University of Wollongong

Research Online

Centre for Statistical & Survey Methodology Working Paper Series

Faculty of Engineering and Information Sciences

2012

Unbiased regression estimation for multi-linked data in the presence of correlated linkage error

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

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

Follow this and additional works at: https://ro.uow.edu.au/cssmwp

Recommended Citation

Kim, G. and Chambers, Ray, Unbiased regression estimation for multi-linked data in the presence of correlated linkage error, Centre for Statistical and Survey Methodology, University of Wollongong, Working Paper 3-12, 2012, 24.
https://ro.uow.edu.au/cssmwp/90

Research Online is the open access institutional repository for the University of Wollongong. For further information contact the UOW Library: research-pubs@uow.edu.au

Centre for Statistical and Survey Methodology

The University of Wollongong

Working Paper

03-12

Unbiased Regression Estimation for Multi-Linked Data in the Presence of Correlated Linkage Error

Gunky Kim and Raymond Chambers

Copyright © 2008 by the Centre for Statistical & Survey Methodology, UOW. Work in progress, no part of this paper may be reproduced without permission from the Centre.

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

Unbiased Regression Estimation for Multi-Linked Data in the Presence of Correlated Linkage Error

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

Abstract

Linkage errors can occur when probability-based methods are used to link records from two or more distinct data sets corresponding to the same target population. Recent research on methods for modifying standard methods of regression analysis to allow for these errors assumes that when more than two linked data sets are used to generate the data for this analysis, the linkage errors in these different data sets are independent. In this paper we extend these results to accommodate the more realistic scenario of dependent linkage errors. Our simulation results show that an incorrect assumption of independent linkage errors can lead to insufficient linkage error bias correction, while an approach that allows for correlated linkage errors appears to fully correct this bias.

key words: probabilistic record linkage; correlated linkage errors; linear regression; estimating equations.

1 Introduction

Probabilistic data linkage is widely used when direct measurement is impossible or extremely costly. One important application is where different data sets relating to the same individuals at different points in time are 'multi-linked' to provide a synthetic longitudinal data record for each individual. However, even with a unique identifier, there exists the possibility that linkage errors in the merged data could lead to such a longitudinal record being actually made up of data items from different individuals. These linkage errors will lead to bias and loss of efficiency in regression modelling using the merged data set. Kim and Chambers

(2012b) describe methods for correcting the bias due to linkage errors when multiple data sets are probabilistically multi-linked. These methods assume independent pairwise linkage errors. A more realistic scenario, however, is to allow dependent pairwise linkage errors, in the sense that it is more likely that if the records corresponding to two different individuals in data sets A and B are incorrectly linked, then it is quite likely that the records for the same two individuals in data sets A and C will also be incorrectly linked. In this paper we show how the bias due to correlated linkage errors in the resulting merged data set can be corrected. Our methods are based on the inference framework described in Chambers (2009), and we focus on the situation where the merged data set is obtained by linking three separate data sources via two possibly dependent linkage operations. These data sources could represent different registers for the same population at different points in time or they could correspond to where a survey sample is linked to two separate population registers, one contemporaneous with the survey and the other containing historical information.

1.1 Technical background and assumptions

For notational simplicity we denote conditioning by a subscript in what follows, so the conditional expectation $E(\boldsymbol{y}|\boldsymbol{X})$ is written $E_{\boldsymbol{X}}(\boldsymbol{y})$ and so on. Suppose that we are interested in fitting a regression model of the form $E_{\boldsymbol{X}}(\boldsymbol{y}) = f(\boldsymbol{X};\boldsymbol{\beta})$ where f is a known function, but the parameter $\boldsymbol{\beta}$ is to be estimated. Here $\boldsymbol{y}$ denotes the vector of population values of the response variable of interest, and $\boldsymbol{X}$ denotes the corresponding matrix of population values for a set of explanatory variables, which are themselves drawn from multiple sources. In particular, we focus on the situation where the actual values making up $\boldsymbol{y}$ and $\boldsymbol{X}$ are unknown, but probabilistic linkage is used to reconstruct their values using the data in two or more population registers. To fix concepts, we assume throughout this paper that the regression model of interest is the linear model

$$\boldsymbol{y} = \boldsymbol{1}\beta_{\boldsymbol{0}} + \boldsymbol{X}_{\boldsymbol{1}}\beta_{\boldsymbol{1}} + \boldsymbol{X}_{\boldsymbol{2}}\beta_{\boldsymbol{2}} + \boldsymbol{\epsilon} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1}$$

where $\boldsymbol{y}$, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ denote data stored on three separate population registers and $\boldsymbol{1}$ denote the unit vector of order N. However, our development is quite general, and the linear model (1) is easily replaced by a generalized linear model. The model errors $\boldsymbol{\epsilon}$ are assumed to have zero mean and are uncorrelated given $\boldsymbol{X}$, with $\mathrm{Var}_{\boldsymbol{X}}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{I}_N$ where $\boldsymbol{I}_N$ is the identity matrix of order N, the population size. It is assumed that no unique identifier exists, and so these three registers cannot be perfectly linked. Instead the data available to fit this regression model is generated via a probabilistic linkage process, so linkage errors are possible. These mismatches will lead to biased estimation of $\boldsymbol{\beta}$. The aim of this paper is to describe a methodology that can be used to eliminate this bias.

In what follows, we do not distinguish between the population registers that define $\boldsymbol{y}$, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ and the data sets themselves. Also, without loss of generality, we take one of the three data sets to be the 'benchmark' register, i.e. all linkage errors are defined relative to the population ordering defined by it. Following Kim and Chambers (2012b) we focus on the situation where is the benchmark register and there are linkage errors between $\boldsymbol{y}$ and $\boldsymbol{X}_1$ and between $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$.

In common with the development in Kim and Chambers (2012a), we initially consider the situation where the data sets are all population registers with complete linkage. This enables us to develop our notation and general approach in a situation where the basic analytic issues are clear. We then move to the more interesting situation where one data set is a sample, which is linked to two separate population registers. In this case we allow for incomplete linkage. We start by stating our basic assumptions for the first situation, i.e. where the linked data set is constructed by linking three population registers:

1 All registers have complete coverage of the target population and are of size N. In particular, for each distinct population unit there exist unique records in each of $\boldsymbol{y}$, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ that correspond to this population unit.

2 The registers $\boldsymbol{y}$, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ can each be partitioned into Q 'match blocks' or 'm-blocks' such that linkage errors occur only within them. That is, records in distinct m-blocks can never be linked, and the records for any population unit in an m-block are contained in the corresponding m-blocks of each register. We denote quantities associated with the q^{th} m-block by a subscript of q. Thus the M_q records making up the q^{th} m-block within $\boldsymbol{X}_1$ are denoted $\boldsymbol{X}_{1q}$, etc. An individual record i in m-block q is denoted $i \in q$.

3 We have non-informative linkage in the sense that linkage errors within an m-block are independent of any regression errors associated with observations from that m-block.

In many practical situations, a sample $\boldsymbol{s}$ of records from the register $\boldsymbol{X}_1$ is selected, and a subsample of these records is linked to the two separate population registers $\boldsymbol{y}$ and $\boldsymbol{X}_2$. In this situation we make the following additional assumptions:

4 The method of sampling is non-informative within m-blocks for the regression model of interest, in the sense that the same regression model holds for both sampled and non-sampled population units within an m-block. Furthermore, the linked sample records in an m-block are 'linked at random', so the non-informativeness assumption also holds for the linked sample units. Note that this last assumption is a strong one. See Kim and Chambers (2012a).

5 Let $\boldsymbol{y} = (y_i)$, $\boldsymbol{X}_1 = [\boldsymbol{x}_{1i}^T]$ and $\boldsymbol{X}_2 = [\boldsymbol{x}_{2i}^T]$. A consistent estimator of any population quantity of the form $\sum_{i=1}^{N} g(y_i, \boldsymbol{x}_{1i}, \boldsymbol{x}_{2i})$ is its sample-weighted equivalent $\sum_{\boldsymbol{s}} w_i g(y_i, \boldsymbol{x}_{1i}, \boldsymbol{x}_{2i})$, where $\boldsymbol{w}_{\boldsymbol{s}} = (w_i; i \in \boldsymbol{s})$ is a vector of known sample weights.

2 Methodological Development

Fellegi and Sunter (1969) describe an approach to optimal probability-based linkage that is based on maximising the probability of a declared link being correct. Unfortunately, most practical implementations of their approach require one to trade off the number of links made against their accuracy. As a consequence, any implementation of probabilistic linkage will result in unmade linkages or non-linkages as well as linkage errors where linkages are actually made. In what follows, we show that the bias caused by linkage errors can be corrected if we know the probability of correct linkage. In particular, we develop efficient estimators for regression coefficients when three data sources have been probabilistically linked to form the data set used in the analysis. Although our primary interest in this context is where a sample from one register has been independently linked to two other registers, we start by considering the case where three registers are completely linked.

