NIASRA

NATIONAL INSTITUTE FOR APPLIED STATISTICS RESEARCH AUSTRALIA

National Institute for Applied Statistics Research Australia

The University of Wollongong

Working Paper

16-13

Bias Reduction for Correlated Linkage Error

Gunky Kim and Raymond Chambers

National Institute for Applied Statistics Research Australia, University of Wollongong, Wollongong NSW 2522. Phone +61 2 4221 5435, Fax +61 2 4221 4845. Email: anica@uow.edu.au

Bias Reduction for Correlated Linkage Error

Gunky Kim and Raymond Chambers
National Institute for Applied Statistical Research Australia
University of Wollongong

Abstract

Linked data sets are often multi-linked, i.e. they are created by matching records from three or more data sources. In such cases, probability-based methods for record linkage may lead to correlated linkage errors. Furthermore, it is often the case that not all records can be linked, due to the linking procedure not being able to find suitable matches in at least one of the data sources. This can be simply because the data source is a sample, and so does not contain the requisite matching records. More generally, however, the probability algorithm used to create the matches may not be able to find another record that meets the minimum criterion for matching. In this paper we develop methods for carrying out regression analysis using multi-linked data that allow for both correlated linkage error as well as unlinked records. We also investigate the role of auxiliary information in this process, focussing on the situation where marginal distribution information from the data sets being linked is available. Our simulation results show that recently published bias reduction methods based on an assumption of independent linkage errors can lead to insufficient bias correction in the correlated case, and that a modified approach which allows for correlated linkage errors is superior. We also show that auxiliary marginal information about the data sets being linked can help further reduce the bias due to both non-linkage and linkage errors.

key words: probabilistic record linkage; auxiliary population information; multi-linked data; linear regression; estimating equations.

1 Introduction

Probabilistic data linkage is now widely used, especially in areas (e.g. health science) where direct measurement is impossible or extremely costly. For example, Brook *et al.* (2008) claim

that linked health record data sets produced by the Western Australian Data Linkage system (Holman *et al.*, 1999) led to 708 research outputs over the period 1995 - 2003. Following the pioneering work of Fellegi and Sunter (1969), methods for linking records stored on different data bases have been extensively developed, with practical applications including census data (Jaro, 1989) and population health data (Newcombe, 1988). In this paper, we focus on one important application where different data sets relating to the same individuals at different points in time or collected from different institutions are 'multi-linked' to provide a synthetic data record for each individual that covers the different data sets. In particular, the Oxford Record Linkage system, the Scottish Record Linkage system and the Canadian Mortality Data Base all use forms of multi-linking to create linked longitudinal data records.

From an analytic perspective, multi-linking is of no concern when error-free unique identifiers are present in each data set being linked. However, when multi-linking is probabilistically based there is always the possibility of linkage errors in the merged data. Such errors can, for example, lead to a linked longitudinal record being actually made up of data items from different individuals. This is also the case when the identifiers used in linking contain errors. For example, Adams *et al.* (1997) found that linking based on Social Security Number is not adequate, and recommend the use of probabilistic data linkage, even though some incorrect linkage is then unavoidable. A similar result was reported in Rotermann (2009). The Census Data Enhancement project of the Australian Bureau of Statistics aims to link data from the same individuals over a number of censuses in order to create a tool for research into the longitudinal dynamics of the Australian population. An initial test for this project (Bishop and Khoo, 2007) showed that 13% of the test records were incorrectly linked under optimal conditions, i.e. when names and address were used in the matching process. These figures are representative of those obtained in similar Australian studies. Holman *et al.* (1999) reported 87% correct linkage for Western Australia hospital record linkages in 1996-1997, while linked hospital morbidity data in Victoria in 1993-94 showed a 78-86% correct match rate.

Possible bias due to linkage errors is discussed in Gomatam *et al.* (2002), Nitsch *et al.* (2006) and Fair *et al.* (2000), while Krewski *et al.* (2005) explores the impact of linkage errors on statistical inferences in cohort mortality studies. They show that record linkage errors lead to bias and additional variation, and that this increases as linkage errors increase.

Since some degree of linkage error is inevitable in most linked data sets, one needs to adopt

methods that correct the resulting bias when analysing these data. Scheuren and Winkler (1993), Scheuren and Winkler (1997) and Lahiri and Larsen (2005) consider the case of linear regression analysis using data sourced from two probabilistically linked sets of records, and describe methods for correcting bias due to linkage errors. Their results assume that all linkage probabilities associated with the probabilistic matching process are known, and that the linkage is complete, i.e. every unit in the files being matched is linked. However, in many cases the analyst does not have access to the entire set of linkage probabilities associated with the probabilistic matching process, e.g. because of data confidentiality requirements. For example, Kelman *et al.* (2002) states that technicians involved in the linking carried out by the West Australian Data Linkage System are not permitted to take part in any subsequent analysis of the linked data. This confidentiality requirement therefore prohibits researchers using the linked data to access all the linkage probabilities or all the population records used in the probabilistic matching process. In this case the bias correction methods of Scheuren and Winkler (1993), Scheuren and Winkler (1997) and Lahiri and Larsen (2005) cannot be used. In effect, one requires bias-correction methods that can be implemented by a 'secondary analyst', i.e. one who does not know all the linkage probabilities and who does not have access to all the population records used in the matching process.

In previous work (Kim and Chambers, 2012b), we assume a simple first order exchangeable model for linkage errors that depends only on the marginal probability of correct linkage, and use this to describe methods for correcting the bias due to linkage errors when multiple data sets are probabilistically multi-linked. A key assumption in this development is that linkage errors are pairwise independent. A more realistic scenario, however, is where linkage errors are pairwise dependent, in the sense that if the records corresponding to two different individuals in distinct data sets $\mathcal{A}$ and $\mathcal{B}$ are incorrectly linked, then it is likely that the records for the same two individuals in distinct data sets $\mathcal{A}$ and $\mathcal{C}$ 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.

