Centre for Statistical and Survey Methodology

The University of Wollongong

Working Paper

18-10

Fitting Linear Mixed Models Using Linked Data

Klairung Samart 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

FITTING LINEAR MIXED MODELS USING LINKED DATA

Klairung Samart* and Ray Chambers

University of Wollongong

Summary

Probabilistic matching of records from different data sets is often used to create linked data sets for use in research in health, epidemiology, economics, demography and sociology. Clearly, this type of matching can lead to linkage errors, which in turn can lead to bias and increased variability when standard statistical estimation techniques are used with the linked data. In this paper we develop unbiased regression parameter estimates when fitting a linear mixed model to probabilistically linked data. Furthermore, since estimation of variance components is also an important objective when fitting a mixed model, we develop appropriate modifications to standard methods of variance components estimation in order to account for linkage error. In particular, we focus on three widely used methods of variance components estimation: analysis of variance, maximum likelihood and restricted maximum likelihood. Simulation results show that our estimators performed reasonably well compared to the naive weighted least square estimator that just uses the linked data.

Key words: analysis of variance; linear mixed model; linkage error; maximum likelihood; measurement error; record matching; restricted maximum likelihood; weighted least square.

*Author to whom correspondence should be addressed.
Centre for Statistical and Survey Methodology, School of Mathematics and Applied Statistics, University of Wollongong, Wollongong NSW 2522, Australia. e-mail: ks208@uowmail.edu.au

1. Introduction

Linked data sets are particular useful for research in health, epidemiology, economics, demography, sociology and many other scientific areas. To create linked data, probabilistic matching of records from different data sets is often used. Clearly, this type of matching can lead to linkage errors which are a particular type of measurement error and therefore can lead to biased inference unless appropriate steps are taken to control and/or adjust for this bias (Chambers, 2009). Unfortunately, these errors are typically ignored when analysis is undertaken. Although there has been a number of statistical methods for linking data sets, there has been quite little methodological research carried out on the impact of linkage errors on analysis of linked data.

The impact of linkage errors was first raised by Neter et al.(1965). In their study, they found that relatively small linkage error could lead to a substantial bias in estimating the relationship between response errors and true values. Scheuren & Winkler (1993) then investigated the effect of linkage errors on the bias of ordinary least squares estimators of regression coefficients in a standard regression model and proposed a method of adjusting for the bias. However, their estimator is not unbiased in general. The best performance of this estimator still produces a very small bias. Therefore, Lahiri & Larsen (2005) proposed an alternative method for the bias correction which provides an unbiased estimator directly for a transformed regression model. In their simulations, they found that their unbiased method performed very well across a range of situations.

A methodological framework of linked data is mainly developed in Chambers (2009). In his work, appropriate modifications to standard statistical analysis methods are used to ensure that they remain unbiased when applied to probabilistically linked data. A simple linear regression model is fitted to linked data from two registers that each cover the same population. However, the inference used in this work is based on the assumption that all measurements are independent. Obviously, this assumption is relatively unrealistic since measurements are usually made on clusters of correlated statistical units, such as people in a family, patients in a hospital or students in a school, and when analysing such data, linear mixed models are often used. Consequently, based on the inferential framework of

Chambers (2009), we develop methodologies for efficient fitting of linear mixed models to probabilistically linked data. Moreover, estimation of variance components is also of interest since it is an important objective when fitting a mixed model.

The structure of the paper is as follows. In the following section we review the linkage errors model used in Chambers (2009) for fitting a simple linear regression model to linked data. Then in Section 3 we describe a framework for fitting a linear mixed model to linked data. In this section we obtain unbiased estimators of regression coefficients when clustering is accounted for. In Section 4 we describe three methods of variance components estimation: analysis of variance, pseudo-maximum likelihood and pseudo-restricted maximum likelihood. In Section 5 simulation results comparing all estimators are presented. Lastly, all conclusions and discussion for further research are revealed in Section 6 .

2. Linkage errors model

In what follows we assume that there is a population of N units, indexed by $i = 1, ..., N$. For each unit in this population, there are the measurement values of a scalar random variable Y and a vector random variable $\boldsymbol{X}$. The aim is to model the relationship between Y and $\boldsymbol{X}$ in this population, particularly fitting the linear regression model where the regression coefficients are unknown and needs to be estimated. However, the values of Y and $\boldsymbol{X}$ from this population do not exist. Instead, one can obtain such values from two registers that separately contain the population values of Y and $\boldsymbol{X}$ and both registers refer to the same population and have no duplicates. That is, each register consists of N records.

In order to link records from the two registers, one needs a unique identifier in each unit in the population. However, such identifier does not exist. Instead, some form of probability-linking algorithm is used to link records from the two registers. We also assumed that linkage is complete and one to one between the Y and $\boldsymbol{X}$-registers. Clearly, this type of data set can contain linkage errors, i.e. records where the values of Y and $\boldsymbol{X}$ actually come from different population units.

In addition, it is assumed that the linked records can be partitioned into Q distinct blocks such that there is no possibility that linked records in different blocks contain data for the same population unit. This assumption is based on the fact that best probability linking algorithms will only attempt to link records that are similar in some sense. That is, there is a categorical population variable Z that can be obtained from the information on either register and that different blocks correspond to different values of Z. In other words, if a record on one register does not have the same value of Z as the record on the other register, then two records cannot be the same unit in the population. Therefore, Z is referred as a blocking variable such that population units with the same value of Z are in the same block. Furthermore, we also assumed that Z is measured without error on both the Y-register and the $\boldsymbol{X}$-register which leads to an assumption that linkage errors can only occur within the same block.

Without loss of generality, we denote Q as distinct values taken by Z by 1,...,Q and let block q correspond to the M_q population units with $Z = q$ such that $N = \sum_q M_q$.

Let i denote index records in the linked data set. This index is the same for both the $\boldsymbol{X}$-register and the Y-register. y_i^* is used to denote the Y-value from block q on the Y-register that is matched to $\boldsymbol{X}_i$ in block q on the $\boldsymbol{X}$-register i.e. there are M_q linked data pairs $(y_i^*, \boldsymbol{X}_i)$ in block q. Thus, $\boldsymbol{y}_q^*$ is used to denote the vector of order M_q of the linked values y_i^* in block q and $\boldsymbol{X}_q$ as the matrix with rows defined by the values $\boldsymbol{X}_i$ in the same block. Lastly, $\boldsymbol{y}_q$ is denoted to be the unknown vector of the true Y values in block q that are associated with $\boldsymbol{X}_q$.

Since the linkage is assumed to be complete and one to one between the Y and $\boldsymbol{X}$-registers, Chambers (2009) modeled randomness in the outcome of the linkage process via the identity

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

where $\boldsymbol{A}_q = [a_{ij}^q]$ is an unknown random permutation matrix of order M_q i.e. the entries a_{ij}^q of $\boldsymbol{A}_q$ are either zero or one, with a value of one occurring just once in each row and column. In addition, due to an assumption that linkage errors can occur within blocks, then $\boldsymbol{A}_{q_1}$ and $\boldsymbol{A}_{q_2}$ are independently distributed when $q_1 \neq q_2$.

Since some assumptions about the distribution of the $\boldsymbol{A}_q$ will be needed for the inference, Chambers (2009) assumed that linkage is *non-informative* at each level of Z such that the distribution of $\boldsymbol{A}_q$ is independent of $\boldsymbol{y}_q$ given $\boldsymbol{X}_q$. Let

$$\mathrm{E}(\boldsymbol{A}_q \mid \boldsymbol{X}_q) = \boldsymbol{E}_q. \tag{2}$$

By assuming that a linked data is more likely to be correct than incorrect and the probability of correct linkage is the same for all records in a block, Chambers (2009), following the suggestion of Neter et al. (1965), characterized both of these assumptions via an exchangeable linkage errors model, where for each value of q

$$\Pr(\text{correct linkage}) = \Pr(a_{ii}^q = 1) = \lambda_q \tag{3}$$

and, for $i \neq j$,

$$\Pr(\text{incorrect linkage}) = \Pr(a_{ij}^q = 1) = \gamma_q. \tag{4}$$

Given (3) and (4) hold, (2) is then of the form

$$\boldsymbol{E}_q = (\lambda_q - \gamma_q)\boldsymbol{I}_q + \gamma_q \mathbf{1}_q \mathbf{1}_q^\top \tag{5}$$

where $\boldsymbol{I}_q$ is the identity matrix of order M_q and $\mathbf{1}_q$ denotes a vector of ones of length M_q. Since $\mathbf{1}_q^\top \boldsymbol{A}_q = \mathbf{1}_q^\top$ and $\boldsymbol{A}_q \mathbf{1}_q = \mathbf{1}_q$ thus, $\mathbf{1}_q^\top \boldsymbol{E}_q = \mathbf{1}_q^\top$ and $\boldsymbol{E}_q \mathbf{1}_q = \mathbf{1}_q$. That is, (5) implies

$$\begin{aligned} \lambda_q + (M_q - 1)\gamma_q &= 1 \\ \gamma_q &= \frac{1 - \lambda_q}{M_q - 1}. \end{aligned} \tag{6}$$

Obviously, λ_q is the key parameter to completely specify the first order properties of the linkage mechanism under the model (5). All of these properties and models will be used throughout this paper for the theory development.

