1-1-2014

Linear regression with nested errors using probability-linked data

Klairung Samart
Prince of Songkla University, ks208@uow.edu.au

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

Linear regression with nested errors using probability-linked data

Abstract

Probabilistic matching of records is widely used to create linked data sets for use in health science, epidemiological, economic, demographic and sociological research. 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 to be used when fitting a linear model with nested errors to probabilistically linked data. Since estimation of variance components is typically an important objective when fitting such a model, we also 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 perform reasonably well when compared to standard estimation methods that ignore linkage errors. 2014 Australian Statistical Publishing Association Inc. Published by Wiley Publishing Asia Pty Ltd.

Keywords

linear, nested, errors, probability, linked, data, regression

Disciplines

Engineering | Science and Technology Studies

Publication Details

Samart, K. & Chambers, R. L. (2014). Linear regression with nested errors using probability-linked data. Australian and New Zealand Journal of Statistics, 56 (1), 27-46.

LINEAR REGRESSION WITH NESTED ERRORS USING PROBABILITY-LINKED DATA

KLAIRUNG SAMART[1,2*] AND RAY CHAMBERS[2]

Prince of Songkla University and University of Wollongong

Summary

Probabilistic matching of records is widely used to create linked data sets for use in health science, epidemiological, economic, demographic and sociological research. 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 to be used when fitting a linear model with nested errors to probabilistically linked data. Since estimation of variance components is typically an important objective when fitting such a model, we also 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 perform reasonably well when compared to standard estimation methods that ignore linkage errors.

Key words: analysis of variance; linkage error; maximum likelihood; measurement

*Author to whom correspondence should be addressed.

[1]Department of Mathematics and Statistics, Faculty of Science, Prince of Songkla University, Songkhla, 90112, Thailand. e-mail: klairung.s@psu.ac.th

[2]National Institute for Applied Statistics Research Australia, School of Mathematics and Applied Statistics, University of Wollongong, Wollongong NSW 2522, Australia.

Acknowledgements. The research described in this paper was financially supported by the Prince of Songkla University and a University Postgraduate Award from the University of Wollongong.

error; mixed model; record matching; restricted maximum likelihood; weighted least squares.

1. Introduction

Linked data sets, created by probabilistic matching of records, are widely used for research in health, epidemiology, economics, demography, sociology and many other scientific areas. However, probabilistic matching can lead to linkage errors, which is a type of measurement error and 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 of linked data is undertaken. Although there have been a number of statistical methods developed for efficient linkage (see Herzog et al., 2007), there has been comparatively little methodological research carried out on the impact of linkage errors on analysis of linked data.

An early reference is Neter et al. (1965), who found that relatively small amounts of linkage error can lead to a substantial bias when estimating a regression relationship. Scheuren & Winkler (1993, 1997) investigated the effect of linkage errors on the bias of ordinary least squares estimators in a standard linear regression model and proposed a method of adjusting for the bias. However, their estimator is not unbiased in general. Subsequently, Lahiri & Larsen (2005) proposed an alternative unbiased estimator, based on a regression model with transformed covariates. In their simulations, they found that their approach performed very well across a range of situations.

A methodological framework for analysis of linked data was developed in Chambers (2009). Under this approach, appropriate modifications to standard statistical analysis methods are used to ensure that they remain unbiased when applied to probabilistically linked data. However, this development assumes that measurements are mutually independent. This is unrealistic when they correspond to observations from clusters of correlated statistical units, such as members of a family, patients in a hospital or students in a school. Nested error models are often used when analyzing such data. Consequently, in this paper we develop methods for efficient fitting of linear models with nested errors

to probabilistically linked data.

The structure of the paper is as follows. In the following section we review the linkage error model used in Chambers (2009). In Section 3 we then describe a framework for fitting a linear model with nested errors given linked data generated under this linkage error model, and obtain unbiased estimators of regression coefficients for this case. In Section 4 we next describe three methods of variance components estimation using probabilistically linked data: analysis of variance, pseudo-maximum likelihood and pseudo-restricted maximum likelihood. Simulation results that compare the estimators defined in the previous sections are presented in Section 5. Section 6 concludes the paper with a summary of its results and suggestions for further research.

2. The exchangeable linkage errors model

In this section we summarize the linkage error model underpinning the development in Chambers (2009). We assume that there is a population of N units, indexed by $i = 1, ..., N$. For each unit in this population, there exists an observable value 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, and in particular to estimate the coefficients of a linear model for the regression of Y on $\boldsymbol{X}$. However, there is no single database that contains the joint population values of Y and $\boldsymbol{X}$. Instead, there are two population registers, which we denote by register A and register B, that separately contain these values, i.e. register A contains the values of Y and register B contains the values of $\boldsymbol{X}$. Both registers refer to the same population and have no duplicates, so each consists of N records.

Given a unique identifier for each unit in the population, it is straightforward to link the records from the two individual registers to create one joint register. However, such an identifier usually does not exist. Instead, some form of probability-based matching is used to link records from the two registers. We assume that the resulting linkage is complete (i.e. all records are linked) and one to one between register A and register B. However, since the linkage is probabilistic, the linked data set can contain linkage errors,

i.e. records where the values of Y and $\boldsymbol{X}$ that ostensibly belong to the same population unit actually come from different population units.

In most cases, there are common auxiliary variables measured on both registers. These variables are typically used for probability matching, and allow us to assume that the linked records can be partitioned into Q distinct sets or blocks such that there is no possibility that linked records in different blocks contain data for the same population unit. We characterize this situation by defining a categorical variable F such that different blocks correspond to different values of F. In other words, if a record on one register does not have the same value of F as the record on the other register, then the two records cannot correspond to the same unit in the population. An immediate consequence is that only linked records with the same value of F can contain linkage errors, i.e. linkage errors can only occur within a block.

Without loss of generality, we assume that F takes the Q distinct values $1...Q$, and let block q correspond to the M_q population units with $F = q$ so $N = \sum_q M_q$. Let y^*_{iq} denote the Y-value from block q on the A register that is matched to the $\boldsymbol{X}$-value $\boldsymbol{x}_{iq}$ in block q on the B register i.e. there are M_q linked data pairs $(y^*_{iq}, \boldsymbol{x}_{iq})$ in block q. We denote the vector of dimension M_q of the linked values y^*_{iq} in block q by $\boldsymbol{y}^*_q$ and similarly let $\boldsymbol{X}_q$ denote the matrix with rows defined by the values $\boldsymbol{x}_{iq}$ in the same block. Finally, we use $\boldsymbol{y}_q$ to denote the unknown vector of the true Y values y_{iq} in block q that are associated with $\boldsymbol{X}_q$. Note that if linkage is perfect, then $y^*_{iq} = y_{iq}$ so $\boldsymbol{y}^*_q = \boldsymbol{y}_q$.

Since linkage is assumed to be complete and one to one between register A and register B, randomness in the outcome of the linkage process can be modeled via the identity

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

where $\boldsymbol{A}_q = [a^q_{ij}]$ is an unknown random permutation matrix of dimension $M_q \times M_q$. Given that linkage errors can only occur within blocks, it is natural to assume that $\boldsymbol{A}_{q_1}$ and $\boldsymbol{A}_{q_2}$ are independently distributed when $q_1 \neq q_2$. We further assume that linkage is *non-informative* at each level of F in the sense that the distribution of $\boldsymbol{A}_q$ is independent of $\boldsymbol{y}_q$ given $\boldsymbol{X}_q$, and define $\boldsymbol{T}_q = \mathrm{E}(\boldsymbol{A}_q)$

The distribution of linkage errors will depend on the characteristics of the probability-linking method actually used. In many cases, this information will not be available to the data analyst. Consequently, we follow Neter et al. (1965) and model the distribution of linkage errors using an exchangeable linkage errors (ELE) model. Under this model, for each value of q

$$\Pr(\text{correct linkage}) = \Pr(a_{ii}^q = 1) = \lambda_q \tag{2}$$

and, for $i \neq j$,

$$\Pr(\text{incorrect linkage}) = \Pr(a_{ij}^q = 1) = \gamma_q. \tag{3}$$

Given (2) and (3) hold, $\boldsymbol{T}_q$ is then of the form

$$\boldsymbol{T}_q = (\lambda_q - \gamma_q)\boldsymbol{I}_q + \gamma_q \mathbf{1}_q \mathbf{1}_q^\top \tag{4}$$

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 $\boldsymbol{A}_q \mathbf{1}_q = \mathbf{1}_q$, $\mathbf{1}_q^\top \boldsymbol{T}_q = \mathbf{1}_q^\top$ and $\boldsymbol{T}_q \mathbf{1}_q = \mathbf{1}_q$. That is, (4) implies

$$\gamma_q \;=\; \frac{1-\lambda_q}{M_q - 1}.$$

A major advantage of the ELE model is that it only requires one parameter (λ_q) to completely specify the first order properties of the probability-linkage mechanism.