Provided the linkage error model is correctly specified, bias correction methods work well when the linked data set is large and has characteristics that are representative of those of the source data sets. However, this is not the case when the linked data set is small, which can happen when the linking procedure is unable to find matches for a significant number of records. Bias correction methods based on large sample approximations are not adequate in such cases, particularly if the characteristics of the correctly linked, incorrectly linked and unlinked records all differ substantially. Here appropriate bias corrections depend crucially on correct specification of both the linkage error process as well as the process that drives creation of the links in the first place. For example, Nitsch *et al.* (2006) points out that non-linkage, like non-response, can be informative and should not be ignored. Similarly, Bishop and Khoo (2007) note that the largest source of error that they observed when analysing linked Census records was that associated with non-linked records, i.e. records that could not be adequately matched in the probabilistic linkage process. In order to ameliorate this bias, we propose to use population auxiliary information, e.g. population summary statistics like totals or means, as additional information in the bias correction process. Although not likelihood-based, our approach is motivated by the ideas set out in Section 8.5 of Chambers *et al.* (2012).

The structure of the paper is as follows. We introduce our notation and basic assumptions in Subsection 1.1 below. Section 2 then describes the methodology underpinning the suggested bias correction technique. In particular, in Subsections 2.1 - 2.3 we develop this methodology for the case where only linked sample data are available. The extension to the situation where population auxiliary information is also available is then set out in Subsection 2.4. Section 3 contains simulation results and Section 4 concludes the paper with a discussion of potential extensions of our results.

1.1 Notation and assumptions

The assumptions we make here are the same as in Kim and Chambers (2012b). 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 $E_{\boldsymbol{X}}(\boldsymbol{y}) = f(\boldsymbol{X};\boldsymbol{\beta}) = \boldsymbol{f}$ 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 them using the data in two or more population registers. To fix concepts, and for simplicity of exposition, we assume here that the regression model of interest is the linear model

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

where $\boldsymbol{y}$, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ denote data stored on three separate population registers and $\mathbf{1}$ denotes the unit vector of order N, the population size. Note, however, that our subsequent development is quite general, with inference about $\boldsymbol{\beta}$ based on the solution of an unbiased estimating equation, so 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. It is assumed that the linked data needed to fit (1) 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 underpinning $\boldsymbol{y}$, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ and the data sets themselves, with one of the three registers assumed to be the 'benchmark' register, i.e. all linkage errors are defined relative to the way that it orders the population units. Without loss of generality we take $\boldsymbol{X}_1$ to be this benchmark register, so linkage errors arise between $\boldsymbol{y}$ and $\boldsymbol{X}_1$ and between $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. We initially consider the situation where there is complete linkage, i.e. there is one to one matching of all records in each register. 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 complete linking of 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 an m-block. This is equivalent to assuming that records in distinct m-blocks are never linked, and that the records for any population unit in an m-block are contained in that m-block on 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

Optimal probability-based linkage is based on maximising the probability of a declared link being correct. Unfortunately, most practical implementations require one to trade off the number of links made against their accuracy. Any implementation of probabilistic linkage will therefore result in unmade linkages or non-linkages, as well as linkage errors in those cases where linkages are actually made. If the analyst has access to the complete set of joint linkage probabilities as well as all the population records used in the probabilistic matching

process, then he or she can apply the bias correction method described in Lahiri and Larsen (2005). However, this type of information is typically unavailable to researchers not involved in the actual data linkage process. Below we show that the bias caused by linkage errors can be corrected if we know the marginal probability of correct linkage for any particular record. These marginal probabilities can be obtained from the record linkage system without violating confidentiality restrictions or can be estimated using small random 'audit samples' that identify whether or not particular links are correct. Given this information, 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 linked to two other registers, we start by considering the case where three registers are completely linked. This allows us to introduce the basic ideas used in the subsequent methodological development.

2.1 A model for correlated linkage error

In this Subsection 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 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. We use the exchangeable linkage errors (ELE) model of Chambers (2009) to specify $\boldsymbol{T}_{Aq} = E_{\boldsymbol{X}^*}(\boldsymbol{A}_q)$ and $\boldsymbol{T}_{Bq} = E_{\boldsymbol{X}^*}(\boldsymbol{B}_q)$. Under this model, any linked record i in m-block q has the same probability of being correctly linked, and is also equally likely to be incorrectly linked to any other record in the same m-block. A similar model is used in Neter *et al.* (1965). 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, $\lambda_{Aq} = \Pr(\boldsymbol{x}_{1iq}, y_{jq} \text{ correctly linked})$ and $\gamma_{Aq} = \Pr(\boldsymbol{x}_{1iq}, y_{jq} \text{ incorrectly linked}, i \neq j) = (M_q - 1)^{-1}(1 - \lambda_{Aq})$, 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 $\boldsymbol{B}_q$ given $\boldsymbol{A}_q$ we extend the ELE model, assuming 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 this correlated ELE model for $\boldsymbol{A}_q$ and $\boldsymbol{B}_q$,

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

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

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

then under the linear model (1)

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

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

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

where $\boldsymbol{f}_q^* = \boldsymbol{X}_q^*\boldsymbol{\beta}$ and $\boldsymbol{G}_q^* = \boldsymbol{X}_q^{*T}$. Here $\boldsymbol{X}_q^* = (\mathbf{1}_q, \boldsymbol{X}_{1q}, \boldsymbol{X}_{2q}^*)$. 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

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

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*

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

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

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

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

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

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

Note:

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

$$\tilde{\sigma}^2 = N^{-1}\Big[\sum_q (\boldsymbol{y}^*_q - \boldsymbol{f}^{E|A}_q)^T(\boldsymbol{y}^*_q - \boldsymbol{f}^{E|A}_q) - 2\sum_q (\boldsymbol{f}^{E|A}_q)^T(\boldsymbol{I}_q - \boldsymbol{T}_{Aq})\boldsymbol{f}^{E|A}_q\Big]. \qquad (5)$$

   We can therefore estimate $\boldsymbol{V}(\boldsymbol{y}^*_q)$ above by substituting $\hat{\boldsymbol{\beta}}^*$ for $\boldsymbol{\beta}$ in the definitions of $\boldsymbol{f}^{E|A}_q$, $\boldsymbol{f}^*_{2q}$ 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 $\boldsymbol{G}^*_q = (\boldsymbol{T}_{Aq}\boldsymbol{X}^{E|A}_q)^T$. The optimal weighting function, i.e. the one that minimises the asymptotic variance of $\hat{\boldsymbol{\beta}}^*$ (Godambe, 1960), depends on the unknown model parameters and is given by

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

An iterative approach to calculating $\boldsymbol{G}_q^*$, combining the estimator (5) of σ^2 with this optimal weighting function specification, 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, in which case estimates of these probabilities are required. 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 Corollary to Theorem 1. Its proof is in the Appendix.