3. Estimation of regression coefficients

Fitting a simple linear regression model to linked data is described in Chambers (2009). In this section we consider the situation where the group structure is allowed for and thus the linear mixed model is the focus of inference. In addition to what we describe in Section 2, here we have a grouping variable F which is taken by $1, ..., G$ and let group g correspond to the N_g population units with $F = g$ such that $M_q = \sum_g n_{qg}$ and $N_g = \sum_q n_{qg}$ where n_{qg} is the number of population in block q group g. The linear mixed model is then given by

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}$$

where $\boldsymbol{X}$ and $\boldsymbol{Z}$ are the design matrices,

$$\mathrm{E}\begin{bmatrix}\boldsymbol{u}\\ \boldsymbol{e}\end{bmatrix} = \begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{0}\end{bmatrix} \quad \text{and} \quad \mathrm{Cov}\begin{bmatrix}\boldsymbol{u}\\ \boldsymbol{e}\end{bmatrix} = \begin{bmatrix}\boldsymbol{\zeta} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{R}\end{bmatrix}$$

where $\boldsymbol{\zeta} = \sigma_u^2 \boldsymbol{I}_G$ and $\boldsymbol{R} = \sigma_e^2 \boldsymbol{I}_N$. Thus, the variance-covariance matrix of Y is of the form

$$\boldsymbol{V} = \boldsymbol{Z}\boldsymbol{\zeta}\boldsymbol{Z}^\top + \boldsymbol{R} = \sigma_u^2 \boldsymbol{Z}\boldsymbol{Z}^\top + \sigma_e^2 \boldsymbol{I}_N$$

and the population values of Y and $\boldsymbol{X}$ in each block satisfy

$$\mathrm{E}_X(\boldsymbol{y}_q) = \boldsymbol{X}_q\boldsymbol{\beta} = \boldsymbol{f}_q \tag{7}$$

$$\mathrm{Var}_X(\boldsymbol{y}_q) = \sigma_u^2 \boldsymbol{Z}_q\boldsymbol{Z}_q^\top + \sigma_e^2 \boldsymbol{I}_q \tag{8}$$

$$\mathrm{Cov}_X(\boldsymbol{y}_q, \boldsymbol{y}_r) = \sigma_u^2 \boldsymbol{Z}_q\boldsymbol{Z}_r^\top \tag{9}$$

where r is another block index, σ_u^2 is between-group variance, and σ_e^2 is within-group variance. Note that apart from the regression parameter $\boldsymbol{\beta}$ in (7), which is the target of inference, variance components σ_u^2 and σ_e^2 in (8) are also included as unknown parameters which need to be estimated. Given the $\boldsymbol{y}_r$ and $\boldsymbol{X}_q$, the naive linked data weighted least squares (WLS) estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}^{*} = \left(\sum_{q} \sum_{r} \boldsymbol{X}_{q}^{\top} \boldsymbol{W}_{qr} \boldsymbol{X}_{r} \right)^{-1} \left(\sum_{q} \sum_{r} \boldsymbol{X}_{q}^{\top} \boldsymbol{W}_{qr} \boldsymbol{y}_{r}^{*} \right) \tag{10}$$

where $\boldsymbol{W} = \boldsymbol{V}^{-1}$ is a weight matrix and that $\boldsymbol{W}_{qr}$ is a partitioned matrix of matrix $\boldsymbol{W}$.

We see that under the linkage error model (1), the naive WLS estimator (10) based on the linked data set is biased since

$$\mathrm{E}_{X}(\hat{\boldsymbol{\beta}}^{*}) = \left(\sum_{q} \sum_{r} \boldsymbol{X}_{q}^{\top} \boldsymbol{W}_{qr} \boldsymbol{X}_{r} \right)^{-1} \left(\sum_{q} \sum_{r} \boldsymbol{X}_{q}^{\top} \boldsymbol{W}_{qr} \boldsymbol{E}_{r} \boldsymbol{X}_{r} \right) \boldsymbol{\beta} = \boldsymbol{D} \boldsymbol{\beta}. \tag{11}$$

Given $\boldsymbol{E}_{r}$ and $\boldsymbol{W}_{qr}$ are known and the inverse of $\boldsymbol{D}$ in (11) exists, Chambers (2009) suggested an unbiased estimator using a ratio-type correction for the bias in the naive estimator of $\boldsymbol{\beta}$,

$$\hat{\boldsymbol{\beta}}_{R} = \boldsymbol{D}^{-1} \hat{\boldsymbol{\beta}}^{*} = \left(\sum_{q} \sum_{r} \boldsymbol{X}_{q}^{\top} \boldsymbol{W}_{qr} \boldsymbol{E}_{r} \boldsymbol{X}_{r} \right)^{-1} \left(\sum_{q} \sum_{r} \boldsymbol{X}_{q}^{\top} \boldsymbol{W}_{qr} \boldsymbol{y}_{r}^{*} \right) \tag{12}$$

where $\sum_{q} \sum_{r} \boldsymbol{X}_{q}^{\top} \boldsymbol{W}_{qr} \boldsymbol{E}_{r} \boldsymbol{X}_{r}$ is of full rank.

Alternatively, since $\boldsymbol{y}_{q}^{*} = \boldsymbol{A}_{q} \boldsymbol{y}_{q}$, and $\boldsymbol{A}_{q}$ and $\boldsymbol{y}_{q}$ are independently distributed given $\boldsymbol{X}_{q}$ it follows

$$\mathrm{E}_{X}(\boldsymbol{y}_{q}^{*}) = \mathrm{E}_{X}(\boldsymbol{A}_{q}) \mathrm{E}_{X}(\boldsymbol{y}_{q}) = \boldsymbol{E}_{q} \boldsymbol{X}_{q} \boldsymbol{\beta} = \boldsymbol{H}_{q} \boldsymbol{\beta}. \tag{13}$$

We see that the $\boldsymbol{y}_{q}^{*}$ are also in a linear form with regression coefficient $\boldsymbol{\beta}$ but with a modified set of explanatory variables $\boldsymbol{H}_{q}$ in block q. According to Lahiri & Larsen (2005) and Chambers (2009), $\boldsymbol{\beta}$ can be estimated by using the WLS estimator for this situation as

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{A} &= \left(\sum_{q} \sum_{r} \boldsymbol{H}_{q}^{\top} \boldsymbol{W}_{qr} \boldsymbol{H}_{r} \right)^{-1} \left(\sum_{q} \sum_{r} \boldsymbol{H}_{q}^{\top} \boldsymbol{W}_{qr} \boldsymbol{y}_{r}^{*} \right) \\ &= \left(\sum_{q} \sum_{r} \boldsymbol{X}_{q}^{\top} \boldsymbol{E}_{q}^{\top} \boldsymbol{W}_{qr} \boldsymbol{E}_{r} \boldsymbol{X}_{r} \right)^{-1} \left(\sum_{q} \sum_{r} \boldsymbol{X}_{q}^{\top} \boldsymbol{E}_{q}^{\top} \boldsymbol{W}_{qr} \boldsymbol{y}_{r}^{*} \right). \end{aligned} \tag{14}$$