2.1 A model for correlated linkage error

In this sub-section and the next we assume that all three linked data sets are registers and linkage is complete, i.e. linkage is one to one and onto. We use a superscript of * to denote quantities defined using the linked data and, following Kim and Chambers (2012b), we model the relationship between the true, but unobserved, values of $\boldsymbol{y}$ and $\boldsymbol{X}_2$ and the observed linked values $\boldsymbol{y}^*$ and $\boldsymbol{X}_2^*$ within m-block q by writing

$$\boldsymbol{y}_q^* = \boldsymbol{A}_q \boldsymbol{Y}_q \text{ and } \boldsymbol{X}_{2q}^* = \boldsymbol{B}_q \boldsymbol{X}_{2q}$$

where $\boldsymbol{A}_q$ and $\boldsymbol{B}_q$ are unobserved random permutation matrices of order M_q that characterise the outcomes of the two linkage processes in m-block q. In particular, we put

$$\boldsymbol{T}_{Aq} = E_{\boldsymbol{X}^*}(\boldsymbol{A}_q) \text{ and } \boldsymbol{T}_{Bq} = E_{\boldsymbol{X}^*}(\boldsymbol{B}_q).$$

We follow Kim and Chambers (2012b) and use the exchangeable linkage errors (ELE) model of Chambers (2009) to define $\boldsymbol{T}_{Aq}$ and $\boldsymbol{T}_{Bq}$. That is, for any linked record i in m-block q, it is equally likely that it is correctly linked to itself or incorrectly linked to any other record in the same m-block. This leads to the representations

$$\boldsymbol{T}_{Aq} = (\lambda_{Aq} - \gamma_{Aq})\boldsymbol{I}_q + \gamma_{Aq}\mathbf{1}_q\mathbf{1}_q^T$$

and

$$\boldsymbol{T}_{Bq} = (\lambda_{Bq} - \gamma_{Bq})\boldsymbol{I}_q + \gamma_{Bq}\mathbf{1}_q\mathbf{1}_q^T$$

where, for any two distinct records i and j in m-block q, $\lambda_{Aq} = \Pr(\boldsymbol{x}_{1iq}, y_{jq}$ correctly linked) and $\gamma_{Aq} = \Pr(\boldsymbol{x}_{1iq}, y_{jq}$ incorrectly linked, $i \neq j) = (M_q - 1)^{-1}(1 - \lambda_{Aq})$, with λ_{Bq} and γ_{Bq} defined similarly.

Kim and Chambers (2012b) assume that $\boldsymbol{A}_q$ and $\boldsymbol{B}_q$ are independently distributed. However, it is more realistic to assume that if the records corresponding to two different individuals in data sets $\boldsymbol{X}_1$ and $\boldsymbol{y}$ are incorrectly linked, then it is quite likely that the records for the same two individuals in data sets $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ will also be incorrectly linked, i.e. $\boldsymbol{B}_q$ and $\boldsymbol{A}_q$ are dependent random matrices. Let $\boldsymbol{A}_q = [a_{ij}^q]$ and $\boldsymbol{B}_q = [b_{ij}^q]$. In order to model the conditional distribution of given we further assume that

$$\phi_q = \Pr(\boldsymbol{x}_{1iq}, \boldsymbol{x}_{2iq} \text{ correctly linked \& } \boldsymbol{x}_{1iq}, y_{iq} \text{ correctly linked})$$

does not depend on i. Put $\lambda_{B|Aq} = \lambda_{Aq}^{-1}\phi_q$. Under the correlated ELE model for $\boldsymbol{A}_q$ and $\boldsymbol{B}_q$ defined above, it can then be shown that

$$\boldsymbol{T}_{B|Aq} = E_{\boldsymbol{X}^*}(\boldsymbol{B}_q|\boldsymbol{A}_q) = (\lambda_{B|Aq} - \gamma_{B|Aq})\boldsymbol{I}_q + \gamma_{B|Aq}\mathbf{1}_q\mathbf{1}_q^T$$

where $\gamma_{B|Aq} = (M_q - 1)^{-1}(1 - \lambda_{B|Aq})$. It follows that

$$\boldsymbol{X}_q^{E|A} = E_{\boldsymbol{X}^*}\Big(\big[\boldsymbol{1}_q, \boldsymbol{X}_{1q}, \boldsymbol{B}_q^T\boldsymbol{X}_{2q}^*\big]\big|\boldsymbol{A}_q\Big) = \big[\boldsymbol{1}_q, \boldsymbol{X}_{1q}, \boldsymbol{T}_{B|Aq}\boldsymbol{X}_{2q}^*\big] \tag{2}$$

and so

$$E_{\boldsymbol{X}^*}(\boldsymbol{y}_q^*) = E_{\boldsymbol{X}^*}(\boldsymbol{A}_q\boldsymbol{y}_q) = E_{\boldsymbol{X}^*}(\boldsymbol{A}_q)E_{\boldsymbol{X}^*}(\boldsymbol{y}_q|\boldsymbol{A}_q) = \boldsymbol{T}_{Aq}\boldsymbol{X}_q^{E|A}\boldsymbol{\beta}. \tag{3}$$

2.2 Unbiased regression estimation under correlated linkage error

We focus on the situation where the aim is to estimate the parameter $\boldsymbol{\beta}$ of the linear regression model of interest using an adjusted unbiased estimating function. When both $\boldsymbol{y}_q$ and $\boldsymbol{X}_q$ are available this is the function of the form $\boldsymbol{H}(\boldsymbol{\beta}) = \sum_q \boldsymbol{G}_q(\boldsymbol{y}_q - \boldsymbol{f}_q)$ where $\boldsymbol{f}_q = E_{\boldsymbol{X}}(\boldsymbol{y}_q) = \boldsymbol{X}_q\boldsymbol{\beta}$ and $\boldsymbol{G}_q$ is a weighting function that depends on $\boldsymbol{X}_q$ but not on $\boldsymbol{y}_q$. However, we do not observe $\boldsymbol{y}_q$ or $\boldsymbol{X}_q$. Instead, their linked versions $\boldsymbol{y}_q^*$ and $\boldsymbol{X}_q^*$ are observed. A naive estimating function based on $\boldsymbol{H}(\boldsymbol{\beta})$ then takes the form

$$\boldsymbol{H}^*(\boldsymbol{\beta}) = \sum_q \boldsymbol{G}_q^*(\boldsymbol{y}_q^* - \boldsymbol{f}_q^*)$$

where $\boldsymbol{f}_q^* = \boldsymbol{X}_q^*\boldsymbol{\beta}$ and $\boldsymbol{G}_q^* = \boldsymbol{X}_q^{*T}$. Here $\boldsymbol{X}_q^* = (\boldsymbol{1}_q, \boldsymbol{X}_{1q}, \boldsymbol{X}_{2q}^*)$. The *naive estimator* is defined by solving $\boldsymbol{H}^*(\boldsymbol{\beta}) = 0$. This estimator is biased since $E_{\boldsymbol{X}^*}(\boldsymbol{y}_q^*) = \boldsymbol{T}_{Aq}\boldsymbol{f}_q^{E|A} \neq \boldsymbol{f}_q^*$, where $\boldsymbol{f}_q^{E|A} = \boldsymbol{X}_q^{E|A}\boldsymbol{\beta}$. On the other hand, using (2) and (3), we see that an unbiased estimating function based on the linked data is of the form

$$\boldsymbol{H}^*(\boldsymbol{\beta}) = \sum_q \boldsymbol{G}_q^*(\boldsymbol{y}_q^* - \boldsymbol{T}_{Aq}\boldsymbol{f}_q^{E|A}) \tag{4}$$

and so an unbiased estimator of $\boldsymbol{\beta}$ can be defined as the solution $\hat{\boldsymbol{\beta}}^*$ to the estimating equation defined by setting (4) to zero. The following Theorem extends Theorem 1 of Kim and Chambers (2012b) and develops the asymptotic variance of $\hat{\boldsymbol{\beta}}^*$ under correlated linkage error. Its proof is in the Appendix.

Theorem 1. *Let* $\boldsymbol{f}_{2q}^* = (f_{2iq}^*; i \in q) = \boldsymbol{X}_{2q}^*\beta_2$ *and let* $\hat{\boldsymbol{\beta}}^*$ *denote the solution to setting (4) to zero. The asymptotic variance of* $\hat{\boldsymbol{\beta}}^*$ *is then*

$$\boldsymbol{V}(\hat{\boldsymbol{\beta}}^*) = \Big[\sum_q \boldsymbol{G}_q^*\boldsymbol{T}_{Aq}\boldsymbol{X}_q^{E|A}\Big]^{-1}\Big[\sum_q \boldsymbol{G}_q^*\boldsymbol{V}(\boldsymbol{y}_q^*)\boldsymbol{G}_q^{*T}\Big]\Big(\Big[\sum_q \boldsymbol{G}_q^*\boldsymbol{T}_{Aq}\boldsymbol{X}_q^{E|A}\Big]^{-1}\Big)^T$$

where $\boldsymbol{V}(\boldsymbol{y}_q^*) = \sigma^2\boldsymbol{I}_q + \boldsymbol{V}_{Aq} + \boldsymbol{V}_{Cq}$. *Here*

$$\boldsymbol{V}_{Aq} = (1-\lambda_{Aq})\,\mathrm{diag}\left[\left\{\lambda_{Aq}(f_{iq}^{E|A} - \bar{f}_q^{E|A})^2 + \bar{f}_q^{E|A(2)} - (\bar{f}_q^{E|A})^2\right\}; i \in q\right]$$

where $\boldsymbol{f}_q^{E|A} = (f_{iq}^{E|A}; i \in q)$, $\bar{f}_q^{E|A} = M_q^{-1}\sum_{i\in q} f_{iq}^{E|A}$ *and* $\bar{f}_q^{E|A(2)} = M_q^{-1}\sum_{i\in q}(f_{iq}^{E|A})^2$. *Similarly*

$$\boldsymbol{V}_{Cq} = (1-\lambda_{B|Aq})\,\mathrm{diag}\left[(M_q-1)^{-1}\left\{(\lambda_{Aq}M_q-1)d_i + M_q(1-\lambda_{Aq})\bar{d}_q\right\}; i \in q\right]$$

with $d_i = \lambda_{B|Aq}(f_{2iq}^* - \bar{f}_{2q}^*)^2 + \bar{f}_{2q}^{*(2)} - (\bar{f}_{2q}^*)^2$, $\bar{f}_{2q}^* = M_q^{-1}\sum_{i\in q} f_{2iq}^*$ *and* $\bar{f}_{2q}^{*(2)} = M_q^{-1}\sum_{i\in q}(f_{2iq}^*)^2$.