Corollary 2. *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 $\boldsymbol{D} = \Big[\sum_q \boldsymbol{G}_q^*\boldsymbol{T}_{Aq}\boldsymbol{X}_q^{E|A}\Big]^{-1}$, $\boldsymbol{J}_q = [J_{ijq}] = Cov(\hat{\lambda}_{Aq}, \hat{\phi}_q)$ *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 calculated using the number of incorrect linkages observed in a random 'audit sample' of records in m-block q taken from the linked data base $\big[\boldsymbol{y}^*\ \boldsymbol{X}_1\ \boldsymbol{X}_2^*\big]$.

2.3 Incomplete sample to registers linkage

We now consider the more realistic case where 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 $\boldsymbol{X}_{1suq}$ (i.e. we ignore partial linkages). 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

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

Under the ELE model, this becomes

$$\boldsymbol{H}_{sl}^*(\boldsymbol{\beta}) = \sum_q \boldsymbol{G}_{slq}^* \Big[\boldsymbol{y}_{slq}^* - \Big(\frac{\lambda_{Aq} M_q - 1}{M_q - 1}\Big) \boldsymbol{f}_{slq}^{E|A} - \Big(\frac{1 - \lambda_{Aq}}{M_q - 1}\Big) \mathbf{1}_{slq} \mathbf{1}_q^T \boldsymbol{f}_q^{E|A}\Big]. \tag{6}$$

This modified estimating function depends on the value of $\mathbf{1}_q^T \boldsymbol{f}_q^{E|A}$, which is a population, rather than a sample, quantity. Given assumptions 4 and 5 at the end of section 1, we can estimate $\mathbf{1}_q^T \boldsymbol{f}_q^{E|A}$ using the weighted sample estimate $\tilde{\boldsymbol{w}}_{slq}^T \boldsymbol{f}_{slq}^{E|A}$, where $\tilde{\boldsymbol{w}}_{slq} = M_{sq} M_{slq}^{-1} \boldsymbol{w}_{slq}$. 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 (6) can be replaced by

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

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, (7) requires further approximation, since the linear regression model assumption, together with (2), implies

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

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 (7) with

$$\tilde{\boldsymbol{H}}_{sl}^*(\boldsymbol{\beta}) = \sum_q \boldsymbol{G}_{slq}^* \big(\boldsymbol{y}_{slq}^* - \tilde{\boldsymbol{T}}_{(sl)Aq} \tilde{\boldsymbol{f}}_{slq}^{E|A} \big) \tag{8}$$

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 (8) to zero. We do not provide its proof since it is along the same lines as that of Theorem 1 and Corollary 2. We also use the same notation as in these results.

Theorem 3. *Let $\hat{\boldsymbol{\beta}}_s^*$ denote the solution to setting the modified estimating function (8) to zero. Given the assumptions (4) and (5) at the end of section 1 as well as those implicit in Theorem 1 and Corollary 2, 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* $\bar{f}_{2slq}^{*(2)} = M_{slq}^{-1}\sum_{i\in slq}(f_{2iq}^*)^2$.

2.4 Calibrating to population summary information

The target of inference for the estimating equation approach used in this paper is the parameter $\boldsymbol{\beta}$ that is the solution to setting the expected value of the estimating function (4) to zero, where this expectation is with respect to the stochastic process that defines the regression of $\boldsymbol{y}$ on $\boldsymbol{X}$ as well as the random processes that underpin the linkage. When the number of linked sample units is small, the large sample approximations used to justify replacing (4) by (8) can be inaccurate. In such cases the solution to the estimating equation

defined by setting (8) to zero can be biased for $\boldsymbol{\beta}$. Similarly, the weighting used to adjust for non-linkage may not be effective in small samples, which can also lead to (8) being biased. In both cases, this bias can be viewed as symptomatic of a lack of representativeness of the linked sample units. In this Subsection we therefore explore how summary information about the linked registers can be used to 'calibrate' inference based on (8), thereby reducing its potential small sample bias. In particular, we assume that population average values are known for each m-block, noting that release of such summary information will typically not contravene the confidentiality of the data held in the linked registers.

Let $\bar{\boldsymbol{X}}_q$ and $\bar{\boldsymbol{y}}_q$ respectively denote the known mean values of the m-block components $\boldsymbol{X}_q$ and $\boldsymbol{y}_q$ of the registers $\boldsymbol{X}$ and $\boldsymbol{y}$. This information can then be used in two ways. The first is to remove the need for the approximations used in going from (6) to (8). This follows from noting that when $\bar{\boldsymbol{X}}_q$ is known, and the ELE model holds, then $\bar{\boldsymbol{X}}_q^{E|A} = \bar{\boldsymbol{X}}_q$. Furthermore, in this case $\boldsymbol{X}_{slq}^{E|A} = \big(\mathbf{1}_{slq}, \boldsymbol{X}_{1slq}, \boldsymbol{T}_{(sl)B|Aq}\boldsymbol{X}_{2q}^*\big)$ where

$$\boldsymbol{T}_{(sl)B|Aq}\boldsymbol{X}_{2q}^* = (\lambda_{B|Aq} - \gamma_{B|Aq})\boldsymbol{X}_{2slq}^* + M_q\gamma_{B|Aq}\mathbf{1}_{slq}\bar{\boldsymbol{X}}_{2q},$$

with $\gamma_{B|Aq} = (1 - \lambda_{B|Aq})/(M_q - 1)$. Consequently, rather than using the approximation (8), we can replace (6) by

$$\boldsymbol{H}_{sl}^*(\boldsymbol{\beta}) = \sum_q \boldsymbol{G}_{slq}^*\Big[\boldsymbol{y}_{slq}^* - \big\{(\lambda_{Aq} - \gamma_{Aq})\boldsymbol{X}_{slq}^{E|A}\boldsymbol{\beta} + M_q\gamma_{Aq}\mathbf{1}_{slq}\bar{\boldsymbol{X}}_q^T\boldsymbol{\beta}\big\}\Big].$$

