_{r=1}^{R} |\hat{\beta}^{(r)} - \beta|}{R}, \; ASD(\hat{\beta}) = \frac{\sum_{r=1}^{R} (\hat{\beta}^{(r)} - \beta)^2}{R},$$

where $\hat{\beta}^{(r)}$ is value of $\hat{\beta}$ calculated based on simulation run r. We can also measure improvement of an estimator $\hat{\beta}$ over $\hat{\beta}_N$ with respect to AAD and ASD by relative per cent improvement (RPI). The formulas are shown in the following:

$$RPI_{AAD}(\hat{\beta}) = \frac{AAD(\hat{\beta}_N) - AAD(\hat{\beta})}{AAD(\hat{\beta}_N)} \times 100\%, \; RPI_{ASD}(\hat{\beta}) = \frac{ASD(\hat{\beta}_N) - ASD(\hat{\beta})}{ASD(\hat{\beta}_N)} \times 100\%.$$

A 95% confidence interval for β is obtained as

$$\hat{\beta} \pm 1.96\sqrt{\text{var}_J(\hat{\beta})},$$

where $\text{var}_J(\hat{\beta})$ is the jackknife variance estimate. The coverage rate is defined as the percentage of times that the previous 95% confidence interval covers the true value of β among the R simulation runs.

The Monte Carlo estimates of the bias, variance and mean squared error of an estimator $\hat{\beta}$ are respectively given by

$$\text{bias}(\hat{\beta}) = \frac{1}{R}\sum_{r=1}^{R} \hat{\beta}^{(r)} - \beta, \; \text{var}(\hat{\beta}) = \frac{1}{R-1}\sum_{r=1}^{R} \left(\hat{\beta}^{(r)} - \bar{\hat{\beta}}\right)^2, \; \text{mse}(\hat{\beta}) = \frac{1}{R}\sum_{r=1}^{R} (\hat{\beta}^{(r)} - \beta)^2,$$

where $\bar{\hat{\beta}} = \frac{1}{R}\sum_{r=1}^{R} \hat{\beta}^{(r)}$. The Monte Carlo estimates of the bias, variance and mean squared error of the relative deviation $(\hat{\beta} - \beta)/\beta$ are denoted by $\text{R.bias}(\hat{\beta})$, $\text{R.var}(\hat{\beta})$ and $\text{R.mse}(\hat{\beta})$, respectively. They are given by

$$\text{R.bias}(\hat{\beta}) = \frac{\text{bias}(\hat{\beta})}{\beta}, \; \text{R.var}(\hat{\beta}) = \text{var}(\hat{\beta}), \; \text{R.mse}(\hat{\beta}) = \frac{\text{mse}(\hat{\beta})}{\beta^2}.$$

To assess the accuracy of the Monte Carlo estimates, the Monte Carlo standard deviation is used. It is defined as the standard deviation of a Monte Carlo estimate estimator, that is, for a Monte Carlo estimator $\hat{\theta}$,

