Centre for Statistical and Survey Methodology

The University of Wollongong

Working Paper

10-11

Regression Analysis under Probabilistic Mulit-Linkage

Gunky Kim and Ray Chambers

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

Regression Analysis under Probabilistic Multi-Linkage

Gunky Kim and Raymond Chambers
Centre for Statistical and Survey Methodology,
University of Wollongong, NSW, 2522, Australia

Abstract

Linkage errors can occur when probability-based methods are used to link records from two distinct data sets corresponding to the same target population. Current approaches to modifying standard methods of regression analysis to allow for these errors only deal with the case of two linked data sets and assume that the linkage process is complete, i.e. all records on the two data sets are linked. This paper extends these ideas to accommodate the situation when more than two data sets are probabilistically linked and the linkage is incomplete.

1. Introduction

Data linkage is now an important research tool in many areas of scientific research. For example Wilkins, Shields & Rothermann (2009) describe an analysis that models the relationship between an individual's probability of hospitalization and length of time spent subsequently in hospital and his/her smoking status using a linked data set obtained by merging data collected in the Canadian Community Health Survey and data held in Statistics Canada's Hospital Person-Oriented Information database. In Australia, Brook, Rosman & Holman (2008) claim that linked health record data sets produced by the Western Australia Data Linkage Unit over the period 1995 - 2003 were used in 708 research outputs, comprising journal articles, reports, presentations, conference proceedings and theses. Thus, Moorin & Holman (2008) use merged data sets from the Western Australia mortality register and the hospital morbidity data system to explore patterns of health expenditure for in-patient care in the last 3 years of life. Similarly, Zhao, Connors, Wright & Guthridge (2008) use data obtained by linking a primary care chronic disease register with hospital inpatient databases to determine the 2005 prevalence rates of chronic diseases for the remote indigenous population of the Northern Territory of Australia. In all of these linkage applications, different data sets relating to the same individuals at different points in time are linked to provide a longitudinal data record for each individual, thus permitting longitudinal analysis.

However, the use of probabilistic linkage raises issues about potential biases induced by linkage errors. In particular, there is always the possibility that linkage errors in the merged data could lead to a longitudinal record ostensibly relating to a single individual being actually made up of a composite of data items from different individuals. Furthermore, such errors are not confined to probabilistic linkage, since even if a unique identifier is thought to exist, and is used in the linkage process, there can still be linkage errors in the merged data sets. For example, Adams *et al*. (1997) found that use of the Social Security Number in the US is not adequate for complete linkage. Consequently, they use probabilistic linkage in their study. A similar situation is reported in Rotermann (2009).

The Census Data Enhancement project of the Australian Bureau of Statistics (ABS) 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. Initial development of this project included a test of the quality of the proposed linkage process, based on records

from a sample data set being linked with those on the census database. The results of this test are reported in Bishop & Khoo (2007), who state that 87 per cent of the test records were correctly linked when names and address were used in the matching process. These figures are representative of those obtained in similar Australian studies. For example, Holman, Bass, Rouse & Hobbs (1998) show that a linkage procedure carried out in Western Australia in 1996-97 provided 87 per cent correct linkage, while linked hospital morbidity data in Victoria in 1993-1994 showed a 78-86 per cent correct match rate. Clearly, these correct match rates will be lower when names and addresses are not used in the matching process. This last scenario is the one of greatest interest as far as the ABS Census Data Enhancement Project is concerned, since confidentiality restrictions mean that the actual linkage will be carried out without name and address information. An obvious consequence of this increased error will be an increase in the bias and a resulting loss of efficiency when the linked Census records are subject to longitudinal modelling.

Neter, Maynes & Ramanathan (1965) demonstrate that even a small amount of mismatching can lead to significant response errors, and Scheuren & Winkler (1993, 1997), Lahiri & Larsen (2005) and Chambers (2009) investigate methods for eliminating the resulting bias in the context of regression analysis based on data from two probabilistically linked data sets. However, these approaches cannot be used when the linked data are the result of probabilistically linking more than two data sets. In this paper we extend the results in Chambers (2009) and Kim & Chambers (2009) to the situation where there are more than two linked data sets, including the practically important case where at least one of these data sets is a sample from the underlying population and where linkage is incomplete. To fix things, we consider the case of three linked data sources. It is straightforward to extend our results to where more than three data sets are linked.

1.1 Technical background and assumptions

For notational simplicity we denote conditioning by a subscript in what follows, so the conditional expectation $E(\mathbf{Y} \mid \mathbf{X})$ is written $E_X(\mathbf{Y})$ and so on. Suppose that we are interested in fitting a regression model of the form $E_X(\mathbf{Y}) = f(\mathbf{X};\theta)$, where f is a known function, but the parameter θ is to be estimated. Here $\mathbf{Y}$ denotes the vector of population values of the response variable of interest, and $\mathbf{X}$ 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 of $\mathbf{Y}$ and $\mathbf{X}$ are unknown, but probabilistic linkage is used to reconstruct their values using the data in two or more population registers. Throughout this paper we shall assume that the regression model of interest is the linear model

$$\mathbf{Y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \epsilon = \mathbf{X}\beta + \epsilon$$

where $\mathbf{Y}$, $\mathbf{X}_1$ and $\mathbf{X}_2$ denote data obtained from three separate population registers. The model errors are assumed to have zero mean and are uncorrelated given $\mathbf{X}$, with $Var_X(\epsilon) = \sigma^2\mathbf{I}_N$ where $\mathbf{I}_N$ 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 β. The aim of this paper is to describe a methodology that can be used to eliminate this bias.

Without loss of generality, we take one of the three registers to be the 'benchmark' register, i.e. all linkage errors are defined relative to it, in the sense that this register is separately

linked to each of the other two registers. There are then four different linkage error scenarios that can then be defined:

- **Case 1:** $\mathbf{X}_1$ is the benchmark register and the only linkage errors are those arising from the linkage of $\mathbf{X}_1$ and $\mathbf{Y}$. That is, $\mathbf{X}_1$ and $\mathbf{X}_2$ are perfectly linked.
- **Case 2:** $\mathbf{X}_1$ is the benchmark register and the only linkage errors are those arising from the linkage of $\mathbf{X}_1$ and $\mathbf{X}_2$. That is, $\mathbf{X}_1$ and $\mathbf{Y}$ are perfectly linked.
- **Case 3:** $\mathbf{Y}$ is the benchmark register and there are linkage errors between $\mathbf{Y}$ and $\mathbf{X}_1$ and between $\mathbf{Y}$ and $\mathbf{X}_2$.
- **Case 4:** $\mathbf{X}_1$ is the benchmark register and there are linkage errors between $\mathbf{Y}$ and $\mathbf{X}_1$ and between $\mathbf{X}_1$ and $\mathbf{X}_2$.