Secondly, when $\bar{\boldsymbol{X}}_q$ and $\bar{\boldsymbol{y}}_q$ are known, we can use the implied information about the average residual defined by the non-sample records to add a further, calibrating, term to this estimating function. In order to do this, we define $\bar{\boldsymbol{y}}_{rq}^*$ and $\bar{\boldsymbol{X}}_{rq}^*$ via the identities

$$(M_q - M_{slq})\bar{\boldsymbol{y}}_{rq}^* = M_q\bar{\boldsymbol{y}}_q - M_{slq}\bar{\boldsymbol{y}}_{slq}^*$$

and

$$(M_q - M_{slq})\bar{\boldsymbol{X}}_{rq}^* = M_q(1 - M_{slq}\gamma_{Aq})\bar{\boldsymbol{X}}_q - M_{slq}(\lambda_{Aq} - \gamma_{Aq})\bar{\boldsymbol{X}}_{slq}^{E|A}$$

respectively. Note that $E_{\boldsymbol{X}^*}\big\{\bar{\boldsymbol{y}}_{rq}^* - (\bar{\boldsymbol{X}}_{rq}^*)^T\boldsymbol{\beta}\big\} = 0$. The calibrated estimation function is then

$$\boldsymbol{H}_{sl:cal}^*(\boldsymbol{\beta}) = \sum_q \Big(\boldsymbol{G}_{slq}^*\Big[\boldsymbol{y}_{slq}^* - \big\{(\lambda_{Aq} - \gamma_{Aq})\boldsymbol{I}_{slq}\boldsymbol{X}_{slq}^{E|A}\boldsymbol{\beta} + M_q\gamma_{Aq}\bar{\boldsymbol{X}}_q\boldsymbol{\beta}\big\}\Big] + (M_q - M_{slq})\bar{\boldsymbol{G}}_{rq}^*\big\{\bar{\boldsymbol{y}}_{rq}^* - \bar{\boldsymbol{X}}_{rq}^*\boldsymbol{\beta}\big\}\Big)$$

or equivalently

$$\boldsymbol{H}^*_{sl:cal}(\boldsymbol{\beta}) = \sum_q \Big(\boldsymbol{G}^*_{slq}\Big[\boldsymbol{y}^*_{slq} - \big\{(\lambda_{Aq} - \gamma_{Aq})\boldsymbol{X}^{E|A}_{slq}\boldsymbol{\beta} + M_q\gamma_{Aq}\mathbf{1}_{slq}\bar{\boldsymbol{X}}_q\boldsymbol{\beta}\big\}\Big] + \bar{\boldsymbol{G}}^*_{rq}\big[\big\{M_q\bar{\boldsymbol{y}}_q - M_{slq}\bar{\boldsymbol{y}}^*_{slq}\big\} - \big\{M_q(1 - M_{slq}\gamma_{Aq})\bar{\boldsymbol{X}}_q - M_{slq}(\lambda_{Aq} - \gamma_{Aq})\bar{\boldsymbol{X}}^{E|A}_{slq}\big\}\boldsymbol{\beta}\big]\Big). \tag{9}$$

There are a number of ways the weighting matrix $\bar{\boldsymbol{G}}^*_{rq}$ in (9) can be specified. For example, the analogy with optimal weighting for an estimating equation suggests

$$\bar{\boldsymbol{G}}^*_{rq} = (\bar{\boldsymbol{X}}^*_{rq})^T\big(\bar{\boldsymbol{V}}(\boldsymbol{y}^*_{slq})\big)^{-1}$$

where $\bar{\boldsymbol{V}}(\boldsymbol{y}^*_{slq})$ is the mean of the diagonal terms in $\boldsymbol{V}(\boldsymbol{y}^*_{slq})$. On the other hand, a more robust specification that avoids the need to estimate $\boldsymbol{V}(\boldsymbol{y}^*_{slq})$ is $\bar{\boldsymbol{G}}^*_{rq} = (\bar{\boldsymbol{X}}^*_{rq})^T$.

The following Theorem sets out the asymptotic variance of the solution to setting (9) to zero. The notation used is the same as that in previous results, and its proof is set out in the Appendix.

Theorem 4. *Let $\hat{\boldsymbol{\beta}}^*_{sl:cal}$ denote the solution to setting the calibrated estimating function (9) to zero. Suppose that λ_{Aq} and $\lambda_{B|Aq}$ are known. Then the asymptotic variance of $\hat{\boldsymbol{\beta}}^*_{sl:cal}$ is*

$$\boldsymbol{V}(\hat{\boldsymbol{\beta}}^*_{sl:cal}) = \boldsymbol{D}_{sl:cal}\Big[\sum_q \Big(\boldsymbol{G}^*_{slq}\boldsymbol{V}(\boldsymbol{y}^*_{slq})\boldsymbol{G}^{*T}_{slq} - 2\boldsymbol{G}^*_{slq}\,\mathrm{diag}(\boldsymbol{V}(\boldsymbol{y}^*_{slq}))\mathbf{1}_{slq}(\bar{\boldsymbol{G}}^*_{rq})^T + \bar{\boldsymbol{G}}^*_{rq}m_q\bar{\boldsymbol{V}}(\boldsymbol{y}^*_{slq})(\bar{\boldsymbol{G}}^*_{rq})^T\Big)\Big]\boldsymbol{D}^T_{sl:cal}$$

where

$$\boldsymbol{D}_{sl:cal} = \Big[\sum_q \Big\{\boldsymbol{G}^*_{slq}\big[(\lambda_{Aq} - \gamma_{Aq})\boldsymbol{X}^{E|A}_{slq} + M_q\gamma_{Aq}\mathbf{1}_{slq}\bar{\boldsymbol{X}}_q\big] + \bar{\boldsymbol{G}}^*_{rq}\big[M_q(1 - M_{slq}\gamma_{Aq})\bar{\boldsymbol{X}}_q - M_{slq}(\lambda_{Aq} - \gamma_{Aq})\bar{\boldsymbol{X}}^{E|A}_{slq}\big]\Big\}\Big]^{-1}.$$