Nevertheless, this estimator would be optimal if the regression errors under (13) were

homoscedastic. Clearly, this generally does not hold, since the variances of the regression errors defined by the linked data vary between blocks. That is

$$\begin{aligned} \mathrm{Var}_X(\boldsymbol{y}_q^*) &= \mathrm{E}_X\left\{\mathrm{Var}_{AX}(\boldsymbol{y}_q^*)\right\} + \mathrm{Var}_X\left\{\mathrm{E}_{AX}\left(\boldsymbol{y}_q^*\right)\right\} \\ &= \sigma_u^2 \mathrm{E}_X\left(\boldsymbol{A}_q \boldsymbol{Z}_q \boldsymbol{Z}_q^\top \boldsymbol{A}_q^\top\right) + \sigma_e^2 \boldsymbol{I}_q + \boldsymbol{V}_q \\ &= \sigma_u^2\ \boldsymbol{K}_q + \sigma_e^2 \boldsymbol{I}_q + \boldsymbol{V}_q \end{aligned} \tag{15}$$

where $\boldsymbol{V}_q$ was approximated by Chambers (2009) that

$$\boldsymbol{V}_q \approx \mathrm{diag}\left[(1-\lambda_q)\left\{\lambda_q(f_i - \bar{f}_q)^2 + \bar{f}_q^{\,(2)} - \bar{f}_q^2\right\}\right]$$

where $\boldsymbol{f}_q = (f_i)$ and $\bar{f}_q,\ \bar{f}_q^{\,(2)}$ denote the block q averages of the components of $\boldsymbol{f}_q$ and their squares respectively. The approximation of $\boldsymbol{K}_q$ is however developed similarly to $\boldsymbol{V}_q$. We define $\boldsymbol{K}_q = [k_{ij}]$ as

$$\boldsymbol{K}_q = \begin{cases} \lambda + \frac{(1-\lambda)}{M_q - 1}(G_q n_{qh} - 1), & \text{if } i = j \\ \\ \left\{\lambda + (n_{qh}-2)\frac{(1-\lambda)}{M_q-1}\right\}^2 & \\ +(n_{qh}-1)\left\{\frac{(1-\lambda)}{M_q-1}\right\}^2\left\{1 + (G_q - 1)n_{qh}\right\}, & \text{if } i \neq j; \\ & i, j \in g;\ g = h \\ \\ (n_{qh}-1)\frac{(1-\lambda)}{M_q-1}\left\{2\lambda + \frac{(1-\lambda)}{M_q-1}(G_q n_{qh} - 2)\right\}, & \text{if } i \neq j;\ i \in g \\ & j \in h;\ g \neq h. \end{cases}$$

where G_q is number of groups in block q. Also, using the law of total covariance, the covariance between $\boldsymbol{y}_q^*$ and $\boldsymbol{y}_r^*$ is then

$$\begin{aligned}\mathrm{Cov}_X(\boldsymbol{y}_q^*, \boldsymbol{y}_r^*) &= \mathrm{Cov}_X(\boldsymbol{A}_q\boldsymbol{y}_q, \boldsymbol{A}_r\boldsymbol{y}_r) \\ &= \mathrm{E}_X\left\{\boldsymbol{A}_q\mathrm{Cov}_X(\boldsymbol{y}_q, \boldsymbol{y}_r)\boldsymbol{A}_r^\top\right\} + \mathrm{Cov}_X(\boldsymbol{A}_q\boldsymbol{f}_q, \boldsymbol{A}_r\boldsymbol{f}_r) \\ &= \sigma_u^2\left(\boldsymbol{E}_q\boldsymbol{Z}_q\boldsymbol{Z}_r^\top\boldsymbol{E}_r^\top\right).\end{aligned}$$

Thus, the best linear unbiased estimator (BLUE) for β given these data is

$$\begin{aligned}\hat{\boldsymbol{\beta}}_C &= \left(\sum_q\sum_r \boldsymbol{H}_q^\top\boldsymbol{\Sigma}_{qr}\boldsymbol{H}_r\right)^{-1}\left(\sum_q\sum_r \boldsymbol{H}_q^\top\boldsymbol{\Sigma}_{qr}\boldsymbol{y}_r^*\right) \\ &= \left(\sum_q\sum_r \boldsymbol{X}_q^\top\boldsymbol{E}_q^\top\boldsymbol{\Sigma}_{qr}\boldsymbol{E}_r\boldsymbol{X}_r\right)^{-1}\left(\sum_q\sum_r \boldsymbol{X}_q^\top\boldsymbol{E}_q^\top\boldsymbol{\Sigma}_{qr}\boldsymbol{y}_r^*\right) \qquad (16)\end{aligned}$$

where $\boldsymbol{\Sigma} = \mathrm{Var}^{-1}(\boldsymbol{y}^*)$ and that $\boldsymbol{\Sigma}_{qr}$ is a partitioned matrix of matrix $\boldsymbol{\Sigma}$.

4. Variance component estimation

When fitting a mixed model, estimation of variance components is also an important objective. In this section we develop appropriate modifications to standard methods of variance components estimation in order to account for linkage error. In particular, we focus on three widely used methods of variance components estimation: Analysis of Variance (ANOVA), pseudo-maximum likelihood (pseudo-ML) and pseudo-restricted maximum likelihood (pseudo-REML). The details of each method are described in Section 4.1, 4.2 and 4.3, respectively.

4.1. Analysis of Variance (ANOVA)

Historically, ANOVA is the starting point of methods of estimating variance components (Searle et al., 2006). This method is based on deriving the expected values of sum of squares between groups (SSA) and sum of squares within groups (SSE) from the definitions. One then equates observed and expected values and solves for estimators.

The two sum of squares that are the basis of ANOVA of the linked data are

$$\text{SSA} \quad = \boldsymbol{y}^{*\top} \boldsymbol{B}\ \boldsymbol{y}^*, \quad ; \boldsymbol{B} = \{b_{ij}\} = \begin{cases} \frac{1}{N_g} - \frac{1}{N}, & \text{if } i, j \in g \\ -\frac{1}{N}, & \text{if } i \in g,\ j \in h,\ g \neq h \end{cases}$$

$$\text{SSE} \quad = \boldsymbol{y}^{*\top} \boldsymbol{C}\ \boldsymbol{y}^*, \quad ; \boldsymbol{C} = \{c_{ij}\} = \begin{cases} 1 - \frac{1}{N_g}, & \text{if } i, j \in g,\ i = j \\ -\frac{1}{N_g}, & \text{if } i, j \in g,\ i \neq j \\ 0, & \text{otherwise.} \end{cases}$$

We then need to obtain $\mathrm{E}_X(\text{SSA})$ and $\mathrm{E}_X(\text{SSE})$ and equate them to the observed values mentioned above. The details of method of deriving the expected values of the two sum of squares are illustrated in Appendix I which yields the variance components estimators

$$\hat{\sigma}_e^2 = \frac{mc - na}{bc - da} \tag{17}$$

and

$$\hat{\sigma}_u^2 = \frac{m - \hat{\sigma}_e^2 b}{a} \tag{18}$$

where

$$\begin{aligned} a &= \sum_q \mathrm{tr}\left(\boldsymbol{B}_{qq}\boldsymbol{K}_q\right) + \sum_q \sum_{r \neq q} \mathrm{tr}\left(\boldsymbol{E}_r \boldsymbol{Z}_r \boldsymbol{Z}_q^\top \boldsymbol{E}_q^\top \boldsymbol{B}_{qr}\right) \\ b &= \sum_q \mathrm{tr}\left(\boldsymbol{B}_{qq}\right) \\ c &= \sum_q \mathrm{tr}\left(\boldsymbol{C}_{qq}\boldsymbol{K}_q\right) + \sum_q \sum_{r \neq q} \mathrm{tr}\left(\boldsymbol{E}_r \boldsymbol{Z}_r \boldsymbol{Z}_q^\top \boldsymbol{E}_q^\top \boldsymbol{C}_{qr}\right) \\ d &= \sum_q \mathrm{tr}\left(\boldsymbol{C}_{qq}\right) \end{aligned}$$

$$m = \text{SSA} - \sum_q \left\{ \text{tr}\left(\boldsymbol{B}_{qq}\boldsymbol{V}_q\right) + \boldsymbol{f}_q^\top \boldsymbol{E}_q^\top \boldsymbol{B}_{qq}\boldsymbol{E}_q\boldsymbol{f}_q \right\} - \sum_q \sum_{r\neq q} \boldsymbol{f}_q^\top \boldsymbol{E}_q^\top \boldsymbol{B}_{qr}\boldsymbol{E}_r\boldsymbol{f}_r$$

$$n = \text{SSE} - \sum_q \left\{ \text{tr}\left(\boldsymbol{C}_{qq}\boldsymbol{V}_q\right) + \boldsymbol{f}_q^\top \boldsymbol{E}_q^\top \boldsymbol{C}_{qq}\boldsymbol{E}_q\boldsymbol{f}_q \right\} - \sum_q \sum_{r\neq q} \boldsymbol{f}_q^\top \boldsymbol{E}_q^\top \boldsymbol{C}_{qr}\boldsymbol{E}_r\boldsymbol{f}_r$$

