2009

Regression analysis under incomplete linkage

G. Kim
University of Wollongong, gkim@uow.edu.au

R. Chambers
University of Wollongong, ray@uow.edu.au

Centre for Statistical and Survey Methodology

The University of Wollongong

Working Paper

17-09

Regression analysis under incomplete linkage

Gunky Kim and Raymond Chambers

Centre for Statistical & Survey Methodology, University of Wollongong, Wollongong NSW 2522. Phone +61 2 4221 5435, Fax +61 2 4221 4845. Email: anica@uow.edu.au

Regression analysis under incomplete linkage

Gunky Kim and Raymond Chambers
Centre for Statistical and Survey Methodology
University of Wollongong

Summary

Most probability-based methods used to link records from two distinct data sets corresponding to the same target population do not lead to perfect linkage, i.e. there are linkage errors in the merged data. Chambers (2008) describes modifications to standard methods of regression analysis that can be used with such imperfectly linked data. However, these methods assume that the linkage process is complete, i.e. all records on the two data sets are linked. This paper extends these ideas to regression analysis using data that have been incompletely linked, and in particular to the situation where one of the data sets being linked is a sample from the target population and the other is a register, i.e. it covers the entire target population.

Key words: Record matching; linkage errors; linear regression; logistic regression; estimating equations; measurement error.

1 Introduction

Methods for probabilistically linking data relating to the same target population, but stored on different data bases, have been extensively developed over the last three decades. In particular, there are now a number of software packages for computerized record linkage that can deal with the large data sets that arise in practical applications, e.g. census data (Jaro

(1989)) and population health data (Newcombe (1988)). Since a linked data file allows much more powerful analysis than the individual contributing data sets, data linkage has become an important research tool in areas such as health, business, economics and sociology, where there can be a considerable increase in the capacity for more efficient and more cost effective analysis using the information in a linked database compared with the separate component databases. To illustrate, the Census Data Enhancement project of the Australian Bureau of Statistics aims to develop a Statistical Longitudinal Census Dataset by linking data from the same individuals over a number of censuses. It is expected that this linked data set will provide a powerful tool for future research into the longitudinal dynamics of the Australian population.

Following the pioneering work of Fellegi and Sunter (1969), probabilistic matching has become a major tool in data linkage. In this case, the aim is to optimize the linkage process by minimizing the number of linked records that correspond to potentially incorrect linkages. However, it is important to note that minimizing the incidence of linkage errors does not in itself imply that there are no incorrectly linked records in the linked data set.

Neter *et al.* (1965) show that even a small amount of mismatching can result in significant response error. Scheuren and Winkler (1993), Scheuren and Winkler (1997) and Lahiri and Larsen (2005) have investigated methods for correcting the bias induced by this error in the context of linear regression analysis. Chambers (2008) extends this work to the general regression case and develops new methods to bias-correct estimated regression parameters when linkage is complete, i.e. when there are no records that are not (at least potentially) linked. However, the reality is that there are many situations where not all the records in the contributing data bases can be linked. For example, it is often the case that one contributing data base corresponds to a sample, while the other covers the entire population of interest, i.e. is a register. Furthermore, even in such cases typically not all sample records can be linked to the register. In this situation, direct application of the methods in Chambers (2008) is inappropriate. Our main contribution in this article is to extend the approach of Chambers (2008) to accommodate this situation.

1.1 Backgrounds and Assumptions

Suppose that there exist two distinct data sets $\boldsymbol{y}$ and $\boldsymbol{X}$ where each value of $\boldsymbol{y}$ depends on a value of $\boldsymbol{X}$ via a known functional form. In particular, we are interested in fitting a regression model of the form $E(\boldsymbol{y}|\boldsymbol{X}) = g(\boldsymbol{X};\boldsymbol{\theta})$, where g is known but the parameter $\boldsymbol{\theta}$ is unknown. Estimation of $\boldsymbol{\theta}$ is straightforward when correctly linked values of $\boldsymbol{y}$ and $\boldsymbol{X}$ are available. However, due to linkage errors, the values of $\boldsymbol{y}$ are not all observable. Instead, one can observe $\boldsymbol{y}^*$, the values that are linked to the values of $\boldsymbol{X}$. If there are no linkage errors, then $\boldsymbol{y}^*$ will be the same as $\boldsymbol{y}$, but if there are errors, then $\boldsymbol{y}^*$ won't be the same as $\boldsymbol{y}$. Estimating $\boldsymbol{\theta}$ by substituting $\boldsymbol{y}^*$ for $\boldsymbol{y}$ can therefore lead to bias. Chambers (2008) suggests a number of different methods to correct this bias when g corresponds to either linear or logistic regression under the assumption that all records are linked and linkage is one to one between $\boldsymbol{y}$ and $\boldsymbol{X}$. This reference also considers the situation where $\boldsymbol{X}$ corresponds to a sample, whereas $\boldsymbol{y}$ covers the entire population of interest. That is, $\boldsymbol{X}$ is incomplete but all records from $\boldsymbol{X}$ are linked with records in $\boldsymbol{y}$. In this paper, we extend this idea to accommodate the situation where some of the records in $\boldsymbol{y}$ and $\boldsymbol{X}$ cannot be linked.

The notation and assumptions used in the rest of this paper are set out below:

1. The total number of units in the population of interest, and hence the number of records making up both $\boldsymbol{y}$ and $\boldsymbol{X}$, is N. However, we only observe a sample s of n records from $\boldsymbol{X}$. Furthermore the method of sampling is non-informative given $\boldsymbol{X}$ so that the population relationship $E(\boldsymbol{y}|\boldsymbol{X}) = g(\boldsymbol{X};\boldsymbol{\theta})$ also holds for the correctly linked sampled records from $\boldsymbol{X}$. Here g is arbitrary, but the linear and logistic specifications are of particular interest.

2. Let $\boldsymbol{X}_s$ denote the n sampled records from $\boldsymbol{X}$, noting that not all of these records can be linked to records in $\boldsymbol{y}$.

3. The records making up $\boldsymbol{X}$ can be partitioned into Q distinct and non-overlapping blocks[1].We refer to these as "m-blocks" in what follows, and note that linkage errors only occur within m-blocks, in the sense that records in distinct m-blocks can never be linked. The records from $\boldsymbol{X}$ that make up the q^{th} m-block is denoted $\boldsymbol{X}_q$.

[1]See Chambers (2008) for a more detailed discussion concerning the concept of a block. Essentially blocks serve to post-stratify the linkage errors

4. The random variables corresponding to whether records in $\boldsymbol{X}_s$ can be linked or not and the random variables defining whether records in $\boldsymbol{X}$ are sampled are mutually independent. As a consequence, the regression model that holds for the correctly linked sample records also holds for the non-linked records.

Much of the notation in what follows can be found in Chambers (2008), and so is used without further explanation. Modifications to this notation that are necessary for the extension of the theory set out in that reference are also kept as intuitive as possible.

Assuming that there are Q different m-blocks, one then has $N = \sum_{q=1}^{Q} M_q$, where M_q is the number of records making up $\boldsymbol{X}_q$. By construction, a record from $\boldsymbol{X}_q$ can only be matched to a record in the corresponding m-block $\boldsymbol{y}_q$ of $\boldsymbol{y}$. If we assume that all the records in the q^{th} m-block are linked, then, following Chambers (2008), we can model the outcome of the linkage process by the equation

$$\boldsymbol{y}_q^* = A_q \boldsymbol{y}_q \tag{1}$$

where A_q is an unknown random permutation matrix of order M_q. Further, we can then define

$$E(A_q|\boldsymbol{X}_q) = E_q. \tag{2}$$

When some records are not linked, A_q is no longer a permutation matrix. Let $\boldsymbol{X}_{sq}$ be the set of sampled records in $\boldsymbol{X}_q$. Then $\boldsymbol{X}_{sq}$ can be divided into two groups, defined by $\boldsymbol{X}_{slq}$ which is the set of sampled records in $\boldsymbol{X}_q$ that are linked to $\boldsymbol{y}_q^*$, and $\boldsymbol{X}_{suq}$ which is the set of sampled records in $\boldsymbol{X}_q$ that are not linked to $\boldsymbol{y}_q^*$. Further, let $\boldsymbol{X}_{rq} := \boldsymbol{X}_q - \boldsymbol{X}_{sq}$ denote the set of non-sampled records in $\boldsymbol{X}_q$. Then this also can be partitioned into $\boldsymbol{X}_{rlq}$ and $\boldsymbol{X}_{ruq}$, the set of non-sampled records that can be linked to $\boldsymbol{y}_q^*$ and the set of non-sampled records that cannot be linked to $\boldsymbol{y}_q^*$ respectively. Under the one to one linkage assumption, $\boldsymbol{y}_q^*$ can therefore also be theoretically divided into four groups, namely $\boldsymbol{y}_{slq}^*$, $\boldsymbol{y}_{suq}^*$, $\boldsymbol{y}_{rlq}^*$ and $\boldsymbol{y}_{ruq}^*$. Thus, (1) can be modified to allow non-linkage of sampled records by writing

$$\boldsymbol{y}_q^* = \begin{pmatrix} \boldsymbol{y}_{slq}^* \\ \boldsymbol{y}_{suq}^* \\ \boldsymbol{y}_{rlq}^* \\ \boldsymbol{y}_{ruq}^* \end{pmatrix} = \begin{pmatrix} A_{slsl,q} & A_{slsu,q} & A_{slrl,q} & A_{slru,q} \\ A_{susl,q} & A_{susu,q} & A_{surl,q} & A_{suru,q} \\ A_{rlsl,q} & A_{rlsu,q} & A_{rlrl,q} & A_{rlru,q} \\ A_{rusl,q} & A_{rusu,q} & A_{rurl,q} & A_{ruru,q} \end{pmatrix} \begin{pmatrix} \boldsymbol{y}_{slq} \\ \boldsymbol{y}_{suq} \\ \boldsymbol{y}_{rlq} \\ \boldsymbol{y}_{ruq} \end{pmatrix} = A_q \boldsymbol{y}_q. \tag{3}$$