When λ_{Aq} and $\lambda_{B|Aq}$ are unknown, and replaced by unbiased estimates, this asymptotic variance is

$$\boldsymbol{V}(\hat{\boldsymbol{\beta}}^*_{sl:cal}) = \boldsymbol{D}_{sl:cal}\Big[\sum_q \Big(\boldsymbol{\Gamma}_{slq} + \sum_{i=1}^{2}\sum_{j=1}^{2}\big(\partial_i\boldsymbol{H}^*_{ws}\big)J_{ijq}\big(\partial_j\boldsymbol{H}^*_{ws}\big)^T\Big)\Big]\boldsymbol{D}^T_{sl:cal}$$

where

$$\boldsymbol{\Gamma}_{slq} = \boldsymbol{G}^*_{slq}\boldsymbol{V}(\boldsymbol{y}^*_{slq})\boldsymbol{G}^{*T}_{slq} - 2\boldsymbol{G}^*_{slq}\,\mathrm{diag}(\boldsymbol{V}(\boldsymbol{y}^*_{slq}))\mathbf{1}_{slq}(\bar{\boldsymbol{G}}^*_{rq})^T + \bar{\boldsymbol{G}}^*_{rq}m_q\bar{\boldsymbol{V}}(\boldsymbol{y}^*_{slq})(\bar{\boldsymbol{G}}^*_{rq})^T.$$

3 Simulation study

In this section we describe the results of a Monte Carlo simulation that was used to compare the performances of estimators of $\boldsymbol{\beta}$ under both incomplete linkage and correlated linkage errors in the case where no population summary information is available and also when this information is available. The population regression 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 errors ϵ_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 $\lambda_{B|A3} = 0.91$.

We considered the case where these probabilities are known as well as the case where they were estimated from independent audit samples. These audit samples were defined by taking independent random samples 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 to y and x_1 to x_2 and y audit sample records respectively.

The linkage setup for which results are presented in this paper is the sample to registers linking case considered in Subsection 2.3, with three m-blocks each of size 2500, and with 500 records in each m-block randomly assigned as unlinkable. Independent random samples were then selected in each m-block, independently of whether population units were linkable or not. We considered three sample size scenarios. In the first, a sample of size 1000 was taken from 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. The results for this scenario are discussed in Subsection 3.1 below. The remaining two sample size scenarios were chosen in order to investigate the gains from calibration to external register information. In this case, independent samples of size 200 and of size 30 were taken from each m-block. Results for these two small sample scenarios are set out in Subsection 3.2.

The estimation methods for the parameters of the population regression model that were used in the first scenario are set out below. Methods based on the estimating function (8) substituted estimates for unknown parameters when calculating the weighting matrix $\boldsymbol{G}^*_{slq}$.

ST The naive OLS estimator based on the linked sample data;

Aind The solution to (8), with $\boldsymbol{G}^*_{slq} = \left(\tilde{\boldsymbol{T}}_{(sl)Aq}\tilde{\boldsymbol{X}}^{E|A}_{slq}\right)^T$, but with $\lambda_{B|Aq} = \lambda_{Bq}$, incorrectly assuming that the two linkage processes ($\boldsymbol{y}$ to $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ to $\boldsymbol{X}_1$) are uncorrelated. Note that this choice of weighting matrix is related to the Lahiri and Larsen (2005) method of bias correction;

Cind The solution to (8) with plug in approximations to the optimal weights $\boldsymbol{G}^*_{slq} = \left(\tilde{\boldsymbol{T}}_{(sl)Aq}\tilde{\boldsymbol{X}}^{E|A}_{slq}\right)^T\boldsymbol{V}^{-1}(\boldsymbol{y}^*_{slq})$ where $\boldsymbol{V}(\boldsymbol{y}^*_{slq}) = \sigma^2\boldsymbol{I}_{slq} + \tilde{\boldsymbol{V}}_{(sl)Aq} + \tilde{\boldsymbol{V}}_{(sl)Cq}$. Again, we incorrectly assume that the two linkage processes are uncorrelated.

Acor The same estimator as Aind, but now allowing for correlated linkage errors;

Ccor The same estimator as Cind, but allowing for correlated linkage errors.

In the second and third scenarios, the focus was on the small sample gains due to inclusion of population summary information in the estimating equation for the regression parameter. In this case, Acor and Ccor were again computed, as well as an alternative to Ccor, denoted Ccor2. This differed from Ccor in the method used to calculate the variance term in the optimal weights, using instead the heteroskedasticity-robust squared residual,

$$\boldsymbol{V}(\boldsymbol{y}^*_{slq}) = \boldsymbol{S}(\boldsymbol{\beta})^T\boldsymbol{S}(\boldsymbol{\beta}) \tag{10}$$

where

$$\boldsymbol{S}(\boldsymbol{\beta}) = \boldsymbol{y}^*_{slq} - \left\{(\lambda_{Aq} - \gamma_{Aq})\boldsymbol{X}^{E|A}_{slq}\boldsymbol{\beta} + M_q\gamma_{Aq}\mathbf{1}_{slq}\bar{\boldsymbol{X}}_q\boldsymbol{\beta}\right\}.$$

An advantage of calculating the variance term in the optimal weights using (10) is that it reduces the compounding effect of the approximation errors that result when this variance term is estimated as in Ccor, i.e. by plugging parameter estimates into the different components of the asymptotic variance expression displayed in Theorem 3. Calibrated versions of Acor, Ccor and Ccor2, computed using (9), were also calculated. These are denoted Acor:cal, Ccor:cal and Ccor2:cal respectively in Subsection 3.2.

3.1 Simulation results for the case where there is no population summary information

Table 1 shows the Monte Carlo relative bias and RMSE for the different estimators as well as the actual coverages of normal theory-based nominal 95 per cent confidence intervals based on plug in estimates of the asymptotic variance calculated using the expression in Theorem 3. We see immediately that although Aind and Cind reduce the bias in estimation of the regression model coefficients relative to that of the naive estimator ST, these methods still have a substantial bias as far as estimation of the coefficients associated with y and x_2 are concerned. Note, however, that this bias vanishes once we allow for correlated linkage errors via Acor and Ccor. This is intuitively reasonable since it is estimation of these parameters that is most affected by the linkage errors. Also, as one expects, there is an increase in variability when the linkage model parameters are estimated from the audit sample. Considering the RMSE results, it is clear that use of an optimal weighting system (Cind or Ccor) represents the best option when basing estimation on (8). This is similar to the findings in Kim and Chambers (2012a) and Kim and Chambers (2012b).

We observe that the coverage rates defined by Ccor are consistently higher than 95%, indicating that the plug in estimator of the asymptotic variance of this estimator is biased upwards. This is not the case for Acor and does not appear to happen when only two data sets are linked, see Chambers (2009) and Kim and Chambers (2012a). Since this phenomenon also occurs when the linkage probabilities are specified (rather than estimated from the audit subsample), it is almost certainly due to estimation errors associated with the use of the estimator (5) of the regression error variance σ^2 when estimating this asymptotic variance.