Note that if linkage is perfect , $c = 0$ i.e. $\hat{\sigma}_e^2 = n/d$ and $\hat{\sigma}_u^2 = (m - \hat{\sigma}_e^2 b)/a$

where

$$a = N - \frac{\sum_g N_g^2}{N}$$

$$b = G - 1 \text{ is degrees of freedom of SSA}$$

$$d = N - G \text{ is degrees of freedom of SSE}$$

$$m = \text{SSA} - \sum_q \sum_r \boldsymbol{f}_q^\top \boldsymbol{B}_{qr}\boldsymbol{f}_r$$

$$n = \text{SSE} - \sum_q \sum_r \boldsymbol{f}_q^\top \boldsymbol{C}_{qr}\boldsymbol{f}_r$$

However, the ANOVA estimates in (17) and (18) can be negative. According to Searle et al. (2006), there is nothing in the ANOVA method of estimation that will prevent a negative estimate. A more serious alternative would be to use a method of estimation that explicitly excludes the possibility of negative estimates. Such methods are maximum likelihood (ML) and restricted maximum likelihood (REML) which will be detailed in the next two sections.

4.2. Pseudo-Maximum Likelihood (Pseudo-ML)

One of the most well-established and well-respected statistical methods used for fitting a statistical model to data, and providing estimates for the model's parameters is maximum likelihood estimation. Here we use this method as an alternative approach to constructing efficient estimators of $\boldsymbol{\beta}, \sigma_u^2$ and σ_e^2 given the linked data.

Unlike the ANOVA method of estimation, one of the basic requirements of ML estimations is that we have to assume an underlying probability distribution for the data. In this data, the multivariate normal distribution is assumed. However, in general, there are no analytical expressions for the variance component estimators obtained by using ML. These have to be done numerically. In this section, we will use the method of scoring as an algorithm to obtain the estimators.

We assume that $\boldsymbol{y}^* \sim N(\boldsymbol{Ef}, \boldsymbol{\Sigma})$, therefore the likelihood function is given by

$$L = (2\pi)^{-N/2}|\boldsymbol{\Sigma}|^{-1/2}\exp\left\{-\frac{1}{2}(\boldsymbol{y}^* - \boldsymbol{Ef})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{y}^* - \boldsymbol{Ef})\right\}$$

and the log-likelihood function is denoted by l:

$$l = -\frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}(\boldsymbol{y}^* - \boldsymbol{Ef})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{y}^* - \boldsymbol{Ef}). \tag{19}$$

To maximize l, we differentiate (19), first with respect to $\boldsymbol{\beta}$ which yields

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \boldsymbol{l}_{\boldsymbol{\beta}} \;\; = \;\; \boldsymbol{X}^\top \boldsymbol{E}^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{y}^* - \boldsymbol{Ef}). \tag{20}$$

where the differentiation of $\boldsymbol{\Sigma}$ with respect to $\boldsymbol{\beta}$ has been ignored.

Second, differentiating (19) with respect to σ_u^2 gives

$$\frac{\partial l}{\partial \sigma_u^2} = \boldsymbol{l}_{\sigma_u^2} = -\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_u\right) + \frac{1}{2}(\boldsymbol{y}^* - \boldsymbol{Ef})^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_u\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}^* - \boldsymbol{Ef}) \tag{21}$$

where $\boldsymbol{\Sigma}_u = \partial\boldsymbol{\Sigma}/\partial\sigma_u^2$. Finally, differentiating (19) with respect to σ_e^2 gives

$$\frac{\partial l}{\partial \sigma_e^2} = \boldsymbol{l}_{\sigma_e^2} \;\; = \;\; -\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\right) + \frac{1}{2}(\boldsymbol{y}^* - \boldsymbol{Ef})^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}^* - \boldsymbol{Ef}). \tag{22}$$

The ML estimators for $\boldsymbol{\beta}, \sigma_u^2$ and σ_e^2 are defined by setting (20), (21) and (22) to zero and solving for these parameters. Since $\boldsymbol{\Sigma}$ is a function of $\boldsymbol{\beta}, \sigma_u^2$ and σ_e^2 this needs to be done numerically. A numerical method commonly used for maximizing nonlinear functions is

the Newton-Raphson method. However, to avoid the heavy computational burden of the second-derivative matrix, another method that has been used is the method of scoring in which the Hessian is replaced by its expected value (Searle et al., 2006).

Let $\boldsymbol{\theta}$ denote all the parameters to be estimated, i.e., $\boldsymbol{\theta}^\top = (\boldsymbol{\beta}^\top \ \sigma_u^2 \ \sigma_e^2)$. The method of scoring thus uses an iteration scheme defined by

$$\boldsymbol{\theta}^{(m+1)} = \boldsymbol{\theta}^{(m)} + \{\boldsymbol{I}(\boldsymbol{\theta}^{(m)})\}^{-1} \frac{\partial l}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}^{(m)}},$$

where $\boldsymbol{I}(\boldsymbol{\theta}^{(m)})$ is the information matrix calculated using $\boldsymbol{\theta} = \boldsymbol{\theta}^{(m)}$.

We now develop $\boldsymbol{I}(\boldsymbol{\theta})$ for $\boldsymbol{\beta}, \sigma_u^2$ and σ_e^2.

Following (20),

$$\begin{aligned}
\frac{\partial^2 l}{\partial \boldsymbol{\beta} \ \partial \boldsymbol{\beta}^\top} = \boldsymbol{l}_{\boldsymbol{\beta}\boldsymbol{\beta}} &= -\boldsymbol{X}^\top \boldsymbol{E}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{E} \boldsymbol{X} \\
\frac{\partial^2 l}{\partial \boldsymbol{\beta} \ \partial \sigma_u^2} = \boldsymbol{l}_{\boldsymbol{\beta}\sigma_u^2} &= -\boldsymbol{X}^\top \boldsymbol{E}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma}_u \boldsymbol{\Sigma}^{-1} (\boldsymbol{y}^* - \boldsymbol{E}\boldsymbol{f}) \\
\frac{\partial^2 l}{\partial \boldsymbol{\beta} \ \partial \sigma_e^2} = \boldsymbol{l}_{\boldsymbol{\beta}\sigma_e^2} &= -\boldsymbol{X}^\top \boldsymbol{E}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma}^{-1} (\boldsymbol{y}^* - \boldsymbol{E}\boldsymbol{f})
\end{aligned}$$

Furthermore, following (21) and (22) we have

$$\begin{aligned}
\frac{\partial^2 l}{\partial \sigma_u^2 \ \partial \sigma_u^2} = \boldsymbol{l}_{\sigma_u^2 \sigma_u^2} &= \frac{1}{2} \mathrm{tr}\left(\boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma}_u \boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma}_u\right) - (\boldsymbol{y}^* - \boldsymbol{E}\boldsymbol{f})^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma}_u \boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma}_u \boldsymbol{\Sigma}^{-1} (\boldsymbol{y}^* - \boldsymbol{E}\boldsymbol{f}) \\
\frac{\partial^2 l}{\partial \sigma_u^2 \ \partial \sigma_e^2} = \boldsymbol{l}_{\sigma_u^2 \sigma_e^2} &= \frac{1}{2} \mathrm{tr}\left(\boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma}_u\right) - (\boldsymbol{y}^* - \boldsymbol{E}\boldsymbol{f})^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma}_u \boldsymbol{\Sigma}^{-1} (\boldsymbol{y}^* - \boldsymbol{E}\boldsymbol{f})
\end{aligned}$$

and

$$\frac{\partial^2 l}{\partial \sigma_e^2 \ \partial \sigma_e^2} = \boldsymbol{l}_{\sigma_e^2 \sigma_e^2} = \frac{1}{2} \mathrm{tr}\left(\boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma}^{-1}\right) - (\boldsymbol{y}^* - \boldsymbol{E}\boldsymbol{f})^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma}^{-1} (\boldsymbol{y}^* - \boldsymbol{E}\boldsymbol{f}).$$