Similarly, (2) can be modified as

$$E(A_q|\boldsymbol{X}_q) = E_q = \begin{pmatrix} E_{slsl,q} & E_{slsu,q} & E_{slrl,q} & E_{slru,q} \\ E_{susl,q} & E_{susu,q} & E_{surl,q} & E_{suru,q} \\ E_{rlsl,q} & E_{rlsu,q} & E_{rlrl,q} & E_{rlru,q} \\ E_{rusl,q} & E_{rusu,q} & E_{rurl,q} & E_{ruru,q} \end{pmatrix}. \tag{4}$$

2 The adjusted estimating function approach

In the previous section, we showed how we can partition $\boldsymbol{y}_q$, $\boldsymbol{y}_q^*$ and the permutation matrix A_q according to the partition of $\boldsymbol{X}_q$ defined by $\boldsymbol{X}_{slq}$, $\boldsymbol{X}_{suq}$, $\boldsymbol{X}_{rlq}$ and $\boldsymbol{X}_{ruq}$. These partitions play an important role in estimation of $\boldsymbol{\theta}$ when some records in $\boldsymbol{X}_{sq}$ cannot be linked to any records in $\boldsymbol{y}_q$, which is the main theme of this section.

In what follows, we modify the adjusted estimating function approach used in Chambers (2008) to accommodate non-linked sample records. First, however, we briefly consider the case of complete and one to one linkage in order to introduce notation and ideas from Chambers (2008) that are necessary for our development. We then explain how we modify this approach to accommodate non-linked sample data.

2.1 The estimating function approach under one to one complete linkage

A more detailed explanation of the development this subsection can be found in Chambers (2008). Given that $E(\boldsymbol{y}|\boldsymbol{X}) = g(\boldsymbol{X};\boldsymbol{\theta})$, we assume that the $\boldsymbol{\theta}$ can be estimated by solving

$$\boldsymbol{H}(\boldsymbol{\theta}) = \mathbf{0},$$

where $\boldsymbol{H}(\boldsymbol{\theta})$ is an unbiased estimating function. That is, it satisfies $E_X\big[\boldsymbol{H}(\boldsymbol{\theta}_0)\big] = \mathbf{0}$ when $\boldsymbol{\theta}_0$ is the true value of $\boldsymbol{\theta}$. Let ∂_θ be the partial differentiation operator with respect to $\boldsymbol{\theta}$. Suppose that $\hat{\boldsymbol{\theta}}$ satisfies $\boldsymbol{H}(\hat{\boldsymbol{\theta}}) = \mathbf{0}$. Then, under regularity conditions that ensure sufficient smoothness for valid Taylor expansion,

$$\mathbf{0} = \boldsymbol{H}(\hat{\boldsymbol{\theta}}) \approx \boldsymbol{H}(\boldsymbol{\theta_0}) + (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0})\partial_\theta \boldsymbol{H}(\boldsymbol{\theta_0}).$$

If $\boldsymbol{H}(\boldsymbol{\theta})$ is an unbiased estimating function and $\partial_\theta \boldsymbol{H}(\boldsymbol{\theta}_0)$ is non-singular, then one has asymptotic unbiasedness since

$$E_X\big[\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\big] \approx -\big[\partial_\theta \boldsymbol{H}(\boldsymbol{\theta}_0)\big]^{-1} E_X\big[\boldsymbol{H}(\boldsymbol{\theta}_0)\big] = \mathbf{0}.$$

Note that the corresponding asymptotic variance of $\hat{\boldsymbol{\theta}}$ is given by

$$\mathrm{Var}_X(\hat{\boldsymbol{\theta}}) \approx \big[\partial_\theta \boldsymbol{H}(\boldsymbol{\theta}_0)\big]^{-1} \mathrm{Var}_X\big[\boldsymbol{H}(\boldsymbol{\theta}_0)\big] \Big(\big[\partial_\theta \boldsymbol{H}(\boldsymbol{\theta}_0)\big]^{-1}\Big)^T. \tag{5}$$

In Chambers (2008), the estimating function is assumed to be of the form

$$\boldsymbol{H}(\boldsymbol{\theta}) = \sum_q \boldsymbol{G}_q(\boldsymbol{\theta})\big\{\boldsymbol{y}_q - \boldsymbol{f}_q\big\}, \tag{6}$$

where $\boldsymbol{f}_q = E_X(\boldsymbol{y}_q)$ and $\boldsymbol{G}_q(\boldsymbol{\theta})$ is a function of $\boldsymbol{\theta}$ and $\boldsymbol{X}_q$ but not of $\boldsymbol{y}_q$. However, since $\boldsymbol{y}_q$ is not observable, a naive estimator $\hat{\boldsymbol{\theta}}^*$ of $\boldsymbol{\theta}_0$ can be defined by solving the equation

$$\boldsymbol{H}^*(\hat{\boldsymbol{\theta}}^*) = \sum_q \boldsymbol{G}_q(\hat{\boldsymbol{\theta}}^*)\big\{\boldsymbol{y}_q^* - \boldsymbol{f}_q(\hat{\boldsymbol{\theta}}^*)\big\} = \mathbf{0}. \tag{7}$$

This naive estimator assumes no linkage errors. Hence, in general,

$$E_X\big[\boldsymbol{H}^*(\boldsymbol{\theta}_0)\big] = \sum_q \boldsymbol{G}_q(\boldsymbol{\theta}_0)\big\{\big(E_q - I_q\big)\boldsymbol{f}_q(\boldsymbol{\theta}_0)\big\} \neq \mathbf{0}. \tag{8}$$

An unbiased version of this estimating function is

$$\boldsymbol{H}^*_{adj}(\hat{\boldsymbol{\theta}}^*) = \boldsymbol{H}^*(\boldsymbol{\theta}) - \sum_q \boldsymbol{G}_q(\boldsymbol{\theta})\big\{\big(E_q - I_q\big)\boldsymbol{f}_q(\boldsymbol{\theta})\big\} = \sum_q \boldsymbol{G}_q(\boldsymbol{\theta})\big\{\boldsymbol{y}_q^* - E_q\boldsymbol{f}_q(\boldsymbol{\theta})\big\}. \tag{9}$$

The bias-adjusted estimator $\hat{\boldsymbol{\theta}}^*_{adj}$ for $\boldsymbol{\theta}$ is defined as the solution of

$$\boldsymbol{H}^*_{adj}(\hat{\boldsymbol{\theta}}^*_{adj}) = \mathbf{0}.$$

The asymptotic variance of $\hat{\boldsymbol{\theta}}^*_{adj}$ is then of the form

$$\mathrm{Var}_X(\hat{\boldsymbol{\theta}}^*_{adj}) \approx \big[\partial_\theta \boldsymbol{H}^*_{adj}(\boldsymbol{\theta}_0)\big]^{-1} \mathrm{Var}_X\big[\boldsymbol{H}^*_{adj}(\boldsymbol{\theta}_0)\big] \Big(\big[\partial_\theta \boldsymbol{H}^*_{adj}(\boldsymbol{\theta}_0)\big]^{-1}\Big)^T. \tag{10}$$

This estimating function approach is effective and easy to implement. Also, by using different functions for $\boldsymbol{G}_q$, one can define a variety of different estimators. For example, using the standard 'hat' notation to indicate an estimate, we can define (see Chambers (2008) for a more detailed development):

1. The Naive estimator: $\boldsymbol{G}_q = \boldsymbol{X}_q^T$.

2. The Lahiri and Larsen estimator: $\boldsymbol{G}_q = \boldsymbol{X}_q^T \hat{E}_q^T$.

3. The EBLUE (empirical best linear unbiased estimator): $\boldsymbol{G}_q = \boldsymbol{X}_q^T \hat{E}_q^T (\hat{\sigma}^2 \boldsymbol{I}_q + \hat{V}_q)^{-1}$. Here $\sigma^2 = \mathrm{Var}_X(y)$ and $V_q = \mathrm{Var}_X(E_X[A_q y_q | A_q])$.