Note:

1. Given $\boldsymbol{T}_{Aq}$, $\boldsymbol{T}_{B|Aq}$ and $\boldsymbol{f}_q^{E|A}$, an unbiased estimator of σ^2 is

   $$\tilde{\sigma}^2 = N^{-1}\Big[\sum_q(\boldsymbol{y}_q^* - \boldsymbol{f}_q^{E|A})^T(\boldsymbol{y}_q^* - \boldsymbol{f}_q^{E|A}) - 2\sum_q(\boldsymbol{f}_q^{E|A})^T(\boldsymbol{I}_q - \boldsymbol{T}_{Aq})\boldsymbol{f}_q^{E|A}\Big].$$

   We can therefore estimate $\boldsymbol{V}(\boldsymbol{y}_q^*)$ above by substituting $\hat{\boldsymbol{\beta}}^*$ for $\boldsymbol{\beta}$ in the definitions of $\boldsymbol{f}_q^{E|A}$, $\boldsymbol{f}_{2q}^*$ and $\tilde{\sigma}^2$. An estimator of the asymptotic variance $\boldsymbol{V}(\hat{\boldsymbol{\beta}}^*)$ of $\hat{\boldsymbol{\beta}}^*$ follows directly.

2. The value of $\hat{\boldsymbol{\beta}}^*$ depends on choice of the weighting function $\boldsymbol{G}_q^*$. A popular choice is $\boldsymbol{G}_q^* = (\boldsymbol{X}_q^*)^T$. However, there are alternative choices. For example, Lahiri and Larsen (2005) develop an adjusted estimator for $\boldsymbol{\beta}$ that, when placed in an estimating equation framework, corresponds to setting $\boldsymbol{G}_q^* = (\boldsymbol{T}_{Aq}\boldsymbol{X}_q^{E|A})^T$. The optimal weighting function, i.e. the one that minimises the asymptotic variance of $\hat{\boldsymbol{\beta}}^*$ (see Godambe (1960)), depends on the unknown model parameters and is given by

   $$\boldsymbol{G}_q^* = \Big(\frac{\partial}{\partial\boldsymbol{\beta}}\big[E_{\boldsymbol{X}^*}(\boldsymbol{y}_q^*)\big]\Big)^T\Big(\boldsymbol{V}(\boldsymbol{y}_q^*)\Big)^{-1} = \Big(\boldsymbol{X}_q^*\Big)^T\Big(\sigma^2\boldsymbol{I}_q + \boldsymbol{V}_{Aq} + \boldsymbol{V}_{Cq}\Big)^{-1}.$$

   This suggests that an iterative approach to weighting should lead to an efficient adjusted estimator $\hat{\boldsymbol{\beta}}^*$. Simulation studies in the next section compare the performances of the estimators defined by these alternative choices.

The development so far has assumed that the correct linkage probabilities λ_{Aq} and $\lambda_{B|Aq}$ are known. This will not be the case in practice, and estimates of these probabilities will need to be used. The actual asymptotic variance of $\hat{\boldsymbol{\beta}}^*$ then also depends on the additional variability induced by this estimation process, as we show in the following Lemma to Theorem 1.

Lemma 1. *When* λ_{Aq} *and* ϕ_q *are unknown and are replaced by* m*-block-specific consistent estimators* $\hat{\lambda}_{Aq}$ *and* $\hat{\phi}_q$*, the asymptotic variance of* $\hat{\boldsymbol{\beta}}^*$ *becomes*

$$\boldsymbol{V}(\hat{\boldsymbol{\beta}}^*) = \boldsymbol{D}\Big[\sum_q \Big(\boldsymbol{G}_q^* \boldsymbol{V}(\boldsymbol{y}_q^*)\boldsymbol{G}_q^{*T} + \sum_{i=1}^{2}\sum_{j=1}^{2}\big(\partial_i \boldsymbol{H}^*\big) J_{ijq}\big(\partial_j \boldsymbol{H}^*\big)^T\Big)\Big]\boldsymbol{D}^T$$

where $\boldsymbol{D} = \left[\sum_q \boldsymbol{G}_q^* \boldsymbol{T}_{Aq} \boldsymbol{X}_q^{E|A}\right]^{-1}$*,* $\boldsymbol{J}_q = [J_{ijq}] = Cov(\hat{\lambda}_{Aq}, \hat{\phi}_q)$ *and*

$$\partial_i \boldsymbol{H}^* = \big(\partial_i \boldsymbol{G}_q^*\big)\big(\boldsymbol{y}_q^* - \boldsymbol{T}_{Aq}\boldsymbol{f}_q^{E|A}\big) - \boldsymbol{G}_q^*\Big[\big(\partial_i \boldsymbol{T}_{Aq}\big)\boldsymbol{f}_q^{E|A} + \boldsymbol{T}_{Aq}\big(\partial_i \boldsymbol{f}_q^{E|A}\big)\Big].$$

Here $\partial_1 = \partial/\partial\lambda_A$ *and* $\partial_2 = \partial/\partial\phi_q$*.*

Observe that as in Chambers (2009) and Kim and Chambers (2012b), the estimated parameters $\hat{\lambda}_{Aq}$ and $\hat{\phi}_q$ can be based on a random 'audit sample' of records in m-block q of the linked data base $\big[\boldsymbol{y}^* \ \boldsymbol{X}_1 \ \boldsymbol{X}_2^*\big]$.

2.3 Incomplete sample to registers linkage

We finally consider the more realistic case when a sample s of n records from the benchmark register $\boldsymbol{X}_1$ is taken and an attempt is made to link these records to the $\boldsymbol{y}$ and $\boldsymbol{X}_2$ registers. However, this linkage is incomplete, i.e. there are some records in the sample s that cannot be linked, either to records in the $\boldsymbol{X}_2$ register or to records in the $\boldsymbol{y}$ register, or both. Note that assumption 4 at the end of section 1 applies here, so whether a record in $\boldsymbol{X}_{1q}$ is sampled or not has nothing to do with whether it can be linked to a record in $\boldsymbol{y}_q$ or one in $\boldsymbol{X}_{2q}$ (or both) and furthermore, actual linkage is then a random event.

Let $\boldsymbol{X}_{1sq}$ be the set of the sample records from $\boldsymbol{X}_{1q}$. Also let $\boldsymbol{X}_{1slq}$ be the set of sample records in $\boldsymbol{X}_{1sq}$ that are linked to both $\boldsymbol{X}_2$ and to $\boldsymbol{y}$. The set of sample records in $\boldsymbol{X}_{1sq}$ that cannot be linked in this way are denoted by $\boldsymbol{X}_{1suq}$. Similarly, $\boldsymbol{X}_{1rq}$ denotes the set of non-sample records in $\boldsymbol{X}_{1q}$. Following Kim and Chambers (2012b), we assume that that there exists, at least in theory, a corresponding set of decompositions of the set of non-sample records. In particular, $\boldsymbol{X}_{1rlq}$ represents the set of non-sample records that are potentially 'linkable' to both $\boldsymbol{X}_2$ and $\boldsymbol{y}$. The remaining non-sampled 'unlinkable' records are denoted

$\boldsymbol{X}_{1ruq}$. It immediately follows that the following partitions exist:

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

where

$$E(\boldsymbol{A}_q | \boldsymbol{X}^*_q) = \boldsymbol{T}_{Aq} = \begin{pmatrix} \boldsymbol{T}_{(sl)Aq} \\ \boldsymbol{T}_{(su)Aq} \\ \boldsymbol{T}_{(rl)Aq} \\ \boldsymbol{T}_{(ru)Aq} \end{pmatrix} = \begin{pmatrix} \boldsymbol{T}_{(slsl)Aq} & \boldsymbol{T}_{(slsu)Aq} & \boldsymbol{T}_{(slrl)Aq} & \boldsymbol{T}_{(slru)Aq} \\ \boldsymbol{T}_{(susl)Aq} & \boldsymbol{T}_{(susu)Aq} & \boldsymbol{T}_{(surl)Aq} & \boldsymbol{T}_{(suru)Aq} \\ \boldsymbol{T}_{(rlsl)Aq} & \boldsymbol{T}_{(rlsu)Aq} & \boldsymbol{T}_{(rlrl)Aq} & \boldsymbol{T}_{(rlru)Aq} \\ \boldsymbol{T}_{(rusl)Aq} & \boldsymbol{T}_{(rusu)Aq} & \boldsymbol{T}_{(rurl)Aq} & \boldsymbol{T}_{(ruru)Aq} \end{pmatrix}.$$

Further, because $\boldsymbol{X}^*_2$ can be similarly partitioned into $\boldsymbol{X}^*_{2slq}$, $\boldsymbol{X}^*_{2suq}$, $\boldsymbol{X}^*_{2rlq}$ and $\boldsymbol{X}^*_{2ruq}$, one has

$$\boldsymbol{T}_{B|Aq} = \begin{pmatrix} \boldsymbol{T}_{(sl)B|Aq} \\ \boldsymbol{T}_{(su)B|Aq} \\ \boldsymbol{T}_{(rl)B|Aq} \\ \boldsymbol{T}_{(ru)B|Aq} \end{pmatrix} = \begin{pmatrix} \boldsymbol{T}_{(slsl)B|Aq} & \boldsymbol{T}_{(slsu)B|Aq} & \boldsymbol{T}_{(slrl)B|Aq} & \boldsymbol{T}_{(slru)B|Aq} \\ \boldsymbol{T}_{(susl)B|Aq} & \boldsymbol{T}_{(susu)B|Aq} & \boldsymbol{T}_{(surl)B|Aq} & \boldsymbol{T}_{(suru)B|Aq} \\ \boldsymbol{T}_{(rlsl)B|Aq} & \boldsymbol{T}_{(rlsu)B|Aq} & \boldsymbol{T}_{(rlrl)B|Aq} & \boldsymbol{T}_{(rlruB|A)q} \\ \boldsymbol{T}_{(rusl)B|Aq} & \boldsymbol{T}_{(rusu)B|Aq} & \boldsymbol{T}_{(rurl)B|Aq} & \boldsymbol{T}_{(ruru)B|Aq} \end{pmatrix}.$$