$$SD_{MC}(\hat{\theta}) = \sqrt{\text{var}(\hat{\theta})},$$

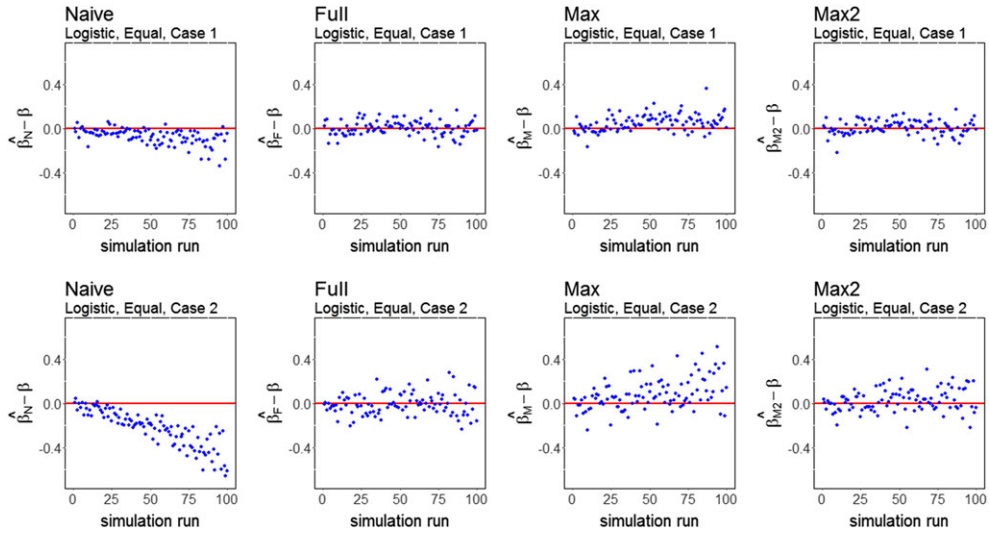

Figure 1. *Simulation results for logistic regression under Case 1 and Case 2 in the equal scenario: Scatter plots for deviations of $\hat{\beta}_N$, $\hat{\beta}_F$, $\hat{\beta}_M$ and $\hat{\beta}_{M2}$ from true values of β over 100 simulation runs. [Colour figure can be viewed at wileyonlinelibrary.com]*

where $\text{var}(\hat{\theta})$ is the estimated variance of $\hat{\theta}$. For example, the Monte Carlo standard deviations of the estimated bias and mean squared error are given by

$$SD_{MC}\left(\text{bias}(\hat{\beta})\right) = \sqrt{\frac{\text{var}(\hat{\beta})}{R}}, \quad SD_{MC}(\text{mse}(\hat{\beta})) = \sqrt{\frac{\text{var}[(\hat{\beta}-\beta)^2]}{R}}.$$

5.4 Simulation Results

In this part, we compare the performances of different estimators of regression coefficient β in a simple logistic model. In each simulation run, values of a scalar-independent variable x are randomly and independently selected from $N(0,1)$, and the corresponding values of y are given by

$$P(y_{ij}=1|x_{ij}) = \mu(\beta) = \frac{\exp(\beta x_{ij})}{1+\exp(\beta x_{ij})}, \; j=1,\ldots,n_i, \; i=1,\ldots,m,$$

where β is randomly selected from the uniform distribution $U(0,1)$.

Figure 1 displays scatter plots for deviations of different estimates from the true values of β over 100 simulation runs in Case 1 and Case 2. A red, solid and horizontal reference line with intercept 0 is plotted. If an estimator performs well, the deviations of its corresponding estimates should gather closely and evenly around the reference line. Recall that the true values of β range from 0 to 1 across the 100 simulation runs. In order to investigate whether the true values affect the performance of the four different estimators, the simulation runs are sorted in an increasing order based on the true values of β; that is, the true value of β used in the first (last) run is the smallest (largest) among all the true values for the 100 simulation runs.

Based on the result, we can clearly see that the deviations for $\hat{\beta}_N$ are mostly below the reference line and get further from the line as the true value of β increases. The phenomenon is especially obvious in Case 2, where the linkage errors are more likely to occur. It implies that $\hat{\beta}_N$ usually underestimates β, and the degree of underestimation expands as the true value of

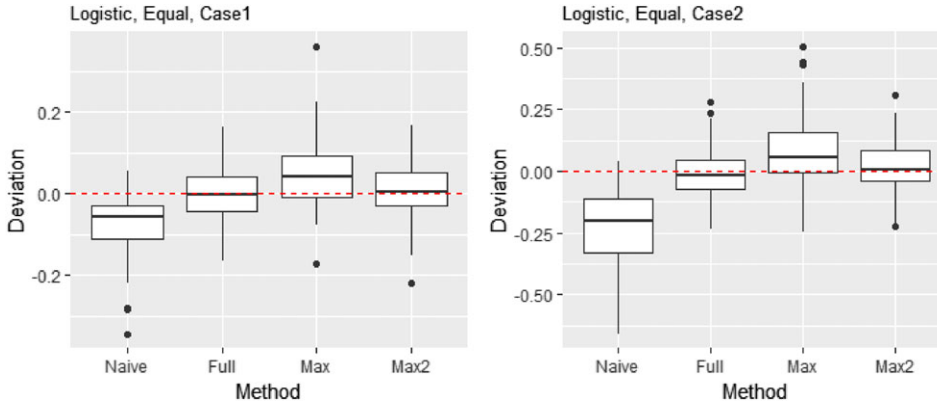

Figure 2. *Simulation results for logistic regression under Case 1 and Case 2 in the equal scenario: Box plots for deviations of $\hat{\beta}_N$, $\hat{\beta}_F$, $\hat{\beta}_M$ and $\hat{\beta}_{M2}$ from the true value of β over 100 simulation runs. [Colour figure can be viewed at wileyonlinelibrary.com]*

β increases. That is because the linkage errors weaken the correlation between y and x, which introduces a bias towards zero when estimating the regression coefficient.