However, this estimating function approach is based on the assumption that the populations underlying $\boldsymbol{X}$ and $\boldsymbol{y}$ are the same and that linkage is one to one and complete, so that A is a permutation matrix. When some of the records in $\boldsymbol{X}$ cannot be linked to $\boldsymbol{y}$, then A is no longer a permutation matrix. The next subsection therefore extends the adjustment approach described in this subsection to accommodate the incomplete linkage case, using the partitions defined in the previous section.

2.2 The estimating function approach with incompletely linked sample records

The aim of this subsection is to extend the estimating function approach described in Chambers (2008) to accommodate incomplete linkage of sample records.

An immediate consequence of the incomplete linkage of sampled records is that instead of observing $\boldsymbol{y}_q^*$, we observe $\boldsymbol{y}_{slq}^*$. If the size of $\boldsymbol{y}_{slq}^*$ is small, the estimating function approach can be ineffective. Two possible reasons for this are

1. small sample bias, or

2. the distribution of $\boldsymbol{y}_{slq}^*$ may be different from that of $\boldsymbol{y}_{sq}$.

We will investigate both these cases using simulation in the next section where we examine the efficiency of the adjusted estimating function approach developed below. For the time being, however, we assume that the size of $\boldsymbol{y}_{slq}^*$ is not small. Furthermore, in this subsection we assume that the distribution of $\boldsymbol{y}_{slq}^*$ given $\boldsymbol{X}_{sq}$ is the same as that of $\boldsymbol{y}_{sq}$ given $\boldsymbol{X}_{sq}$ (i.e. we have non-informative incomplete linkage). We consider a situation where this assumption can be relaxed later in this section.

Since we only observe $\boldsymbol{y}_{slq}^*$, a modified version of the estimating function that ignores the linkage errors is of the form

$$\boldsymbol{H}_{sl}^*(\boldsymbol{\theta}) = \sum_q \boldsymbol{G}_{slq}(\boldsymbol{\theta}) \big\{ \boldsymbol{y}_{slq}^* - \boldsymbol{f}_{slq}(\boldsymbol{\theta}) \big\}, \tag{11}$$

where

$$\boldsymbol{y}^*_{slq} = A_{sq}\boldsymbol{y}_q$$

and

$$A_{sq} = \begin{pmatrix} A_{slsl,q} & A_{slsu,q} & A_{slrl,q} & A_{slru,q} \end{pmatrix}.$$

A similar partition exists for the expected value of this matrix,

$$E_{sq} = E_X\big[A_{sq}\big] = \begin{pmatrix} E_{slsl,q} & E_{slsu,q} & E_{slrl,q} & E_{slru,q} \end{pmatrix}.$$

Correcting for the bias caused by linkage errors then leads us to an estimating function of the form

$$\begin{aligned}
\boldsymbol{H}^*_{adj,sl}(\boldsymbol{\theta}) &= \boldsymbol{H}^*_{sl}(\boldsymbol{\theta}) - E_X\big[\boldsymbol{H}^*_{sl}(\boldsymbol{\theta})\big] \\
&= \sum_q \boldsymbol{G}_{slq}(\boldsymbol{\theta})\big\{\boldsymbol{y}^*_{slq} - E_X\big[\boldsymbol{y}^*_{slq}\big]\big\} \\
&= \sum_q \boldsymbol{G}_{slq}(\boldsymbol{\theta})\big\{\boldsymbol{y}^*_{slq} - E_{sq}\boldsymbol{f}_q(\boldsymbol{\theta})\big\} \\
&= \sum_q \boldsymbol{G}_{slq}(\boldsymbol{\theta})\big\{\boldsymbol{y}^*_{slq} - E_{slsl,q}\boldsymbol{f}_{slq}(\boldsymbol{\theta})\big\} \\
&\quad - \sum_q \boldsymbol{G}_{slq}(\boldsymbol{\theta})\big\{E_{slsu,q}\boldsymbol{f}_{suq}(\boldsymbol{\theta}) + E_{slrl,q}\boldsymbol{f}_{rlq}(\boldsymbol{\theta}) + E_{slru,q}\boldsymbol{f}_{ruq}(\boldsymbol{\theta})\big\}.
\end{aligned} \tag{12}$$

In order to proceed further we need to specify the distribution of the linkage errors. We adapt the exchangeable linkage error model defined in Chambers (2008). That is, for the q^{th} m-block we assume that

$$Pr(\text{correct linkage}) = Pr(a^q_{ii} = 1) = \lambda_q, \tag{13}$$

and, for $i \neq j$,

$$Pr(\text{incorrect linkage}) = Pr(a^q_{ij} = 1) = \gamma_q. \tag{14}$$

It follows that

$$E_q = (\lambda_q - \gamma_q)\boldsymbol{I}_q + \gamma_q \mathbf{1}_q\mathbf{1}_q^T, \tag{15}$$

where

$$\lambda_q + (M_q - 1)\gamma_q = 1. \tag{16}$$

Under the exchangeable linkage error model, one can therefore write

$$\begin{aligned} E_{slsl,q} &= \Big[\frac{\lambda_q M_q - 1}{M_q - 1}\Big]\boldsymbol{I}_{slq} + \Big[\frac{1-\lambda_q}{M_q-1}\Big]\mathbf{1}_{slq}\mathbf{1}_{slq}^T, \\ E_{slsu,q} &= \Big[\frac{1-\lambda_q}{M_q-1}\Big]\mathbf{1}_{slq}\mathbf{1}_{suq}^T, \\ E_{slrl,q} &= \Big[\frac{1-\lambda_q}{M_q-1}\Big]\mathbf{1}_{slq}\mathbf{1}_{rlq}^T, \\ E_{slru,q} &= \Big[\frac{1-\lambda_q}{M_q-1}\Big]\mathbf{1}_{slq}\mathbf{1}_{ruq}^T \end{aligned} \tag{17}$$

and so

$$\begin{aligned} &\sum_q \boldsymbol{G}_{slq}(\boldsymbol{\theta})\big\{E_{slsu,q}\boldsymbol{f}_{suq}(\boldsymbol{\theta}) + E_{slrl,q}\boldsymbol{f}_{rlq}(\boldsymbol{\theta}) + E_{slru,q}\boldsymbol{f}_{ruq}(\boldsymbol{\theta})\big\} \\ &\quad= \sum_q \Big(\frac{1-\lambda_q}{M_q-1}\Big)\boldsymbol{G}_{slq}(\boldsymbol{\theta})\mathbf{1}_{slq}\Big[\mathbf{1}_{suq}^T\boldsymbol{f}_{suq}(\boldsymbol{\theta}) + \mathbf{1}_{rlq}^T\boldsymbol{f}_{rlq}(\boldsymbol{\theta}) + \mathbf{1}_{ruq}^T\boldsymbol{f}_{ruq}(\boldsymbol{\theta})\Big] \\ &\quad= \sum_q \Big(\frac{1-\lambda_q}{M_q-1}\Big)\boldsymbol{G}_{slq}(\boldsymbol{\theta})\mathbf{1}_{slq}\Big[\mathbf{1}_q^T\boldsymbol{f}_q(\boldsymbol{\theta}) - \mathbf{1}_{slq}^T\boldsymbol{f}_{slq}(\boldsymbol{\theta})\Big]. \end{aligned}$$

Since the distribution of $\boldsymbol{y}_{slq}$ is assumed to be the same as that of $\boldsymbol{y}_{sq}$ and sampling is non-informative, the unknown population sum $\mathbf{1}_q^T\boldsymbol{f}_q(\theta)$ can be approximated by the weighted sample sum $\boldsymbol{w}_{slq}^T\boldsymbol{f}_{slq}(\theta)$[2]. Using (12), we then define a modified estimating function that allows for incomplete linkage of the form

$$\begin{aligned} \boldsymbol{H}_{wsl}^{adj}(\boldsymbol{\theta}) &= \sum_q \boldsymbol{G}_{slq}(\boldsymbol{\theta})\big\{\boldsymbol{y}_{slq}^* - E_{slsl,q}\boldsymbol{f}_{slq}(\boldsymbol{\theta})\big\} \\ &\quad - \sum_q \Big(\frac{1-\lambda_q}{M_q-1}\Big)\boldsymbol{G}_{slq}(\boldsymbol{\theta})\mathbf{1}_{slq}\Big[\boldsymbol{w}_{slq}^T\boldsymbol{f}_{slq}(\boldsymbol{\theta}) - \mathbf{1}_{slq}^T\boldsymbol{f}_{slq}(\boldsymbol{\theta})\Big] \\ &= \sum_q \boldsymbol{G}_{slq}(\boldsymbol{\theta})\big\{\boldsymbol{y}_{slq}^* - \tilde{E}_{slsl,q}\boldsymbol{f}_{slq}(\boldsymbol{\theta})\big\}, \end{aligned} \tag{18}$$

where

$$\tilde{E}_{slsl,q} = \Big[\frac{\lambda_q M_q - 1}{M_q - 1}\Big]\boldsymbol{I}_{slq} + \Big[\frac{1-\lambda_q}{M_q-1}\Big]\mathbf{1}_{slq}\boldsymbol{w}_{slq}^T.$$

[2]In general, the definition of these sample weights will depend on the method of sampling. Here however, we just use simple expansion weights $\boldsymbol{w}_{slq} = (\frac{M_q}{m_{slq}})\mathbf{1}_{slq}$, where m_{slq} is the number of linked sample records, and M_q is the total population number in the $q^{th}m$-block.