3. Estimation of regression coefficients

In this section we consider the situation where a two level linear model with nested errors is the focus of inference. We therefore introduce an auxiliary grouping variable Z which takes values $1, ..., G$, and let group g correspond to the N_g population units with $Z = g$ such that $M_q = \sum_g n_{qg}$ and $N_g = \sum_q n_{qg}$ where n_{qg} is the number of population units in block q and group g. That is, we allow distinct units within the same group to be independently linked (correctly or incorrectly) in different blocks. We assume throughout that the values of Z in the linked data are correct, i.e. this variable is stored on register B. The two level linear model for the regression of Y on $\boldsymbol{X}$ in the population is then

given by

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

where $\boldsymbol{Y}_B$ is the vector of true values y_i of Y associated with the records on the B register, $\boldsymbol{X}_B$ is the matrix whose rows correspond to the values $\boldsymbol{x}_i$ of $\boldsymbol{X}$ on the B register, and $\boldsymbol{Z}_B$ is the matrix that identifies the group to which each record in the B register belongs. The vector $\boldsymbol{u} = \{u_g\}$ is a vector of random group effects, while $\boldsymbol{e}$ is a vector of random individual effects, with

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

where $\boldsymbol{\zeta} = \sigma_u^2\boldsymbol{I}_G$ and $\boldsymbol{R} = \sigma_e^2\boldsymbol{I}_N$. The between-group variance σ_u^2 and the within-group variance σ_e^2 are the variance components of the linear mixed model, and the variance-covariance matrix of $\boldsymbol{Y}_B$ is of the form

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

The values of Y and $\boldsymbol{X}$ in each block then satisfy

$$\begin{aligned}\mathrm{E}_X(\boldsymbol{y}_q) &= \boldsymbol{X}_q\boldsymbol{\beta} = \boldsymbol{f}_q\\ \mathrm{var}_X(\boldsymbol{y}_q) &= \sigma_u^2\boldsymbol{Z}_q\boldsymbol{Z}_q^\top + \sigma_e^2\boldsymbol{I}_q\\ \mathrm{cov}_X(\boldsymbol{y}_q, \boldsymbol{y}_r) &= \sigma_u^2\boldsymbol{Z}_q\boldsymbol{Z}_r^\top\end{aligned}$$

where the subscript X denotes conditioning on the value $\boldsymbol{X}_q$ and r is another block index. The naive linked data weighted least squares (WLS) estimator of $\boldsymbol{\beta}$ is then

$$\hat{\boldsymbol{\beta}}^* = \left(\sum_q\sum_r \boldsymbol{X}_q^\top\boldsymbol{W}_{qr}\boldsymbol{X}_r\right)^{-1}\left(\sum_q\sum_r \boldsymbol{X}_q^\top\boldsymbol{W}_{qr}\boldsymbol{y}_r^*\right) \tag{5}$$

where $\boldsymbol{W} = \boldsymbol{V}^{-1}$ and $\boldsymbol{W}_{qr}$ is the component of $\boldsymbol{W}_B$ corresponding to block $F = q$ and block $F = r$.

It is straightforward to see that under the linkage error model (1), the naive WLS estimator (5) based on the linked data set is biased since

$$\mathrm{E}_X(\hat{\boldsymbol{\beta}}^*) = \left(\sum_q \sum_r \boldsymbol{X}_q^\top \boldsymbol{W}_{qr} \boldsymbol{X}_r\right)^{-1} \left(\sum_q \sum_r \boldsymbol{X}_q^\top \boldsymbol{W}_{qr} \boldsymbol{T}_r \boldsymbol{X}_r\right) \boldsymbol{\beta} = \boldsymbol{D}\boldsymbol{\beta}. \tag{6}$$

Given $\boldsymbol{T}_r$ and $\boldsymbol{W}_{qr}$ are known and the inverse of $\boldsymbol{D}$ in (6) exists, Chambers (2009) suggests an unbiased estimator using a ratio-type correction for the bias in the naive estimator of $\boldsymbol{\beta}$,

$$\hat{\boldsymbol{\beta}}_R = \boldsymbol{D}^{-1}\hat{\boldsymbol{\beta}}^* = \left(\sum_q \sum_r \boldsymbol{X}_q^\top \boldsymbol{W}_{qr} \boldsymbol{T}_r \boldsymbol{X}_r\right)^{-1} \left(\sum_q \sum_r \boldsymbol{X}_q^\top \boldsymbol{W}_{qr} \boldsymbol{y}_r^*\right) \tag{7}$$

so long as $\sum_q \sum_r \boldsymbol{X}_q^\top \boldsymbol{W}_{qr} \boldsymbol{T}_r \boldsymbol{X}_r$ 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 that

$$\mathrm{E}_X(\boldsymbol{y}_q^*) = \mathrm{E}_X(\boldsymbol{A}_q)\mathrm{E}_X(\boldsymbol{y}_q) = \boldsymbol{T}_q \boldsymbol{X}_q \boldsymbol{\beta} = \boldsymbol{H}_q \boldsymbol{\beta}.$$

We see that the $\boldsymbol{y}_q^*$ can also be modelled linearly, with regression coefficient $\boldsymbol{\beta}$ but with a modified set of explanatory variables $\boldsymbol{H}_q$ in block q. Following Lahiri & Larsen (2005), an alternative estimator of $\boldsymbol{\beta}$ in this case is therefore

$$\begin{aligned} \hat{\boldsymbol{\beta}}_A &= \left(\sum_q \sum_r \boldsymbol{H}_q^\top \boldsymbol{W}_{qr} \boldsymbol{H}_r\right)^{-1} \left(\sum_q \sum_r \boldsymbol{H}_q^\top \boldsymbol{W}_{qr} \boldsymbol{y}_r^*\right) \\ &= \left(\sum_q \sum_r \boldsymbol{X}_q^\top \boldsymbol{T}_q^\top \boldsymbol{W}_{qr} \boldsymbol{T}_r \boldsymbol{X}_r\right)^{-1} \left(\sum_q \sum_r \boldsymbol{X}_q^\top \boldsymbol{T}_q^\top \boldsymbol{W}_{qr} \boldsymbol{y}_r^*\right). \end{aligned} \tag{8}$$

This estimator is not optimal since the variances of the regression errors defined by the

linked data vary between blocks. That is

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

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

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

where f_i denote components of $\boldsymbol{f}_q$ and $\bar{f}_q,\ \bar{f}_q^{\,(2)}$ denote the block q averages of the components of $\boldsymbol{f}_q$ and their squares respectively. A similar approximation of $\boldsymbol{K}_q$ can be developed. Defining $\boldsymbol{K}_q = [k_{ij}]$, this approximation is given by

$$k_{ij} = \begin{cases}
\lambda + \frac{(1-\lambda)}{M_q - 1}(G_q M_{qh} - 1), & \text{if } i = j \\
\left\{\lambda + (n_{qh} - 2)\frac{(1-\lambda)}{M_q - 1}\right\}^2 & \\
+(M_{qh} - 1)\left\{\frac{(1-\lambda)}{M_q - 1}\right\}^2 \left\{1 + (G_q - 1)M_{qh}\right\}, & \text{if } i \neq j \text{ and } i, j \text{ are in the same group} \\
(M_{qh} - 1)\frac{(1-\lambda)}{M_q - 1}\left\{2\lambda + \frac{(1-\lambda)}{M_q - 1}(G_q M_{qh} - 2)\right\}, & \text{if } i \neq j \text{ and } i, j \text{ are not in the same group}
\end{cases}$$

where G_q is number of groups in block q and M_{qh} is number of population units in block q and group h. Also, the covariance between $\boldsymbol{y}_q^*$ and $\boldsymbol{y}_r^*$ is then

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

Thus, the best linear unbiased estimator (BLUE) for β given these data can be approxi-

mated by

$$\begin{aligned}\hat{\boldsymbol{\beta}}_C &= \left(\sum_q \sum_r \boldsymbol{H}_q^\top \boldsymbol{\Sigma}_{qr}^{-1} \boldsymbol{H}_r\right)^{-1} \left(\sum_q \sum_r \boldsymbol{H}_q^\top \boldsymbol{\Sigma}_{qr}^{-1} \boldsymbol{y}_r^*\right) \\ &= \left(\sum_q \sum_r \boldsymbol{X}_q^\top \boldsymbol{T}_q^\top \boldsymbol{\Sigma}_{qr}^{-1} \boldsymbol{T}_r \boldsymbol{X}_r\right)^{-1} \left(\sum_q \sum_r \boldsymbol{X}_q^\top \boldsymbol{T}_q^\top \boldsymbol{\Sigma}_{qr}^{-1} \boldsymbol{y}_r^*\right) \qquad (9)\end{aligned}$$

where $\boldsymbol{\Sigma} = \mathrm{var}_X(\boldsymbol{y}_B^*)$ and $\boldsymbol{\Sigma}_{qr}^{-1}$ is the component of $\boldsymbol{\Sigma}^{-1}$ corresponding to block $F = q$ and block $F = r$.