Since in Case 1 there are no linkage errors between $\mathbf{X}_1$ and $\mathbf{X}_2$, it is equivalent to the situation considered in Chambers (2009) and Kim & Chambers (2009). In this article we focus on Case 4. See Kim & Chambers (2010) for the corresponding development for Cases 2 and 3.

In common with the development in Chambers (2009) and Kim and Chambers (2009), we make the following assumptions in this paper:

1. All registers have complete coverage of the target population and are of size *N*. In particular, there is a unique record in each of $\mathbf{Y}$, $\mathbf{X}_1$ and $\mathbf{X}_2$ that corresponds to the same population unit.
2. $\mathbf{Y}$, $\mathbf{X}_1$ and $\mathbf{X}_2$ 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. 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 $\mathbf{X}_1$ are denoted $\mathbf{X}_{1q}$ and so on.
3. Not all records in $\mathbf{X}_1$ can be linked. However, this 'non-linkage' is at random, so the same regression model holds for the linked and non-linked records. Note that this is a strong assumption. See Kim and Chambers (2009).
4. Linkage errors within a *m*-block are independent of any regression errors associated with observations from that *m*-block.
5. The benchmark register $\mathbf{X}_1$ does not need to be fully observed. If only a sample $\mathbf{X}_{1s}$ of population units on this register are observed, then the analyst has access to appropriate sample weights $\mathbf{w}_s$ that can be used to define consistent estimators of population quantities given the data for these sampled units.

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 will develop efficient

estimators for regression coefficients in the presence of linkage error, given that more than two data sources have been independently 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 will start by considering the case where three registers are completely linked.

2.1 Complete register to registers linkage

In this sub-section we assume that all the linked data sets are registers and linkage is complete, i.e. linkage is one to one and onto between them. Following Chambers (2009) we use a superscript of * to denote quantities defined using the linked data. In particular, we model the relationship between the true, but unobserved, values of $\mathbf{Y}$ and $\mathbf{X}_2$ and the observed linked values $\mathbf{Y}^*$ and $\mathbf{X}_2^*$ within *m*-block *q* by writing

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

where $\mathbf{A}_q$ and $\mathbf{B}_q$ are unobserved random permutation matrices that characterise the outcomes of the two independent linkage processes in *m*-block *q*. Note that one then has

$$\mathbf{X}_q = (\mathbf{1}_q, \mathbf{X}_{1q}, \mathbf{X}_{2q}) = (\mathbf{1}_q, \mathbf{X}_{1q}, \mathbf{B}_q^T \mathbf{X}_{2q}^*) \quad (1)$$

where $\mathbf{1}_q$ is the unit vector of size M_q. Under assumption 4 above, the distributions of $\mathbf{A}_q$ and $\mathbf{B}_q$ are independent of the distributions of the values making up $\mathbf{Y}_q$ and $\mathbf{X}_{2q}$. Hence

$$E_{\mathbf{X}^*}(\mathbf{X}_{2q}) = E_{\mathbf{X}^*}(\mathbf{B}_q^T)\mathbf{X}_{2q}^* = \mathbf{T}_{Bq}\mathbf{X}_{2q}^* \ .$$

A convenient model for the distributions of $\mathbf{A}_q$ and $\mathbf{B}_q$ is the *exchangeable linkage errors* (ELE) model. This is useful in the practically important situation where the person carrying out the linking and the person analysing the linked data are not the same, and confidentiality restrictions do not allow the release of all the information that was used in the linkage process. For example, in the Western Australian Diabetes Linkage Project, the people involved in creating the linkage key files were not permitted to take part in the analysis of the linked data. See Kelman, Bass & Holman (2002). Under the ELE model,

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

where

$$\lambda_{Bq} = \Pr(\text{correct linkage between } \mathbf{X}_{1q} \text{ and } \mathbf{X}_{2q})$$

and

$$\gamma_{Bq} = \Pr(\text{incorrect linkage between } \mathbf{X}_{1q} \text{ and } \mathbf{X}_{2q})$$

Furthermore, under this model, we have

$$\mathbf{X}_q^E = E_{\mathbf{X}^*}(\mathbf{X}_q) = E_{\mathbf{X}^*}\left[\left(\mathbf{1}_q, \mathbf{X}_{1q}, \mathbf{X}_{2q}\right)\right] = \left(\mathbf{1}_q, \mathbf{X}_{1q}, \mathbf{T}_{Bq}\mathbf{X}_{2q}^*\right). \quad (2)$$

It follows that

$$E_{\mathbf{X}^*}(\mathbf{Y}_q^*) = E_{\mathbf{X}^*}(\mathbf{A}_q\mathbf{Y}_q) = E_{\mathbf{X}^*}(\mathbf{A}_q)E_{\mathbf{X}^*}(\mathbf{Y}_q) = \mathbf{T}_{Aq}E_{\mathbf{X}^*}(\mathbf{Y}_q) = \mathbf{T}_{Aq}\mathbf{X}_q^E\beta \quad (3)$$

where, under the ELE model,

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

with

$$\lambda_{Aq} = \Pr(\text{correct linkage between } \mathbf{X}_{1q} \text{ and } \mathbf{Y}_q)$$

$$\gamma_{Aq} = \Pr(\text{incorrect linkage between } \mathbf{X}_{1q} \text{ and } \mathbf{Y}_q).$$

We now describe a method of estimating the parameter β of the linear regression model of interest using an adjusted unbiased estimating function. An unbiased estimating function for β when both $\mathbf{Y}_q$ and $\mathbf{X}_q$ are available is $\mathbf{H}(\beta) = \sum_q \mathbf{G}_q \left(\mathbf{Y}_q - \mathbf{f}_q \right)$ where $\mathbf{f}_q = E_X(\mathbf{Y}_q) = \mathbf{X}_q \beta$ and $\mathbf{G}_q$ is a weighting function that depends on $\mathbf{X}_q$ but not on $\mathbf{Y}_q$. However, we do not observe $\mathbf{Y}_q$ or $\mathbf{X}_q$. Instead, their linked versions $\mathbf{Y}_q^*$ and $\mathbf{X}_q^*$ are observed. A naive estimating function based on $\mathbf{H}(\beta)$ then takes the form

$$\mathbf{H}^*(\beta) = \sum_q \mathbf{G}_q^* \left(\mathbf{Y}_q^* - \mathbf{f}_q^* \right)$$

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

$$\mathbf{H}^*(\beta) = \sum_q \mathbf{G}_q^* \left(\mathbf{Y}_q^* - \mathbf{T}_{Aq} \mathbf{f}_q^E \right) \tag{4}$$

and a corresponding estimator of β can be defined as the solution $\hat{\beta}^*$ to the estimating equation defined by setting (4) to zero. The following Theorem states the asymptotic variance of $\hat{\beta}^*$. Its proof is in the Appendix.