In contrast, all of our proposed estimators seem to correct the linkage bias to varying extents under the two cases, with $\hat{\beta}_N$ and $\hat{\beta}_{M2}$ being the most efficient, and $\hat{\beta}_M$ being the least efficient. The performance of $\hat{\beta}_N$ and $\hat{\beta}_{M2}$ are quite consistent for the varying values of β, while $\hat{\beta}_M$ appears to overestimate β a little bit, and the degree of overestimation intensifies as β increases. However, $\hat{\beta}_M$ still performs much better than $\hat{\beta}_N$, especially in Case 2. In general, our proposed estimators perform better than $\hat{\beta}_N$ based on the visual inspection. This is probably because they take account of the linkage errors in the linked data.

We can also compare the four estimators based on their overall performance using the two box plots of their corresponding deviations across the 100 simulation runs shown in Figure 2, one for Case 1, and the other for Case 2. Note that the solid black dots shown at the end of a box plot are the outliers of the estimates. The overall performance can be measured by the median of deviations over the 100 simulation runs, which is indicated by the segment inside the rectangle box. It can also be measured by the interquartile range of deviations, which is indicated by the extent to which the central rectangle box spans. Based on the result, we can see that the medians of $\hat{\beta}_F$ and $\hat{\beta}_{M2}$ (almost) overlap with the red, dashed, horizontal reference line with intercept zero in both cases. Numerically, the medians for $\hat{\beta}_N$, $\hat{\beta}_F$, $\hat{\beta}_M$ and $\hat{\beta}_{M2}$ are -0.058, -0.002, -0.040 and 0.004 in Case 1, and -0.202, -0.015, 0.054 and 0.008 in Case 2. Besides, the interquartile ranges of $\hat{\beta}_F$ and $\hat{\beta}_{M2}$ are almost the same and are the smallest two among the four estimators in both cases. In summary, our proposed estimators performs better than $\hat{\beta}_N$ in both cases based on the overall performance of their corresponding estimates, with $\hat{\beta}_F$ and $\hat{\beta}_{M2}$ being the best two.

Figure 3 shows heatmaps of 100 absolute deviations and squared deviations of each estimates from true values of β in a simple logistic model under two cases. The darker the colour, the smaller is the absolute deviation. We can clearly see that our proposed estimators perform much better than $\hat{\beta}_N$, especially in Case 2. Table 3 displays the AAD and ASD of estimates for β, as well as the RPI over $\hat{\beta}_N$ under Case 1 and Case 2 in the equal scenario. Values of AAD and ASD are shown in black, and values of RPI are shown in blue. Under both cases, $\hat{\beta}_N$ has the largest values of AAD and ASD among the four estimators, implying that it performs the worst in estimating the regression coefficient in a simple logistic model; $\hat{\beta}_M$ performs better, but not as well as $\hat{\beta}_F$ and $\hat{\beta}_{M2}$. We can also see that values of AAD and ASD increase under Case 2