In taking expected values of all derivatives obtained above we use $\mathrm{E}(\boldsymbol{y}^*) = \boldsymbol{E}\boldsymbol{f}$ and hence

$\mathrm{E}(\boldsymbol{y}^* - \boldsymbol{E}\boldsymbol{f}) = \mathbf{0}$, and

$$\mathrm{E}(\boldsymbol{y}^* - \boldsymbol{E}\boldsymbol{f})^\top \boldsymbol{C}(\boldsymbol{y}^* - \boldsymbol{E}\boldsymbol{f}) = \mathrm{tr}(\boldsymbol{C}\boldsymbol{\Sigma}) \quad \text{for non-stochastic } \boldsymbol{C}$$

gives the information matrix as

$$\begin{aligned}
\boldsymbol{I}\begin{bmatrix} \boldsymbol{\beta} \\ \sigma_u^2 \\ \sigma_e^2 \end{bmatrix} &= -\mathrm{E}\begin{bmatrix} \boldsymbol{l}_{\boldsymbol{\beta}\boldsymbol{\beta}} & \boldsymbol{l}_{\boldsymbol{\beta}\sigma_u^2} & \boldsymbol{l}_{\boldsymbol{\beta}\sigma_e^2} \\ \boldsymbol{l}_{\sigma_u^2\boldsymbol{\beta}} & \boldsymbol{l}_{\sigma_u^2\sigma_u^2} & \boldsymbol{l}_{\sigma_u^2\sigma_e^2} \\ \boldsymbol{l}_{\sigma_e^2\boldsymbol{\beta}} & \boldsymbol{l}_{\sigma_e^2\sigma_u^2} & \boldsymbol{l}_{\sigma_e^2\sigma_e^2} \end{bmatrix} \\
&= \frac{1}{2}\begin{bmatrix} 2\boldsymbol{X}^\top \boldsymbol{E}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{E}\boldsymbol{X} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_u\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_u\right) & \mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_u\right) \\ \mathbf{0} & \mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_u\right) & \mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}^{-1}\right) \end{bmatrix}. \qquad (23)
\end{aligned}$$

However, the variance component estimators obtained by this method are usually biased. Therefore, restricted maximum likelihood (REML) is an alternative maximum likelihood procedure which maximizes the likelihood of linear combinations of elements of $\mathbf{y}^*$ (McCulloch & Searle, 2001).

4.3. Pseudo-Restricted Maximum Likelihood (Pseudo-REML)

One criticism of the ML method is that in estimating variance components it takes no account of the degrees of freedom that are involved in estimating fixed effects (Searle et al., 2006). Also, the variance component estimators obtained by solving the likelihood equations are not, in general, in good agreement with those obtained by ANOVA methods, and they are generally biased, unlike the ANOVA estimators (Harville 1977, Searle et al. 2006).

The property of ML estimation not taking account of the degrees of freedom used for estimating fixed effects when estimating variance components is overcome by a method

known as restricted maximum likelihood (REML) (Searle et al., 2006). The procedure of this method is that rather than using $\boldsymbol{y}^*$ directly, REML is based on linear combinations of elements of $\boldsymbol{y}^*$, chosen in such a way that those combinations do not contain any fixed effects, no matter what their value. That is, with a set of values $\boldsymbol{s}^\top\boldsymbol{y}^*$, vectors $\boldsymbol{s}^\top$ are chosen so that $\mathrm{E}(\boldsymbol{s}^\top\boldsymbol{y}^*) = \boldsymbol{s}^\top\boldsymbol{E}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{0}$.

Hence

$$\boldsymbol{s}^\top\boldsymbol{E}\boldsymbol{X} = \boldsymbol{0}. \tag{24}$$

However, in terms of the linked data in this situation, variance of $\mathbf{y}^*$ is implicitly a function of $\boldsymbol{\beta}$. Therefore, we call this method "pseudo-REML" as we use the same procedure as REML except our variance still contains fixed effects.

With $\boldsymbol{E}\boldsymbol{X}$ of order $N \times p$ of rank r, there are only $N - r$ linearly independent vectors $\boldsymbol{s}^\top$ satisfying (24) (Searle et al., 2006). Using a set of such $N - r$ linearly independent vectors $\boldsymbol{s}^\top$ as rows of $\boldsymbol{S}^\top$, we then have $\boldsymbol{S}^\top\boldsymbol{y}^*$ where $\boldsymbol{S}^\top$ is a $(N-r)\times N$ matrix whose rows are any $N-r$ linearly independent rows of the matrix $\boldsymbol{I}-\boldsymbol{E}\boldsymbol{X}\{(\boldsymbol{E}\boldsymbol{X})^\top(\boldsymbol{E}\boldsymbol{X})\}^{-1}(\boldsymbol{E}\boldsymbol{X})^\top$.

With $\boldsymbol{y}^* \sim N(\boldsymbol{E}\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{\Sigma})$ we have, for $\boldsymbol{S}^\top\boldsymbol{E}\boldsymbol{X} = \boldsymbol{0}$

$$\boldsymbol{S}^\top\boldsymbol{y}^* \sim N(\boldsymbol{0}, \boldsymbol{S}^\top\boldsymbol{\Sigma}\boldsymbol{S}).$$

With l_R being the log likelihood function of $\boldsymbol{S}^\top\boldsymbol{y}^*$ define

$$l_R = -\frac{1}{2}(N-r)\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{S}^\top\boldsymbol{\Sigma}\boldsymbol{S}| - \frac{1}{2}\boldsymbol{y}^{*\top}\boldsymbol{M}\boldsymbol{y}^* \quad \text{where } \boldsymbol{M} = \boldsymbol{S}(\boldsymbol{S}^\top\boldsymbol{\Sigma}\boldsymbol{S})^{-1}\boldsymbol{S}^\top.$$

Note that,

$$\begin{aligned}
\frac{\partial \boldsymbol{M}}{\partial \sigma_u^2} &= \frac{\partial}{\partial \sigma_u^2}\boldsymbol{S}(\boldsymbol{S}^\top\boldsymbol{\Sigma}\boldsymbol{S})^{-1}\boldsymbol{S}^\top \\
&= -\boldsymbol{S}(\boldsymbol{S}^\top\boldsymbol{\Sigma}\boldsymbol{S})^{-1}\boldsymbol{S}^\top\boldsymbol{\Sigma}_u\boldsymbol{S}(\boldsymbol{S}^\top\boldsymbol{\Sigma}\boldsymbol{S})^{-1}\boldsymbol{S}^\top = -\boldsymbol{M}\boldsymbol{\Sigma}_u\boldsymbol{M} \\
\frac{\partial \boldsymbol{M}}{\partial \sigma_e^2} &= -\boldsymbol{M}\boldsymbol{M}.
\end{aligned}$$