Theorem 1. Let $\mathbf{f}_{2q}^* = \left(f_{2iq}^*; i \in q \right) = \mathbf{X}_{2q}^* \beta_2$. The asymptotic variance of $\hat{\beta}^*$ is then

$$V(\hat{\beta}^*) = \left[\sum_q \mathbf{G}_q^* \mathbf{T}_{Aq} \mathbf{X}_q^E \right]^{-1} \left[\sum_q \mathbf{G}_q^* V(\mathbf{Y}_q^*) \mathbf{G}_q^{*T} \right] \left(\left[\sum_q \mathbf{G}_q^* \mathbf{T}_{Aq} \mathbf{X}_q^E \right]^{-1} \right)^T$$

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

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

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

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

where $d_i = \lambda_{Bq} (f_{2iq}^* - \bar{f}_{2q}^*)^2 + \bar{f}_{2q}^{*(2)} - (\bar{f}_{2q}^*)^2$, $\bar{f}_{2q}^* = M_q^{-1} \sum_{i \in q} f_{2iq}^*$ and $\bar{f}_{2q}^{*(2)} = M_q^{-1} \sum_{i \in q} \left(f_{2iq}^* \right)^2$.

Note:

1. Given the values of $\mathbf{T}_{Aq}$, $\mathbf{T}_{Bq}$ and $\mathbf{f}_q^E$, an unbiased estimator of σ^2 is

$$\tilde{\sigma}^2 = N^{-1} \left[\sum_q \left(\mathbf{Y}_q^* - \mathbf{f}_q^E \right)^T \left(\mathbf{Y}_q^* - \mathbf{f}_q^E \right) - 2 \sum_q \left(\mathbf{f}_q^E \right)^T \left(\mathbf{I}_q - \mathbf{T}_{Aq} \right) \mathbf{f}_q^E \right].$$

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

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

$$\mathbf{G}_q^* = \left(\partial_\theta E_{X^*}(\mathbf{Y}_q^*)\right)^T \left(V(\mathbf{Y}_q^*)\right)^{-1} = \left(\mathbf{T}_{Aq}\mathbf{X}_q^E\right)^T \left(\sigma^2\mathbf{I}_q + \mathbf{V}_{Aq} + \mathbf{V}_{Cq}\right)^{-1}.$$

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

3. The development so far has assumed that the correct linkage probabilities λ_{Aq} and λ_{Bq} 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 $\hat{\beta}^*$ then depends also on the additional variability induced by this estimation process. If the estimators $\hat{\lambda}_{Aq}$ and $\hat{\lambda}_{Bq}$ of these probabilities are uncorrelated (e.g. if they are obtained from independent audit samples randomly selected from each *m*-block of the linked data), then the result in Theorem 1 can be extended to

$$V(\hat{\beta}^*) = \mathbf{W}\left[\sum_q \mathbf{G}_q^* \left\{V(\mathbf{Y}_q^*) + \mathbf{J}_{Aq}V(\hat{\lambda}_{Aq})\mathbf{J}_{Aq}^T + \mathbf{J}_{Bq}V(\hat{\lambda}_{Bq})\mathbf{J}_{Bq}^T\right\}\mathbf{G}_q^{*T}\right]\mathbf{W}^T$$

where $V(\hat{\lambda}_{Aq})$, $V(\hat{\lambda}_{Bq})$ are the variances of $\hat{\lambda}_{Aq}$, $\hat{\lambda}_{Bq}$ respectively, $\mathbf{W} = \left[\sum_q \mathbf{G}_q^*\mathbf{T}_{Aq}\mathbf{X}_q^E\right]^{-1}$, $\mathbf{J}_{Aq} = \left\{(M_q-1)^{-1}(M_q\mathbf{I}_q - \mathbf{1}_q\mathbf{1}_q^T)\right\}\mathbf{f}_q^E$ and $\mathbf{J}_{Bq} = \mathbf{T}_{Aq}\left\{(M_q-1)^{-1}(M_q\mathbf{I}_q - \mathbf{1}_q\mathbf{1}_q^T)\right\}\mathbf{f}_{2q}^*$.

2.2 Incomplete sample to registers linkage

We now consider the more realistic case when a sample $\mathbf{s}$ of records from the benchmark data set $\mathbf{X}_1$ are linked to the registers $\mathbf{Y}$ and $\mathbf{X}_2$, and this linkage is incomplete - i.e. there are some records in the sample $\mathbf{s}$ that cannot be linked, either to records in $\mathbf{X}_2$ register or to records in the $\mathbf{Y}$ register, or both. When linkage is incomplete, $\mathbf{A}_q$ and $\mathbf{B}_q$ are not permutation matrices because the entries for some rows in these matrices are all zero due to this non-linkage. However, we can still use the ideas introduced in the previous subsection.