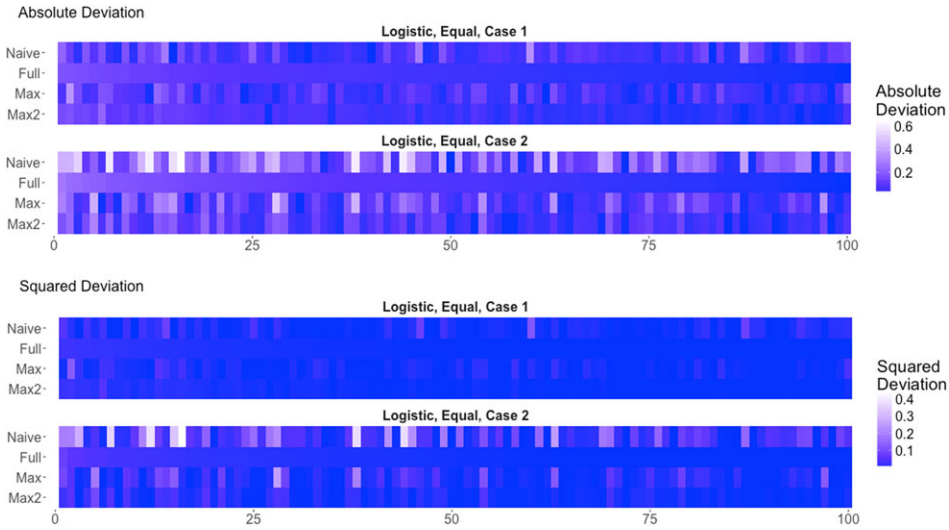

Figure 3. *Simulation results for logistic regression under Case 1 and Case 2 in the equal scenario: Heatmap for absolute deviations (top 2) and squared deviations (bottom 2) of estimators of regression coefficient β. [Colour figure can be viewed at wileyonlinelibrary.com]*

Table 3. *Simulation results for logistic regression under Case 1 and Case 2 in the equal scenario: Average absolute deviations (AAD) and average squared deviations (ASD) of $\hat{\beta}_N$, $\hat{\beta}_F$, $\hat{\beta}_M$ and $\hat{\beta}_{M2}$ of regression coefficient β. [Colour table can be viewed at wileyonlinelibrary.com]*

Estimator	Case 1		Case 2	
	AAD	ASD	AAD	ASD
$\hat{\beta}_N$	0.0811	0.0112	0.2320	0.0811
$\hat{\beta}_F$	0.0527	0.0045	0.0755	0.0098
	35.01%	59.78%	66.61%	87.90%
$\hat{\beta}_M$	0.0681	0.0080	0.1215	0.0269
	16.02%	28.60%	47.65%	66.78%
$\hat{\beta}_{M2}$	0.0517	0.0043	0.0796	0.0106
	36.24%	61.28%	65.70%	86.99%

The per cent relative improvement of the proposed estimator over naive estimator is shown in blue.

when compared with Case 1. This is as expected because Case 2 has more difficult simulation conditions (less matching fields, larger block sizes, small probabilities of agreement among matches and larger probabilities of agreement among mismatches), resulting in more linkage errors. However, our proposed estimators improved more over $\hat{\beta}_N$ in Case 2 then in Case 1, indicated by the larger values of RPI under Case 2. It shows that our proposed estimator would be especially useful when linkage errors are more likely to occur.

In order to further evaluate the performances of these four estimators, another set of 100 simulation runs for logistic regression under Case 1 in equal scenario is performed with β fixed at 0.5. Box plot of deviations and relative deviations of different estimators from the true value of β is shown in Figure 4. Table 4 gives Monte Carlo estimates of bias, relative bias, mean squared error, relative mean squared error, length and coverage of nominal 95% confidence

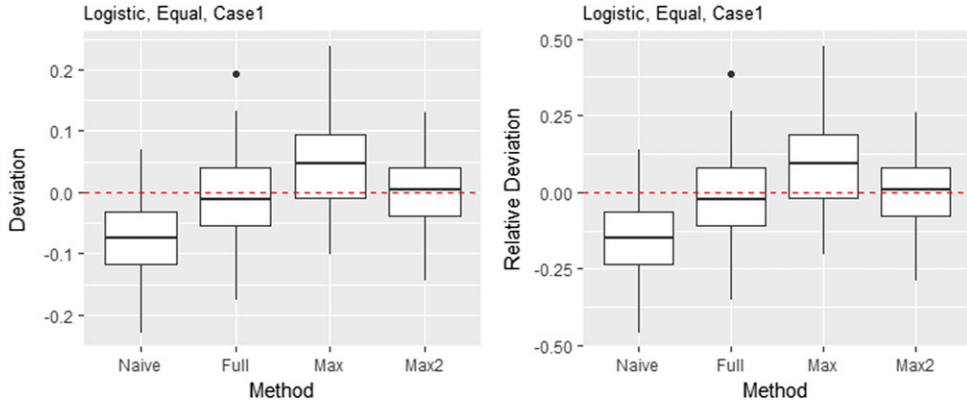

Figure 4. *Simulation results for logistic regression under Case 1 in the equal scenario: Box plots of deviations and relative deviations of $\hat{\beta}_N$, $\hat{\beta}_F$, $\hat{\beta}_M$ and $\hat{\beta}_{M2}$ from the true value of β over 100 simulation runs. The value of β is fixed at 0.5 for each simulation run. [Colour figure can be viewed at wileyonlinelibrary.com]*