2.3 An asymptotic variance estimator

We now derive a variance estimator for the $\hat{\boldsymbol{\theta}}$ based on the estimating function approach developed in the previous subsection. Suppose that $\hat{\boldsymbol{\theta}}$ is the solution of $\boldsymbol{H}^{adj}_{wsl}(\boldsymbol{\theta}) = \mathbf{0}$ defined in (18). Then the asymptotic variance of $\hat{\boldsymbol{\theta}}$ is

$$\mathrm{Var}_X(\hat{\boldsymbol{\theta}}) \approx \left[\partial_\theta \boldsymbol{H}^{adj}_{wsl}(\boldsymbol{\theta}_0)\right]^{-1} \mathrm{Var}_X\left[\boldsymbol{H}^{adj}_{wsl}(\boldsymbol{\theta}_0)\right] \left(\left[\partial_\theta \boldsymbol{H}^{adj}_{wsl}(\boldsymbol{\theta}_0)\right]^{-1}\right)^T. \tag{19}$$

Consequently, an estimator of this asymptotic variance is

$$\hat{V}_X(\hat{\boldsymbol{\theta}}) = \left[\partial_\theta \boldsymbol{H}^{adj}_{wsl}(\hat{\boldsymbol{\theta}})\right]^{-1} \hat{V}_X\left[\boldsymbol{H}^{adj}_{wsl}(\hat{\boldsymbol{\theta}})\right] \left(\left[\partial_\theta \boldsymbol{H}^{adj}_{wsl}(\hat{\boldsymbol{\theta}})\right]^{-1}\right)^T. \tag{20}$$

To calculate $\hat{V}_X(\hat{\boldsymbol{\theta}})$, we need to first evaluate the terms $\partial_\theta \boldsymbol{H}^{adj}_{wsl}(\hat{\boldsymbol{\theta}})$ and $\hat{V}_X\left[\boldsymbol{H}^{adj}_{wsl}(\hat{\boldsymbol{\theta}})\right]$. In general, $\boldsymbol{G}_{slq}$ depends on $\boldsymbol{\theta}$. However, in this paper, we only consider the case where $\boldsymbol{G}_{slq}$ is independent of $\boldsymbol{\theta}$[3]. Thus,

$$\partial_\theta \boldsymbol{H}^{adj}_{wsl}(\hat{\boldsymbol{\theta}}) = \sum_q \boldsymbol{G}_{slq} \tilde{E}_{slsl,q} \partial_\theta \boldsymbol{f}_{slq}(\hat{\boldsymbol{\theta}}). \tag{21}$$

Further,

$$\mathrm{Var}_X\left[\boldsymbol{H}^{adj}_{wsl}(\boldsymbol{\theta}_0)\right] = \sum_q \boldsymbol{G}_{slq} \mathrm{Var}_X(\boldsymbol{y}^*_{slq}) \boldsymbol{G}^T_{slq}. \tag{22}$$

where

$$\begin{aligned} \boldsymbol{\Sigma}_{slq} &= \mathrm{Var}_X(\boldsymbol{y}^*_{slq}) \\ &= \mathrm{Var}_X(A_{sq}\boldsymbol{y}_q) \\ &= E_X\left[\mathrm{Var}_{AX}(A_{sq}\boldsymbol{y}_q)\right] + \mathrm{Var}_X\left[E_{AX}(A_{sq}\boldsymbol{y}_q)\right]. \end{aligned} \tag{23}$$

In order to evaluate $\boldsymbol{\Sigma}_{slq}$, we first consider $\boldsymbol{\Sigma}_q = \mathrm{Var}_X(\boldsymbol{y}^*_q)$. Let $\boldsymbol{D}_q = \mathrm{Var}_X(\boldsymbol{y}_q)$, and suppose that, for $i \neq j$, $\mathrm{cov}_X(y_i, y_j) = 0$. Then $\boldsymbol{D}_q$ can be written as

$$\boldsymbol{D}_q = \mathrm{diag}\left[d^q_i; i \in \{1, \ldots, M_q\}\right].$$

Let

$$A_q = \left[a^q_{ij}; i, j \in \{1, \ldots, M_q\}\right]$$

and

$$E_q = \left[e^q_{ij}; i, j \in \{1, \ldots, M_q\}\right].$$

[3]For the linear and logistic regression cases we only consider the functional forms for $\boldsymbol{G}_{slq}$ that appear in Chambers (2008). These are independent of $\boldsymbol{\theta}$.

Then

$$\begin{aligned}\boldsymbol{\Sigma}_q &= \mathrm{Var}_X(\boldsymbol{y}_q^*) \\ &= \mathrm{Var}_X(A_q\boldsymbol{y}_q) \\ &= E_X\big[\mathrm{Var}_{AX}(A_q\boldsymbol{y}_q)\big] + \mathrm{Var}_X\big[E_{AX}(A_q\boldsymbol{y}_q)\big].\end{aligned} \tag{24}$$

Note that

$$E_X\big[\mathrm{Var}_{AX}(A_q\boldsymbol{y}_q)\big] = E_x\big[A_q\boldsymbol{D}_qA_q^T\big].$$

Since $\boldsymbol{D}_q$ is a diagonal matrix, $A_qA_q^T = \boldsymbol{I}_q$ and $(a_{ij}^q)^2 = a_{ij}^q,\ \forall\ i,j \in \{1,\ldots,M_q\}$, we can write

$$\begin{aligned}A_q\boldsymbol{D}_qA_q^T &= \mathrm{diag}\big[\sum_{j=1}^{M_q}(a_{ij}^q)^2d_j^q; i \in \{1,\ldots,M_q\}\big] \\ &= \mathrm{diag}\big[\sum_{j=1}^{M_q}a_{ij}^qd_j^q; i \in \{1,\ldots,M_q\}\big].\end{aligned} \tag{25}$$

Hence,

$$\begin{aligned}E_X\big[A_q\boldsymbol{D}_qA_q^T\big] &= E_X\Big[\mathrm{diag}\big[\sum_{j=1}^{M_q}a_{ij}^qd_j^q; i \in \{1,\ldots,M_q\}\big]\Big] \\ &= \mathrm{diag}\Big[\big(\lambda_q - \frac{1-\lambda_q}{M_q-1}\big)d_i^q + \frac{M_q(1-\lambda_q)}{M_q-1}\bar{d}^q; i \in \{1,\ldots,M_q\}\Big],\end{aligned} \tag{26}$$

where

$$\bar{d}^q = M_q^{-1}\sum_{j=1}^{M_q}d_j^q.$$

Then, by (25) and (26), it can be seen that

$$\begin{aligned}E_X\big[\mathrm{Var}_{AX}(A_{sq}\boldsymbol{y}_q)\big] &= E_X\Big[\mathrm{diag}\big[\sum_{j=1}^{M_q}a_{ij}^qd_j^q; i \in \{1,\ldots,m_{slq}\}\big]\Big] \\ &= \mathrm{diag}\Big[\big(\lambda_q - \frac{1-\lambda_q}{M_q-1}\big)d_i^q + \frac{M_q(1-\lambda_q)}{M_q-1}\bar{d}^q; i \in \{1,\ldots,m_{slq}\}\Big].\end{aligned} \tag{27}$$

However, $\bar{d}^q$ is not observable and so we replace it by the sample mean

$$\bar{d}_{sl}^q = m_{slq}^{-1}\sum_{j=1}^{m_{slq}}d_j^q.$$

That is, we have the approximation

$$E_X\big[\mathrm{Var}_{AX}(A_{sq}\boldsymbol{y}_q)\big] \approx \mathrm{diag}\Big[\big(\lambda_q - \frac{1-\lambda_q}{M_q-1}\big)d_i^q + \frac{M_q(1-\lambda_q)}{M_q-1}\bar{d}_{sl}^q; i \in \{1,\ldots,m_{slq}\}\Big]. \tag{28}$$