For the information matrix we need first and second derivatives of l_R :

$$\frac{\partial l_R}{\partial \sigma_u^2} = \boldsymbol{l}_{\sigma_u^2} = -\frac{1}{2}\mathrm{tr}\left(\boldsymbol{M}\boldsymbol{\Sigma}_u\right) + \frac{1}{2}\boldsymbol{y}^{*\top}\boldsymbol{M}\boldsymbol{\Sigma}_u\boldsymbol{M}\boldsymbol{y}^*$$

$$\frac{\partial^2 l_R}{\partial \sigma_u^2\, \partial \sigma_u^2} = \boldsymbol{l}_{\sigma_u^2\sigma_u^2} = \frac{1}{2}\mathrm{tr}\left(\boldsymbol{M}\boldsymbol{\Sigma}_u\boldsymbol{M}\boldsymbol{\Sigma}_u\right) - \boldsymbol{y}^{*\top}\boldsymbol{M}\boldsymbol{\Sigma}_u\boldsymbol{M}\boldsymbol{\Sigma}_u\boldsymbol{M}\boldsymbol{y}^*$$

$$\frac{\partial l_R}{\partial \sigma_e^2} = \boldsymbol{l}_{\sigma_e^2} = -\frac{1}{2}\mathrm{tr}\left(\boldsymbol{M}\right) + \frac{1}{2}\boldsymbol{y}^{*\top}\boldsymbol{M}\boldsymbol{M}\boldsymbol{y}^*$$

$$\frac{\partial^2 l_R}{\partial \sigma_e^2\, \partial \sigma_e^2} = \boldsymbol{l}_{\sigma_e^2\sigma_e^2} = \frac{1}{2}\mathrm{tr}\left(\boldsymbol{M}\boldsymbol{M}\right) - \boldsymbol{y}^{*\top}\boldsymbol{M}\boldsymbol{M}\boldsymbol{M}\boldsymbol{y}^*.$$

In Appendix II we show that taking expected values of second derivatives of l_R gives the information matrix as

$$\therefore \quad \mathbf{I}\begin{bmatrix}\sigma_u^2\\ \sigma_e^2\end{bmatrix} = -\mathrm{E}\begin{bmatrix}\boldsymbol{l}_{\sigma_u^2\sigma_u^2} & \boldsymbol{l}_{\sigma_u^2\sigma_e^2}\\ \boldsymbol{l}_{\sigma_e^2\sigma_u^2} & \boldsymbol{l}_{\sigma_e^2\sigma_e^2}\end{bmatrix}$$

$$= \frac{1}{2}\begin{bmatrix}\mathrm{tr}\left(\boldsymbol{M}\boldsymbol{\Sigma}_u\boldsymbol{M}\boldsymbol{\Sigma}_u\right) & \mathrm{tr}\left(\boldsymbol{M}\boldsymbol{M}\boldsymbol{\Sigma}_u\right)\\ \mathrm{tr}\left(\boldsymbol{M}\boldsymbol{M}\boldsymbol{\Sigma}_u\right) & \mathrm{tr}\left(\boldsymbol{M}\boldsymbol{M}\right)\end{bmatrix}. \tag{25}$$

5. Simulation studies

In the previous sections we have theoretically shown the various bias adjusted estimators of the coefficient parameters and variance components in the linear mixed model. In this section we illustrate simulation results comparing them under repeated application of probabilistic linkage based on the exchangeable linkage error model defined by equations (1) - (5).

In a simulation, data were generated for a population of size $N = 800$ which consists of four blocks of size 200 in each block. Also, each block is made up of 50 groups with four subjects in each group. Values of X were then independently drawn from the uniform

distribution over [0,1] with corresponding values of Y given by

$$y_{ig} = 2 + 4x_{ig} + u_g + e_{ig}$$

where e_{ig} were independently drawn from the $N(0, 3)$ distribution and u_g were independently drawn from the $N(0, 1)$ distribution. Then true data pairs (y_{ig}, x_{ig}) were randomly allocated to blocks and groups. Next, linked data pairs (y^*_{ig}, x_{ig}) were generated by using the exchangeable linkage errors model defined by (1) - (5) with correct linkage probabilities $\lambda_1 = 1,\ \lambda_2 = 0.95,\ \lambda_3 = 0.85$ and $\lambda_4 = 0.75$. Note that all links for block 1 were assumed to be correct, while those for blocks 2, 3 and 4 were assumed to have some errors. Here we present simulation results for two scenarios. The first set of results were obtained from known linkage probabilities. The second set of results were obtained from estimated linkage probabilities by taking random samples of $m_q = 25$ linked pairs from each of block 2, 3 and 4 and checking to see whether these sampled links were correct. Following Chambers (2009), the estimate of λ_q is calculated as

$$\hat{\lambda}_q = \min\left\{m_q^{-1}(m_q - 0.5), \max(M_q^{-1}, l_q)\right\}$$

where l_q is the proportion of correctly linked pairs identified in the audit sample in block q.

A total of 800 independent simulations were carried out. Table 1 illustrates the relative biases and relative root mean squared errors of the coefficient estimators described in Section 3 whereas Table 2 reveals the relative biases and relative root mean squared errors of variance components estimators described in Section 4. The WLS estimator based on perfectly linked data, TR, as well as the naive WLS estimator, were obtained using the default settings of the `lme` function in the R software package. Note that variance components estimators obtained using ANOVA method are functions of $\boldsymbol{\beta}$. The estimator R, A, C and B presented in Table 2 represent the ANOVA estimators obtained using those coefficient estimators. The actual coverages of the nominal 95% confidence intervals for all of the model parameters are then illustrated in Table 3.

The results set out in Table 1 show that the naive WLS estimator that just used the linked data was clearly biased. Since linkage error is a particular type of measurement error, this bias attenuated the estimate of the slope parameter and exaggerated that of the intercept. On the other hand, all four of the adjusted estimators corrected this bias in which the estimator C, the empirical BLUE from (16), was the most efficient. Under Scenario 2 where linkage probabilities were estimated by taking small audit samples, we see that the results were also in a similar way of those under Scenario 1 except that the estimator MLE turned out to be the most efficient in this case.

TABLE 1
Simulation results for the coefficient parameters of the linear mixed model

Estimator	Relative Bias		Relative RMSE	
	Intercept	Slope	Intercept	Slope
Scenario 1: Linkage Probabilities Correctly Specified				
TR	0.04	0.03	18.48	19.07
Naïve	10.99	-10.92	24.27	29.12
R	-0.18	0.25	19.99	21.70
A	-0.14	0.21	19.98	21.64
C	-0.11	0.18	20.01	21.65
MLE	-0.24	0.31	20.00	21.65
Scenario 2: Linkage Probabilities Estimated From Audit Sample				
TR	0.29	-0.33	17.77	18.12
Naïve	11.46	-11.50	24.27	29.65
R	-0.12	0.07	19.53	21.22
A	0.22	-0.26	19.31	20.92
C	0.31	-0.35	19.26	20.89
MLE	0.19	-0.23	19.27	20.90

Investigation of the results displayed in Table 2 shows that the naive variance components estimators that just used the linked data were also biased. As expected, the estimator obtained using the ML approach was slightly biased since degrees of freedom for fixed effects did not get taken into account. All of the remaining adjusted estimators were essentially unbiased in which the estimator REML was the most efficient. The results under Scenario 2 were also in the same direction of those under Scenario 1.

The results displayed in Table 3 show that variance estimators that allowed for the extra variability induced by estimation of these parameters led to confidence intervals with good coverage properties.

TABLE 2
Simulation results for the variance components of the linear mixed model

Estimator	Relative Bias		Relative RMSE	
	Between-Group	Within-Group	Between-Group	Within-Group
Scenario 1: Linkage Probabilities Correctly Specified				
TR	1.07	0.04	31.30	15.49
Naïve	-20.82	5.50	34.69	23.27
R	0.50	-0.13	37.17	17.21
A	0.50	-0.11	37.17	17.13
C	0.50	-0.10	37.17	17.13
MLE	-2.84	-0.07	34.76	16.93
REML	0.93	-0.02	35.32	16.95
Scenario 2: Linkage Probabilities Estimated From Audit Sample				
TR	-1.38	0.13	29.71	15.51
Naïve	-22.30	5.53	34.64	23.63
R	-0.01	-0.30	36.78	17.94
A	-0.01	-0.19	36.78	17.78
C	-0.01	-0.16	36.78	17.77
MLE	-5.04	0.05	33.75	17.46
REML	-1.34	0.09	34.05	17.49

TABLE 3
The actual coverages of the nominal 95% confidence intervals for the model parameters

Estimator	Coverage			
	Intercept	Slope	Between-Group	Within-Group
Scenario 1: Linkage Probabilities Correctly Specified				
TR	95.4	94.5	96.9	94.5
Naïve	84.6	78.9	98.2	82.5
R	95.4	95.4	93.8	95.9
A	95.2	94.6	93.8	95.9
C	95.1	94.5	93.8	96.1
MLE	94.9	94.5	92.2	94.1
REML	-	-	94.6	94.4
Scenario 2: Linkage Probabilities Estimated From Audit Sample				
TR	95.8	95.9	97.1	94.8
Naïve	86.5	78.6	98.2	79.9
R	95.4	96.5	92.4	95.4
A	95.4	96.1	92.4	95.6
C	95.2	96.1	92.4	95.5
MLE	94.6	95.5	91.9	94.4
REML	-	-	93.4	94.4

6. Closing remarks

In this paper we have shown how to develop the inferential framework of Chambers (2009) to obtain unbiased regression parameter estimates when fitting a linear mixed model to probabilistically linked data. Moreover, since estimation of variance components is also an important objective when fitting a mixed model, we have appropriately modified standard methods of variance components estimation in order to account for linkage error. Particularly, we focus on three widely used methods of variance components estimation: ANOVA, psuedo-ML and psuedo-REML. Our simulation results indicate that all the methods developed in this paper work reasonably well in terms of correcting bias induced by linkage error. However, they also show an evidence of increases in variability due to application of linear mixed models to the linked data.