Variance estimators for $\hat{\boldsymbol{\beta}}_R$, $\hat{\boldsymbol{\beta}}_A$ and $\hat{\boldsymbol{\beta}}_C$ can be defined using first order approximations to solutions of estimating equations. These estimators are derived in Appendix III.

4. Estimation of variance components

We now develop appropriate modifications to three standard methods of variance components estimation in order to account for linkage error. These are the method of moments, typically referred to as the analysis of variance (ANOVA) method, the maximum likelihood (ML) method, and the restricted maximum likelihood (REML) method. The details of the modified version of each method are set out in Sections 4.1, 4.2 and 4.3, respectively. Note that all population quantities referred to in this Section are ordered as in the B register, so we drop the B subscript.

4.1. Analysis of variance (ANOVA)

Historically, ANOVA is the starting point for estimation of variance components (Searle et al., 2006). The method is based on equating the between groups sum of squares (SSA) and the within groups sum of squares (SSE) with their expected values under the nested error model of interest. The two sums of squares that are the basis of

ANOVA for the linked data are

$$\text{SSA} = \boldsymbol{y}^{*\top} \boldsymbol{B}\, \boldsymbol{y}^* \quad ; \quad \boldsymbol{B} = [b_{ij}] \quad \text{where} \quad b_{ij} = \begin{cases} \frac{1}{N_g} - \frac{1}{N}, & \text{if } i,j \text{ are in the same group} \\ -\frac{1}{N}, & \text{if } i,j \text{ are not in the same group} \end{cases}$$

$$\text{SSE} = \boldsymbol{y}^{*\top} \boldsymbol{C}\, \boldsymbol{y}^* \quad ; \quad \boldsymbol{C} = [c_{ij}] \quad \text{where} \quad c_{ij} = \begin{cases} 1 - \frac{1}{N_g}, & \text{if } i = j \\ -\frac{1}{N_g}, & \text{if } i \neq j \text{ and i, j are in the same group} \\ 0, & \text{otherwise.} \end{cases}$$

The expected values of these two sum of squares are derived in Appendix I. When $\mathrm{E}_X(\text{SSA})$ and $\mathrm{E}_X(\text{SSE})$ are equated to their observed values we obtain the variance components estimators

$$\hat{\sigma}_e^2 = \frac{mc - na}{bc - da} \tag{10}$$

and

$$\hat{\sigma}_u^2 = \frac{m - \hat{\sigma}_e^2 b}{a} \tag{11}$$

where

$$\begin{aligned}
a &= \sum_q \mathrm{tr}\left(\boldsymbol{B}_{qq}\boldsymbol{K}_q\right) + \sum_q \sum_{r\neq q} \mathrm{tr}\left(\boldsymbol{T}_r \boldsymbol{Z}_r \boldsymbol{Z}_q^\top \boldsymbol{T}_q^\top \boldsymbol{B}_{qr}\right) \\
b &= \sum_q \mathrm{tr}\left(\boldsymbol{B}_{qq}\right) \\
c &= \sum_q \mathrm{tr}\left(\boldsymbol{C}_{qq}\boldsymbol{K}_q\right) + \sum_q \sum_{r\neq q} \mathrm{tr}\left(\boldsymbol{T}_r \boldsymbol{Z}_r \boldsymbol{Z}_q^\top \boldsymbol{T}_q^\top \boldsymbol{C}_{qr}\right) \\
d &= \sum_q \mathrm{tr}\left(\boldsymbol{C}_{qq}\right) \\
m &= \text{SSA} - \sum_q \left\{\mathrm{tr}\left(\boldsymbol{B}_{qq}\boldsymbol{V}_q\right) + \boldsymbol{f}_q^\top \boldsymbol{T}_q^\top \boldsymbol{B}_{qq} \boldsymbol{T}_q \boldsymbol{f}_q\right\} - \sum_q \sum_{r\neq q} \boldsymbol{f}_q^\top \boldsymbol{T}_q^\top \boldsymbol{B}_{qr} \boldsymbol{T}_r \boldsymbol{f}_r \\
n &= \text{SSE} - \sum_q \left\{\mathrm{tr}\left(\boldsymbol{C}_{qq}\boldsymbol{V}_q\right) + \boldsymbol{f}_q^\top \boldsymbol{T}_q^\top \boldsymbol{C}_{qq} \boldsymbol{T}_q \boldsymbol{f}_q\right\} - \sum_q \sum_{r\neq q} \boldsymbol{f}_q^\top \boldsymbol{T}_q^\top \boldsymbol{C}_{qr} \boldsymbol{T}_r \boldsymbol{f}_r.
\end{aligned}$$

Estimators of the large sample variances of $\hat{\sigma}_e^2$ and $\hat{\sigma}_u^2$ defined by (10) and (11) are derived in Appendix III.

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

$$\begin{aligned} a &= N - \frac{\sum_g N_g^2}{N} \\ b &= G - 1 \text{ is degrees of freedom of SSA} \\ d &= N - G \text{ is degrees of freedom of SSE} \\ m &= \text{SSA} - \sum_q \sum_r \boldsymbol{f}_q^\top \boldsymbol{B}_{qr} \boldsymbol{f}_r \\ n &= \text{SSE} - \sum_q \sum_r \boldsymbol{f}_q^\top \boldsymbol{C}_{qr} \boldsymbol{f}_r. \end{aligned}$$

The ANOVA estimates in (10) and (11) can be negative. Consequently, it is usually better 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).

4.2. Pseudo maximum likelihood (Pseudo-ML)

Unlike the ANOVA method of estimation, a basic requirement of ML estimation is that the probability distribution of the data is known. We follow the usual convention of assuming multivariate normality. That is, we assume that $\boldsymbol{y}^* \sim N(\boldsymbol{T}\boldsymbol{f}, \boldsymbol{\Sigma})$. The log-likelihood function is then

$$l = -\frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}(\boldsymbol{y}^* - \boldsymbol{T}\boldsymbol{f})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{y}^* - \boldsymbol{T}\boldsymbol{f}). \tag{12}$$

In what follows, we assume that $\boldsymbol{\Sigma}$ is fixed. Differentiating (12) with respect to $\boldsymbol{\beta}$ then yields

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \boldsymbol{X}^\top \boldsymbol{T}^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{y}^* - \boldsymbol{T}\boldsymbol{f}). \tag{13}$$

Similarly, differentiating (12) with respect to σ_u^2 leads to

$$\frac{\partial l}{\partial \sigma_u^2} = -\frac{1}{2}\text{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_u\right) + \frac{1}{2}(\boldsymbol{y}^* - \boldsymbol{T}\boldsymbol{f})^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_u\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}^* - \boldsymbol{T}\boldsymbol{f}) \quad (14)$$

where $\boldsymbol{\Sigma}_u = \partial\boldsymbol{\Sigma}/\partial\sigma_u^2$. Finally, differentiating (12) with respect to σ_e^2 gives

$$\frac{\partial l}{\partial \sigma_e^2} = -\frac{1}{2}\text{tr}\left(\boldsymbol{\Sigma}^{-1}\right) + \frac{1}{2}(\boldsymbol{y}^* - \boldsymbol{T}\boldsymbol{f})^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}^* - \boldsymbol{T}\boldsymbol{f}). \quad (15)$$

The pseudo-ML estimators for $\boldsymbol{\beta}, \sigma_u^2$ and σ_e^2 are defined by setting the derivatives (13), (14) and (15) to zero and solving for these parameters. Note that we refer to the resulting estimators as pseudo-ML because their estimating functions, which are defined by these derivatives, are based on the assumption that $\boldsymbol{\Sigma}$ is a known matrix. However, in reality this matrix is a function of $\boldsymbol{\beta}, \sigma_u^2$ and σ_e^2, and so analytic solutions to these estimating equations do not exist. We therefore now describe how the method of scoring (Searle et al., 2006) can be used to solve them.