Since both the sampling and linking processes are assumed to be non-informative within m-blocks, the estimating function for $\boldsymbol{\beta}$ based on the linked sample data is

$$\begin{aligned} \boldsymbol{H}^*_{sl}(\boldsymbol{\beta}) &= \sum_q \boldsymbol{G}^*_{slq}\big(\boldsymbol{y}^*_{slq} - \boldsymbol{T}_{(sl)Aq}\boldsymbol{f}^{E|A}_q\big) \\ &= \sum_q \boldsymbol{G}^*_{slq}\big(\boldsymbol{y}^*_{slq} - \boldsymbol{T}_{(slsl)Aq}\boldsymbol{f}^{E|A}_{slq} - \boldsymbol{T}_{(slsu)Aq}\boldsymbol{f}^{E|A}_{suq} - \boldsymbol{T}_{(slrl)Aq}\boldsymbol{f}^{E|A}_{rlq} - \boldsymbol{T}_{(slru)Aq}\boldsymbol{f}^{E|A}_{ruq}\big). \end{aligned} \tag{5}$$

Under the ELE model , (5) becomes

$$\boldsymbol{H}^*_{sl}(\boldsymbol{\beta}) = \sum_q \boldsymbol{G}^*_{slq}\Big[\boldsymbol{y}^*_{slq} - \Big(\frac{\lambda_{Aq}M_q - 1}{M_q - 1}\Big)\boldsymbol{f}^{E|A}_{slq} - \Big(\frac{1-\lambda_{Aq}}{M_q - 1}\Big)\boldsymbol{1}_{slq}\boldsymbol{1}^T_q\boldsymbol{f}^{E|A}_q\Big].$$

This modified estimating function depends on the value of $\boldsymbol{1}^T_q\boldsymbol{f}^{E|A}_q$, which is a population, rather than a sample, quantity. Given assumptions 4 and 5 at the end of section 1, we can estimate $\boldsymbol{1}^T_q\boldsymbol{f}^{E|A}_q$ using the weighted sample estimate $\tilde{\boldsymbol{w}}^T_{slq}\boldsymbol{f}^{E|A}_{slq}$, where $\tilde{\boldsymbol{w}}_{slq} = M_{sq}M^{-1}_{slq}\boldsymbol{w}_{slq}$. Here $\boldsymbol{w}_{slq}$ denotes the vector of sampling weights associated with the M_{slq} linked sample records in the q^{th} m-block, while M_{sq} is the total number of sampled records in this block. In the special case where $\boldsymbol{X}_{1sq}$ corresponds to an equal probability sample from $\boldsymbol{X}_{1q}$, $\tilde{\boldsymbol{w}}_{slq} = M_q M_{slq}^{-1} \mathbf{1}_{slq}$, where M_q is the number of records in q^{th} m-block. It immediately follows that (5) can be replaced by

$$\boldsymbol{H}_{sl}^{*}(\boldsymbol{\beta}) = \sum_{q} \boldsymbol{G}_{slq}^{*}\big(\boldsymbol{y}_{slq}^{*} - \tilde{\boldsymbol{T}}_{(sl)Aq} \boldsymbol{f}_{slq}^{E|A}\big) \tag{6}$$

where

$$\tilde{\boldsymbol{T}}_{(sl)Aq} = (M_q - 1)^{-1}\Big\{(\lambda_{Aq} M_q - 1)\boldsymbol{I}_{slq} + (1 - \lambda_{Aq})\mathbf{1}_{slq}\tilde{\boldsymbol{w}}_{slq}^{T}\Big\}.$$

Unfortunately, there is still an issue with use of (6) since, by (2),

$$\boldsymbol{f}_{slq}^{E|A} = \big(\mathbf{1}_{slq}, \boldsymbol{X}_{1slq}, \boldsymbol{T}_{(sl)B|Aq}\boldsymbol{X}_{2q}^{*}\big)\boldsymbol{\beta}$$

where

$$\boldsymbol{T}_{(sl)B|Aq}\boldsymbol{X}_{2q}^{*} = \boldsymbol{T}_{(slsl)B|Aq}\boldsymbol{X}_{2slq}^{*} + \boldsymbol{T}_{(slsu)B|Aq}\boldsymbol{X}_{2suq}^{*} + \boldsymbol{T}_{(slrl)B|Aq}\boldsymbol{X}_{2rlq}^{*} + \boldsymbol{T}_{(slru)B|Aq}\boldsymbol{X}_{2ruq}^{*}$$

and the last three terms on the right hand side in the preceding identity are dependent on the unlinked sample and non-sample (linked and unlinked) values in $\boldsymbol{X}_2$, which are unknown. The same argument used to justify sample weighting above then leads to $\boldsymbol{T}_{(sl)B|Aq}\boldsymbol{X}_{2q}^{*}$ being approximated by $\tilde{\boldsymbol{T}}_{(sl)B|Aq}\boldsymbol{X}_{2slq}^{*}$ where

$$\tilde{\boldsymbol{T}}_{(sl)B|Aq} = (M_q - 1)^{-1}\Big\{(\lambda_{B|Aq} M_q - 1)\boldsymbol{I}_{slq} + (1 - \lambda_{B|Aq})\mathbf{1}_{slq}\tilde{\boldsymbol{w}}_{slq}^{T}\Big\}.$$

That is, the final form of the estimating function that can be used in this case replaces (6) with

$$\tilde{\boldsymbol{H}}_{sl}^{*}(\boldsymbol{\beta}) = \sum_{q} \boldsymbol{G}_{slq}^{*}\big(\boldsymbol{y}_{slq}^{*} - \tilde{\boldsymbol{T}}_{(sl)Aq} \tilde{\boldsymbol{f}}_{slq}^{E|A}\big) \tag{7}$$

where $\tilde{\boldsymbol{f}}_{slq}^{E|A} = \big(\mathbf{1}_{slq}, \boldsymbol{X}_{1slq}, \tilde{\boldsymbol{T}}_{(sl)B|Aq}\boldsymbol{X}_{2q}^{*}\big)\boldsymbol{\beta} = \tilde{\boldsymbol{X}}_{slq}^{E|A}\boldsymbol{\beta}$.

As in the previous sub-section, the development so far has assumed that the probabilities λ_{Aq} and $\lambda_{B|Aq}$ are known. In practice, these will be unknown and replaced by the values of suitable estimators $\hat{\lambda}_{Aq}$ and $\hat{\lambda}_{B|Aq}$ respectively. The following Theorem sets out the form of the asymptotic variance for the solution $\hat{\boldsymbol{\beta}}_{s}^{*}$ to setting (7) to zero. Its proof is along the same lines as that of Theorem 1 and Lemma 1. The notation used there is used again without further comment.

Theorem 2. *Let $\hat{\boldsymbol{\beta}}_s^*$ denote the solution to setting the modified estimating function (7) to zero. Given the assumptions (4) and (5) at the end of section 1 as well as those implicit in Theorem 1 and Lemma 1, the asymptotic variance of $\hat{\boldsymbol{\beta}}_s^*$ is then*

$$\boldsymbol{V}(\hat{\boldsymbol{\beta}}_s^*) = \tilde{\boldsymbol{D}}_{sl}\Big[\sum_q \Big(\boldsymbol{G}_{slq}^* \boldsymbol{V}(\boldsymbol{y}_{slq}^*)\boldsymbol{G}_{slq}^{*T} + \sum_{i=1}^{2}\sum_{j=1}^{2}\big(\partial_i \tilde{\boldsymbol{H}}^*_{slq}\big) J_{ijq}\big(\partial_j \tilde{\boldsymbol{H}}^*_{slq}\big)^T\Big)\Big]\tilde{\boldsymbol{D}}_{sl}^T$$

where

$$\tilde{\boldsymbol{D}}_{sl} = \Big[\sum_q \boldsymbol{G}_{slq}^* \tilde{\boldsymbol{T}}_{(sl)Aq}\tilde{\boldsymbol{X}}_{slq}^{E|A}\Big]^{-1},$$

$$\boldsymbol{V}(\boldsymbol{y}_{slq}^*) = \sigma^2 \boldsymbol{I}_{slq} + \tilde{\boldsymbol{V}}_{(sl)Aq} + \tilde{\boldsymbol{V}}_{(sl)Cq}$$

and

$$\partial_i \tilde{\boldsymbol{H}}^*_{slq} = \big(\partial_i \boldsymbol{G}_{slq}^*\big)\big(\boldsymbol{y}_{slq}^* - \tilde{\boldsymbol{T}}_{(sl)Aq}\tilde{\boldsymbol{f}}_{slq}^{E|A}\big) - \boldsymbol{G}_{slq}^*\Big[\big(\partial_i \tilde{\boldsymbol{T}}_{(sl)Aq}\big)\tilde{\boldsymbol{f}}_{slq}^{E|A} + \tilde{\boldsymbol{T}}_{(sl)Aq}\big(\partial_i \tilde{\boldsymbol{f}}_{slq}^{E|A}\big)\Big].$$

Here

$$\tilde{\boldsymbol{V}}_{Aq} = (1-\lambda_{Aq})\,\mathrm{diag}\Big[\big\{\lambda_{Aq}(\tilde{f}_{iq}^{E|A} - \overline{\tilde{f}_{slq}^{E|A}})^2 + \overline{\tilde{f}_{slq}^{E|A(2)}} - (\overline{\tilde{f}_{slq}^{E|A}})^2\big\}; i \in slq\Big]$$

where $\overline{\tilde{f}_{slq}^{E|A}} = M_{slq}^{-1}\sum_{i\in slq}\tilde{f}_{iq}^{E|A}$ *and* $\overline{\tilde{f}_{slq}^{E|A(2)}} = M_{slq}^{-1}\sum_{i\in slq}(\tilde{f}_{iq}^{E|A})^2$. *Similarly*

$$\tilde{\boldsymbol{V}}_{(sl)Cq} = (1-\lambda_{B|Aq})\,\mathrm{diag}\Big[(M_q - 1)^{-1}\big\{(\lambda_{Aq}M_q - 1)\tilde{d}_i + M_q(1-\lambda_{Aq})\overline{\tilde{d}_q}\big\}; i \in q\Big]$$

with $\tilde{d}_i = \lambda_{B|Aq}(f_{2iq}^* - \bar{f}_{2slq}^*)^2 + \bar{f}_{2slq}^{*(2)} - (\bar{f}_{2slq}^*)^2$, $\bar{f}_{2slq}^* = M_{slq}^{-1}\sum_{i\in slq} f_{2iq}^*$ *and* $\bar{f}_{2slq}^{*(2)} = M_{slq}^{-1}\sum_{i\in slq}(f_{2iq}^*)^2$.