Table 4. *Simulation results for logistic regression under Case 1 in the equal scenario: Monte Carlo estimates of bias, relative bias, variance, relative variance, mean squared error, relative mean squared error, length and coverage rate (C.R.) of the nominal 95% confidence intervals of β by different methods over 100 simulation runs. [Colour table can be viewed at wileyonlinelibrary.com]*

Method	Bias	R.bias	Var	R.var	MSE	R.MSE	Length	C.R.
Naive	-0.0731	-0.1462	0.0044	0.0175	0.0097	0.0387	0.1869	63%
	0.0066	0.0132			0.0011	0.0045	0.0022	0.0485
Full	-0.0096	-0.0192	0.0046	0.0185	0.0047	0.0187	0.2429	93%
	0.0068	0.0136			0.0007	0.0027	0.0048	0.0256
Max	0.0430	0.0859	0.0042	0.0167	0.0060	0.0239	0.2388	90%
	0.0065	0.0129			0.0009	0.0035	0.0051	0.0302
Max2	-0.0007	-0.0014	0.0037	0.0148	0.0037	0.0147	0.2259	95%
	0.0061	0.0122			0.0005	0.0019	0.0034	0.0219

Value of β is set to 0.5 for each simulation run. The corresponding Monte Carlo standard deviations are show in blue.

intervals of β for each method. The standard errors of these estimates are shown in blue. The negative values of bias and relative bias of $\hat{\beta}_N$ implies that it underestimates values of β, and the other three estimators correct this bias, with $\hat{\beta}_{M2}$ and $\hat{\beta}_F$ being the most efficient. The correctness of bias and relative bias also lead to the decrease of mean squared error and relative mean squared error. In terms of mean squared error and relative mean squared error, $\hat{\beta}_{M2}$ performs the best, followed closely by $\hat{\beta}_F$. We can also see that the coverage rates of confidence intervals produced by $\hat{\beta}_{M2}$ and $\hat{\beta}_F$ and their jackknife variances are very close to their desired nominal level, while those produced by $\hat{\beta}_N$ is lower than the desired nominal level.

In summary, a total of 300 simulation runs (three sets of $R = 100$ runs) are performed to investigate the performance of different estimators under various conditions. The first two sets are performed under Case 1 and Case 2, respectively, and allow the true value of β to vary across simulation runs. The third one is performed under Case 1, and the values of β are fixed at 0.5. The results are quite consistent across the three set of 100 simulations, showing that $\hat{\beta}_F$ and $\hat{\beta}_{M2}$ perform more efficiently than the others. A larger number of simulation runs can better demonstrate the validity of our proposed method. However, we found it difficult to increase

the number because the method is computationally intensive in estimating variance with the standard jackknife method introduced in Section 4. The method requires us to re-estimate ψ for 100 jackknife replicates, resulting 101 iterations (one for the full estimate, 100 for the replicate estimates) for each simulation run. In order to reduce the computational burden, we recommend to use the simplified version of the jackknife method, by using the full estimate $\hat{\psi}$ of ψ in each iteration. We realise that the simplified jackknife method would underestimate the variance because it does not take account of the uncertainty of $\hat{\psi}$. The simulation study reported in Han (2018) shows that there is not much difference in the variances estimated by the two methods, implying that the variability of $\hat{\psi}$ is negligible. This could be due to the fact that the number of comparison vectors (that is, $\sum_{b=1}^{B} N_b n_b$) is large, which is usually the case in practice. In terms of β estimation alone, the simulation is not quite computationally intensive.