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

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

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

We now develop the expression for $\bar{\boldsymbol{I}}(\boldsymbol{\theta})$. From (13),

$$\frac{\partial^2 l}{\partial \boldsymbol{\beta}\, \partial \boldsymbol{\beta}^\top} = -\boldsymbol{X}^\top \boldsymbol{T}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{T}\boldsymbol{X}$$

$$\frac{\partial^2 l}{\partial \boldsymbol{\beta}\, \partial \sigma_u^2} = -\boldsymbol{X}^\top \boldsymbol{T}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_u\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}^* - \boldsymbol{T}\boldsymbol{f})$$

$$\frac{\partial^2 l}{\partial \boldsymbol{\beta}\, \partial \sigma_e^2} = -\boldsymbol{X}^\top \boldsymbol{T}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}^* - \boldsymbol{T}\boldsymbol{f}).$$

Furthermore, from (14) and (15) we have

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

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

and

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

The expected information matrix is obtained by taking the expected values of the derivatives set out above, noting that $\mathrm{E}(\boldsymbol{y}^*) = \boldsymbol{T}\boldsymbol{f}$ and hence $\mathrm{E}(\boldsymbol{y}^* - \boldsymbol{T}\boldsymbol{f}) = \boldsymbol{0}$. Also, $\mathrm{E}(\boldsymbol{y}^* - \boldsymbol{T}\boldsymbol{f})^\top\boldsymbol{G}(\boldsymbol{y}^* - \boldsymbol{T}\boldsymbol{f}) = \mathrm{tr}(\boldsymbol{G}\boldsymbol{\Sigma})$ for non-stochastic $\boldsymbol{G}$. It follows that

$$\bar{\boldsymbol{I}} = \begin{bmatrix} \bar{\boldsymbol{I}}(\boldsymbol{\beta}) & \boldsymbol{0} \\ \boldsymbol{0} & \bar{\boldsymbol{I}}(\sigma_u^2, \sigma_e^2) \end{bmatrix}$$

where $\bar{\boldsymbol{I}}(\boldsymbol{\beta}) = \boldsymbol{X}^\top\boldsymbol{T}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{T}\boldsymbol{X}$ and

$$\bar{\boldsymbol{I}}(\sigma_u^2, \sigma_e^2) = \frac{1}{2}\begin{bmatrix} \mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_u\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_u\right) & \mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_u\right) \\ \mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_u\right) & \mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}^{-1}\right) \end{bmatrix}. \tag{16}$$

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 generally biased, unlike the ANOVA estimators (Harville 1977, Searle et al. 2006).

The first criticism above is overcome by using restricted maximum likelihood (REML) (Searle et al., 2006). Rather than using $\boldsymbol{y}^*$ directly, REML uses ML estimating equations

based on a modified response variable defined by a linear combination $\boldsymbol{s}^\top \boldsymbol{y}^*$ of elements of $\boldsymbol{y}^*$, chosen in such a way that the distribution of this combination does not depend on the fixed effects in the model. In particular, the vector $\boldsymbol{s}$ is chosen so that $\mathrm{E}(\boldsymbol{s}^\top \boldsymbol{y}^*) = \boldsymbol{s}^\top \boldsymbol{T}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{0}$, i.e.

$$\boldsymbol{s}^\top \boldsymbol{T}\boldsymbol{X} = \boldsymbol{0}. \tag{17}$$

Note that strict application of the REML approach requires that the distribution of $\boldsymbol{s}^\top \boldsymbol{y}^*$ does not depend on $\boldsymbol{\beta}$. However, when we use linked data, the variance of $\boldsymbol{y}^*$ is implicitly a function of this parameter. Consequently, we refer to this method as "pseudo-REML" since it is based on application of standard REML arguments, ignoring the fact that the variance still depends on the fixed effects in the model.

When $\boldsymbol{T}\boldsymbol{X}$ of order $N \times p$ has rank r, there are $N - r$ linearly independent vectors $\boldsymbol{s}^\top$ satisfying (17) (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 can form $\boldsymbol{S}^\top \boldsymbol{y}^*$ where $\boldsymbol{S}^\top$ is a $(N - r) \times N$ matrix whose rows are $N - r$ linearly independent rows of the matrix $\boldsymbol{I} - \boldsymbol{T}\boldsymbol{X}\{(\boldsymbol{T}\boldsymbol{X})^\top(\boldsymbol{T}\boldsymbol{X})\}^{-1}(\boldsymbol{T}\boldsymbol{X})^\top$.

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

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

Let l_R be the log likelihood for the variance effects defined by $\boldsymbol{S}^\top \boldsymbol{y}^*$. That is

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

The REML estimating function for $\boldsymbol{\beta}$ is unchanged from the corresponding ML estimating function (13). However the ML estimating functions (14) and (15) for the variance components σ_u^2 and σ_e^2 are now replaced by alternative REML estimating functions obtained

by differentiating l_R. In this context, we note that

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

The REML estimating equations are defined by setting (13) and the REML estimating functions for σ_u^2 and σ_e^2 to zero. As before, we use the method of scoring to solve these equations. In order to define the expected information matrix in this case, we need the first and second derivatives of l_R :

$$\begin{aligned}\frac{\partial l_R}{\partial \sigma_u^2} &= -\frac{1}{2}\mathrm{tr}\left(\boldsymbol{M}\boldsymbol{\Sigma}_u\right) + \frac{1}{2}\boldsymbol{y}^{*\top}\boldsymbol{M}\boldsymbol{\Sigma}_u\boldsymbol{M}\boldsymbol{y}^* \\ \frac{\partial^2 l_R}{\partial \sigma_u^2\,\partial \sigma_u^2} &= \frac{1}{2}\mathrm{tr}\left(\boldsymbol{M}\boldsymbol{\Sigma}_u\boldsymbol{M}\boldsymbol{\Sigma}_u\right) - \boldsymbol{y}^{*\top}\boldsymbol{M}\boldsymbol{\Sigma}_u\boldsymbol{M}\boldsymbol{\Sigma}_u\boldsymbol{M}\boldsymbol{y}^*\end{aligned}$$

$$\begin{aligned}\frac{\partial l_R}{\partial \sigma_e^2} &= -\frac{1}{2}\mathrm{tr}\left(\boldsymbol{M}\right) + \frac{1}{2}\boldsymbol{y}^{*\top}\boldsymbol{M}\boldsymbol{M}\boldsymbol{y}^* \\ \frac{\partial^2 l_R}{\partial \sigma_e^2\,\partial \sigma_e^2} &= \frac{1}{2}\mathrm{tr}\left(\boldsymbol{M}\boldsymbol{M}\right) - \boldsymbol{y}^{*\top}\boldsymbol{M}\boldsymbol{M}\boldsymbol{M}\boldsymbol{y}^*.\end{aligned}$$

The expected values of these second derivatives of l_R are developed in Appendix II. It follows that the component $\bar{\boldsymbol{I}}(\sigma_u^2, \sigma_e^2)$ of the observed information matrix is then given by

$$\bar{\boldsymbol{I}}(\sigma_u^2, \sigma_e^2) = \frac{1}{2}\begin{bmatrix} \mathrm{tr}\left(\boldsymbol{M}\boldsymbol{\Sigma}_u\boldsymbol{M}\boldsymbol{\Sigma}_u\right) & \mathrm{tr}\left(\boldsymbol{M}\boldsymbol{M}\boldsymbol{\Sigma}_u\right) \\ \mathrm{tr}\left(\boldsymbol{M}\boldsymbol{M}\boldsymbol{\Sigma}_u\right) & \mathrm{tr}\left(\boldsymbol{M}\boldsymbol{M}\right) \end{bmatrix}. \tag{18}$$

Estimators of the large sample variances of either the pseudo-ML or pseudo-REML versions of $\hat{\sigma}_e^2$ and $\hat{\sigma}_u^2$ can be defined using standard large sample approximations based on the expected information i.e. $\mathrm{var}(\hat{\sigma}_u^2, \hat{\sigma}_e^2) \simeq \left[\bar{\boldsymbol{I}}(\hat{\sigma}_u^2, \hat{\sigma}_e^2)\right]^{-1}$ where $\mathrm{var}(\hat{\sigma}_u^2, \hat{\sigma}_e^2)$ is the 2×2

matrix of variances and covariances of the variance components estimates and $\bar{\boldsymbol{I}}(\sigma_u^2, \sigma_e^2)$ is given in (16) for ML and (18) for REML.