3 Simulation Results

3.1 Specification of the simulation study

We used Monte Carlo simulation to compare the performances of the estimating function-based estimators defined by different choices of the weighting function in (4) and (7). The data model used in the simulation was

$$y_i = 1 + 5x_{1i} + 8x_{2i} + \epsilon_i.$$

The values x_{1i} were drawn from the standard normal distribution and the values x_{2i} were drawn from a normal distribution with a mean of 2 and a variance of 4, while the errorsϵ_i were independently drawn from the standard normal distribution.

The population was generated as three m-blocks, with linkage errors generated according to the correlated ELE model. In particular, the probabilities of correct linkage between $\boldsymbol{y}_q$ and $\boldsymbol{X}_{1q}$ were set to $\lambda_{A1} = 1$, $\lambda_{A2} = 0.95$ and $\lambda_{A3} = 0.85$, the probabilities of correct linkage between $\boldsymbol{X}_{1q}$ and $\boldsymbol{X}_{2q}$ were set to $\lambda_{B1} = 1$, $\lambda_{B2} = 0.85$ and $\lambda_{B3} = 0.8$, and the joint correct linkage probabilities ϕ_q were set to $\phi_2 = 0.845$ and $\phi_3 = 0.77$. Note that with these choices we then had $\lambda_{B|A2} = 0.89$ and $\lambda_{B|A3} = 0.91$.

Kim and Chambers (2012b) already showed that the estimating equation with ELE model corrects the bias due to linkage errors in multi-linked data under the assumption that the linkage errors in $\boldsymbol{B}_q$ is not depend on the linkage errors in $\boldsymbol{A}_q$. Our simulation will show that the estimation equation in Kim and Chambers (2012b) is not enough when the linkage errors in $\boldsymbol{B}_q$ and $\boldsymbol{A}_q$ are correlated. To do that, we will generate the data sets $\boldsymbol{y}$, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ in which the linkage errors in $\boldsymbol{B}_q$ and $\boldsymbol{A}_q$ are correlated. Then we will estimate the parameter $\boldsymbol{\beta} = (1, 5, 8)^T$ using the method in Kim and Chambers (2012b) as well as the method describe in this article to show that the incorrect independent assumption between $\boldsymbol{B}_q$ and $\boldsymbol{A}_q$ cause some bias in estimation of $\boldsymbol{\beta}$. There should be some bias as long as $\lambda_{Bq} \neq \lambda_{B|Aq}$.

We considered the case where these probabilities are known as well as the case where they were estimated from audit samples. These audit samples were defined by taking independent random samples of size 200 in the m-blocks corresponding to q= 2 and $q = 3$ and then estimating λ_{Aq} and ϕ_q as the proportion of correctly linked $x_1 - y$ and $x_1 - x_2 - y$ audit sample records respectively.

Two linkage scenarios were examined in the simulations.

- Scenario 1: This is the register to registers linking case considered in sections 2.1 and 2.2, with three m-blocks each of size 1000.
- Scenario 2: This is the sample to registers linking case considered in section 2.3, with three m-blocks each of size 2500. 500 records in each m-block were randomly assigned to be unlinkable. An independent random sample of size 1000 was then selected in each m-block, so that, on average, 800 of the sampled records were able to be linked (not necessarily correctly) to both registers in each simulation.

Three methods of estimating the regression parameter $\boldsymbol{\beta} = (1, 5, 8)^T$ were considered:
Scenario 1

ST The naive OLS estimator based on the linked data;

R-cor The solution to (4) with $\boldsymbol{G}_q^* = \left(\boldsymbol{X}_q^{E|A}\right)^T$ - the ratio type estimator;

A-cor The solution to (4) with $\boldsymbol{G}_q^* = \left(\boldsymbol{T}_{Aq}\boldsymbol{X}_q^{E|A}\right)^T$ - the implied Lahiri-Larsen estimator;

C-cor The solution to (4) with $\boldsymbol{G}_q^* = \left(\boldsymbol{T}_{Aq}\boldsymbol{X}_q^{E|A}\right)^T\left(\sigma^2\boldsymbol{I}_q + \boldsymbol{V}_{Aq}\boldsymbol{V}_{Cq}\right)^{-1}$ - the implied efficient estimator;

R-ind The ratio type estimator, incorrect independent assumption between $\boldsymbol{B}_q$ and $\boldsymbol{A}_q$;

A-ind The implied Lahiri-Larsen estimator, incorrect independent assumption;

C-ind The implied efficient estimator, incorrect independent assumption.

Scenario 2

ST The naive OLS estimator based on the linked data;

R-cor The solution to (7) with $\boldsymbol{G}_{slq}^* = \left(\tilde{\boldsymbol{X}}_q^{E|A}\right)^T$ - the ratio type estimator;

A-cor The solution to (7) with $\boldsymbol{G}_{slq}^* = \left(\tilde{\boldsymbol{T}}_{(sl)Aq}\tilde{\boldsymbol{X}}_{slq}^{E|A}\right)^T$ - the implied Lahiri-Larsen estimator;

C-cor The solution to (7) with $\boldsymbol{G}_{slq}^* = \left(\tilde{\boldsymbol{T}}_{(sl)Aq}\tilde{\boldsymbol{X}}_{slq}^{E|A}\right)^T\left(\sigma^2\boldsymbol{I}_{slq} + \tilde{\boldsymbol{V}}_{(sl)Aq}\tilde{\boldsymbol{V}}_{(sl)Cq}\right)^{-1}$ - the implied efficient estimator;

R-ind The ratio type estimator, incorrect independent assumption between $\boldsymbol{B}_q$ and $\boldsymbol{A}_q$;

A-ind The implied Lahiri-Larsen estimator, incorrect independent assumption;

C-ind The implied efficient estimator, incorrect independent assumption.

3.2 Simulation results

The main objective of this study is to show that, when the linked data set contains linkage errors inevitably due to probabilistic record linkage process and these linkage errors are correlated, our methods adjust bias due to linkage errors. In doing so, we examine the effect of the correlation measure, ϕ in this setting. In our previous study, we developed some error correction methods under the assumption that the linkage errors between $\boldsymbol{X}_1$ and $\boldsymbol{X}_2^*$ are independent to the linkage errors between $\boldsymbol{X}_1$ and $\boldsymbol{y}^*$. First thing we want to check is whether our methods from previous study can also adjust the bias even though we ignored the effect of ϕ. Therefore, in our simulation, we also added the simulation results using the methods from our previous study. The methods from our previous study can be obtained by replacing $\lambda_{B|Aq}$ withλ_{Bq} . For more details, see Kim and Chambers (2012b). With this in mind, let us explain the figure files. ST from the left hand side represents the results of nave estimator. Following R-ind, A-ind and C-ind after ST from the left side represent the results when we assume that the linkage errors are independent, while R-cor, A-cor and C-cor represent the results when we consider the correlation measures.

The two scenarios were independently simulated 1000 times and the estimates of $\boldsymbol{\beta}$ (based on ST, A-ind, R-ind and C-ind as well as R-cor, A-cor and C-cor) calculated using the linked data generated in each simulation. Table 1 shows the relative bias and RMSE for these estimators as well as the actual coverage of nominal 95 per cent confidence intervals based on estimates of the asymptotic variances shown in Theorems 1 and 2. Clearly, the estimators R-cor, A-cor and C-cor correct the bias due to incorrect linkages and their correlation measure, with the implied efficient estimator C-cor generally outperforming the Lahiri-Larsen estimator A-cor and the ratio type estimator R-cor in terms of relative root mean squared error. However, these results are not true if we ignore the correlation measure as we can see the biases from R-ind, A-ind and C-ind estimators. Note that R-ind, A-ind and C-ind estimators do not produce the biases for β_1, the coefficient for $\boldsymbol{X}_1$. The reason for this is that λ_{Aq} are correctly specified for R-ind, A-ind and C-ind estimators. However, because, in R-ind, A-ind and C-ind estimators, $\lambda_{B|Aq}$ are mis-specified as λ_{Bq}, R-ind, A-ind and C-ind estimators produce the biases for β_2 as well as β_0.

Tables 1, 2 here.

It is noteworthy that coverage rates for the estimators R-cor, A-cor and C-cor are consistently higher than 95%, indicating that the estimators of the asymptotic variances of these estimators are biased upwards. This does not appear to happen when only two data sets are linked, see Chambers (2009) and Kim and Chambers (2012a).

Figures 1-3 show box plots of the distributions of estimation errors underpinning the results shown in Tables 1 and 2. These distributions are for Scenario 2. The corresponding results for Scenario 1 were very similar. The overall superiority of method C-cor, as well as the increase in variability when the correct linkage probabilities are estimated, is clear.

Figure 1 here.

Figure 2 here.

Figure 3 here.

References

Chambers, R. (2009). Regression analysis of probability-linked data. Research series, Official Statistics 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**, 1183–1210.

Godambe, V. (1960). An optimum property of regular maximum likelihood estimation. *The Annals of Mathematical Statistics*, **31**, 1208–1211.

Kim, G. and Chambers, R. (2012a). Regression analysis under incomplete linkage. *Computational Statistics and Data Analysis*, **56**, 2756–2770.