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

'unlinkable' records are denoted $\mathbf{X}_{1ruq}$. It immediately follows that the following partitions exist:

$$\mathbf{Y}_q^* = \begin{pmatrix} \mathbf{Y}_{slq}^* \\ \mathbf{Y}_{suq}^* \\ \mathbf{Y}_{rlq}^* \\ \mathbf{Y}_{ruq}^* \end{pmatrix} = \begin{bmatrix} \mathbf{A}_{slsl,q} & \mathbf{A}_{slsu,q} & \mathbf{A}_{slrl,q} & \mathbf{A}_{slru,q} \\ \mathbf{A}_{susl,q} & \mathbf{A}_{susu,q} & \mathbf{A}_{surl,q} & \mathbf{A}_{suru,q} \\ \mathbf{A}_{rlsl,q} & \mathbf{A}_{rlsu,q} & \mathbf{A}_{rlrl,q} & \mathbf{A}_{rlru,q} \\ \mathbf{A}_{rusl,q} & \mathbf{A}_{rusu,q} & \mathbf{A}_{rurl,q} & \mathbf{A}_{ruru,q} \end{bmatrix} \begin{pmatrix} \mathbf{Y}_{slq} \\ \mathbf{Y}_{suq} \\ \mathbf{Y}_{rlq} \\ \mathbf{Y}_{ruq} \end{pmatrix} = \mathbf{A}_q \mathbf{Y}_q$$

where

$$E(\mathbf{A}_q \mid \mathbf{X}_q^*) = \mathbf{T}_{Aq} = \begin{bmatrix} \mathbf{T}_{(slsl)Aq} & \mathbf{T}_{(slsu)Aq} & \mathbf{T}_{(slrl)Aq} & \mathbf{T}_{(slru)Aq} \\ \mathbf{T}_{(susl)Aq} & \mathbf{T}_{(susu)Aq} & \mathbf{T}_{(surl)Aq} & \mathbf{T}_{(suru)Aq} \\ \mathbf{T}_{(rlsl)Aq} & \mathbf{T}_{(rlsu)Aq} & \mathbf{T}_{(rlrl)Aq} & \mathbf{T}_{(rlru)Aq} \\ \mathbf{T}_{(rusl)Aq} & \mathbf{T}_{(rusu)Aq} & \mathbf{T}_{(rurl)Aq} & \mathbf{T}_{(ruru)Aq} \end{bmatrix}.$$

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

$$E(\mathbf{B}_q \mid \mathbf{X}_q^*) = \mathbf{T}_{Bq} = \begin{bmatrix} \mathbf{T}_{(sl)Bq} \\ \mathbf{T}_{(su)Bq} \\ \mathbf{T}_{(rl)Bq} \\ \mathbf{T}_{(ru)Bq} \end{bmatrix} = \begin{bmatrix} \mathbf{T}_{(slsl)Bq} & \mathbf{T}_{(slsu)Bq} & \mathbf{T}_{(slrl)Bq} & \mathbf{T}_{(slru)Bq} \\ \mathbf{T}_{(susl)Bq} & \mathbf{T}_{(susu)Bq} & \mathbf{T}_{(surl)Bq} & \mathbf{T}_{(suru)Bq} \\ \mathbf{T}_{(rlsl)Bq} & \mathbf{T}_{(rlsu)Bq} & \mathbf{T}_{(rlrl)Bq} & \mathbf{T}_{(rlru)Bq} \\ \mathbf{T}_{(rusl)Bq} & \mathbf{T}_{(rusu)Bq} & \mathbf{T}_{(rurl)Bq} & \mathbf{T}_{(ruru)Bq} \end{bmatrix}.$$

The corresponding estimating function for β based on the linked sample data is

$$\begin{aligned} \mathbf{H}_{sl}^*(\beta) &= \sum\nolimits_q \mathbf{G}_{slq}^* \left(\mathbf{Y}_{slq}^* - \mathbf{T}_{(sl)Aq} \mathbf{f}_q^E \right) \\ &= \sum\nolimits_q \mathbf{G}_{slq}^* \left(\mathbf{Y}_{slq}^* - \mathbf{T}_{(slsl)Aq} \mathbf{f}_{slq}^E - \mathbf{T}_{(slsu)Aq} \mathbf{f}_{suq}^E - \mathbf{T}_{(slrl)Aq} \mathbf{f}_{rlq}^E - \mathbf{T}_{(slru)Aq} \mathbf{f}_{ruq}^E \right). \end{aligned} \tag{5}$$

Under the ELE model

$$\mathbf{T}_{(slsl)Aq} = \left(M_q - 1 \right)^{-1} \left\{ \left(\lambda_{Aq} M_q - 1 \right) \mathbf{I}_{slq} + \left(1 - \lambda_{Aq} \right) \mathbf{1}_{slq} \mathbf{1}_{slq}^T \right\}$$

$$\mathbf{T}_{(slsu)Aq} = \left(M_q - 1 \right)^{-1} \left(1 - \lambda_{Aq} \right) \mathbf{1}_{slq} \mathbf{1}_{suq}^T$$

$$\mathbf{T}_{(slrl)Aq} = \left(M_q - 1 \right)^{-1} \left(1 - \lambda_{Aq} \right) \mathbf{1}_{slq} \mathbf{1}_{rlq}^T$$

$$\mathbf{T}_{(slru)Aq} = \left(M_q - 1 \right)^{-1} \left(1 - \lambda_{Aq} \right) \mathbf{1}_{slq} \mathbf{1}_{ruq}^T .$$

In this case (5) becomes

$$\mathbf{H}_{sl}^*(\beta) = \sum\nolimits_q \mathbf{G}_{slq}^* \left\{ \mathbf{Y}_{slq}^* - \left(\frac{\lambda_{Aq} M_q - 1}{M_q - 1} \right) \mathbf{f}_{slq}^E - \left(\frac{1 - \lambda_{Aq}}{M_q - 1} \right) \mathbf{1}_{slq} \mathbf{1}_q^T \mathbf{f}_q^E \right\}.$$

The main problem with calculating the value of this modified estimating function is calculating the value of $\mathbf{1}_q^T \mathbf{f}_q^E$. This is a population, rather than a sample, quantity. If we

assume that the distribution of the values in $\mathbf{X}_{1slq}$ is the same as that of the values in $\mathbf{X}_{1sq}$, then we can approximate this population quantity by the weighted sample estimate $\tilde{\mathbf{w}}_{slq}^T \mathbf{f}_{slq}^E$, where $\tilde{\mathbf{w}}_{slq} = M_{sq} M_{slq}^{-1} \mathbf{w}_{slq}$. Here M_{slq} is the number of linked sample records in the q^{th} m-block, while M_{sq} is the total number of sampled records in this block. It immediately follows that the estimating function (5) can then be approximated by

$$\mathbf{H}_{sl}^*(\beta) = \sum\nolimits_q \mathbf{G}_{slq}^* \left\{ \mathbf{Y}_{slq}^* - \tilde{\mathbf{T}}_{(sl)Aq} \mathbf{f}_{slq}^E \right\} \tag{6}$$

where

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

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

$$\mathbf{f}_{slq}^E = \left(\mathbf{1}_{slq}, \mathbf{X}_{1slq}, \mathbf{T}_{(sl)Bq} \mathbf{X}_{2q}^* \right) \beta$$