5. Simulation results

This section contains results from a small scale simulation study that illustrates the comparative performances of the parameter estimators described in previous sections, given an exchangeable linkage error (ELE) model. The simulations themselves are based on a simple balanced two level population structure. In particular, in each simulation we generated a population of size $N = 800$ made up of 50 equal-sized groups, so that each group consisted of 16 units. Population units were then randomly allocated to four equal-sized blocks, each of size 200, such that each group contained an equal number of units (4) from each block thus ensuring that the distributions of Y and $\boldsymbol{X}$ were the same in each block. We note that more complex distributions of block sizes and different covariate distributions within the different groups will usually prevail in realistic settings. However, the purpose here is to demonstrate the comparative bias-correcting properties of the different estimators rather than to evaluate their stability and efficiency under realistic population structures and linkage scenarios.

Values of X were 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 the e_{ig} were independently drawn from the $N(0, 3)$ distribution and the u_g were independently drawn from the $N(0, 1)$ distribution. The true data pairs (y_{ig}, x_{ig}) were then randomly allocated to blocks and groups. Next, linked data pairs (y_{ig}^*, x_{ig}) were generated by using the ELE model defined by (1) - (4) with correct linkage probabilities $\lambda_1 = 1,\ \lambda_2 = 0.95,\ \lambda_3 = 0.85$ and $\lambda_4 = 0.75$. That is, all links for block 1 were assumed to be correct, while those for blocks 2, 3 and 4 were assumed to have some errors. This ELE model allows a record in a block with λ less than one to be potentially matched to

any record located in the same block irrespective of the group status of the record. That is, the group identifier is not a component of the blocking variable and hence not part of the linkage process. This ensures that the non-informative linking assumption holds in the simulations.

We present simulation results for two scenarios. The first corresponds to known linkage probabilities. In the second, these probabilities were estimated by taking random audit samples of $m_q = 25$ linked pairs from each of blocks $2-4$ and checking to see how many of these sampled links were correct. Following Chambers (2009), the estimate of λ_q was then 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. Variance estimators were adjusted for the extra variability induced by this estimation of λ_q using the approach described in Chambers (2009). The details of this approach are described in Appendix III.

A total of 800 independent simulations were sufficient to illustrate the different bias and variance properties of each estimator. Table 1 and Figure 1 show the relative biases and relative root mean squared errors of the regression coefficient estimators described in Section 3, and Table 2 and Figure 2 show the relative biases and relative root mean squared errors of the variance components estimators described in Section 4. The WLS estimator TR based on perfectly linked data and the naive WLS estimator based on the actual linked data were obtained using the default settings of the `lme` function in the R software package. The estimators R, A and C denote the bias-corrected estimators (7), (8) and (9) respectively. Note that variance components estimators obtained using the ANOVA method are functions of $\boldsymbol{\beta}$, and so were evaluated using the bias-corrected options (R, A and C) for this parameter. These different ANOVA estimators are denoted by R, A and C suffixes in Tables 2 and 3. The actual coverages of the nominal 95% confidence intervals for all the model parameters are shown in Table 3.

The results set out in Table 1 show that the naive WLS estimator that assumes the

data are perfectly linked is clearly biased. Since linkage error is a particular type of measurement error, this bias attenuates the estimate of the slope parameter and exaggerates that of the intercept. On the other hand, all five of the adjusted estimators correct this bias, with the REML estimator being the most efficient. The results are unchanged under Scenario 2 where linkage probabilities were estimated by taking small audit samples.

[TABLE 1 ABOUT HERE]

[FIGURE 1 ABOUT HERE]

The results displayed in Table 2 show that the naive variance components estimators that treat the linkage as perfect are also biased. As expected, the estimator obtained using the ML approach is slightly biased. All of the remaining adjusted estimators are essentially unbiased, with REML being the most efficient. Again, the results under Scenario 2 are in the same direction as those under Scenario 1.

[TABLE 2 ABOUT HERE]

[FIGURE 2 ABOUT HERE]

Finally, we note that the results displayed in Table 3 show that variance estimators that allow for the extra variability induced by estimation of the correct linkage probabilities lead to confidence intervals with good coverage properties.

[TABLE 3 ABOUT HERE]

6. Summary and Further Research

In this paper we show how one can extend the inferential framework of Chambers (2009) to obtain unbiased estimators of the regression parameters when fitting a two level linear model to probabilistically linked data assuming an exchangeable linkage errors model. We also show how three standard methods of estimation for the variance components of the linear mixed model (ANOVA, maximum likelihood and REML) can

be modified in order to make them approximately unbiased under this model. Our simulation results indicate that all the methods developed in this paper work reasonably well in terms of correcting biases induced by linkage errors. However, they also show evidence of increased variability due to the use of bias correction.

An important area of application of two level models using linked data is where registers are linked over time to create data sets suitable for fitting longitudinal models with random individual effects. Further research extending the methodology described in this paper to this situation is ongoing. An important aspect of this research is that it addresses the issue of linkage errors in the model grouping structure - something not considered in this paper. Another issue concerns the assumption of an exchangeable linkage errors model. Although a convenient first approximation, most realistic linkage applications involve multiple linkage operations and so will possess a more complex error structure. Further research into correcting linkage error bias under alternative linkage error models is therefore necessary.

A limitation of the research reported in this paper is that the simulation results were based on a small sample size and a relatively simple two-level population structure. This was because the simulation study was designed to illustrate the performances of the different bias-correction methods, rather than to provide an extensive comparison for a variety of population structures and linkage error situations. A larger study is desirable, especially to test the robustness of our methods to failure of key assumptions (e.g. non-informative linkage), but will require considerable computational resources. These are issues that will be considered in our further research in this area.

Appendices

I. ANOVA estimation

We first develop an expression for $\mathrm{E}_X(\mathrm{SSA})$. The corresponding expression for $\mathrm{E}_X(\mathrm{SSE})$ follows similarly.

$$\mathrm{E}_X(\mathrm{SSA}) = \sum_q \mathrm{E}_X\left(\boldsymbol{y}_q^{*\top}\boldsymbol{B}_{qq}\boldsymbol{y}_q^*\right) + \sum_q\sum_{r\neq q}\mathrm{E}_X\left(\boldsymbol{y}_q^{*\top}\boldsymbol{B}_{qr}\boldsymbol{y}_r^*\right).$$

Consider the first term on the right hand side above. This can be written

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

The second term on the right hand side of the expression for $\mathrm{E}_X(\mathrm{SSA})$ can be expanded similarly:

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

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

$$\mathrm{E}_X(\mathrm{SSA}) = \sigma_u^2 a + \sigma_e^2 b + m_0$$

where

$$a = \sum_q \mathrm{tr}\left(\boldsymbol{B}_{qq}\boldsymbol{K}_q\right) + \sum_q \sum_{r\neq q} \mathrm{tr}\left(\boldsymbol{T}_r\boldsymbol{Z}_r\boldsymbol{Z}_q^\top\boldsymbol{T}_q^\top\boldsymbol{B}_{qr}\right))$$

$$b = \sum_q \mathrm{tr}\left(\boldsymbol{B}_{qq}\right)$$

and

$$m_0 = \sum_q \left\{\mathrm{tr}\left(\boldsymbol{B}_{qq}\boldsymbol{V}_q\right) + \boldsymbol{f}_q^\top\boldsymbol{T}_q^\top\boldsymbol{B}_{qq}\boldsymbol{T}_q\boldsymbol{f}_q\right\} + \sum_q \sum_{r\neq q} \boldsymbol{f}_q^\top\boldsymbol{T}_q^\top\boldsymbol{B}_{qr}\boldsymbol{T}_r\boldsymbol{f}_r.$$

Similarly,

$$\mathrm{E}_X(\mathrm{SSE}) = \sigma_u^2 c + \sigma_e^2 d + n_0$$

where

$$c = \sum_q \mathrm{tr}\left(\boldsymbol{C}_{qq}\boldsymbol{K}_q\right) + \sum_q \sum_{r\neq q} \mathrm{tr}\left(\boldsymbol{T}_r\boldsymbol{Z}_r\boldsymbol{Z}_q^\top\boldsymbol{T}_q^\top\boldsymbol{C}_{qr}\right)$$

$$d = \sum_q \mathrm{tr}\left(\boldsymbol{C}_{qq}\right)$$

and

$$n_0 = \sum_q \left\{\mathrm{tr}\left(\boldsymbol{C}_{qq}\boldsymbol{V}_q\right) + \boldsymbol{f}_q^\top\boldsymbol{T}_q^\top\boldsymbol{C}_{qq}\boldsymbol{T}_q\boldsymbol{f}_q\right\} + \sum_q \sum_{r\neq q} \boldsymbol{f}_q^\top\boldsymbol{T}_q^\top\boldsymbol{C}_{qr}\boldsymbol{T}_r\boldsymbol{f}_r.$$

Replacing $\mathrm{E}_X(\mathrm{SSA})$ and $\mathrm{E}_X(\mathrm{SSE})$ by SSA and SSE respectively in these two equations and solving for σ_u^2 and σ_e^2 then leads to the estimators

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

and

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

where $m = \mathrm{SSA} - m_0$ and $n = \mathrm{SSE} - n_0$.