Although the theoretical results described in the previous sections are well developed, there are a lot of issues that still need investigation such as application these methods to real life linked data, the characteristics of the linkage situation, and application to longitudinal modeling.

Appendix

I. ANOVA Estimation

If the linkage is not perfect, then we have

$$\begin{aligned}
\text{SSA} &= \boldsymbol{y}^{*'}\boldsymbol{B}\boldsymbol{y}^{*} \\
\text{E}_X(\text{SSA}) &= \text{E}_X\left(\sum_q \boldsymbol{y}_q^{*'}\boldsymbol{B}_{qq}\boldsymbol{y}_q^{*}\right) + \text{E}_X\left(\sum_q\sum_{r\neq q} \boldsymbol{y}_q^{*'}\boldsymbol{B}_{qr}\boldsymbol{y}_r^{*}\right) \\
&= \sum_q \text{E}_X\left(\boldsymbol{y}_q^{*'}\boldsymbol{B}_{qq}\boldsymbol{y}_q^{*}\right) + \sum_q\sum_{r\neq q} \text{E}_X\left(\boldsymbol{y}_q^{*'}\boldsymbol{B}_{qr}\boldsymbol{y}_r^{*}\right).
\end{aligned}$$

Now, we consider the first term of $\mathrm{E}_X(\mathrm{SSA})$.

$$\begin{aligned}
\sum_q \mathrm{E}_X\left(\boldsymbol{y}_q^{*'}\boldsymbol{B}_{qq}\boldsymbol{y}_q^*\right) &= \sum_q \mathrm{tr}\left\{\boldsymbol{B}_{qq}Var(\mathbf{y}_q^*)\right\} + \sum_q \boldsymbol{f}_q'\boldsymbol{E}_q'\boldsymbol{B}_{qq}\boldsymbol{E}_q\boldsymbol{f}_q \\
&= \sum_q \mathrm{tr}\left\{\boldsymbol{B}_{qq}(\sigma_u^2\boldsymbol{K}_q + \sigma_e^2\boldsymbol{I}_q + \boldsymbol{V}_q)\right\} + \sum_q \mathrm{f}_q'\boldsymbol{E}_q'\boldsymbol{B}_{qq}\boldsymbol{E}_q\boldsymbol{f}_q \\
&= \sigma_u^2\sum_q \mathrm{tr}\left(\boldsymbol{B}_{qq}\boldsymbol{K}_q\right) + \sigma_e^2\sum_q \mathrm{tr}\left(\boldsymbol{B}_{qq}\right) \\
&\quad + \sum_q \mathrm{tr}\left\{(\boldsymbol{B}_{qq}\boldsymbol{V}_q) + \boldsymbol{f}_q'\boldsymbol{E}_q'\boldsymbol{B}_{qq}\boldsymbol{E}_q\boldsymbol{f}_q\right\}
\end{aligned}$$

The second term of $\mathrm{E}_X(\mathrm{SSA})$ is then given by

$$\begin{aligned}
\sum_q\sum_{r\neq q} \mathrm{E}_X\left(\boldsymbol{y}_q^{*'}\boldsymbol{B}_{qr}\boldsymbol{y}_r^*\right) &= \sum_q\sum_{r\neq q} \mathrm{E}_X\left\{\mathrm{E}_{X|y_r^*}\left(\boldsymbol{y}_q^{*'}\boldsymbol{B}_{qr}\boldsymbol{y}_r^*\right)\right\} \\
&= \sum_q\sum_{r\neq q} \mathrm{E}_X\left[\left\{\boldsymbol{E}_q\boldsymbol{f}_q + \boldsymbol{\Sigma}_{qr}\boldsymbol{\Sigma}_{rr}^{-1}(\boldsymbol{y}_r^* - \boldsymbol{E}_r\boldsymbol{f}_r)\right\}'\boldsymbol{B}_{qr}\boldsymbol{y}_r^*\right] \\
&= \sum_q\sum_{r\neq q} \mathrm{E}_X\left[\left\{\boldsymbol{f}_q'\boldsymbol{E}_q' + (\boldsymbol{y}_r^{*'} - \boldsymbol{f}_r'\boldsymbol{E}_r')\boldsymbol{\Sigma}_{rr}^{-1}\boldsymbol{\Sigma}_{qr}'\right\}\boldsymbol{B}_{qr}\boldsymbol{y}_r^*\right] \\
&= \sum_q\sum_{r\neq q} \mathrm{E}_X\left(\boldsymbol{f}_q'\boldsymbol{E}_q'\boldsymbol{B}_{qr}\boldsymbol{y}_r^*\right) + \sum_q\sum_{r\neq q} \mathrm{E}_X\left(\boldsymbol{y}_r^{*'}\boldsymbol{\Sigma}_{rr}^{-1}\boldsymbol{\Sigma}_{qr}'\boldsymbol{B}_{qr}\boldsymbol{y}_r^*\right) \\
&\quad - \sum_q\sum_{r\neq q} \mathrm{E}_X\left(\boldsymbol{f}_r'\boldsymbol{E}_r'\boldsymbol{\Sigma}_{rr}^{-1}\boldsymbol{\Sigma}_{qr}'\boldsymbol{B}_{qr}\boldsymbol{y}_r^*\right) \\
&= \sum_q\sum_{r\neq q} \boldsymbol{f}_q'\boldsymbol{E}_q'\boldsymbol{B}_{qr}\boldsymbol{E}_r\boldsymbol{f}_r + \sum_q\sum_{r\neq q} \mathrm{tr}\left(\boldsymbol{\Sigma}_{rr}^{-1}\boldsymbol{\Sigma}_{qr}'\boldsymbol{B}_{qr}\boldsymbol{\Sigma}_{rr}\right) \\
&\quad + \sum_q\sum_{r\neq q} \boldsymbol{f}_r'\boldsymbol{E}_r'\boldsymbol{\Sigma}_{rr}^{-1}\boldsymbol{\Sigma}_{qr}'\boldsymbol{B}_{qr}\boldsymbol{E}_r\boldsymbol{f}_r - \sum_q\sum_{r\neq q} \boldsymbol{f}_r'\boldsymbol{E}_r'\boldsymbol{\Sigma}_{rr}^{-1}\boldsymbol{\Sigma}_{qr}'\boldsymbol{B}_{qr}\boldsymbol{E}_r\boldsymbol{f}_r \\
&= \sum_q\sum_{r\neq q} \boldsymbol{f}_q'\boldsymbol{E}_q'\boldsymbol{B}_{qr}\boldsymbol{E}_r\boldsymbol{f}_r + \sigma_u^2\sum_q\sum_{r\neq q} \mathrm{tr}\left(\boldsymbol{E}_r\boldsymbol{Z}_r\boldsymbol{Z}_q'\boldsymbol{E}_q'\boldsymbol{B}_{qr}\right)
\end{aligned}$$

where $\boldsymbol{\Sigma}_{qr}$ is the covariance between $\boldsymbol{y}_q^*$ and $\boldsymbol{y}_r^*$.