Also, it is shown in Chambers (2008) that

$$\mathrm{Var}_X\big[E_{AX}(A_q\boldsymbol{y}_q)\big] \approx \mathrm{diag}\Big[(1-\lambda_q)\big\{\lambda_q(f_i-\bar{f}_q)^2+\bar{f}_q^{(2)}-(\bar{f}_q)^2\big\}; i\in\{1,\ldots,M_q\}\Big],$$

where $\bar{f}_q = M_q^{-1}\sum_{k=1}^{M_q} f_k$ and $\bar{f}_q^{(2)} = M_q^{-1}\sum_{k=1}^{M_q} f_k^2$. In the case of $\mathrm{Var}_X\big[E_{AX}(A_{sq}\boldsymbol{y}_q)\big]$, this approximation becomes

$$\mathrm{Var}_X\big[E_{AX}(A_{sq}\boldsymbol{y}_q)\big] \approx \mathrm{diag}\Big[(1-\lambda_q)\big\{\lambda_q(f_i-\bar{f}_{slq})^2+\bar{f}_{slq}^{(2)}-(\bar{f}_{slq})^2\big\}; i\in\{1,\ldots,m_{slq}\}\Big], \quad (29)$$

where

$$\bar{f}_{slq} = m_{slq}^{-1}\sum_{k=1}^{m_{slq}} f_k$$

and

$$\bar{f}_{slq}^{(2)} = m_{slq}^{-1}\sum_{k=1}^{m_{slq}} f_k^2.$$

Therefore, by (23), (28) and (29)

$$\boldsymbol{\Sigma}_{slq} \approx \mathrm{diag}\Big[\big(\lambda_q-\frac{1-\lambda_q}{M_q-1}\big)d_i^q+\frac{M_q(1-\lambda_q)}{M_q-1}\bar{d}_{sl}^q+(1-\lambda_q)\big\{\lambda_q(f_i-\bar{f}_{slq})^2+\bar{f}_{slq}^{(2)}-(\bar{f}_{slq})^2\big\}\Big] \quad (30)$$

for all $i\in\{1,\ldots,m_{slq}\}$. It follows that one can calculate $\hat{V}_X(\hat{\boldsymbol{\theta}})$ by first evaluating (30), (22) and (21) and then substituting these values into (20).

2.4 The estimating function approach under non-ignorable linkage

In this subsection we consider a special (and extreme) case of linking, where the conditional distribution of $\boldsymbol{y}_{slq}^*$ given $\boldsymbol{X}_{sq}$ is very different from that of the corresponding conditional distribution of $\boldsymbol{y}_{suq}^*$. Note however that we continue to assume that the sampling process itself is non-informative, so the conditional distribution of $\boldsymbol{y}_q$ given $\boldsymbol{X}_q$ is the same as that of $\boldsymbol{y}_{sq}$ given $\boldsymbol{X}_{sq}$.

The linking model that we assume here is based on a linear regression relationship,

$$\boldsymbol{y}_{sq} = \boldsymbol{X}_{sq}\boldsymbol{\beta}+\boldsymbol{e}_{sq},$$

where the errors e_{sq} are drawn from the $N(0,\sigma^2)$ distribution. In particular, we consider the case where a disproportionate number of linked sample records, i.e. those defining $\boldsymbol{y}_{slq}^*$,

correspond to positive errors under this linear regression model. As a consequence, we expect that the mean of $\boldsymbol{y}^*_{slq}$ will be larger than that of $\boldsymbol{y}^*_{suq}$. However, since the regression errors $\boldsymbol{e}_{sq}$ are distributed as $N(0, \sigma^2)$, we see that although $\boldsymbol{f}_{sq} = E_X(\boldsymbol{y}_{sq})$ is the same as in the case of ignorable linking, clearly $E_X(\boldsymbol{y}_{slq}) \neq E_X(\boldsymbol{y}_{sq})$. One way of dealing with this problem is to reduce the contribution of $\boldsymbol{y}^*_{slq}$ to the estimating function by introducing 'linkage' weights that ensure that the weighted contribution of $\boldsymbol{y}^*_{slq}$ is the same as that of $\boldsymbol{y}^*_{suq}$. To do this we need to know the distribution of the signs of the regression errors for the linked records[4].

Let p_{slq} denote the proportion of records making up of $\boldsymbol{y}^*_{slq}$ that have positive regression errors, and put

$$\boldsymbol{y}^*_{slq} = \begin{pmatrix} \boldsymbol{y}^{*+}_{slq} \\ \boldsymbol{y}^{*-}_{slq} \end{pmatrix},$$

where $\boldsymbol{y}^{*+}_{slq}$ is the subset of $\boldsymbol{y}^*_{slq}$ defined by those linked sample records with positive regression errors, and $\boldsymbol{y}^{*-}_{slq}$ denotes the remaining linked sample records, i.e. those with negative regression errors. Further, note that, by (18)

$$\boldsymbol{H}^{adj}_{wsl}(\theta) = \sum_q \boldsymbol{G}_{slq}(\theta)\big\{\boldsymbol{y}^*_{slq} - \tilde{E}_{slsl,q}\boldsymbol{f}_{slq}(\theta)\big\}.$$

However, this estimating function is only unbiased when $E_X(\boldsymbol{y}^{*+}_{slq}) = E_X(\boldsymbol{y}^{*-}_{slq})$, which is not true in this case. Consequently, we put $w^+_{slq} = \frac{0.5}{p_{slq}}$ and $w^-_{slq} = \frac{0.5}{1-p_{slq}}$ and define a weighted version of y^*_{slq} of the form

$$\boldsymbol{y}^*_{wslq} = \begin{pmatrix} w^+_{slq}\boldsymbol{y}^{*+}_{slq} \\ w^-_{slq}\boldsymbol{y}^{*-}_{slq} \end{pmatrix}. \tag{31}$$

The estimating function that should be used in this case is then

$$\boldsymbol{H}^{adj}_{wsl2}(\theta) = \sum_q \boldsymbol{G}_{slq}(\theta)\big\{\boldsymbol{y}^*_{wslq} - \tilde{E}_{slsl,q}\boldsymbol{f}_{slq}(\theta)\big\}, \tag{32}$$

where

$$\tilde{E}_{slsl,q} = \Big[\frac{\lambda_q M_q - 1}{M_q - 1}\Big]\boldsymbol{I}_{slq} + \Big[\frac{1-\lambda_q}{M_q - 1}\Big]\boldsymbol{1}_{slq}\boldsymbol{w}^T_{slq}.$$

In the next section we use simulation to explore the performance of this weighted estimating function.

[4]Clearly this is very strong assumption. However, it does allow us to investigate a bias correction method for this situation. We aim to relax this assumption in future research.

3 Simulation results for incomplete sample to register linkage

In this section we use simulation to investigate the relative performances of different estimators based on linked data where the linkage is from sample to register, and there is both linkage error as well as incomplete linkage. Estimators are compared in terms of their relative bias, relative root mean squared error and coverage rate for nominal 95% confidence intervals. The simulation designs used are similar to those in Chambers (2008), allowing us to compare the performances of the estimators for the incomplete linkage case with those for the complete linkage case. Box plots showing the distributions of the percentage relative errors generated by different estimators are in the Figures.

3.1 Linear regression with random non-linking

Population values were generated for a linear regression model of the form

$$Y = 1 + 5X + e,$$

where values of the explanatory variable X were drawn from the uniform distribution over [0,1] and values of the regression error e were drawn from the $N(0, 1)$ distribution. The actual data pairs (y_i, x_i) were randomly allocated to three m-blocks, and the linked data pairs (y_i^*, x_i) then generated according to an exchangeable linkage error model with specified probabilities of correct linkage. Finally, non-links were created by randomly selecting records from each m-block. In particular:

- The population size was $N = 4000$, divided into three m-blocks, of sizes 2000, 1000 and 1000 respectively.
- The linked data pairs (y_i^*, x_i) were generated using independent exchangeable linkage error mechanisms in each m-block, with specified probabilities of correct linkage.
- Following this linkage process, 1000 of the 2000 links created in the first m-block were randomly assigned to a 'non-linkable' status. This process was repeated in the second m-block (500 non-links) and the third m-block (600 non-links).

- Sample records were independently selected in each of the m-blocks via simple random sampling without replacement, and with m-block sample sizes of $n_1 = 500$, $n_2 = 300$ and $n_3 = 200$ respectively.
- The sample data used in estimation consisted of the 'linkable' sample records in each of these m-blocks.

Two scenarios for the probabilities of correct linkage in the three m-blocks were simulated. These were:

- Scenario 1: $\lambda_1 = 1,\ \lambda_2 = 0.95$ and $\lambda_3 = 0.75$.
- Scenario 2: $\lambda_1 = 0.95,\ \lambda_2 = 0.75$ and $\lambda_3 = 1$.

In practice, the probabilities λ_q need to be either specified or estimated in some way. Following Chambers (2008) we considered two options in this regard:

1. We assumed that we knew the true values of the λ_q.
2. We estimated the value of λ_q in those m-blocks where linkage is not perfect using the methodology described in Chambers (2008). In this case our estimates were based on the correctly linked/incorrectly linked status of 25 randomly sub-sampled linked records in sample in each such m-block.

Three estimators of the intercept and slope coefficients of the linear regression model [5] were calculated. They are:

1. the naive OLS estimator (ST),
2. the Lahiri-Larsen estimator (A) and
3. the empirical BLUE (C).