II. Expectations used in pseudo-REML information matrix

$$\begin{aligned}
-\mathrm{E}\left(\frac{\partial^2 l_R}{\partial\sigma_u^2\,\partial\sigma_u^2}\right) &= -\frac{1}{2}\mathrm{tr}\left(\boldsymbol{M}\boldsymbol{\Sigma}_u\boldsymbol{M}\boldsymbol{\Sigma}_u\right) + \mathrm{tr}\left(\boldsymbol{M}\boldsymbol{\Sigma}_u\boldsymbol{M}\boldsymbol{\Sigma}_u\boldsymbol{M}\boldsymbol{\Sigma}\right) \\
&\quad + \boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{T}^\top\boldsymbol{M}\boldsymbol{\Sigma}_u\boldsymbol{M}\boldsymbol{\Sigma}_u\boldsymbol{M}\boldsymbol{T}\boldsymbol{X}\boldsymbol{\beta} \\
&= -\frac{1}{2}\mathrm{tr}\left(\boldsymbol{M}\boldsymbol{\Sigma}_u\boldsymbol{M}\boldsymbol{\Sigma}_u\right) + \mathrm{tr}\left(\boldsymbol{\Sigma}_u\boldsymbol{M}\boldsymbol{\Sigma}_u\boldsymbol{M}\boldsymbol{\Sigma}\boldsymbol{M}\right) \\
&= \frac{1}{2}\mathrm{tr}\left(\boldsymbol{M}\boldsymbol{\Sigma}_u\boldsymbol{M}\boldsymbol{\Sigma}_u\right)
\end{aligned}$$

where the second equality follows because $\boldsymbol{MTX} = \mathbf{0}$ and the third equality follows because $\boldsymbol{M\Sigma M} = \boldsymbol{M}$.

$$\begin{aligned}
-\mathrm{E}\left(\frac{\partial^2 l_R}{\partial\sigma_u^2\,\partial\sigma_e^2}\right) &= -\frac{1}{2}\mathrm{tr}\left(\boldsymbol{MM\Sigma}_u\right) + \mathrm{tr}\left(\boldsymbol{MM\Sigma}_u\boldsymbol{M\Sigma}\right) + \boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{T}^\top\boldsymbol{MM\Sigma}_u\boldsymbol{MTX}\boldsymbol{\beta} \\
&= \frac{1}{2}\mathrm{tr}\left(\boldsymbol{MM\Sigma}_u\right) \\
-\mathrm{E}\left(\frac{\partial^2 l_R}{\partial\sigma_e^2\,\partial\sigma_e^2}\right) &= -\frac{1}{2}\mathrm{tr}\left(\boldsymbol{MM}\right) + \mathrm{tr}\left(\boldsymbol{MMM\Sigma}\right) + \boldsymbol{\beta}^\top\boldsymbol{X}^\top\boldsymbol{T}^\top\boldsymbol{MMMTX}\boldsymbol{\beta} \\
&= \frac{1}{2}\mathrm{tr}\left(\boldsymbol{MM}\right).
\end{aligned}$$

III. Variance estimation

Given the values of the variance components, the estimators $\hat{\boldsymbol{\beta}}_R$, $\hat{\boldsymbol{\beta}}_A$ and $\hat{\boldsymbol{\beta}}_C$ can all be represented as the solutions to estimating equations. Consequently, we follow the approach described in Chambers (2009) in order to define large sample estimators of the variances of these regression parameter estimators. In particular, we note that any of these estimators is defined by solving an equation of the form $\boldsymbol{H}(\boldsymbol{\beta}) = \mathbf{0}$ where $\boldsymbol{H}(\boldsymbol{\beta})$ is a $p-$dimensional unbiased estimating function for the regression parameter $\boldsymbol{\beta}$. Let λ denote the vector defined by the block-specific values of λ_r. The general form of the unbiased estimating function used by all three regression parameter estimators above is then

$$\boldsymbol{H}^*(\boldsymbol{\beta},\lambda) = \sum_q\sum_r \boldsymbol{D}_{qr}(\boldsymbol{\beta})\{\boldsymbol{y}_r^* - \boldsymbol{T}_r(\lambda)\boldsymbol{f}_r(\boldsymbol{\beta})\} = \sum_q\sum_r \boldsymbol{U}_{qr}(\boldsymbol{\beta},\lambda_r)$$

which is a function of both $\boldsymbol{\beta}$ and λ. Using a first order Taylor series approximation,

$$\begin{aligned}
\mathbf{0} = \boldsymbol{H}^*(\hat{\boldsymbol{\beta}},\hat{\lambda}) &\approx \boldsymbol{H}^*(\boldsymbol{\beta}_0,\lambda_0) + \partial_{\boldsymbol{\beta}}\boldsymbol{H}^*(\boldsymbol{\beta}_0,\lambda_0)(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0) + \partial_\lambda\boldsymbol{H}^*(\boldsymbol{\beta}_0,\lambda_0)(\hat{\lambda}-\lambda_0) \\
\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 &= -(\partial_{\boldsymbol{\beta}}\boldsymbol{H}_0^*)^{-1}\left\{\boldsymbol{H}_0^* + \partial_\lambda\boldsymbol{H}_0^*(\hat{\lambda}-\lambda_0)\right\} \\
\hat{\boldsymbol{\beta}} &= \boldsymbol{\beta}_0 - (\partial_{\boldsymbol{\beta}}\boldsymbol{H}_0^*)^{-1}\left\{\boldsymbol{H}_0^* + \partial_\lambda\boldsymbol{H}_0^*(\hat{\lambda}-\lambda_0)\right\}
\end{aligned}$$

where $\boldsymbol{H}_0^* = \boldsymbol{H}^*(\boldsymbol{\beta}_0, \lambda_0)$. Note that $\boldsymbol{\beta}_0$ and λ_0 denote the true values of these parameters with $\hat{\boldsymbol{\beta}}$ and $\hat{\lambda}$ denote their corresponding estimators. We can then approximate the variance of $\hat{\boldsymbol{\beta}}$ by

$$\begin{aligned} \mathrm{var}_X(\hat{\boldsymbol{\beta}}) &\approx (\partial_{\boldsymbol{\beta}} \boldsymbol{H}_0^*)^{-1} \mathrm{var}_X \left\{ \boldsymbol{H}_0^* + \partial_\lambda \boldsymbol{H}_0^* (\hat{\lambda} - \lambda_0) \right\} \left\{ (\partial_{\boldsymbol{\beta}} \boldsymbol{H}_0^*)^{-1} \right\}^\top \\ &= (\partial_{\boldsymbol{\beta}} \boldsymbol{H}_0^*)^{-1} \left\{ \mathrm{var}_X \left(\boldsymbol{H}_0^* \right) + \left(\partial_\lambda \boldsymbol{H}_0^* \right) \mathrm{var}_X(\hat{\lambda}) \left(\partial_\lambda \boldsymbol{H}_0^* \right)^\top \right\} \left\{ (\partial_\theta \boldsymbol{H}_0^*)^{-1} \right\}^\top . \end{aligned}$$