6 Concluding Remarks

We present a general integrated model that incorporates linkage errors resulting from a record linkage process. The proposed methodology corrects the bias in the standard statistical analysis due to linkage errors by exploiting record linkage process data. This can lead to substantial cost reduction incurred due to taking an additional sample from the linked data required in secondary data analysis for the purpose of gaining information about the error rates in the record linkage procedure. The method proposed can be extended to the analysis of complex survey data from a finite population. Simulation results are encouraging and demonstrate the superiority of the proposed methods over the standard method that does not correct for biases due to linkage errors. Our proposed methodology does not use information from a sample that is used for secondary data analysis for bias reduction. We plan to extend the proposed method to incorporate the evaluation sample data, if available, to improve our methodology.

As mentioned in Section 2, the linkage mechanism can be classified into three different categories based on the conditional distribution of the matching status indicator l given the data y, γ, X. The method proposed by Chambers (2009) is an example of applying LCAR, because the probability $P(l_{ij}|y, \gamma, X)$ does not depend on y, γ, X. Scheuren and Winkler (1997), Lahiri and Larsen (2005) and our proposed method are several examples of the application of LAR. Up to date, little research has been performed under the assumption of LNAR, where the linkage mechanism depends on the response variable y. It is possible that an estimator derived under the LCAR or LAR assumption would be actually biased under the LNAR assumption. This is a good research area to pursue in the future.

Chambers and Tzavidis (2006), Sinha and Rao (2009), Chambers *et al.* (2014), Gershunskaya and Lahiri (2018) and others proposed small area estimation methods that are relatively insensitive to a model misspecification or to the presence of outliers. Recently, Fabrizi *et al.* (2018) discussed extensions of small area estimation methods in the presence of outliers. However, their proposed method was built under the exchangeable linkage error model. It is of interest to extend our methodology in order to make the methodology robust against model misspecification and outliers. We will leave that for future research.

Acknowledgements

The authors thank an anonymous referee and editors for a few constructive suggestions that led to improvement of an earlier version of the article. The research of the second author was supported in part by the National Science Foundation Grant Number SES-1758808.

References

Alvey, M. & Jamerson, N. (1997). Record linkage techniques – 1997. In *Proceedings of an International Workshop and Exposition*, Arlington, VA, pp. 20–21.

Armstrong, J. B. & Mayda, J. E. (1993). Model-based estimaton of record linkage error rates. *Surv. Methodol.*, **19**, 137–147.

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *Int. Stat. Rev.*, **51**, 279–292.

Chambers, R. (2009). *Regression analysis of probability-linked data*. Statistics New Zealand.

Chambers, R., Chandra, H., Salvati, N. & Tzavidis, N. (2014). Outlier robust small area estimation. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **76**(1), 47–69.

Chambers, R., Chipperfield, J., Davis, W. & Kovacevic, M. (2009). Inference based on estimating equations and probability-linked data, centre for statistical and survey methodology. Working Paper 18-09, University of Wollongong p36.

Chambers, R. & Kim, G. (2016). Secondary analysis of linked data. In *Methodological Developments in Data Linkage*, Eds. K. Harron, H. Goldstein & C. Dibben. Chichester: Wiley, pp. 83–108.

Chambers, R. & Tzavidis, N. (2006). M-quantile models for small area estimation. *Biometrika*, **93**, 255–268.

Chipperfield, J. O., Bishop, G. R. & Campbell, P. (2011). Maximum likelihood estimation for contingency tables and logistic regression with incorrectly linked data. *Surv. Methodol.*, **37**, 13–24.

Chipperfield, J. O. & Chambers, R. L. (2015). Using the bootstrap to account for linkage errors when analysing probabilistically linked categorical data. *J. Off. Stat.*, **31**, 397–414.

D'Orazio, M., Zio, M. D. & Scanu, M. (2006). *Statistical Matching Theory and Practice*: John Wiley.

Dasylva, A. (2014). Design-based estimation with record-linked administrative files. In *Proceedings of the 2014 International Methodology Symposium*, Ottawa, pp. 2014.

Dempster, A., Laird, N. & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B*, **39**, 1–38.

Fabrizi, E., Salvati, N. & Chambers, R. (2018). *Robust small area estimation with linked data*. presented at the small area estimation conference, Shanghai, China.

Fellegi, I. P. & Sunter, A. B. (1969). A theory for record linkage. *J. Am. Stat. Assoc.*, **64**, 1183–1210.