where

$$\mathbf{T}_{(sl)Bq} \mathbf{X}_{2q}^* = \mathbf{T}_{(slsl)Bq} \mathbf{X}_{(sl)2q}^* + \mathbf{T}_{(slsu)Bq} \mathbf{X}_{(su)2q}^* + \mathbf{T}_{(slrl)Bq} \mathbf{X}_{(rl)2q}^* + \mathbf{T}_{(slru)Bq} \mathbf{X}_{(ru)2q}^*$$

and, under the ELE model,

$$\mathbf{T}_{(slsl)Bq} = \left(M_q - 1 \right)^{-1} \left\{ \left(\lambda_{Bq} M_q - 1 \right) \mathbf{I}_{slq} + \left(1 - \lambda_{Bq} \right) \mathbf{1}_{slq} \mathbf{1}_{slq}^T \right\}$$

$$\mathbf{T}_{(slsu)Bq} = \left(M_q - 1 \right)^{-1} \left(1 - \lambda_{Bq} \right) \mathbf{1}_{slq} \mathbf{1}_{suq}^T$$

$$\mathbf{T}_{(slrl)Bq} = \left(M_q - 1 \right)^{-1} \left(1 - \lambda_{Bq} \right) \mathbf{1}_{slq} \mathbf{1}_{rlq}^T$$

$$\mathbf{T}_{(slru)Bq} = \left(M_q - 1 \right)^{-1} \left(1 - \lambda_{Bq} \right) \mathbf{1}_{slq} \mathbf{1}_{ruq}^T .$$

If we now also assume that the distribution of the values defining each column of $\mathbf{X}_{2slq}^*$ is the same as that of the corresponding column of $\mathbf{X}_{2sq}^*$, then the same argument used to justify sample weighting above leads to $\mathbf{T}_{(sl)Bq} \mathbf{X}_{2q}^*$ being approximated by $\tilde{\mathbf{T}}_{(sl)Bq} \mathbf{X}_{2slq}^*$ where

$$\tilde{\mathbf{T}}_{(sl)Bq} = \left(M_q - 1 \right)^{-1} \left\{ \left(\lambda_{Bq} M_q - 1 \right) \mathbf{I}_{slq} + \left(1 - \lambda_{Bq} \right) \mathbf{1}_{slq} \tilde{\mathbf{w}}_{slq}^T \right\}.$$

That is, the final form of the estimating function that can be used in this case is

$$\mathbf{H}_{sl}^*(\beta) = \sum\nolimits_q \mathbf{G}_{slq}^* \left\{ \mathbf{Y}_{slq}^* - \tilde{\mathbf{T}}_{(sl)Aq} \tilde{\mathbf{f}}_{slq}^E \right\} \tag{7}$$

where $\tilde{\mathbf{f}}_{slq}^E = \left(\mathbf{1}_{slq}, \mathbf{X}_{1slq}, \tilde{\mathbf{T}}_{(sl)Bq} \mathbf{X}_{2slq}^* \right) \beta = \tilde{\mathbf{X}}_{slq}^E \beta$.

As in the previous section, the development so far has assumed that the linkage probabilities λ_{Aq} and λ_{Bq} are known. In practice, these will be unknown and replaced by the values of suitable estimators $\hat{\lambda}_{Aq}$ and $\hat{\lambda}_{Bq}$, with variances $V(\hat{\lambda}_{Aq})$ and $V(\hat{\lambda}_{Bq})$ respectively. The following Theorem sets out the form of the asymptotic variance for the solution $\hat{\beta}_s^*$ to setting (7) to zero. Its proof is along the same lines as that of Theorem 1.

Theorem 2. Under the assumption of non-informative non-linkage, i.e. when the distributions of the values in $\mathbf{Y}^*_{slq}$ and $\mathbf{X}^*_{2slq}$ are the same as those in $\mathbf{Y}^*_{sq}$ and $\mathbf{X}^*_{2sq}$, the asymptotic variance of $\hat{\beta}^*_s$ is

$$V(\hat{\beta}^*_s) = \tilde{\mathbf{W}}_{sl}\left[\sum_q \mathbf{G}^*_{slq}\left\{V(\mathbf{Y}^*_{slq}) + \mathbf{J}_{(sl)Aq}V(\hat{\lambda}_{Aq})\mathbf{J}^T_{(sl)Aq} + \mathbf{J}_{(sl)Bq}V(\hat{\lambda}_{Bq})\mathbf{J}^T_{(sl)Bq}\right\}\mathbf{G}^{*T}_{slq}\right]\tilde{\mathbf{W}}^T_{sl}$$

where $V(\mathbf{Y}^*_{slq}) = \sigma^2\mathbf{I}_{slq} + \tilde{\mathbf{V}}_{(sl)Aq} + \tilde{\mathbf{V}}_{(sl)Cq}$, $\mathbf{J}_{(sl)Aq} = \left\{(M_q - 1)^{-1}(M_q\mathbf{I}_{slq} - \mathbf{1}_{slq}\tilde{\mathbf{w}}^T_{slq})\right\}\tilde{\mathbf{f}}^E_{slq}$, $\mathbf{J}_{(sl)Bq} = \tilde{\mathbf{T}}_{(sl)Aq}\left\{(M_q - 1)^{-1}(M_q\mathbf{I}_{slq} - \mathbf{1}_{slq}\tilde{\mathbf{w}}^T_{slq})\right\}\tilde{\mathbf{f}}^*_{(sl)2q}$ and $\tilde{\mathbf{W}}_{sl} = \left[\sum_q \mathbf{G}^*_{slq}\tilde{\mathbf{T}}_{(sl)Aq}\tilde{\mathbf{X}}^E_{slq}\right]^{-1}$, with $\tilde{\mathbf{f}}^*_{(sl)2q} = \tilde{\mathbf{X}}^*_{(sl)2q}\beta_2$. Here

$$\tilde{\mathbf{V}}_{(sl)Aq} = \left(1 - \lambda_{Aq}\right)diag\left[\left\{\lambda_{Aq}(\tilde{f}^E_{islq} - \bar{\tilde{f}}^E_{slq})^2 + \bar{\tilde{f}}^{E(2)}_{slq} - (\bar{\tilde{f}}^E_{slq})^2\right\}; i \in slq\right]$$