An ultimate cluster variance estimator can be used for $\mathrm{var}_X\left(\boldsymbol{H}_0^*\right)$. This is based on the representation

$$\begin{aligned} \mathrm{var}_X\left(\boldsymbol{H}_0^*\right) &= \mathrm{var}_X \left[\sum_q \sum_r \boldsymbol{D}_{qr}(\boldsymbol{\beta}_0) \{ \boldsymbol{y}_r^* - \boldsymbol{T}_r(\lambda_0) \boldsymbol{f}_r(\boldsymbol{\beta}_0) \} \right] \\ &= \mathrm{var}_X \left[\sum_g \left\{ \sum_q \sum_r \boldsymbol{D}_{qr}^g(\boldsymbol{\beta}_0) \boldsymbol{y}_r^{g*} \right\} \right] = \mathrm{var}_X(\boldsymbol{H}_{0C}^*) \end{aligned}$$

where

$$\boldsymbol{H}_{0C}^* = \sum_g \left\{ \sum_q \sum_r \boldsymbol{D}_{qr}^g(\boldsymbol{\beta}_0) \boldsymbol{y}_r^{g*} \right\} = \frac{1}{G} \sum_g \boldsymbol{H}_{0Cg}^*$$

and $\boldsymbol{H}_{0Cg}^* = G \sum_q \sum_r \boldsymbol{D}_{qr}^g(\boldsymbol{\beta}_0) \boldsymbol{y}_r^{g*}$. Note that the $\boldsymbol{H}_{0Cg}^*$ are mutually uncorrelated. A variance estimator for $\boldsymbol{H}_{0C}^*$ can therefore be written in the form

$$\begin{aligned} \hat{\mathrm{V}}_X\left(\boldsymbol{H}_{0C}^*\right) &= \frac{1}{G^2} \sum_g \hat{\mathrm{V}}_X\left(\boldsymbol{H}_{0Cg}^*\right) \\ &= \frac{1}{G(G-1)} \sum_g (\boldsymbol{H}_{0Cg}^* - \boldsymbol{H}_{0C}^*)(\boldsymbol{H}_{0Cg}^* - \boldsymbol{H}_{0C}^*)^\top . \end{aligned}$$

Put $\boldsymbol{U}_{0qr} = \boldsymbol{U}_{qr}(\boldsymbol{\beta}_0, \lambda_{0r})$. Then

$$\partial_{\lambda_r} \boldsymbol{U}_{0qr} = \partial_{\lambda_r} \boldsymbol{D}_{0qr} \{ \boldsymbol{y}_r^* - \boldsymbol{T}_r(\lambda_r) \boldsymbol{f}_{0r} \} = -\boldsymbol{D}_{0qr} \partial_{\lambda_r} \{ \boldsymbol{T}_r(\lambda_r) \} \boldsymbol{f}_{0r}$$

where $\boldsymbol{f}_{0r} = \boldsymbol{f}_r(\boldsymbol{\beta}_0)$.

Under the exchangeable linkage errors model (4)

$$\begin{aligned}\boldsymbol{T}_r(\lambda_r) &= \left\{\lambda_r - \frac{(1-\lambda_r)}{M_r-1}\right\}\boldsymbol{I}_r + \frac{(1-\lambda_r)}{M_r-1}\mathbf{1}_r\mathbf{1}_r^\top \\ &= (M_r-1)^{-1}\{(\lambda_r M_r - 1)\boldsymbol{I}_r + (1-\lambda_r)\mathbf{1}_r\mathbf{1}_r^\top\}\end{aligned}$$

so

$$\partial_{\lambda_r}\{\boldsymbol{T}_r(\lambda_r)\} = (M_r-1)^{-1}(M_r\boldsymbol{I}_r - \mathbf{1}_r\mathbf{1}_r^\top)$$

and hence

$$\partial_{\lambda_r}\boldsymbol{U}_{0qr} = -(M_r-1)^{-1}\boldsymbol{D}_{0qr}(M_r\boldsymbol{I}_r - \mathbf{1}_r\mathbf{1}_r^\top)\boldsymbol{f}_{0r}.$$

That is, we have the approximation

$$\begin{aligned}\mathrm{var}_X(\hat{\boldsymbol{\beta}}) \approx & \left(\sum_q\sum_r \partial_\theta \boldsymbol{U}_{0qr}\right)^{-1}\left\{\hat{\mathrm{V}}_X\left(\boldsymbol{H}^*_{0C}\right) + \sum_q\sum_r \boldsymbol{D}_{0qr}\Delta_{0r}\boldsymbol{D}_{0qr}^\top\right\} \\ & \left\{\left(\sum_q\sum_r \partial_\theta \boldsymbol{U}_{0qr}\right)^{-1}\right\}^\top\end{aligned}$$

where

$$\begin{aligned}\Delta_{0r} &= (M_r-1)^{-2}\mathrm{var}_X(\hat{\lambda}_r)(M_r\boldsymbol{I}_r - \mathbf{1}_r\mathbf{1}_r^\top)\boldsymbol{f}_{0r}\boldsymbol{f}_{0r}^\top(M_r\boldsymbol{I}_r - \mathbf{1}_r\mathbf{1}_r^\top) \\ &= M_r^2(M_r-1)^{-2}\mathrm{var}_X(\hat{\lambda}_r)(\boldsymbol{f}_{0r} - \mathbf{1}_r\bar{f}_{0r})(\boldsymbol{f}_{0r} - \mathbf{1}_r\bar{f}_{0r})^\top.\end{aligned}$$

It only remains to determine $\mathrm{var}_X(\hat{\lambda}_r)$. If the estimates of the probabilities of correct linkage λ_r are obtained by checking a random audit sample of linked records in each block i.e. the number of correct linkages follows the binomial distribution, then $\mathrm{var}_X(\hat{\lambda}_r) = m_r^{-1}\lambda_{0r}(1-\lambda_{0r})$. The required estimator of $\mathrm{var}_X(\hat{\boldsymbol{\beta}})$ can be obtained by plugging in estimates for unknown quantities in the approximation above. That is, this estimator is

given by

$$\begin{aligned}\hat{\mathrm{V}}_X(\hat{\boldsymbol{\beta}}) &= \left(\sum_q\sum_r \partial_{\boldsymbol{\beta}}\hat{\boldsymbol{U}}_{qr}\right)^{-1}\left\{\hat{\mathrm{V}}_X\left(\hat{\boldsymbol{H}}_C^*\right)+\sum_q\sum_r \hat{\boldsymbol{D}}_{qr}\hat{\Delta}_r\hat{\boldsymbol{D}}_{qr}^\top\right\}\\ &\quad\left\{\left(\sum_q\sum_r \partial_{\boldsymbol{\beta}}\hat{\boldsymbol{U}}_{qr}\right)^{-1}\right\}^\top.\end{aligned}$$

Next, we derive the variance estimators for the ANOVA-based variance components estimators. From (10) and Appendix I

$$\hat{\sigma}_e^2 = \frac{(\mathrm{SSA}-m_0)c-(\mathrm{SSE}-n_0)a}{bc-da}.$$

Thus,

$$\begin{aligned}\mathrm{var}(\hat{\sigma}_e^2) &= \mathrm{var}\left[\frac{(\mathrm{SSA})c-(\mathrm{SSE})a}{bc-da}\right]\\ &= \frac{1}{(bc-da)^2}\mathrm{var}[(\mathrm{SSA})c-(\mathrm{SSE})a]\\ &= \frac{1}{(bc-da)^2}\mathrm{var}({\boldsymbol{y}^*}^\top \boldsymbol{L}\,\boldsymbol{y}^*) \quad \text{where } \boldsymbol{L}=\boldsymbol{B}c-\boldsymbol{C}a.\end{aligned}$$

Since $\boldsymbol{y}^* \sim N(\boldsymbol{T}\boldsymbol{f}, \boldsymbol{\Sigma})$, by using Theorem S4 of Searle et al. (2006) we have

$$\hat{\mathrm{V}}(\hat{\sigma}_e^2) = \frac{1}{(bc-da)^2}\left[2\mathrm{tr}(\boldsymbol{L}\hat{\boldsymbol{\Sigma}}\boldsymbol{L}\hat{\boldsymbol{\Sigma}}) + 4\hat{\boldsymbol{f}}^\top\boldsymbol{T}^\top\boldsymbol{L}\hat{\boldsymbol{\Sigma}}\boldsymbol{L}\boldsymbol{T}\hat{\boldsymbol{f}}\right].$$

The estimator of $\mathrm{var}(\hat{\sigma}_e^2)$ is obtained by plugging in estimates for unknown quantities.