The two scenarios above were independently simulated 1000 times. In each simulation, population and linked sample data were generated and the regression parameters estimated using the three estimators specified above. The performances of these estimators were then

[5] See (Chambers (2008)) for details on these estimators.

compared in terms of relative bias, relative RMSE and the actual coverage rate for nominal 95% confidence intervals. These values are reported in Table 1.

[Table 1 here.]

The results set out in Table 1 display very similar patterns to those reported in Chambers (2008), although the actual levels of relative bias and RMSE are higher. This is most probably due to the different m-block sizes and the consequent greater incidence of incorrect linkages in the current set of simulations. In any case, it is clear that both the Lahiri-Larsen estimator (A) and the EBLUE (C) correct the bias of the naive estimator (ST) in both scenarios, irrespective of whether the actual correct linkage probabilities are known or are estimated. As noted in Chambers (2008), the EBLUE (C) outperforms the Lahiri-Larsen estimator (A).

3.2 Logistic regression with random non-linking

In addition to the estimators ST, A and C considered in the previous subsection, we calculated another bias-corrected estimator (M), based on the same weighting function as the MLE under perfect linkage. See Chambers (2008) for more details about this estimator. We also allowed the distribution of X to vary between m-blocks. In particular, for the m-block with $\lambda = 1$, X values were drawn from the uniform distribution on [5,20], while for the m-block with $\lambda = 0.95$, X values were drawn from the uniform distribution on [-5,5]. Finally, for the remaining m-block with smallest correct linkage probability, X values were drawn from the uniform distribution on [-20,5]. Values of Y were then generated as independently distributed Bernoulli variables with

$$logit\big[\text{pr}(y_i = 1|x_i) = 1 - x_i\big].$$

Simulation results for estimates of the slope parameter of the logistic model are set out in Table 2. These show that the naive estimator (ST) is negatively biased as a consequence of incorrect linkage. However, unlike the complete linkage results obtained in Chambers (2008), we see that in this case the adjusted estimators M, A and C appear to overcompensate for

this bias, while also displaying increased variability. As a consequence, it is hard to see any advantage in using these adjusted estimators in this type of situation (large proportion of unlinked records). Essentially, the main advantage of M, A and C here is that their coverage performance remains superior to that of ST. As an aside, we note that the EBLUE-type adjusted estimator C continues to be superior to the adjusted estimators M and A.

[Table 2 here.]

3.3 Linear regression with non-ignorable linkage

In the simulations described so far, the probability of non-linkage has been the same for all records in an m-block, and the random variable corresponding to whether a record is linked or not has been distributed independently of the regression error for that record. That is, whether a record is linked or not has been ignorable, with the only effect of non-linkage being a reduction in the number of sampled records contributing data to the analysis. In this subsection we investigate the situation where linkage is non-ignorable, and simulate linked data where the probability of non-linkage depends on the sign of the error in the underlying regression model. In particular, we simulated population data using the same linear regression model as previously,

$$Y = 1 + 5X + e,$$

but now allowed the linking process to over-represent records with positive regression errors. We consider two cases. In the first, 60% of the linked records are randomly drawn from those records with positive errors, while in the second this proportion is increased to 75%. In both cases the simulations were carried out with linkage errors generated under scenario 2, and with estimators based on the actual probabilities of correct linkage. Results from these simulations are displayed in Table 3.

[Table 3 here.]

As one might expect, these show that the main impact of this type of non-ignorable linkage is to upwardly bias estimates of the intercept parameter in the population linear regression model. When the imbalance of positive relative to negative error terms in the linkable records is relatively small (the 60% case) it is clear that estimator C outperforms estimator A and

both are substantially better than ST. However, when this imbalance is reasonable large (the 75% case) we see that A is preferable to C, although both A and C are still clearly better than ST.

3.4 Linear regression with weighting for non-ignorable linkage

In the previous subsection we saw that a larger proportion of records with positive regression errors in $\boldsymbol{y}^*_{slq}$ tended to lead to increased bias when estimating the intercept parameter of the underlying linear regression model. In subsection 2.4 we showed how this bias can be reduced by using weights that adjust for this imbalance. In this subsection we present simulation results that illustrate this weighting approach. In particular, we continued to use scenario 2, but this time with relatively high imbalances between positive regression errors and negative regression errors in the linked data. In particular, we simulated two situations where:

- 75% of linked records in each m-block corresponded to records with positive regression errors,
- 90% of linked records in each m-block corresponded to records with positive regression errors.

[Table 4 here.]

From the results set out in Table 4 we see that the biases in estimators A and C that were evident in Table 3 have been effectively corrected by appropriate weighting. However, when we compare the relative RMSEs in Table 4 with those for scenario 2 in Table 1 we see that the price paid for this decrease in bias is an increase in variance. This is not unexpected, since correcting the bias of an estimator generally increases its variability. We also note that the weighted version of C used in this simulation appears to be slightly more efficient than the corresponding weighted version of A. However, it should be kept in mind that the 'linkage' weights used here assume that we know the probability that a record will be linked. This is unrealistic in practice, and further work is required to investigate how these weighted versions of A and C behave when linkage probabilities are approximate rather than exact.

3.5 Linear regression with small samples and ignorable linkage

So far in this section we have presented simulation results for the estimating function approach when it is applied to comparatively large samples. In this final subsection we use simulation to investigate the performance of these methods with very small samples. In particular, in this last set of simulations we generated incompletely linked sample data using the same linear regression model as in the previous subsections, but in this case these linkages were based on:

- A total of 30 m-blocks, made up of 10 blocks each of size 200 and 20 blocks each of size 100 (i.e. a population of size $N = 4000$). Samples of size 10 were selected in each m-block.
- Probabilities of correct linkage that were defined to be 1 for the first 10 m-blocks, 0.95 for the next 10 m-blocks and 0.75 for the last 10 m-blocks.
- Linkages that were at random within each m-block. This resulted in linked sample sizes within m-blocks that varied from 3 to 5.

From the results set out in Table 5, it is clear that the very small linked samples in each m-block led to the estimators A and C exhibiting substantial biases. This is in contrast to the unbiased results that we obtained for these estimators in subsection 3.1, where the m-block linked sample sizes were much larger.

[Table 5 here.]

One way of avoiding this small sample bias is to merge similar small m-blocks in order to increase the within m-block sample size. For example, we can merge m-blocks with the same value of λ_q. If this leads to a larger sample in the merged m-block, then the biases evident in Table 5 become much smaller.

4 Conclusions and further research

In this paper we extend the adjusted estimating function approach developed in Chambers (2008) to accommodate the sample to register incomplete linkage case, i.e. where some of sample data cannot be linked to the register. Through simulations we show that the adjusted estimating function approach generally leads to unbiased and more efficient estimators

than those (e.g. ST) defined by estimating functions that treat the linkage as perfect. In particular, we consider the situation where the linkage process is not ignorable, so that the distributions of the linked and unlinked sample data are different. We overcome this problem by introducing another weight function (in addition to the usual survey weights) that reflects the probability of linkage, and show that the corresponding weighted version of the adjusted estimating function approach then corrects for the bias induced by non-ignorable linkage. However, these 'linkage weights' depend on knowing the process that governs whether a record is linked or not, which is a strong assumption. In effect, this problem is completely analogous to the one facing an analyst who wishes to compensate for non-ignorable non-response in a data set. Without knowledge of the non-response process, and in particular the response probabilities for the observed data, this adjustment can be problematic. Further research is needed in this area.

One of problems we face in using the adjusted estimating function approach is that it leads to the use of m-block specific plug–in estimators. These work well when the m-block sample sizes are large, but can inefficient otherwise, as the simulations reported in the preceding subsection demonstrate. One way to correct this is to merge m-blocks with similar linkage behaviour (both in terms of linkage probabilities as well as probabilities of correct linkage). However, this method only works when there are many similar m-blocks. Further research on alternative small sample methods for dealing with incompletely linked data, e.g. MLE based on application of the Missing Information Principle, is needed.

Another limitation of the theory outlined in this paper is that it assumes that just two data sets are linked. This is often not the case in practice, where the linked data set used in analysis may be the consequence of multiple linking operations. In such cases, the linkage error structure can become extremely complicated. For example, consider three linked data sets corresponding to the variables Y, X_1 and X_2 that together define the regression function

$$Y = g(X_1, X_2|\theta) + e.$$

In this case there could be mismatches between X_1 and X_2, Y and X_1 and Y and X_2. Further, the mismatches between X_1 and X_2 could be correlated with the mismatches between Y and X_1 and between Y and X_2. Extending the adjusted estimating function approach to this more complicated situation is currently being researched.

References

Chambers, R. (2008). Regression analysis of probability-linked data. *Statisphere*, **4**. http://www.statisphere.govt.nz/official-statistics-research/series/vol-4.htm.

Fellegi, I. P. and Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, **64**(328), 1183–1210.

Jaro, M. A. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical Association*, **84**, 414–420.