Table 1 here.

3.2 Simulation results for the case where there is population summary information

The aim of this Subsection is to evaluate the performance of the calibrated estimating function (9), as well as to assess the effect of using (10) to compute optimal weights for use in (9). Tables 2 and 3 show the Monte Carlo relative bias and RMSE of uncalibrated estimators based on (8) and calibrated estimators based on (9), as well as the actual coverage of normal

theory-based nominal 95 per cent confidence intervals based on plug in estimators of the asymptotic variances of these estimators (see Theorems 3 and 4) for sample sizes of 200 and 30 respectively.

Tables 2 and 3 here.

The results set out in 2 and 3 provide somewhat mixed messages. To start, we see immediately that the heteroskedasticity-robust approach (10) to calculation of optimal weights for use in either (8) or (9) works very well in terms of both reducing bias and stabilising variance at both sample sizes. Secondly, we see that the main beneficiary of the improved efficiency from introduction of the population summary information is estimation of the intercept coefficient β_0, as one would expect. This is consistent with related results on the use of likelihood-based methods for incorporating this type of population summary information in linear regression (Chambers *et al.*, 2012). However, the introduction of this information is not always beneficial. Although the calibrated estimators Acor:cal and Ccor2:cal are never inferior to their uncalibrated versions Acor and Ccor2 respectively, the calibrated version Ccor:cal of the plug in type optimally weighted version of (8) in fact performs substantially worse than its uncalibrated equivalent Ccor when the sample size is 200 (Table 2). This result is reversed, however, at the much smaller sample size of 30 (Table 3). In effect, at the larger sample size the gain in efficiency due to the use of population summary information in Ccor:cal is more than outweighed by the loss of efficiency due to bias associated with the plug in approach to calculating the estimate of $\boldsymbol{V}(\boldsymbol{y}^*_{slq})$ used in weighting. It is only at a much smaller sample size (30) that we see this trade off reversed.

Finally, we note that the coverage performances generally of the normal theory confidence intervals based on the asymptotic variances in Theorems 3 and 4 is reasonable. In the case of the estimators Acor and Acor:cal, calibration tends to make these intervals slightly more conservative in most (but not all) cases. As far as Ccor2 and Ccor2:cal are concerned, calibration also appears to generally increase coverage. In contrast, with Ccor and Ccor:cal we see a decrease in coverage following calibration. However, since Ccor generates very conservative intervals, this decrease actually tends to move coverage back to nominal levels.

4 Summary and future research

We extend the bias correction methods for secondary regression analysis based on multi-linked data described in Kim and Chambers (2012b). In particular, we develop bias-correction methods that can accommodate both correlated linkage errors and unlinked data, as well as make use of population summary information in order to reduce bias and stabilise variability when linked sample sizes are small. Our approach assumes the simple ELE mechanism for linkage errors suggested by Chambers (2009) since this seems most appropriate for the secondary analysis situation. Our results show that the estimator defined by the estimating function (9) that makes use of the summary information and uses a heroskedasticity-robust approximation to the optimal weighting function performs well, even when sample sizes are small. We focus on the linear regression analysis case, but, since we take an estimating equation approach, our methods are easily extended to generalised linear regression modelling.

Future research is necessary on more efficient ways of integrating population summary information into analysis of linked data. In this context, we are currently investigating alternative maximum likelihood solutions based on the ideas set out in Chambers *et al.* (2012). Another important outstanding problem is dealing with informative non-linkage. We have assumed the availability of weights that correct for the bias induced by non-linkage. In a secondary analysis situation, such weights may not be available, or may need to be constructed using the available linked data. Development of theory for this situation remains an open problem.

References

Adams, M. M., Wilson, H. G., Casto, D. L., Berg, C. J., McDermott, J. M., Gaudino, J. A., and McCarthy, B. J. (1997). Constructing reproductive histories by linking vital records. *American Journal of Epidemiology*, **145**, 339–348.

Bishop, G. and Khoo, J. (2007). Methodology of evaluating the quality of probabilistic linking. Technical Report 1351.0.55.018, Australian Bureau of Statistics.

Brook, E. L., Rosman, D. L., and Holman, C. D. J. (2008). Public good through data linkage:

measuring research outputs from the western australian data linkage system. *Australian and New Zealand Journal of Public Health*, **32**, 19–23.

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.

Chambers, R. L., Steel, D. G., Wang, S., and Welsh, A. (2012). *Maximum Likelihood Estimation for Sample Surveys*. CRC Monographs on Statistics and Applied Probability. Chapman and Hall.

Fair, M., Cyr, M., Allen, A. C., Wen, S. W., Guyon, G., and MacDonald, R. C. (2000). An assessment of the validity of a computer system for probabilistic an assessment of the validity of a computer system for probabilistic record linkage of birth and infant death records in canada. *Chronic Diseases in Canada*, **21**(1), 8–13.

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.

Gomatam, S., Carter, R., Ariet, M., and Mitchell, G. (2002). An empirical comparison of record linkage procedures. *STATISTICS IN MEDICINE*, **21**, 1485–1496.

Holman, C. D. J., Bass, A. J., lan L. Rouse, and Hobbs, M. S. (1999). Population-based linkage of health records in western australia: development of a health services research linked database. *Australian and New Zealand Journal of Public Health*, **23**, 453–459.

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.

Kelman, C., Bass, A., and Halman, C. (2002). Research use of linked health data - a best practice protocol. *Australian and New Zealand Journal of Public Health*, **26**, 251–255.

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.

Krewski, D., Dewanji, A., Wang, Y., Bartlett, S., Zielinski, J., and Mallick, R. (2005). The effect of record linkage errors on risk estimates in cohort mortality studies. *Survey Methodology*, **31**, 13–21.

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

Nitsch, D., DeStavola, B. L., Morton, S., and Leon, D. A. (2006). Linkage bias in estimating the association between childhood exposures and propensity to become a mother: An example of simple sensitivity analyses. *Journal of the Royal Statistical Society. Series A*, **169**(3), 493–505.

Rotermann, M. (2009). Evaluation of the coverage of linked canadian community health survey evaluation of the coverage of linked canadian community health survey and hospital inpatient records. *Health reports*, **20**, 45–51.