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

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

A Apendix

A.1 Proof of Theorem 1

We use ∂_β to denote the partial differentiation operator with respect to $\boldsymbol{\beta}$ and adapt standard arguments used to obtain the asymptotic variance of the solution to an unbiased estimating equation. Furthermore, we only consider the case where $\boldsymbol{G}_q^*$ is a function of $\boldsymbol{X}_q^*$. Then, since

$$\partial_\beta \boldsymbol{H}^*(\boldsymbol{\beta}) = -\sum_q \boldsymbol{G}_q^* \boldsymbol{T}_{Aq} \boldsymbol{X}_q^{E|A}$$

we need only to show that in large samples the variance of $\boldsymbol{y}_q^*$ given $\boldsymbol{X}_q^*$ can be approximated by $\boldsymbol{V}(\boldsymbol{y}_q^*) = \sigma^2 \boldsymbol{I}_q + \boldsymbol{V}_{Aq} + \boldsymbol{V}_{Cq}$. Note that

$$\mathrm{Var}_{\boldsymbol{X}^*}(\boldsymbol{y}_q^*) = E_{\boldsymbol{X}^*}\Big\{\mathrm{Var}_{\boldsymbol{X}^*}\big(\boldsymbol{y}_q^*|\boldsymbol{A}_q\big)\Big\} + \mathrm{Var}_{\boldsymbol{X}^*}\Big\{E_{\boldsymbol{X}^*}\big(\boldsymbol{y}_q^*|\boldsymbol{A}_q\big)\Big\}. \tag{8}$$

Then, by (2) and (3),

$$E_{\boldsymbol{X}^*}\big(\boldsymbol{y}_q^*|\boldsymbol{A}_q\big) = \boldsymbol{A}_q E_{\boldsymbol{X}^*}\big(\boldsymbol{y}_q|\boldsymbol{A}_q\big) = \boldsymbol{A}_q \boldsymbol{X}_q^{E|A}\boldsymbol{\beta} = \boldsymbol{A}_q \boldsymbol{f}_q^{E|A}.$$

Hence $\boldsymbol{V}_{Aq} = \mathrm{Var}_{\boldsymbol{X}^*}\Big\{E_{\boldsymbol{X}^*}\big(\boldsymbol{y}_q^*|\boldsymbol{A}_q\big)\Big\} = \mathrm{Var}_{\boldsymbol{X}^*}\Big(\boldsymbol{A}_q \boldsymbol{f}_q^{E|A}\Big)$. A large sample approximation to this variance is set out equation (16) of Chambers (2009), and is given by

$$\boldsymbol{V}_{Aq} = (1-\lambda_{Aq})\,\mathrm{diag}\Big[\big\{\lambda_{Aq}(f_{iq}^{E|A} - \bar{f}_q^{E|A})^2 + \bar{f}_q^{E|A(2)} - (\bar{f}_q^{E|A})^2\big\}; i \in q\Big]. \tag{9}$$

In order to calculate $E_{\boldsymbol{X}^*}\Big\{\mathrm{Var}_{\boldsymbol{X}^*}\big(\boldsymbol{y}_q^*|\boldsymbol{A}_q\big)\Big\}$, we note that

$$\begin{aligned}\mathrm{Var}_{\boldsymbol{X}^*}\big(\boldsymbol{y}_q^*|\boldsymbol{A}_q\big) = \boldsymbol{A}_q\Big[E_{\boldsymbol{X}^*}\Big\{\mathrm{Var}_{\boldsymbol{X}^*}\big(\boldsymbol{y}_q|(\boldsymbol{B}|\boldsymbol{A})_q\big)\Big\}\Big]\boldsymbol{A}_q^T \\ + \boldsymbol{A}_q\Big[\mathrm{Var}_{\boldsymbol{X}^*}\Big\{E_{\boldsymbol{X}^*}\big(\boldsymbol{y}_q|(\boldsymbol{B}|\boldsymbol{A})_q\big)\Big\}\Big]\boldsymbol{A}_q^T.\end{aligned} \tag{10}$$

From (1) we see that

$$\mathrm{Var}_{\boldsymbol{X}^*}\big(\boldsymbol{y}_q|(\boldsymbol{B}|\boldsymbol{A})_q\big) = \sigma^2 \boldsymbol{I}_q.$$

Hence the first terms on the right hand side of (10) is

$$\boldsymbol{A}_q\Big[E_{\boldsymbol{X}^*}\Big\{\mathrm{Var}_{\boldsymbol{X}^*}\big(\boldsymbol{y}_q|(\boldsymbol{B}|\boldsymbol{A})_q\big)\Big\}\Big]\boldsymbol{A}_q^T = \boldsymbol{A}_q \sigma^2 \boldsymbol{I}_q \boldsymbol{A}_q^T = \sigma^2 \boldsymbol{A}_q \boldsymbol{I}_q \boldsymbol{A}_q^T = \sigma^2 \boldsymbol{I}_q. \tag{11}$$

In order to evaluate the second term on the right had side of (10) we note that, given $\boldsymbol{f}_{2q}^* = \boldsymbol{X}_{2q}^* \beta_2$,

$$\boldsymbol{V}_{Bq} = \mathrm{Var}_{\boldsymbol{X}^*}\Big\{E_{\boldsymbol{X}^*}\big[\boldsymbol{y}_q|(\boldsymbol{B}|\boldsymbol{A})_q\big]\Big\} = \mathrm{Var}_{\boldsymbol{X}^*}\Big((\boldsymbol{B}|\boldsymbol{A})_q^T \boldsymbol{f}_{2q}^*\Big)$$

which has the large sample approximation

$$\boldsymbol{V}_{Bq} = (1-\lambda_{B|Aq})\,\mathrm{diag}\Big[\big\{\lambda_{B|Aq}(f^*_{2iq} - \bar{f}^*_{2q})^2 + \bar{f}^{*(2)}_{2q} - (\bar{f}^*_{2q})^2\big\}\Big] = (1-\lambda_{B|Aq})\,\mathrm{diag}\Big[d_i; i \in d\Big].$$

Put $\boldsymbol{V}_{Cq} = E_{\boldsymbol{X}^*}\Big(\boldsymbol{A}_q\Big[\mathrm{Var}_{\boldsymbol{X}^*}\Big\{E_{\boldsymbol{X}^*}\big[\boldsymbol{y}_q|(\boldsymbol{B}|\boldsymbol{A})_q\big]\Big\}\Big]\boldsymbol{A}_q^T\Big)$. Then

$$\boldsymbol{V}_{Cq} = E_{\boldsymbol{X}^*}\Big(\boldsymbol{A}_q(1-\lambda_{B|Aq})\,\mathrm{diag}\Big[d_i; i \in d\Big]\boldsymbol{A}_q^T\Big) = (1-\lambda_{B|Aq})E_{\boldsymbol{X}^*}\Big(\boldsymbol{A}_q\,\mathrm{diag}\Big[d_i; i \in d\Big]\boldsymbol{A}_q^T\Big).$$

Put

$$e^{Aq}_{ij} = \lambda_{Aq}\boldsymbol{I}(i=j) + \frac{1-\lambda_{Aq}}{M_q - 1}\boldsymbol{I}(i \neq j).$$

Then, using similar arguments to that underpinning equations (66)-(67) of Chambers (2009), we can write down the large sample approximation

$$\begin{aligned} E_{\boldsymbol{X}^*}\Big(\boldsymbol{A}_q\,\mathrm{diag}\Big[d_i; i \in d\Big]\boldsymbol{A}_q^T\Big) &= \mathrm{diag}\Big(\sum_{i=1}^{M_q} d_i e^{Aq}_{ij}; i \in q\Big) \\ &= \mathrm{diag}\Big[(M_q - 1)^{-1}\big\{(\lambda_{Aq}M_q - 1)d_i + M_q(1-\lambda_{Aq})\bar{d}_q\big\}; i \in q\Big] \end{aligned}$$

so the corresponding large sample approximation to $\boldsymbol{V}_{Cq}$ is

$$\boldsymbol{V}_{Cq} = (1-\lambda_{B|Aq})E_{\boldsymbol{X}^*}\,\mathrm{diag}\Big[(M_q - 1)^{-1}\big\{(\lambda_{Aq}M_q - 1)d_i + M_q(1-\lambda_{Aq})\bar{d}_q\big\}; i \in q\Big]. \quad (12)$$

Combining (8), (9), (11) and (12), the required result follows immediately. Use of this asymptotic variance result to estimate the variance of $\hat{\boldsymbol{\beta}}^*$ follows directly. All that is required is an unbiased estimator of σ^2 based on the linked data. Here we note that we can write

$$(\boldsymbol{y}^*_q - \boldsymbol{f}^{E|A}_q)^T(\boldsymbol{y}^*_q - \boldsymbol{f}^{E|A}_q) = \boldsymbol{U}_{1q} + \boldsymbol{U}_{2q} + \boldsymbol{U}_{3q},$$

where

$$\begin{aligned} \boldsymbol{U}_{1q} &= \boldsymbol{y}_q^T\boldsymbol{A}_q^T\boldsymbol{A}_q\boldsymbol{y}_q - \boldsymbol{y}_q^T\boldsymbol{f}_q - \boldsymbol{f}_q^T\boldsymbol{y}_q + \boldsymbol{f}_q^T\boldsymbol{f}_q \\ \boldsymbol{U}_{2q} &= \boldsymbol{y}_q^T\boldsymbol{f}_q - \boldsymbol{f}_q^T\boldsymbol{f}_q \\ \boldsymbol{U}_{3q} &= \boldsymbol{f}_q^T\boldsymbol{y}_q - (\boldsymbol{y}^*_q)^T\boldsymbol{f}^{E|A}_q - (\boldsymbol{f}^{E|A}_q)^T\boldsymbol{y}^*_q + (\boldsymbol{f}^{E|A}_q)^T\boldsymbol{f}^{E|A}_q. \end{aligned}$$

Now

$$E_{\boldsymbol{X}^*}\Big(\sum_q \boldsymbol{U}_{1q}\Big) = E_{\boldsymbol{X}^*}\Big(\sum_q(\boldsymbol{y}_q - \boldsymbol{f}_q)^T(\boldsymbol{y}_q - \boldsymbol{f}_q)\Big) = N\sigma^2.$$