Lahiri, P. and Larsen, M. D. (2005). Regression analysis with linked data. *Journal of the American Statistical Association*, **100**(469), 222–230.

Neter, J., Maynes, E. S., and Ramanathan, R. (1965). The effect of mismatching on the measurement of response error. *Journal of the American Statistical Association*, **60**(312), 1005–1027.

Newcombe, H. B. (1988). *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business*. Oxford, U.K.: Oxford University Press.

Scheuren, F. and Winkler, W. E. (1993). Regression analysis of data files that are computer matched. *Survey Methodology*, **19**, 39–58.

Scheuren, F. and Winkler, W. E. (1997). Regression analysis of data files that are computer matched–part ii. *Survey Methodology*, **23**, 157–165.

Tables

Table 1: Simulation results for linear regression and random non- linking. Coverage is for nominal 95% intervals.

Estimator	Relative Bias		Relative RMSE		Coverage	
	λ known	λ unknown	λ known	λ unknown	λ known	λ unknown
Simulation results for the intercept estimator						
Scenario 1: $\lambda_1 = 1,\ \lambda_2 = 0.95$ and $\lambda_3 = 0.75$						
ST	16.35	17.06	19.67	20.22	64.6	61.6
A	0.04	0.73	11.12	10.96	94.7	97.0
C	0.17	0.85	10.70	10.65	94.8	96.2
Scenario 2: $\lambda_1 = 0.95,\ \lambda_2 = 0.75$ and $\lambda_3 = 1$						
ST	35.89	36.17	37.76	38.24	11.4	12.9
A	0.00	0.36	12.64	13.32	95.7	98.8
C	0.12	0.28	11.77	12.40	95.5	97.6
Simulation results for the slope estimator						
Scenario 1: $\lambda_1 = 1,\ \lambda_2 = 0.95$ and $\lambda_3 = 0.75$						
ST	-6.62	-6.71	17.07	17.19	52.9	52.8
A	-0.10	-0.19	8.66	8.50	94.8	97.2
C	-0.15	-0.22	8.32	8.22	94.9	96.496.4
Scenario 2: $\lambda_1 = 0.95,\ \lambda_2 = 0.75$ and $\lambda_3 = 1$						
ST	-14.29	-14.44	33.31	33.70	4.7	5.5
A	0.06	-0.10	10.31	10.59	95.1	99.4
C	-0.01	-0.10	9.54	9.84	95.2	97.7

Table 2: Simulation results for slope estimators in logistic regression and random non-linking. Coverage is for nominal 95% intervals.

Estimator	Relative Bias		Relative RMSE		Coverage	
	λ known	λ unknown	λ known	λ unknown	λ known	λ unknown
Scenario 1: $\lambda_1 = 1,\ \lambda_2 = 0.95$ and $\lambda_3 = 0.75$						
ST	-8.39	-8.59	18.05	17.65	80.1	80.4
M	9.74	11.45	32.98	45.92	95.3	96.9
A	9.77	11.51	32.89	45.90	95.4	96.9
C	8.35	10.94	32.98	69.48	96.0	96.4
Scenario 2: $\lambda_1 = 0.95,\ \lambda_2 = 0.75$ and $\lambda_3 = 1$						
ST	-7.23	-7.00	19.68	19.81	82.6	83.7
M	17.50	17.32	69.17	86.79	95.1	95.7
A	16.37	15.45	57.02	57.95	95.5	95.7
C	11.43	12.86	46.25	46.94	95.5	96.4

Table 3: Simulation results for linear regression with non-ignorable linking. Scenario 2 with known λ. Coverage is for nominal 95% intervals.

Estimator	Relative Bias		Relative RMSE		Coverage	
	Intercept	slope	Intercept	slope	Intercept	slope
Scenario 2: 60% of linked records with positive errors						
ST	27.07	-14.46	29.58	33.71	36.3	6.0
A	-8.86	-0.12	15.85	10.58	88.7	93.5
C	-2.64	-0.16	11.78	10.08	95.5	93.2
Scenario 2: 75% of linked records with positive errors						
ST	41.20	-14.19	42.82	33.17	4.4	5.5
A	5.35	0.16	13.89	10.70	93.1	94.3
C	13.96	0.01	17.70	10.01	81.5	94.0

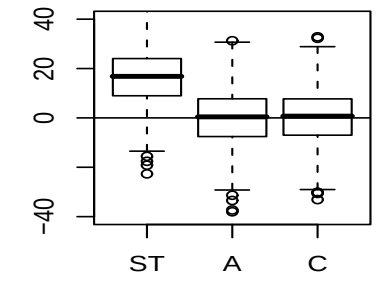
Table 4: Simulation results for linear regression with non-ignorable linking and weighted estimators. Scenario 2 with estimated λ. Coverage is for nominal 95% intervals.

Estimator	Relative Bias		Relative RMSE		Coverage	
	Intercept	slope	Intercept	slope	Intercept	slope
Scenario 2, estimated λ: 75% of linked records with positive errors for all blocks						
ST	34.82	-13.94	37.66	33.94	47.5	35.2
A	-1.20	0.46	15.74	14.79	99.5	98.5
C	-1.00	0.40	14.94	14.18	99.5	98.6
Scenario 2, estimated λ: 90% of linked records with positive errors for all blocks						
ST	36.28	-13.41	46.94	42.72	91.0	82.1
A	-0.12	1.14	34.11	34.74	99.7	97.0
C	-0.09	1.14	33.98	34.66	99.7	96.9

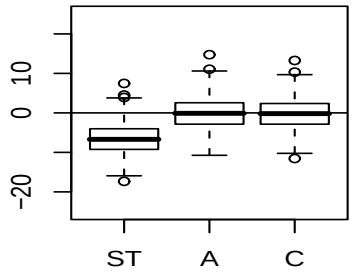
Table 5: Simulation results for linear regression with small linked samples and random non-linking (linked sample sizes between 3 and 5). Coverage is for nominal 95% intervals.

Estimator	Relative Bias		Relative RMSE		Coverage	
	Intercept	slope	Intercept	slope	Intercept	slope
Scenario 2, Known λ						
ST	14.51	-9.95	25.76	28.17	88.8	72.0
A	-9.52	-0.34	24.13	18.15	91.7	94.6
C	-9.88	-0.31	23.35	17.44	91.4	93.5

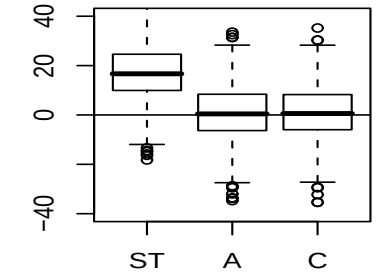
Figures



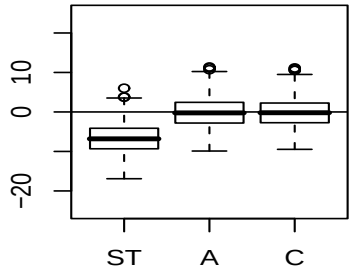
Scenario 1: Known lambda,Intercept



Scenario 1: Known lambda, Slope

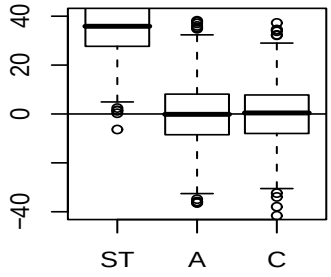


Scenario 1: Unknown lambda, Intercept

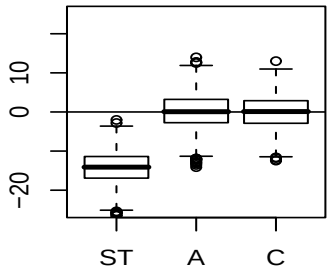


Scenario 1: Unknown lambda, Slope

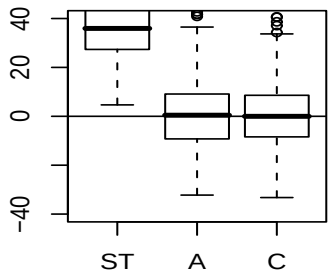
Figure 1: Simulated percentage relative errors for intercept and slope coefficients in linear regression under scenario 1 and random non-linking.



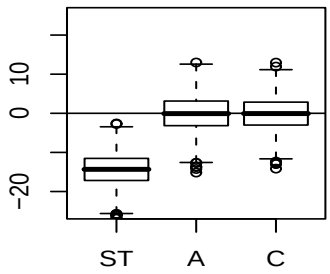
Scenario 2: Known lambda,Intercept



Scenario 2: Known lambda, Slope



Scenario 2: Unknown lambda, Intercept



Scenario 2: Unknown lambda, Slope

Figure 2: Simulated percentage relative errors for intercept and slope coefficients in linear regression under scenario 2 and random non-linking.