Combining the first and second term of $\mathrm{E}_X(\mathrm{SSA})$ gives

$$\begin{aligned}\mathrm{E}_X(\mathrm{SSA}) &= \sigma_u^2\left\{\sum_q \mathrm{tr}\left(\boldsymbol{B}_{qq}\boldsymbol{K}_q\right)+\sum_q\sum_{r\neq q}\mathrm{tr}\left(\boldsymbol{E}_r\boldsymbol{Z}_r\boldsymbol{Z}_q'\boldsymbol{E}_q'\boldsymbol{B}_{qr}\right)\right\}+\sigma_e^2\sum_q\mathrm{tr}\left(\boldsymbol{B}_{qq}\right)\\ &\quad+\sum_q\left\{\mathrm{tr}\left(\boldsymbol{B}_{qq}\boldsymbol{V}_q\right)+\boldsymbol{f}_q'\boldsymbol{E}_q'\boldsymbol{B}_{qq}\boldsymbol{E}_q\boldsymbol{f}_q\right\}+\sum_q\sum_{r\neq q}\boldsymbol{f}_q'\boldsymbol{E}_q'\boldsymbol{B}_{qr}\boldsymbol{E}_r\boldsymbol{f}_r\\ &= \sigma_u^2 a+\sigma_e^2 b+m_0\end{aligned}$$

where

$$\begin{aligned}a &= \sum_q\mathrm{tr}\left(\boldsymbol{B}_{qq}\boldsymbol{K}_q\right)+\sum_q\sum_{r\neq q}\mathrm{tr}\left(\boldsymbol{E}_r\boldsymbol{Z}_r\boldsymbol{Z}_q'\boldsymbol{E}_q'\boldsymbol{B}_{qr}\right)\\ b &= \sum_q\mathrm{tr}\left(\boldsymbol{B}_{qq}\right)\\ m_0 &= \sum_q\left\{\mathrm{tr}\left(\boldsymbol{B}_{qq}\boldsymbol{V}_q\right)+\boldsymbol{f}_q'\boldsymbol{E}_q'\boldsymbol{B}_{qq}\boldsymbol{E}_q\boldsymbol{f}_q\right\}+\sum_q\sum_{r\neq q}\boldsymbol{f}_q'\boldsymbol{E}_q'\boldsymbol{B}_{qr}\boldsymbol{E}_r\boldsymbol{f}_r.\end{aligned}$$

Similarly,

$$\begin{aligned}\mathrm{E}_X(\mathrm{SSE}) &= \sigma_u^2\left\{\sum_q \mathrm{tr}\left(\boldsymbol{C}_{qq}\boldsymbol{K}_q\right)+\sum_q\sum_{r\neq q}\mathrm{tr}\left(\boldsymbol{E}_r\boldsymbol{Z}_r\boldsymbol{Z}_q'\boldsymbol{E}_q'\boldsymbol{C}_{qr}\right)\right\}+\sigma_e^2\sum_q\mathrm{tr}\left(\boldsymbol{C}_{qq}\right)\\ &\quad+\sum_q\left\{\mathrm{tr}\left(\boldsymbol{C}_{qq}\boldsymbol{V}_q\right)+\boldsymbol{f}_q'\boldsymbol{E}_q'\boldsymbol{C}_{qq}\boldsymbol{E}_q\boldsymbol{f}_q\right\}+\sum_q\sum_{r\neq q}\boldsymbol{f}_q'\boldsymbol{E}_q'\boldsymbol{C}_{qr}\boldsymbol{E}_r\boldsymbol{f}_r\\ &= \sigma_u^2 c+\sigma_e^2 d+n_0\end{aligned}$$

where

$$\begin{aligned}c &= \sum_q\mathrm{tr}\left(\boldsymbol{C}_{qq}\boldsymbol{K}_q\right)+\sum_q\sum_{r\neq q}\mathrm{tr}\left(\boldsymbol{E}_r\boldsymbol{Z}_r\boldsymbol{Z}_q'\boldsymbol{E}_q'\boldsymbol{C}_{qr}\right)\\ d &= \sum_q\mathrm{tr}\left(\boldsymbol{C}_{qq}\right)\\ n_0 &= \sum_q\left\{\mathrm{tr}\left(\boldsymbol{C}_{qq}\boldsymbol{V}_q\right)+\boldsymbol{f}_q'\boldsymbol{E}_q'\boldsymbol{C}_{qq}\boldsymbol{E}_q\boldsymbol{f}_q\right\}+\sum_q\sum_{r\neq q}\boldsymbol{f}_q'\boldsymbol{E}_q'\boldsymbol{C}_{qr}\boldsymbol{E}_r\boldsymbol{f}_r.\end{aligned}$$

By solving two linear equations,

$$\hat{\sigma}_u^2 a + \hat{\sigma}_e^2 b = m \quad ; m = \text{SSA} - m_0$$

and

$$\hat{\sigma}_u^2 c + \hat{\sigma}_e^2 d = n \quad ; n = \text{SSE} - n_0$$

it yields the estimators

$$\hat{\sigma}_e^2 = \frac{mc - na}{bc - da}$$

and

$$\hat{\sigma}_u^2 = \frac{m - \hat{\sigma}_e^2 b}{a}.$$

II. Pseudo-REML information matrix derivation

$$\begin{aligned}
-\text{E}\, \boldsymbol{l}_{\sigma_u^2\sigma_u^2} &= -\frac{1}{2}\text{tr}\left(\boldsymbol{M\Sigma}_u\boldsymbol{M\Sigma}_u\right) + \text{tr}\left(\boldsymbol{M\Sigma}_u\boldsymbol{M\Sigma}_u\boldsymbol{M\Sigma}\right) \\
&\quad + \boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{E}^\top\boldsymbol{M\Sigma}_u\boldsymbol{M\Sigma}_u\boldsymbol{MEX}\boldsymbol{\beta} \\
&= -\frac{1}{2}\text{tr}\left(\boldsymbol{M\Sigma}_u\boldsymbol{M\Sigma}_u\right) + \text{tr}\left(\boldsymbol{\Sigma}_u\boldsymbol{M\Sigma}_u\boldsymbol{M\Sigma M}\right) + \boldsymbol{0}, \quad \because\ \boldsymbol{MEX} = \boldsymbol{0} \\
&= \frac{1}{2}\text{tr}\left(\boldsymbol{M\Sigma}_u\boldsymbol{M\Sigma}_u\right), \quad \because\ \boldsymbol{M\Sigma M} = \boldsymbol{M} \\
-\text{E}\, \boldsymbol{l}_{\sigma_u^2\sigma_e^2} &= -\frac{1}{2}\text{tr}\left(\boldsymbol{MM\Sigma}_u\right) + \text{tr}\left(\boldsymbol{MM\Sigma}_u\boldsymbol{M\Sigma}\right) + \boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{E}^\top\boldsymbol{MM\Sigma}_u\boldsymbol{MEX}\boldsymbol{\beta} \\
&= \frac{1}{2}\text{tr}\left(\boldsymbol{MM\Sigma}_u\right) \\
-\text{E}\, \boldsymbol{l}_{\sigma_e^2\sigma_e^2} &= -\frac{1}{2}\text{tr}\left(\boldsymbol{MM}\right) + \text{tr}\left(\boldsymbol{MMM\Sigma}\right) + \boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{E}^\top\boldsymbol{MMMEX}\boldsymbol{\beta} \\
&= \frac{1}{2}\text{tr}\left(\boldsymbol{MM}\right).
\end{aligned}$$

References

[1] Chambers, R. (2009). Regression Analysis Of Probability-Linked Data. *Official Statistics Research Series* 4.

[2] Harville, D. A. (1977). Maximum-likelihood approaches to variance component estimation and to related problems. *J. Amer. Statist. Assoc.* 72, 320-340.

[3] Lahiri, P. & Larsen, M. D. (2005). Regression Analysis With Linked Data. *J. Amer. Statist. Assoc.* 100, 222-230.

[4] McCulloch, C. E. & Searle, S. R. (2001). *Generalized, Linear, and Mixed Models*. New York: John Wiley & Sons.

[5] Neter, J., Maynes, E. S. & Ramanathan, R. (1965). The effect of mismatching on the measurement of response error. *J. Amer. Statist. Assoc.* 60, 1005-1027.

[6] Searle, S. R., Casella, G. & McCulloch, C. E. (2006). *Variance Components*. New York: John Wiley & Sons.

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