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.

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

$$\mathrm{Var}_{\boldsymbol{X}^*}(\boldsymbol{y}_q^*) = E_{\boldsymbol{X}^*}\Big\{\mathrm{Var}_{\boldsymbol{X}^*}\big(\boldsymbol{y}_q^*|\boldsymbol{A}_q\big)\Big\} + \mathrm{Var}_{\boldsymbol{X}^*}\Big\{E_{\boldsymbol{X}^*}\big(\boldsymbol{y}_q^*|\boldsymbol{A}_q\big)\Big\}. \tag{11}$$

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

$$\boldsymbol{V}_{Aq} = (1-\lambda_{Aq})\,\mathrm{diag}\left[\left\{\lambda_{Aq}(f_{iq}^{E|A} - \bar{f}_q^{E|A})^2 + \bar{f}_q^{E|A(2)} - (\bar{f}_q^{E|A})^2\right\}; i \in q\right]. \tag{12}$$

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

$$\begin{aligned}\mathrm{Var}_{\boldsymbol{X}^*}\big(\boldsymbol{y}_q^*|\boldsymbol{A}_q\big) = \boldsymbol{A}_q\Big[E_{\boldsymbol{X}^*}\Big\{\mathrm{Var}_{\boldsymbol{X}^*}\big(\boldsymbol{y}_q|(\boldsymbol{B}|\boldsymbol{A})_q\big)\Big\}\Big]\boldsymbol{A}_q^T \\ + \boldsymbol{A}_q\Big[\mathrm{Var}_{\boldsymbol{X}^*}\Big\{E_{\boldsymbol{X}^*}\big(\boldsymbol{y}_q|(\boldsymbol{B}|\boldsymbol{A})_q\big)\Big\}\Big]\boldsymbol{A}_q^T.\end{aligned} \tag{13}$$

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 (12) is

$$\boldsymbol{A}_q\Big[E_{\boldsymbol{X}^*}\Big\{\mathrm{Var}_{\boldsymbol{X}^*}\big(\boldsymbol{y}_q|(\boldsymbol{B}|\boldsymbol{A})_q\big)\Big\}\Big]\boldsymbol{A}_q^T = \boldsymbol{A}_q\sigma^2\boldsymbol{I}_q\boldsymbol{A}_q^T = \sigma^2\boldsymbol{A}_q\boldsymbol{I}_q\boldsymbol{A}_q^T = \sigma^2\boldsymbol{I}_q. \tag{14}$$

In order to evaluate the second term on the right had side of (11) 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

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

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

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

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

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

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

$$\boldsymbol{V}_{Cq} = (1-\lambda_{B|Aq})E_{\boldsymbol{X}^*}\,\mathrm{diag}\left[(M_q-1)^{-1}\big\{(\lambda_{Aq}M_q-1)d_i+M_q(1-\lambda_{Aq})\bar{d}_q\big\}; i\in q\right]. \quad (15)$$

Combining (11), (12), (14) and (15), 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

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

where

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

Thus,

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

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 Corollary 2

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. This leads us to the large sample approximation

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

where $\boldsymbol{H}_0^*$ denotes $\boldsymbol{H}^*(\boldsymbol{\beta}_0, \lambda_{0,A}, \phi_0)$. Let $\partial_1 = \partial/\partial_{\lambda_A}$ and $\partial_2 = \partial/\partial_{\phi}$. 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 $\boldsymbol{J}_q = [J_{ijq}] = \mathrm{Cov}(\hat{\lambda}_{Aq}, \hat{\phi}_q)$.

A.3 Proof of Theorem 4

When λ_{Aq} and $\lambda_{B|Aq}$ are known, by a first order Taylor series approximation

$$\boldsymbol{V}(\hat{\boldsymbol{\beta}}^*_{sl:cal}) = \left[\partial_{\boldsymbol{\beta}}\boldsymbol{H}^*_0\right]^{-1}\mathrm{Var}_{\boldsymbol{X}^*}(\boldsymbol{H}^*_0)\left(\left[\partial_{\boldsymbol{\beta}}\boldsymbol{H}^*_0\right]^{-1}\right)^T.$$

Let $\boldsymbol{D}_{sl:cal} = -\left[\partial_{\boldsymbol{\beta}}\boldsymbol{H}^*_0\right]^{-1}$. Then it can be seen that

$$\partial_{\boldsymbol{\beta}}\boldsymbol{H}^*_0 = -\sum_q \Big\{\boldsymbol{G}^*_{slq}\big[(\lambda_{Aq}-\gamma_{Aq})\boldsymbol{I}_{slq}\boldsymbol{X}^{E|A}_{slq} + M_q\gamma_{Aq}\mathbf{1}_{slq}\bar{\boldsymbol{X}}_q\big] + \bar{\boldsymbol{G}}^*_{rq}\big[M_q(1-M_{slq}\gamma_{Aq})\bar{\boldsymbol{X}}_q - M_{slq}(\lambda_{Aq}-\gamma_{Aq})\bar{\boldsymbol{X}}^{E|A}_{slq}\big]\Big\}.$$

Also, from its definition,

$$\mathrm{Var}_{\boldsymbol{X}^*}(\boldsymbol{H}^*_0) = \sum_q \Big(\boldsymbol{G}^*_{slq}\boldsymbol{V}(\boldsymbol{y}^*_{slq})\boldsymbol{G}^{*T}_{slq} - 2\boldsymbol{G}^*_{slq}\,\mathrm{diag}(\boldsymbol{V}(\boldsymbol{y}^*_{slq}))\mathbf{1}_{slq}(\bar{\boldsymbol{G}}^*_{rq})^T + \bar{\boldsymbol{G}}^*_{rq}m_q\bar{\boldsymbol{V}}(\boldsymbol{y}^*_{slq})(\bar{\boldsymbol{G}}^*_{rq})^T\Big).$$

Thus,

$$\boldsymbol{V}(\hat{\boldsymbol{\beta}}^*_{sl:cal}) = \boldsymbol{D}_{sl:cal}\Big[\sum_q \Big(\boldsymbol{G}^*_{slq}\boldsymbol{V}(\boldsymbol{y}^*_{slq})\boldsymbol{G}^{*T}_{slq} - 2\boldsymbol{G}^*_{slq}\,\mathrm{diag}(\boldsymbol{V}(\boldsymbol{y}^*_{slq}))\mathbf{1}_{slq}(\bar{\boldsymbol{G}}^*_{rq})^T + \bar{\boldsymbol{G}}^*_{rq}m_q\bar{\boldsymbol{V}}(\boldsymbol{y}^*_{slq})(\bar{\boldsymbol{G}}^*_{rq})^T\Big)\Big]\boldsymbol{D}^T_{sl:cal}$$

as required.