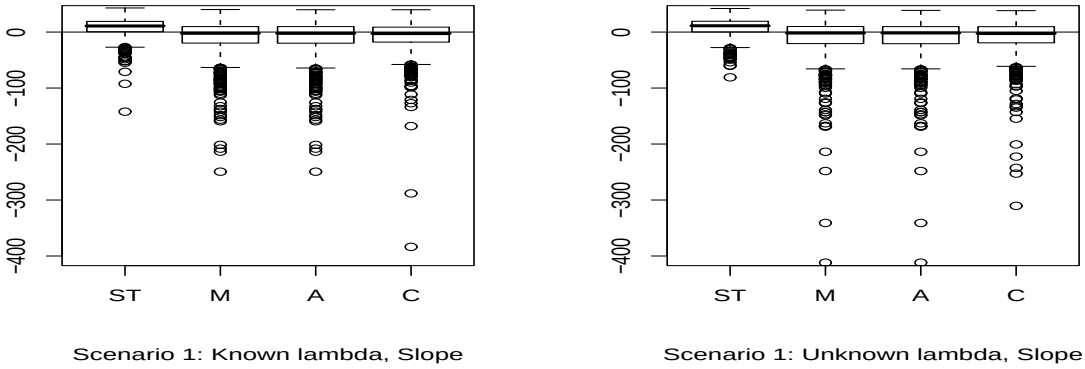


Figure 3: Simulated percentage relative errors for slope coefficient in logistic regression under scenario 1 and random non-linking.

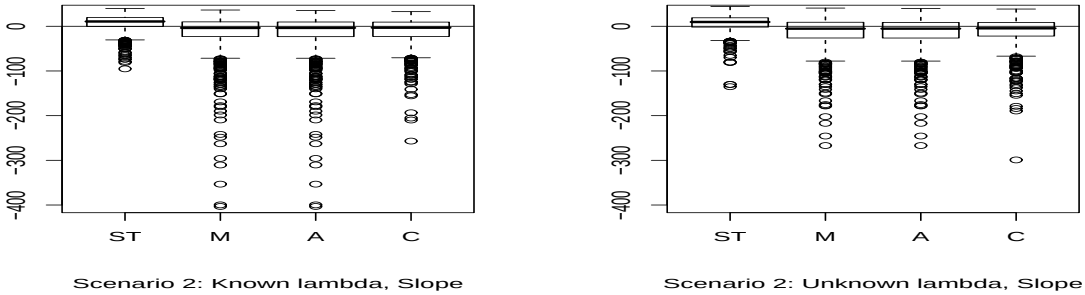


Figure 4: Simulated percentage relative errors for slope coefficient in logistic regression under scenario 2 and random non-linking.

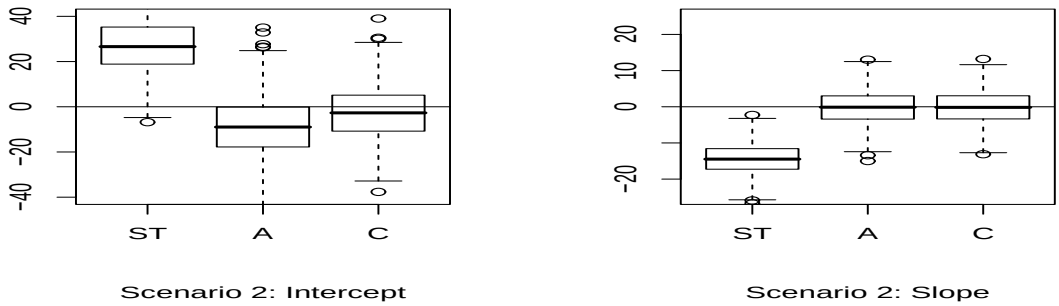


Figure 5: Simulated percentage relative errors for intercept coefficient in linear regression with non-ignorable linking (60% of linked records with positive errors). Scenario 2 with known λ.

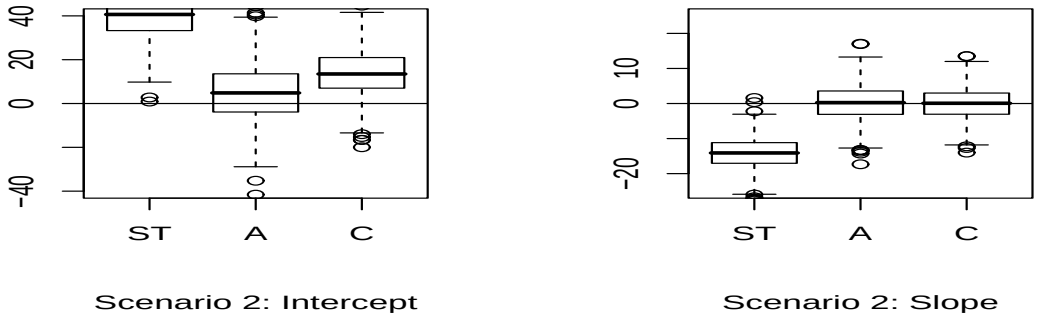
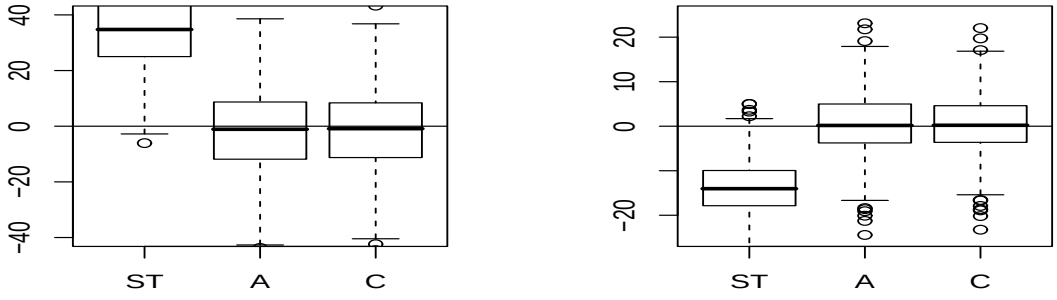
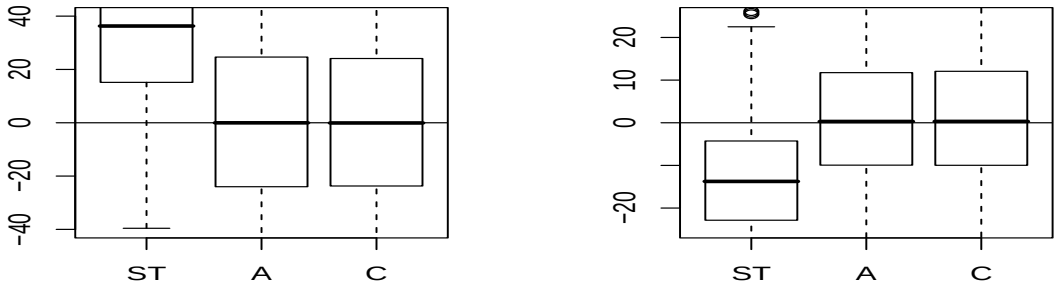


Figure 6: Simulated percentage relative errors for slope coefficient in linear regression with non-ignorable linking (75% of linked records with positive errors). Scenario 2 with known λ.



Scenario 2: Unknown lambda, Intercept Scenario 2: Unknown lambda, Slope

Figure 7: Simulated percentage relative errors for intercept coefficient in linear regression with non-ignorable linking (75% of linked records with positive errors). Scenario 2 with estimated λ. Weighting used to correct for non-ignorable linking.



Scenario 2: Unknown lambda, Intercep Scenario 2: Unknown lambda, Slope

Figure 8: Simulated percentage relative errors for slope coefficient in linear regression with non-ignorable linking (90% of linked records with positive errors). Scenario 2 with estimated λ. Weighting used to correct for non-ignorable linking.

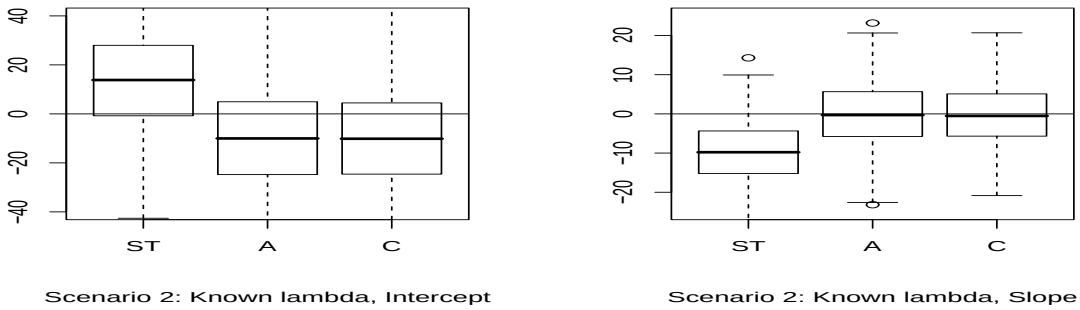


Figure 9: Simulated percentage relative errors for slope coefficient in linear regression under scenario 2 and random non-linking, with linked sample sizes between 3 and 5.