where $\tilde{\mathbf{f}}^E_{slq} = \left(\tilde{f}^E_{islq}; i \in slq\right)$, $\bar{\tilde{f}}^E_{slq} = M^{-1}_{slq}\sum_{i\in slq}\tilde{f}^E_{islq}$ and $\bar{\tilde{f}}^{E(2)}_{slq} = M^{-1}_{slq}\sum_{i\in slq}\left(\tilde{f}^E_{islq}\right)^2$. Similarly

$$\tilde{\mathbf{V}}_{(sl)Cq} = (1 - \lambda_{Bq})\, diag\left[\left(M_q - 1\right)^{-1}\left\{\left(\lambda_{Aq}M_q - 1\right)\tilde{d}_i + M_q(1 - \lambda_{Aq})\bar{\tilde{d}}_{slq}\right\}; i \in slq\right]$$

where $\tilde{d}_i = \lambda_{Bq}(f^*_{i(sl)2q} - \bar{f}^*_{(sl)2q})^2 + \bar{f}^{*(2)}_{(sl)2q} - (\bar{f}^*_{(sl)2q})^2$, $\bar{f}^*_{(sl)2q} = M^{-1}_{slq}\sum_{i\in slq}f^*_{i(sl)2q}$ and $\bar{f}^{*(2)}_{(sl)2q} = M^{-1}_{slq}\sum_{i\in slq}\left(f^*_{i(sl)2q}\right)^2$.

3. Simulation results

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

$$Y_i = 1 + 3X_{1i} + 0.7X_{2i} + \varepsilon_i .$$

The values X_{1i} were drawn from the normal distribution with mean of 2 and a variance of 4, while the errors ε_i were independently drawn from the standard normal distribution. The values X_{2i} were generated as $X_{2i} = 1 + 2Z_i + \gamma_i$ where the values of Z_i were independently drawn from the same distribution as the X_{1i}, and the values γ_i were independently distributed as standard normal.

The population was generated as three *m*-blocks, with linkage errors generated according to the ELE model. In particular, the probabilities of correct linkage between $\mathbf{Y}^*_q$ and $\mathbf{X}_{1q}$ were set to $\lambda_{A1} = 1$, $\lambda_{A2} = 0.95$ and $\lambda_{A3} = 0.75$ while the probabilities of correct linkage between $\mathbf{X}_{1q}$ and $\mathbf{X}^*_{2q}$ were set to $\lambda_{B1} = 1$, $\lambda_{B2} = 0.85$ and $\lambda_{B3} = 0.8$.

We considered the case where these probabilities are known as well as the case where they are estimated from small audit samples taken from each *m*-block. These audit samples were defined by taking a random sample of size 25 in the q^{th} *m*-block for λ_{Aq} and an independent random sample of size 30 in the same block for λ_{Bq}. The estimate of the correct linkage probability in each case was the proportion of correctly linked records in the audit sample.

Two linkage scenarios were examined in the simulations.

- Scenario 1: The register to registers linking case, with three *m*-blocks each of size 500.
- Scenario 2: The sample to registers linking case, with three *m*-blocks each of size 2000. Half of the 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, 500 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 $\beta = (1,3,0.7)^T$ were considered:

Scenario 1

ST The naive OLS estimator based on the linked data;

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

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

Scenario 2

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

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

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

The two scenarios were independently simulated 1000 times and the estimates of β (based on ST, A and C) 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 A and C correct the bias due to incorrect linkage, with the implied efficient estimator C generally outperforming the Lahiri-Larsen estimator A in terms of relative root mean squared error. This is consistent with the results for the two-register complete linkage case reported in Chambers (2009). However, when λ's are unknown, we see that the relative bias of C is larger than that of A.

[Table 1 here.]

It is noteworthy that coverage rates for the estimators A and C 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 Gunky & Chambers (2009).

Figure 1 shows box plots of the distributions of estimation errors underpinning the results shown in Table 1. These distributions are for Scenario 2. The corresponding results for Scenario 1 were very similar. The overall superiority of method C, as well as the increase in variability when the correct linkage probabilities are estimated, is clear.

[Figure 1 here.]

4. Conclusions and further research

In this paper we extend the linkage error adjustment technique for regression analysis initially developed in Chambers (2009) to accommodate the case of sample to registers linkage when the number of linked data sets is greater than two. Our results indicate that the estimation

methods based on the estimating functions (4) and (7) are successful in eliminating the bias induced by linkage errors, provided we know, or are able to unbiasedly estimate, the correct linkage probabilities. They also correct the biases introduced by both sampling and non-linkage via the introduction of appropriate weights, assuming that these processes are non-informative and are independent of one another. However, it is also clear that these bias correction methods generally lead to larger variances.

It is important to note that the assumption of independence between the non-linkage and the linkage error processes is a strong one. In practice, we expect that ability to link a record from $\mathbf{X}_1$ to $\mathbf{Y}$ will increase the probability of being able to link the same record from $\mathbf{X}_1$ to $\mathbf{X}_2$. This implies that the ELE model needs to be extended so that the linkage error matrices $\mathbf{A}_q$ and $\mathbf{B}_q$ are correlated. We are currently investigating this situation.

Bibliography

Adams, M. M., Wilson, H. G., Casto, D. L., Berg, C. J., McDermott, J. M. & Gaudino, J. A. (1997). Constructing reproductive histories by linking vital records. *American Journal of Epidemiology, 145* (4), 339-348.

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

Brook, E. L., Rosman, D. L., & Holman, C. D. (2008). Public good through data linkage: measuring research output from the Western Australian data linkage system. *Australian and New Zealand Journal of Public Health, 32* (1), 19-23.

Chambers, R. (2009). Regression analysis of probability-linked data. *Statisphere*, *4*, Official Statistics Research Series, Statistics New Zealand.

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

Godambe, V. (1960). An optimum property of regular maximum likelihood estimation. *The Annals of Mathematical Statistics, 41* (4), 1208-1211.

Holman, C. D., Bass, A. J., Rouse, I. L. & Hobbs, M. S. (1998). 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* (5), 453-459.

Kelman, C. W., Bass, A. J. & Holman, C. D. (2002). Research use of linked health data - a best practice protocol. *Australian and New Zealand Journal of Public Health, 26* (3), 251-255.

Kim, G. & Chambers, R. (2009). Regression analysis under incomplete linkage. *Working Paper Series 17-09*, Centre for Statistical and Survey Methodology, University of Wollongong.