Fortini, M., Liseo, B., Nuccitelli, A. & Scanu, M. (2000). On Bayesian record linkage, Bayesian methods with applications to science, policy, and official Statistics. In *The Sixth World Meeting of the International Society for Bayesian Analysis*, Ed. E. I. George, pp. 155–164. Benidorm, Spain.

Fortini, M., Nuccitelli, A., Liseo, B. & Scanu, M. (2002). Modeling issues in record linkage: a Bayesian perspective. In *Proc. Amer. Statist. Assoc. Survey Research Meth. Sec*. Joint statistical meetings, At New York City.

Gershunskaya, J. & Lahiri, P. (2018). Robust empirical best small area finite population mean estimation using a mixture model. *Calcutta Stat. Assoc. Bull.*, **69**(2), 183–204.

Gill, L. E. (1997). OX-LINK: the Oxford medical record linkage system demonstration of the PC version. Record linkage techniques - 1997. In *Proc. International Workshop Exposition*. Washington, DC, pp. 15–34.

Gomatam, S. & Larsen, M. D. (2004). Record linkage and counterterrorism. *Chance*, **17**, 25–29.

Han, Y. (2018). *Statistical Inference Using Data from Multiple Files Combined through Record Linkage Diss* College Park: University of Maryland.

Herzog, T. N., Scheuren, F. J. & Winkler, W. E. (2007). *Data Quality and Record Linkage Techniques.* New York, NY: Springer.

Hof, M. H. P. & Zwinderman, A. H. (2012). Methods for analyzing data from probabilistic linkage strategies based on partially identifying variables. *Stat. Med.*, **31**, 4231–4242.

Jaro, M. A. (1989). Advances in record linkage methodology to matching the 1985 census of Tampa, Florida. *J. Am. Stat. Assoc.*, **84**, 414–420.

Jaro, M. A. (1995). Probabilistic linkage of large public health data files. *Statist. Med.*, **14**, 491–498.

Jiang, J., Lahiri, P. & Wan, S.-W. (2002). A unified jackknife theory for empirical best prediction with M-estimation. *Ann. Statist.*, **30**, 1782–1810.

Kandari, N. & Lahiri, P. (2016). Prediction of a function of misclassified binary data. *Statist. Transit. New Ser.*, **17**(3), 429–447.

Kim, G. & Chambers, R. (2012a). Regression analysis under incomplete linkage. *Comput. Stat. Data Anal.*, **56**(9), 2756–2770.

Kim, G. & Chambers, R. (2012b). Regression analysis under probabilistic multi-linkage. *Statist. Neerlandica*, **66**(1), 64–79.

Kim, G. & Chambers, R. (2013). Bias reduction for correlated linkage error. Working Papers Series, NIASRA, University of Wollongong. Wollongong.