Also

$$E_{\boldsymbol{X}^*}\Big(\sum_q \boldsymbol{U}_{2q}\Big) = E_{\boldsymbol{X}^*}\Big((\boldsymbol{y}_q - \boldsymbol{f}_q)^T\boldsymbol{f}_q\Big) = E_{\boldsymbol{X}^*}\Big(\boldsymbol{\epsilon}_q^T\boldsymbol{f}_q\Big) = 0$$

while, after re-arranging terms, we have

$$\boldsymbol{U}_{3q} = \big\{\boldsymbol{y}_q^T \boldsymbol{f}_q^{E|A} - (\boldsymbol{y}_q^*)^T \boldsymbol{f}_q^{E|A}\big\} + \big\{(\boldsymbol{f}_q^{E|A})^T \boldsymbol{f}_q^{E|A} - (\boldsymbol{f}_q^{E|A})^T \boldsymbol{y}_q^*\big\} + \Delta_q,$$

where

$$E_{\boldsymbol{X}^*}\Big(\Delta_q\Big) = E_{\boldsymbol{X}^*}\Big(\big\{\boldsymbol{y}_q^T - (\boldsymbol{f}_q^{E|A})^T\big\}\boldsymbol{f}_q + \big\{(\boldsymbol{f}_q^{E|A})^T - \boldsymbol{y}_q^T\big\}\boldsymbol{f}_q^{E|A}\Big) = 0.$$

Thus,

$$E_{\boldsymbol{X}^*}\Big(\boldsymbol{U}_{3q}\Big) = E_{\boldsymbol{X}^*}\Big[\big\{\boldsymbol{y}_q^T \boldsymbol{f}_q^{E|A} - (\boldsymbol{y}_q^*)^T \boldsymbol{f}_q^{E|A}\big\} + \big\{(\boldsymbol{f}_q^{E|A})^T \boldsymbol{f}_q^{E|A} - (\boldsymbol{f}_q^{E|A})^T \boldsymbol{y}_q^*\big\}\Big] = 2(\boldsymbol{f}_q^{E|A})^T(\boldsymbol{I}_q - \boldsymbol{T}_{Aq})\boldsymbol{f}_q^{E|A}.$$

Hence an unbiased estimator of σ^2 is

$$\hat{\sigma}^2 = N^{-1}\sum_q \Big\{(\boldsymbol{y}_q^* - \hat{\boldsymbol{f}}_q^{E|A})^T(\boldsymbol{y}_q^* - \hat{\boldsymbol{f}}_q^{E|A}) - 2(\hat{\boldsymbol{f}}_q^{E|A})^T(\boldsymbol{I}_q - \boldsymbol{T}_{Aq})\hat{\boldsymbol{f}}_q^{E|A}\Big\}.$$

A.2 Proof of Lemma 1

A first order Taylor series approximation is of the form

$$\begin{aligned} 0 &= \boldsymbol{H}^*(\hat{\boldsymbol{\beta}}, \hat{\lambda}_A, \hat{\phi}) \\ &\approx \boldsymbol{H}^*(\boldsymbol{\beta}_0, \lambda_{0,A}, \phi_0) + \partial_{\boldsymbol{\beta}}\boldsymbol{H}^*(\boldsymbol{\beta}_0, \lambda_{0,A}, \phi_0)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \\ &\quad + \partial_{\lambda_A}\boldsymbol{H}^*(\boldsymbol{\beta}_0, \lambda_{0,A}, \phi_0)(\hat{\lambda}_A - \lambda_{0,A}) + \partial_{\phi}\boldsymbol{H}^*(\boldsymbol{\beta}_0, \lambda_{0,A}, \phi_0)(\hat{\phi} - \phi_0), \end{aligned}$$

where $\boldsymbol{\beta}_0$, $\lambda_{0,A}$ and ϕ_0 denote the true values of $\boldsymbol{\beta}$, λ_A and ϕ respectively. It leads us

$$\begin{aligned} \mathrm{Var}_{\boldsymbol{X}^*}(\hat{\boldsymbol{\beta}}) = \big[\partial_{\boldsymbol{\beta}}\boldsymbol{H}_0^*\big]^{-1}\Big[&\mathrm{Var}_{\boldsymbol{X}^*}(\boldsymbol{H}_0^*) + \big(\partial_{\lambda_A}\boldsymbol{H}_0^*\big)\mathrm{Var}_{\boldsymbol{X}^*}(\hat{\lambda}_A)\big(\partial_{\lambda_A}\boldsymbol{H}_0^*\big)^T + \big(\partial_{\phi}\boldsymbol{H}_0^*\big)\mathrm{Var}_{\boldsymbol{X}^*}(\hat{\phi})\big(\partial_{\phi}\boldsymbol{H}_0^*\big)^T \\ &+ \big(\partial_{\lambda_A}\boldsymbol{H}_0^*\big)\mathrm{Cov}\boldsymbol{X}^*(\hat{\lambda}_A, \hat{\phi})\big(\partial_{\phi}\boldsymbol{H}_0^*\big)^T + \big(\partial_{\phi}\boldsymbol{H}_0^*\big)\mathrm{Cov}\boldsymbol{X}^*(\hat{\phi}, \hat{\lambda}_A)\big(\partial_{\lambda_A}\boldsymbol{H}_0^*\big)^T\Big]\Big(\big[\partial_{\boldsymbol{\beta}}\boldsymbol{H}_0^*\big]^{-1}\Big)^T, \end{aligned}$$

where $\boldsymbol{H}_0^*$ denote $\boldsymbol{H}^*(\boldsymbol{\beta}_0, \lambda_{0,A}, \phi_0)$. Let $\partial_1 = \partial_{\lambda_A}$ and $\partial_2 = \partial_{\phi}$. Then using the definition of $\partial_{\boldsymbol{\beta}}\boldsymbol{H}_0^*$ and $\mathrm{Var}_{\boldsymbol{X}^*}(\boldsymbol{H}_0^*)$ from the proof of Theorem 1, the asymptotic variance of $\hat{\boldsymbol{\beta}}^*$ is

$$\boldsymbol{V}(\hat{\boldsymbol{\beta}}^*) = \boldsymbol{D}\Big[\sum_q \Big(\boldsymbol{G}_q^*\boldsymbol{V}(\boldsymbol{y}_q^*)\boldsymbol{G}_q^{*T} + \sum_{i=1}^2\sum_{j=1}^2 \big(\partial_i\boldsymbol{H}^*\big)J_{ijq}\big(\partial_j\boldsymbol{H}^*\big)^T\Big)\Big]\boldsymbol{D}^T$$

where $\boldsymbol{D} = \Big[\sum_q \boldsymbol{G}_q^*\boldsymbol{T}_{Aq}\boldsymbol{X}_q^{E|A}\Big]^{-1}$ and $\boldsymbol{J}_q = [J_{ijq}] = \mathrm{Cov}(\hat{\lambda}_{Aq}, \hat{\phi}_q)$.

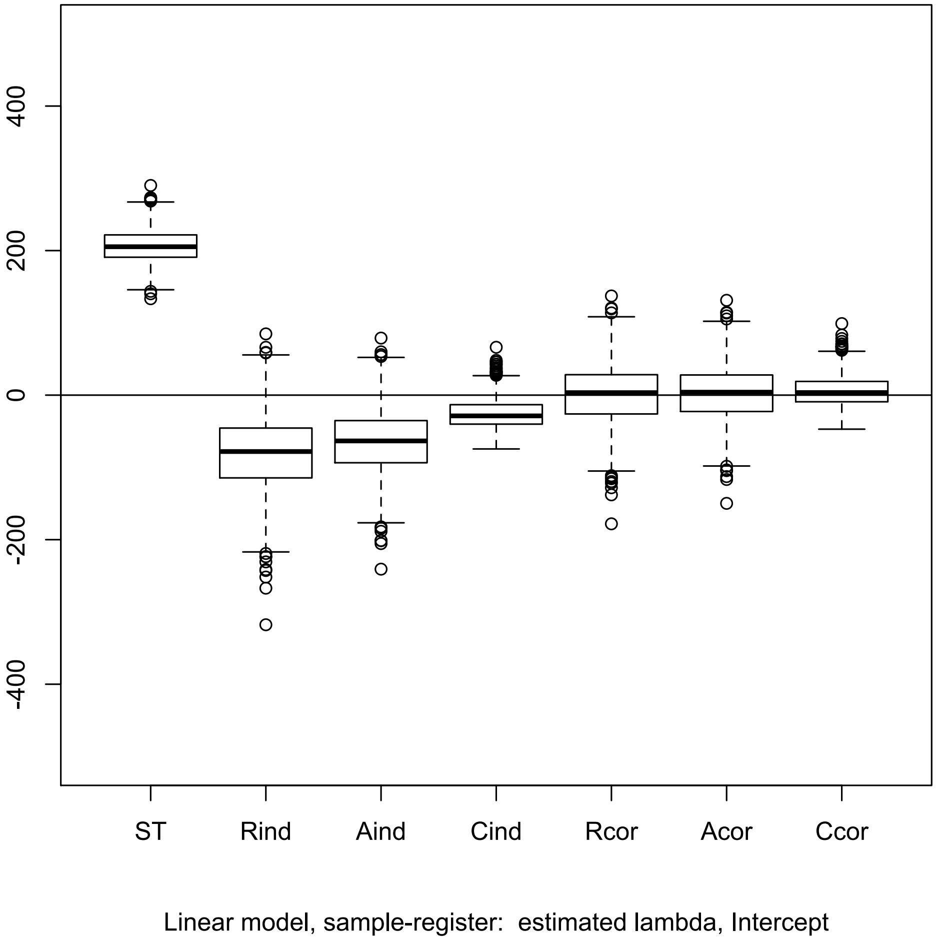


Figure 1: Simulated percentage relative errors for intercept coefficient β_0 in linear regression under scenario 2.