The proof uses similar arguments to those in the proof of Corollary 2 for the case where λ_{Aq} and $\lambda_{B|Aq}$ are unknown and are replaced by estimated values.

Table 1: Monte Carlo relative bias and relative RMSE (both expressed in percentage terms) for parameter estimates based on a sample size of 1000, with an independent audit sample of size 25. Empirical coverages (expressed in percentage terms) of normal theory-based nominal 95 per cent confidence intervals are also shown. These use plug in estimators of the asymptotic variance expression in Theorem 3.

Estimator	Relative Bias		Relative RMSE		Coverage	
	λ known	λ unknown	λ known	λ unknown	λ known	λ unknown
Estimation of β_0						
ST	206.14	206.14	207.43	207.43	0.0	0.0
Aind	-67.96	-64.65	71.87	78.32	22.3	100.0
Cind	-31.45	-25.91	33.46	33.44	78.7	100.0
Acor	1.13	2.65	23.27	37.41	95.6	90.4
Ccor	1.37	5.36	12.99	21.89	99.7	98.2
Estimation of β_1						
ST	-6.69	-6.69	16.35	16.35	37.5	37.5
Aind	-0.03	-0.06	6.77	7.43	97.2	100.0
Cind	-0.49	-0.65	4.36	4.75	99.9	100.0
Acor	-0.03	-0.07	6.69	7.35	96.1	95.4
Ccor	-0.01	-0.21	4.24	4.62	99.7	99.3
Estimation of β_2						
ST	-12.89	-12.89	36.62	36.62	0.0	0.0
Aind	4.25	4.05	12.54	13.74	8.8	100.0
Cind	1.97	1.62	5.74	5.78	48.5	100.0
Acor	-0.07	-0.16	3.56	6.35	93.0	86.7
Ccor	-0.08	-0.33	1.76	3.63	99.6	97.1

Table 2: Monte Carlo relative bias and relative RMSE (both expressed in percentage terms) for parameter estimates based on a sample size of 200, with an independent audit sample of size 25. Empirical coverages (expressed in percentage terms) of normal theory-based nominal 95 per cent confidence intervals are also shown. These use plug in estimators of the asymptotic variance expressions in Theorems 3 and 4.

Estimator	Relative Bias		Relative RMSE		Coverage	
	λ known	λ unknown	λ known	λ unknown	λ known	λ unknown
Estimation of β_0						
Acor	0.43	2.84	55.32	77.37	94.1	83.1
Ccor	2.93	14.15	31.27	49.80	99.8	92.8
Ccor2	-0.58	-0.29	10.15	10.24	93.3	94.0
Acor:cal	0.87	3.17	45.99	69.81	93.1	75.8
Ccor:cal	9.45	20.29	43.14	57.85	96.0	89.9
Ccor2:cal	-0.20	0.08	7.76	7.91	97.8	98.1
Estimation of β_1						
Acor	-0.11	-0.05	14.90	15.86	96.3	94.8
Ccor	-0.18	-0.72	9.59	10.63	99.9	98.4
Ccor2	0.00	-0.01	3.16	3.17	94.8	95.0
Acor:cal	-0.09	-0.03	14.83	15.80	96.5	94.6
Ccor:cal	-0.16	-0.71	15.61	16.43	92.1	90.0
Ccor2:cal	0.00	-0.01	3.15	3.16	94.5	94.8
Estimation of β_2						
Acor	-0.07	-0.21	8.15	12.37	93.2	75.6
Ccor	-0.25	-0.93	4.05	7.80	99.6	89.1
Ccor2	0.01	0.00	1.29	1.31	92.3	91.5
Acor:cal	-0.05	-0.20	8.12	12.33	93.6	75.4
Ccor:cal	-0.59	-1.27	7.55	10.17	90.2	80.8
Ccor2:cal	0.01	-0.01	1.28	1.31	92.5	91.7

Table 3: Monte Carlo relative bias and relative RMSE (both expressed in percentage terms) for parameter estimates based on a sample size of 30, with an independent audit sample of size 25. Empirical coverages (expressed in percentage terms) of normal theory-based nominal 95 per cent confidence intervals are also shown. These use plug in estimators of the asymptotic variance expressions in Theorems 3 and 4.

Estimator	Relative Bias		Relative RMSE		Coverage	
	λ known	λ unknown	λ known	λ unknown	λ known	λ unknown
Estimation of β_0						
Acor	6.46	9.82	151.41	139.47	94.7	94.5
Ccor	23.61	18.69	97.36	90.19	98.2	98.3
Ccor2	-5.56	-3.08	32.82	27.71	90.9	92.4
Acor:cal	3.15	6.63	117.14	102.64	95.1	96.7
Ccor:cal	20.16	16.06	72.27	64.55	99.7	99.7
Ccor2:cal	-5.74	-3.16	27.41	20.98	96.2	97.2
Estimation of β_1						
Acor	-0.73	-0.73	38.50	38.31	96.8	96.8
Ccor	-1.39	-1.23	26.93	26.80	99.1	98.8
Ccor2	-0.08	-0.13	8.59	8.56	93.3	93.0
Acor:cal	-0.70	-0.70	37.46	37.18	97.6	97.1
Ccor:cal	-1.34	-1.19	26.97	26.80	99.0	98.7
Ccor2:cal	-0.03	-0.09	8.45	8.42	94.3	93.7
Estimation of β_2						
Acor	-0.36	-0.58	21.41	18.91	94.9	96.7
Ccor	-1.34	-1.09	12.77	11.40	99.2	98.8
Ccor2	0.35	0.19	4.86	3.76	89.6	90.8
Acor:cal	-0.19	-0.41	20.68	18.13	96.0	97.5
Ccor:cal	-1.25	-1.00	12.76	11.41	98.8	99.0
Ccor2:cal	0.36	0.20	4.84	3.69	90.1	91.4