Kim, G. & Chambers, R. (2010). Regression analysis for longitudinally linked data. *Working Paper Series 22-10*, Centre for Statistical and Survey Methodology, University of Wollongong.

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

Moorin, R. E. & Holman, C. D. (2008). The cost of in-patient case in Western Australia in the last years of life: A population-based data linkage study. *Health Policy, 85*, 380-390.

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

Rotermann, M. (2009). Evaluation of the coverage of linked Canadian Community Health Survey and hospital inpatient records. *Health Reports, 20* (1), pp. 45-51.

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

Scheuren, F. & Winkler, W. E. (1997). Regression analysis of data files that are computer matched - Part II. *Survey Methodology, 23*, 157-165.

Wilkins, K., Shields, M. & Rothermann, M. (2009). Smoker's use of acute care hospitals - a prospective study. *Health Reports, 20* (4), pp. 75-83.

Zhao, Y., Connors, C., Wright, J. & Guthridge, S. (2008). Estimating chronic diesease prevalence among the remote aboriginal population of the northern territory using multiple data sources. *Australian and New Zealand Journal of Public Health, 32* (4), 307-313.

Appendix Proof of Theorem 1

We use ∂_β to denote the partial differentiation operator with respect to β 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 $\mathbf{G}_q^*$ is a function of $\mathbf{X}_q^*$. Then, since

$$\partial_\beta \mathbf{H}^*(\beta) = -\sum\nolimits_q \mathbf{G}_q^* \mathbf{T}_{Aq} \mathbf{X}_q^E$$

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

$$Var_{X^*}(\mathbf{Y}_q^*) = E_{X^*}\left\{Var_{X^*}\left(\mathbf{Y}_q^* \middle| \mathbf{A}_q\right)\right\} + Var_{X^*}\left\{E_{X^*}\left(\mathbf{Y}_q^* \middle| \mathbf{A}_q\right)\right\}. \tag{A1}$$

Then, by (2) and (3),

$$E_{X^*}\left(\mathbf{Y}_q^* \middle| \mathbf{A}_q\right) = \mathbf{A}_q E_{X^*}(\mathbf{Y}_q) = \mathbf{A}_q \mathbf{X}_q^E \beta = \mathbf{A}_q \mathbf{f}_q^E.$$

Hence $\mathbf{V}_{Aq} = Var_{X^*}\left\{E_{X^*}\left(\mathbf{Y}_q^* \middle| \mathbf{A}_q\right)\right\} = Var_{X^*}\left(\mathbf{A}_q \mathbf{f}_q^E\right)$. A large sample approximation to this variance is set out equation (16) of Chambers (2009), and is given by

$$\mathbf{V}_{Aq} = diag\left[(1-\lambda_{Aq})\left\{\lambda_{Aq}\left(f_{iq}^E - \overline{f}_q^E\right)^2 + \overline{f}_q^{E(2)} - (\overline{f}_q^E)^2\right\}\right]. \tag{A2}$$

In order to calculate $E_{X^*}\left\{Var_{X^*}\left(Y_q^* \middle| \mathbf{A}_q\right)\right\}$, we note that independence of $\mathbf{A}_q$ and $\mathbf{B}_q$ allows us to write

$$Var_{X^*}\left(\mathbf{Y}_q^* \middle| \mathbf{A}_q\right) = \mathbf{A}_q\left[E_{X^*}\left\{Var_{X^*}\left(\mathbf{Y}_q \middle| \mathbf{B}_q\right)\right\}\right]\mathbf{A}_q^T + \mathbf{A}_q\left[Var_{X^*}\left\{E_{X^*}\left(\mathbf{Y}_q \middle| \mathbf{B}_q\right)\right\}\right]\mathbf{A}_q^T \tag{A3}$$

From (1) we see that

$$Var_{X^*}\left(\mathbf{Y}_q \middle| \mathbf{B}_q\right) = \sigma^2 \mathbf{I}_q$$

Hence the first term on the right hand side of (A3) is

$$\mathbf{A}_q\left[E_{X^*}\left\{Var_{X^*}\left(\mathbf{Y}_q\middle|\mathbf{B}_q\right)\right\}\right]\mathbf{A}_q^T=\mathbf{A}_q\sigma^2\mathbf{I}_q\mathbf{A}_q^T=\sigma^2\mathbf{A}_q\mathbf{A}_q^T=\sigma^2\mathbf{I}_q\,. \tag{A4}$$

In order to evaluate the second term on the right hand side of (A3) we note that, given $\mathbf{f}_{2q}^*=\mathbf{X}_{2q}^*\beta_2$,

$$\mathbf{V}_{Bq}=Var_{X^*}\left\{E_{X^*}\left(\mathbf{Y}_q\middle|\mathbf{B}_q\right)\right\}=Var_{X^*}\left(\mathbf{B}_q^T\mathbf{f}_{2q}^*\right)$$

which has the large sample approximation

$$\mathbf{V}_{Bq}=diag\left[(1-\lambda_{Bq})\left\{\lambda_{Bq}(f_{iBq}^*-\overline{f}_{Bq}^*)^2+\overline{f}_{Bq}^{*(2)}-(\overline{f}_{Bq}^*)^2\right\}\right]=\left(1-\lambda_{Bq}\right)diag\left[d_i;i\in q\right].$$

Put $\mathbf{V}_{Cq}=E_{X^*}\left[\mathbf{A}_qVar_{X^*}\left\{E_{X^*}\left(\mathbf{Y}_q\middle|\mathbf{B}_q\right)\right\}\mathbf{A}_q^T\right]$. Then

$$\mathbf{V}_{Cq}=E_{X^*}\left[\mathbf{A}_q\left(1-\lambda_{Bq}\right)diag\left[d_i;i\in q\right]\mathbf{A}_q^T\right]=(1-\lambda_{Bq})E_{X^*}\left[\mathbf{A}_qdiag\left[d_i;i\in q\right]\mathbf{A}_q^T\right].$$

Put

$$e_{ij}^{A_q}=\lambda_{Aq}I(i=j)+\frac{1-\lambda_{Aq}}{M_q-1}I(i\neq j)\,.$$

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

$$\begin{aligned}E_{X^*}\left[\mathbf{A}_qdiag\left[d_i;i\in q\right]\mathbf{A}_q^T\right]&=diag\left[\sum_{i=1}^{M_q}d_ie_{ij}^{A_q};i\in q\right]\\&=diag\left[\left(M_q-1\right)^{-1}\left\{(\lambda_{Aq}M_q-1)d_i+M_q(1-\lambda_{Aq})\overline{d}_q\right\};i\in q\right]\end{aligned}$$