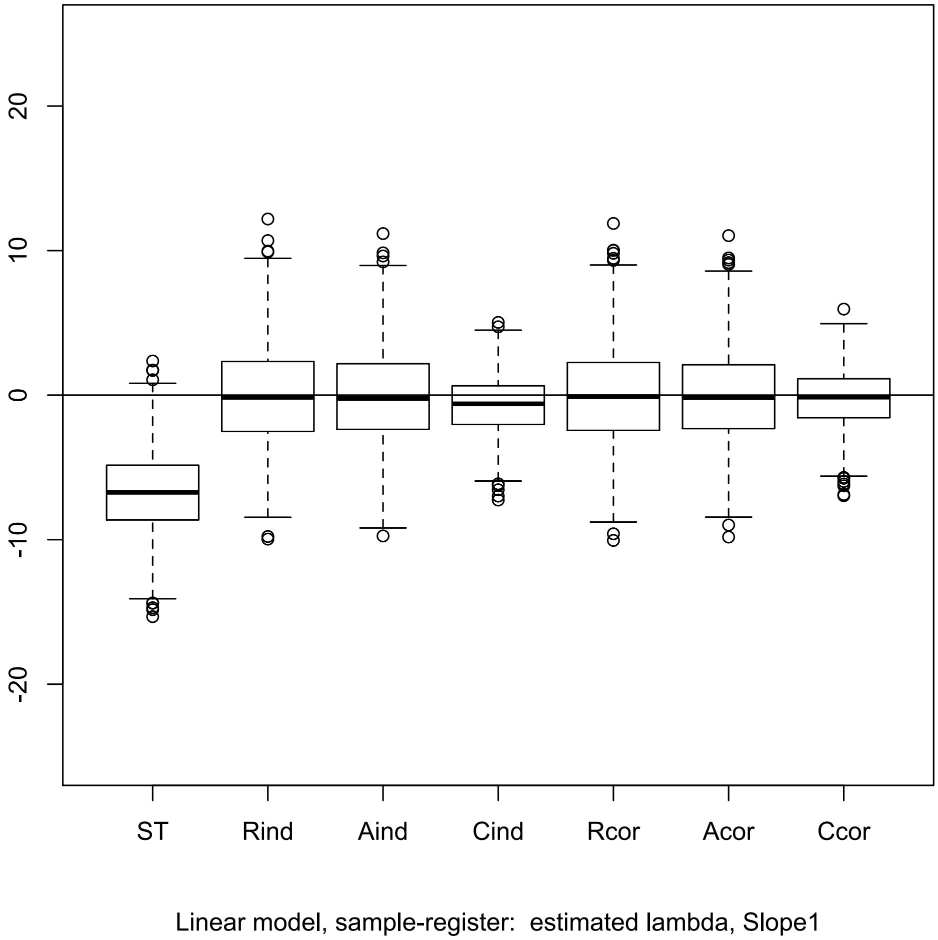
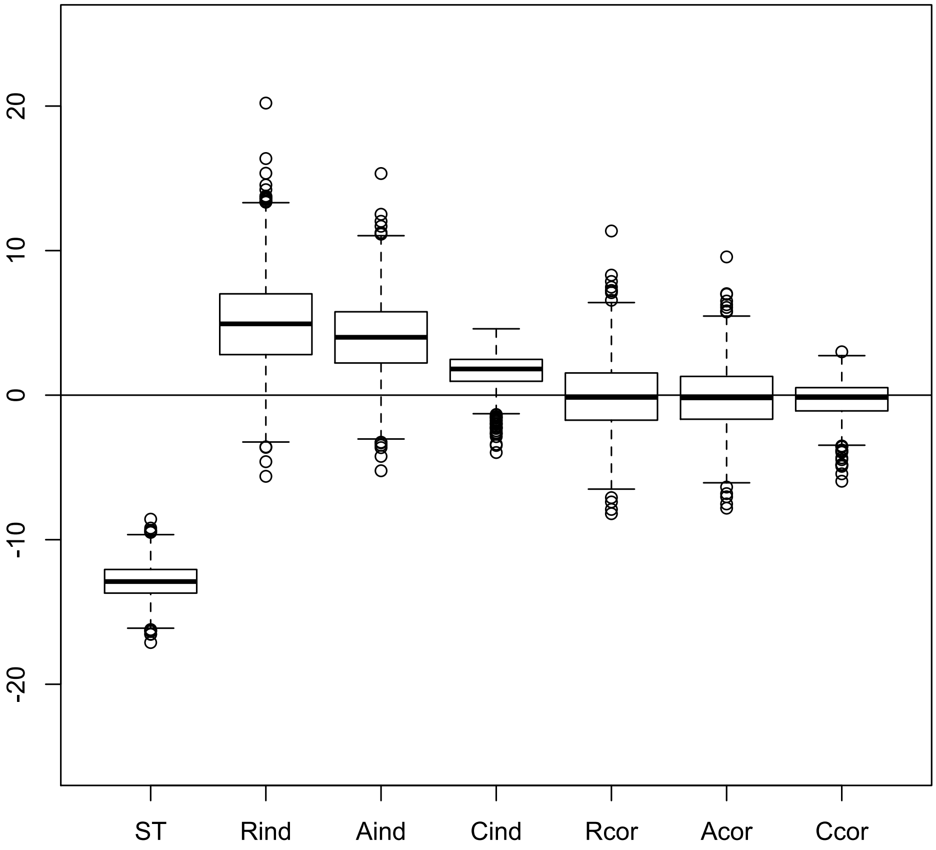


Figure 2: Simulated percentage relative errors for the first slope coefficient $\boldsymbol{\beta}_1$ in linear regression under scenario 2.



Linear model, sample-register: estimated lambda, Slope2

Figure 3: Simulated percentage relative errors for the second slope coefficient $\boldsymbol{\beta}_2$ in linear regression under scenario 2.

Table 1: Relative bias and relative RMSE (both expressed in percentage terms) for parameter estimates investigated in the simulation study when the data sets are all registers. Empirical coverages (expressed in percentage terms) of normal theory-based nominal 95 per cent confidence intervals are also shown. For estimators A and C, these are based on estimators of the asymptotic variances of these estimators as shown in Theorem 1.

Estimator	Relative Bias		Relative RMSE		Coverage	
	λ known	λ unknown	λ known	λ unknown	λ known	λ unknown
Scenario 1- estimation of β_0						
ST	204.77	204.77	205.58	205.58	0.0	0.0
R-ind	-83.08	-82.53	85.78	90.86	8.5	100.0
A-ind	-70.11	-68.35	72.50	75.33	15.3	100.0
C-ind	-32.34	-29.87	33.28	32.86	84.2	100.0
R-cor	-0.91	-0.79	20.07	29.65	98.3	90.0
A-cor	-0.81	-0.08	18.26	26.93	98.9	92.3
C-cor	0.37	2.11	9.56	14.74	100.0	99.3
Scenario 1- estimation of β_1						
ST	-6.73	-6.73	16.43	16.43	38.3	38.3
R-ind	-0.06	-0.09	6.98	7.27	96.6	100.0
A-ind	-0.05	-0.10	6.63	6.90	96.7	100.0
C-ind	-0.51	-0.59	4.29	4.45	99.2	100.0
R-cor	-0.07	-0.10	6.95	7.24	95.8	95.6
A-cor	-0.06	-0.10	6.59	6.87	96.5	96.4
C-cor	-0.03	-0.13	4.16	4.30	99.4	99.3
Scenario 1- estimation of β_2						
ST	-12.79	-12.79	36.32	36.32	0.0	0.0
R-ind	5.19	5.16	15.14	16.05	2.5	100.0
A-ind	4.38	4.27	12.80	13.31	6.5	100.0
C-ind	2.02	1.87	5.84	5.77	43.7	100.0
R-cor	0.06	0.05	3.53	5.23	94.9	80.1
A-cor	0.05	0.00	3.21	4.74	96.1	83.0
C-cor	-0.02	-0.13	1.53	2.52	100.0	97.1

Table 2: Relative bias and relative RMSE (both expressed in percentage terms) for parameter estimates investigated in the simulation study when the data sets are samples. Empirical coverages (expressed in percentage terms) of normal theory-based nominal 95 per cent confidence intervals are also shown. For estimators A and C, these are based on estimators of the asymptotic variances of these estimators as shown in Lemma 1.

Estimator	Relative Bias		Relative RMSE		Coverage	
	λ known	λ unknown	λ known	λ unknown	λ known	λ unknown
			Scenario 2- estimation of β_0			
ST	206.14	206.14	207.43	207.43	0.0	0.0
R-ind	-80.69	-80.25	84.80	96.29	14.0	100.0
A-ind	-67.96	-64.65	71.87	78.32	22.3	100.0
C-ind	-31.45	-25.91	33.46	33.44	78.7	100.0
R-cor	1.16	1.22	24.99	40.96	94.4	100.0
A-cor	1.13	2.65	23.27	37.41	95.6	100.0
C-cor	1.37	5.36	12.99	21.89	99.7	100.0
			Scenario 2- estimation of β_1			
ST	-6.69	-6.69	16.35	16.35	37.5	37.5
R-ind	-0.03	-0.02	7.09	7.83	96.7	100.0
A-ind	-0.03	-0.06	6.77	7.43	97.2	100.0
C-ind	-0.49	-0.65	4.36	4.75	99.9	100.0
R-cor	-0.04	-0.02	7.01	7.74	95.9	100.0
A-cor	-0.03	-0.07	6.69	7.35	96.1	100.0
C-cor	-0.01	-0.21	4.24	4.62	99.7	100.0
			Scenario 2- estimation of β_2			
ST	-12.89	-12.89	36.62	36.62	0.0	0.0
R-ind	5.05	5.02	14.86	16.94	5.4	100.0
A-ind	4.25	4.05	12.54	13.74	8.8	100.0
C-ind	1.97	1.62	5.74	5.78	48.5	100.0
R-cor	-0.07	-0.07	3.91	7.01	91.9	100.0
A-cor	-0.07	-0.16	3.56	6.35	93.0	100.0
C-cor	-0.08	-0.33	1.76	3.63	99.6	100.0