Krewski, D., Dewanji, A., Wang, Y., Bartlett, S., Zielinkski, J. M. & Mallick, R. (2001). Regression analysis with linked data. *Surv. Methodol.*, **31**(1), 13–22.
Lahiri, P. (1996). Final Report on the Health Canada Contract No. 5208 on Record Linkage.
Lahiri, P. & Larsen, M. D. (2005). Regression analysis with linked data. *J. Am. Stat. Assoc.*, **100**(469), 222–230.
Larsen, M. D. (1999a). Multiple imputation analysis of records linkage using mixture models. In *Proc. Statist. Soc. Canada Survey Meth. Sec*, pp. 65–71. Regina, Saskatchewan, Canada.
Larsen, M. D. (2002). Comment on hierarchical Bayesian record linkage. In *Proc. Sec. Bayesian Statist. Sci.*: American Statistical Association meeting, New York City, NY, pp. 1995–2000. CDROM.
Larsen, M. D. & Rubin, D. B. (2001). Iterative automated record linkage using mixture models. *J. Am. Stat. Assoc.*, **96**, 32–34.
Livingston, E. H. & Ko, C. Y. (2005). Effect of diabetes and hypertension on obesity-related mortality. *Surgery*, **137**, 16–25.
McCutcheon, A. L. (1987). *Latent Class Analysis* Newbury Park, CA; London: Sage Publications Inc.
McGlinchy, M. (2004). A Bayesian record linkage methodology for mImputation of missing links. In *Proc. Amer. Statist. Assoc. Survey Research Meth. Sec.*: CDROM, Alexandria, VA.
McLachlan, G. & Peel, D. (2000). *Finite Mixture Models*, 1st ed. New York: Wiley-Interscience.
Meng, X. L. & Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika*, **80**, 267–278.
Neter, J., Maynes, E. & Ramanathan, R. (1965). The effect of mismatching on the measurement of response error. *J. Am. Stat. Assoc.*, **60**, 1005–1027.
Newcombe, H. B., Kennedy, J. M., Axford, S. J. & James, A. (1959). Automatic linkage of vital records. *Science*, **130**, 954–959.
Rao, J. N. K. (2002). Estimating equations for the analysis of survey data using poststratification information. *Sankhyā: The Indian J. Stat.*, **64**, 364–378.
Rässler, S. (2002). *Statistical Matching: A Frequentist Theory, Practical Applications and Alternative Bayesian Approaches,* New York: Springer.
Rauscher, G. H. & Sandler, D. P. (2005). Validating cancer histories in deceased relatives. *Epidemiology*, **16**, 262–265.
Samart, K. & Chambers, R. (2014). Linear regression with nested errors using probability-linked data. *Aust. N. Z. J. Stat.*, **56**(1), 27–46.
Särndal, C.-E., Swensson, B. & Wretman, J. (1992). *Model Assisted Survey Sampling* New York: Springer-Verlag.
Scheuren, F. J & Winkler, W. E. (1991). Regression analysis of data files that are computer matched. In *Proceedings of the Annual Research Conference, the U.S. Census Bureau*, pp. 669–687.
Scheuren, F. J. & Winkler, W. E. (1993). Regression analysis of data files that are computer matched. *Surv. Methodol.*, **19**, 39–58.
Scheuren, F. J. & Winkler, W. E. (1997). Regression analysis of data that are computer matched - Part ii. *Surv. Methodol.*, **23**(2), 157–165.
Sinha, S. K. & Rao, J. N. K. (2009). Robust small area estimation. *Can. J. Stat.*, **37**, 381–399.
Steorts, R. C., Hall, R. & Fienberg, S. E. (2017). A Bayesian approach to graphical record linkage and de-duplication. *J. Am. Stat. Assoc.*, **111**, 1660–1672.
Tancredi, A. & Liseo, B. (2011). A hierarchical Bayesian approach to record linkage and population size problems. *Ann. Appl. Stat.*, **5**, 1553–1585.
Tancredi, A. & Liseo, B. (2015). Regression analysis with linked data: Problems and solutions. *Statistica*, **75**(1), 19–35.
Tancredi, A., Steorts, R. C. & Liseo, B. (2017). A Bayesian approach for deduplication, record linkage and inference with linked data. Working Paper, MEMOTEF, Sapienza Università di Roma, Italy.
Thibaudeau, Y. (1993). The discrimination power of dependency structures in record linkage. *Surv. Methodol.*, **19**, 31–38.
Thompson, D., Kriebel, D., Quinn, M. M., Wegman, D. H. & Eisen, E. A. (2005). Occupational exposure to metaworking fluids and risk of breast cancer among female autoworkers. *Amer. J. Industrial Medicine*, **47**, 153–160.
Winkler, W. E. (1988). Using the EM algorithm for weight computation in the Fellegi-Sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 667–671.
Winkler, W. E. (1989a). Near automatic weight computation in the Fellegi-Sunter model of record linkage. In *Proceedings of the Fifth Census Bureau Annual Research Conference, Washington, DC*, pp. 145–155.
Winkler, W. E. (1990b). String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 354–359.
Winkler, W. E. (1992). Comparative analysis of record linkage decision rules, american statistical association. In *Proceedings of the Section on Survey Research Methods*, pp. 829–834.

Winkler, W. E. (1993). Improved decision rules in the Fellegi-Sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 274–279.
Winkler, W. E. (1994). Advanced methods for record linkage. In *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 467–472.
Winkler, W. E. (1995). Matching and record linkage. In *Business Survey Methods*, Eds. B. G. Cox, D. A. Binder, B. N. Chinnappa, A. Christianson, M. J. Colledge & P. S. Kott. New York: Wiley, pp. 355–384.
Winkler, W. E. (1995). Editing discrete data. In *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 108–113.

[*Received December 2017, accepted August 2018*]