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

$$\mathbf{V}_{Cq}=\left(1-\lambda_{Bq}\right)diag\left[\left(M_q-1\right)^{-1}\left\{(\lambda_{Aq}M_q-1)d_i+M_q(1-\lambda_{Aq})\overline{d}_q\right\};i\in q\right]. \tag{5}$$

Combining (A1), (A2), (A4) and (A5), the required result follows immediately. Use of this asymptotic variance result to estimate the variance of $\hat{\beta}^*$ 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 $(\mathbf{Y}_q^*-\mathbf{f}_q^E)^T(\mathbf{Y}_q^*-\mathbf{f}_q^E)=U_{1q}+U_{2q}+U_{3q}$, where $U_{1q}=\mathbf{Y}_q^T\mathbf{A}_q^T\mathbf{A}_q\mathbf{Y}_q-\mathbf{Y}_q^T\mathbf{f}_q-\mathbf{f}_q^T\mathbf{Y}_q+\mathbf{f}_q^T\mathbf{f}_q$, $U_{2q}=\mathbf{Y}_q^T\mathbf{f}_q-\mathbf{f}_q^T\mathbf{f}_q$ and $U_{3q}=\mathbf{f}_q^T\mathbf{Y}_q-(\mathbf{Y}_q^*)^T\mathbf{f}_q^E-(\mathbf{f}_q^E)^T\mathbf{Y}_q^*+(\mathbf{f}_q^E)^T\mathbf{f}_q^E$.

Now

$$E_{\mathbf{X}^*}\sum\nolimits_qU_{1q}=E_{\mathbf{X}^*}\sum\nolimits_q\left\{\left(\mathbf{Y}_q-\mathbf{f}_q\right)^T\left(\mathbf{Y}_q-\mathbf{f}_q\right)\right\}=N\sigma^2\,.$$

Also

$$E_{\mathbf{X}^*}\left(U_{2q}\right)=E_{\mathbf{X}^*}\left\{\mathbf{f}_q^T\left(\mathbf{Y}_q-\mathbf{f}_q\right)\right\}=E_{\mathbf{X}^*}\left(\mathbf{f}_q^T\epsilon_q\right)=0$$

while, after re-arranging terms, we have

$$U_{3q} = \left\{ \mathbf{Y}_q^T \mathbf{f}_q^E - (\mathbf{Y}_q^*)^T \mathbf{f}_q^E \right\} + \left\{ (\mathbf{f}_q^E)^T \mathbf{f}_q^E - (\mathbf{f}_q^E)^T \mathbf{Y}_q^* \right\} + \left\{ \mathbf{Y}_q^T - (\mathbf{f}_q^E)^T \right\} \mathbf{f}_q + \left\{ (\mathbf{f}_q^E)^T - \mathbf{Y}_q^T \right\} \mathbf{f}_q^E .$$

We note that since $E_{X^*}\left[\left\{ \mathbf{Y}_q^T - (\mathbf{f}_q^E)^T \right\} \mathbf{f}_q \right] = E_{X^*} \left\{ (\mathbf{f}_q^E)^T - \mathbf{Y}_q^T \right\} \mathbf{f}_q^E = 0$,

$$E_{X^*}\left(U_{3q}\right) = E_{X^*}\left[\left\{ \mathbf{Y}_q^T \mathbf{f}_q^E - (\mathbf{Y}_q^*)^T \mathbf{f}_q^E \right\} + \left\{ (\mathbf{f}_q^E)^T \mathbf{f}_q^E - (\mathbf{f}_q^E)^T \mathbf{Y}_q^* \right\} \right] = 2(\mathbf{f}_q^E)^T \left(\mathbf{I}_q - \mathbf{T}_{Aq} \right) \mathbf{f}_q^E .$$

Hence an unbiased estimator of σ^2 is

$$\hat{\sigma}^2 = N^{-1} \sum\nolimits_q \left\{ \left(\mathbf{Y}_q^* - \mathbf{f}_q^E \right)^T \left(\mathbf{Y}_q^* - \mathbf{f}_q^E \right) - 2(\mathbf{f}_q^E)^T \left(\mathbf{I}_q - \mathbf{T}_{Aq} \right) \mathbf{f}_q^E \right\}.$$

Table 1 Relative bias and relative RMSE (both expressed in percentage terms) for parameter estimates investigated in the simulation study. 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 Theorems 1 and 2.

Estimator	Relative Bias		Relative RMSE		Coverage	
	λ known	λ estimated	λ known	λ estimated	λ known	λ estimated
Scenario 1 - estimation of β_0						
ST	128.81	128.81	129.97	129.97	0	0
A	-0.51	3.73	16.52	32.24	98.8	100
C	0.43	7.13	8.02	17.86	99.0	100
Scenario 1 - estimation of β_1						
ST	-9.94	-9.94	17.51	17.51	0	0
A	0.05	-0.35	3.03	5.69	95.6	100
C	-0.07	-0.74	1.38	3.03	97.7	100
Scenario 1 - estimation of β_2						
ST	-19.78	-19.78	17.76	17.76	0	0
A	0.04	-0.48	2.62	4.07	97.3	100
C	-0.03	-0.81	1.36	2.30	97.8	100
Scenario 2 - estimation of β_0						
ST	129.75	129.75	130.98	130.98	0	0
A	0.61	4.25	17.69	33.48	96.0	100
C	0.71	7.13	8.70	18.16	97.9	100
Scenario 2 - estimation of β_1						
ST	-10.08	-10.08	17.71	17.71	0	0
A	-0.08	-0.37	2.86	5.91	96.8	100
C	-0.09	-0.72	1.39	3.13	97.8	100
Scenario 2 - estimation of β_2						
ST	-19.90	-19.90	16.88	16.88	0	0
A	-0.12	-0.67	2.71	4.18	96.6	100
C	-0.07	-0.84	1.38	2.36	98.8	100

Figure 1 Boxplots showing the Monte Carlo distributions of the estimation errors for the model parameters under Scenario 2. Different estimation methods are denoted by ST, A and C prefixes. Left column is for λ known and right column for λ estimated. Top row is for the intercept β_0, middle for the slope parameter β_1 and bottom for the slope parameter β_2.