Finally, from (11),

$$\begin{aligned}\mathrm{var}(\hat{\sigma}_u^2) &= \mathrm{var}\left(\frac{\mathrm{SSA}-\hat{\sigma}_e^2 b}{a}\right)\\ &= \frac{1}{a^2}\left[\mathrm{var}(\mathrm{SSA}) - b^2\mathrm{var}(\hat{\sigma}_e^2)\right]\\ &= \frac{1}{a^2}\left[2\mathrm{tr}(\boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}\boldsymbol{\Sigma}) + 4\boldsymbol{f}^\top\boldsymbol{T}^\top\boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}\boldsymbol{T}\boldsymbol{f} - b^2\mathrm{var}(\hat{\sigma}_e^2)\right].\end{aligned}$$

An estimator of $\text{var}(\hat{\sigma}_u^2)$ is therefore given by

$$\hat{\text{V}}(\hat{\sigma}_u^2) = \frac{1}{a^2}\left[2\text{tr}(\boldsymbol{B}\hat{\boldsymbol{\Sigma}}\boldsymbol{B}\hat{\boldsymbol{\Sigma}}) + 4\hat{\boldsymbol{f}}^\top\boldsymbol{T}^\top\boldsymbol{B}\hat{\boldsymbol{\Sigma}}\boldsymbol{B}\boldsymbol{T}\hat{\boldsymbol{f}} - b^2\hat{\text{V}}(\hat{\sigma}_e^2)\right].$$

References

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

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

Herzog, T. N., Scheuren, F. & Winkler, W. E. (2007). *Data Quality and Record Linkage Techniques*. New York: Springer.

Lahiri, P. & Larsen, M. D. (2005). Regression analysis with linked data. *J. Amer. Statist. Assoc.* 100, 222-230.

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

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.

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.

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

Figure legends

Figure 1 Boxplots of percentage relative errors of coefficient parameters generated by different estimators in linear mixed model simulations. Note that N is the Naïve estimator while RE is the REML estimator.

Figure 2 Boxplots of percentage relative errors of variance components parameters generated by different estimators in linear mixed model simulations. Note that N is the Naïve estimator while RE is the REML estimator.

Table 1
Simulation results for estimators of the regression coefficients of the linear mixed model

Estimator	Relative Bias		Relative RMSE	
	Intercept	Slope	Intercept	Slope
Scenario 1: Linkage Probabilities Correctly Specified				
TR	0.43	-0.29	17.53	18.50
Naïve	11.78	-11.64	24.21	30.05
R	0.70	-0.56	18.83	21.42
A	0.77	-0.62	18.75	21.23
C	0.82	-0.67	18.71	21.19
MLE	0.70	-0.55	18.71	21.21
REML	0.70	-0.55	18.71	21.21
Scenario 2: Linkage Probabilities Estimated From Audit Sample				
TR	0.18	-0.03	18.50	19.13
Naïve	11.42	-11.26	24.83	30.03
R	-0.03	0.18	20.98	23.58
A	0.33	-0.18	20.78	23.19
C	0.42	-0.28	20.73	23.11
MLE	0.30	-0.15	20.74	23.10
REML	0.30	-0.15	20.73	23.10

TABLE 2
Simulation results for estimators of 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	0.79	-0.28	30.46	15.00
Naïve	-20.43	5.18	34.00	22.31
ANOVA-R	1.23	-0.42	36.24	16.77
ANOVA-A	1.24	-0.39	36.24	16.71
ANOVA-C	1.24	-0.38	36.24	16.72
MLE	-2.80	-0.29	33.40	16.43
REML	0.95	-0.24	33.93	16.44
Scenario 2: Linkage Probabilities Estimated From Audit Sample				
TR	1.25	0.19	32.06	15.92
Naïve	-20.59	5.71	34.83	24.43
ANOVA-R	1.33	-0.07	38.38	18.89
ANOVA-A	1.33	0.04	38.38	18.71
ANOVA-C	1.33	0.07	38.38	18.70
MLE	-3.31	0.24	34.95	18.24
REML	0.42	0.29	35.46	18.28

TABLE 3
Actual coverages of nominal 95% confidence intervals for the parameters of the linear mixed model

Estimator	Coverage			
	Intercept	Slope	Between-Group	Within-Group
Scenario 1: Linkage Probabilities Correctly Specified				
TR	96.6	94.1	97.4	96.0
Naïve	84.6	77.1	98.4	85.0
ANOVA-R	96.6	94.9	93.4	97.4
ANOVA-A	96.4	95.2	93.4	97.5
ANOVA-C	96.4	95.4	93.4	97.4
MLE	96.2	95.5	98.0	95.9
REML	96.2	95.5	97.6	95.9
Scenario 2: Linkage Probabilities Estimated From Audit Sample				
TR	94.0	95.0	96.2	93.9
Naïve	85.1	77.2	98.1	78.9
ANOVA-R	94.6	94.9	91.6	94.8
ANOVA-A	94.8	94.8	91.6	94.6
ANOVA-C	94.9	94.4	91.6	94.5
MLE	94.1	93.6	97.6	92.4
REML	94.9	94.6	96.9	92.5

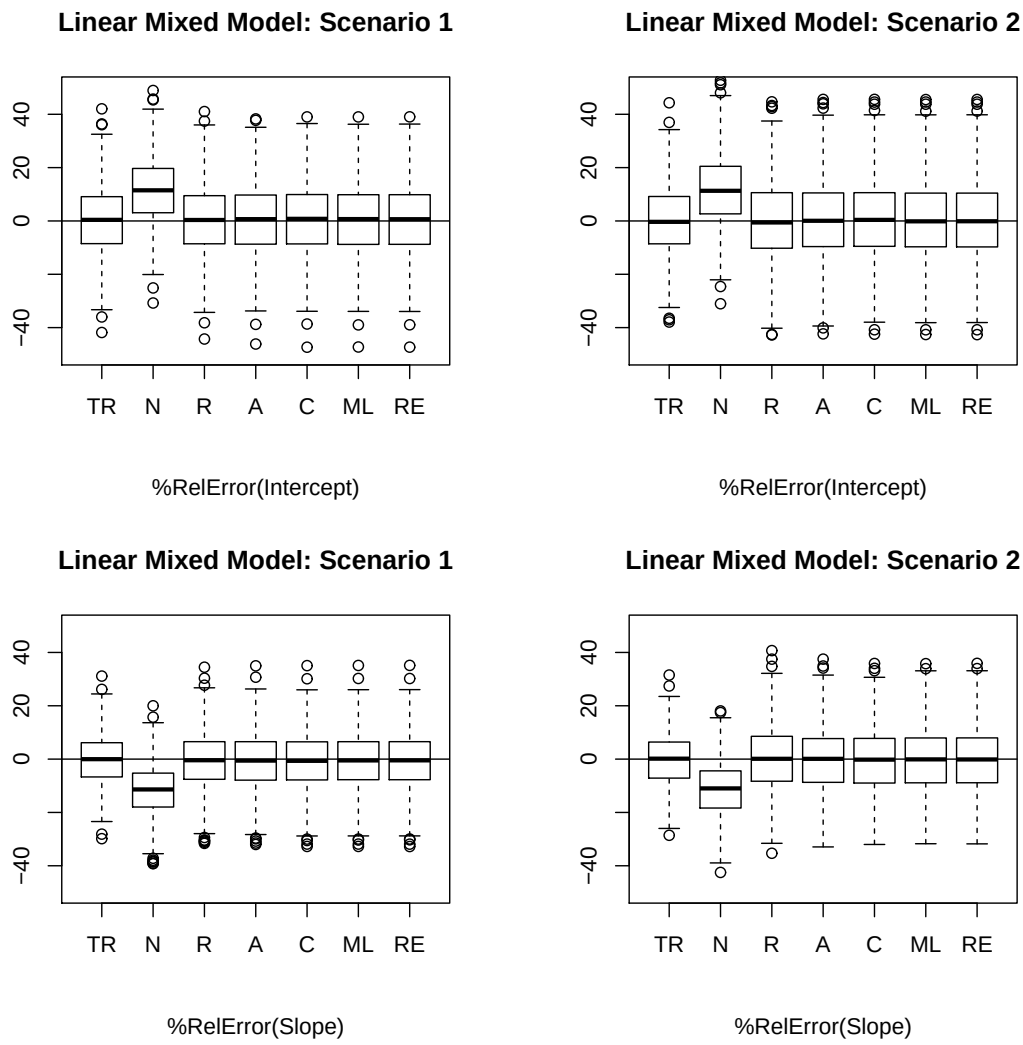
Linear Mixed Model: Scenario 1
Linear Mixed Model: Scenario 2
−40
0
20
40
TR
N
R
A
C
ML
RE
%RelError(Intercept)
%RelError(Intercept)
Linear Mixed Model: Scenario 1
Linear Mixed Model: Scenario 2
%RelError(Slope)
%RelError(Slope)

Figure 1

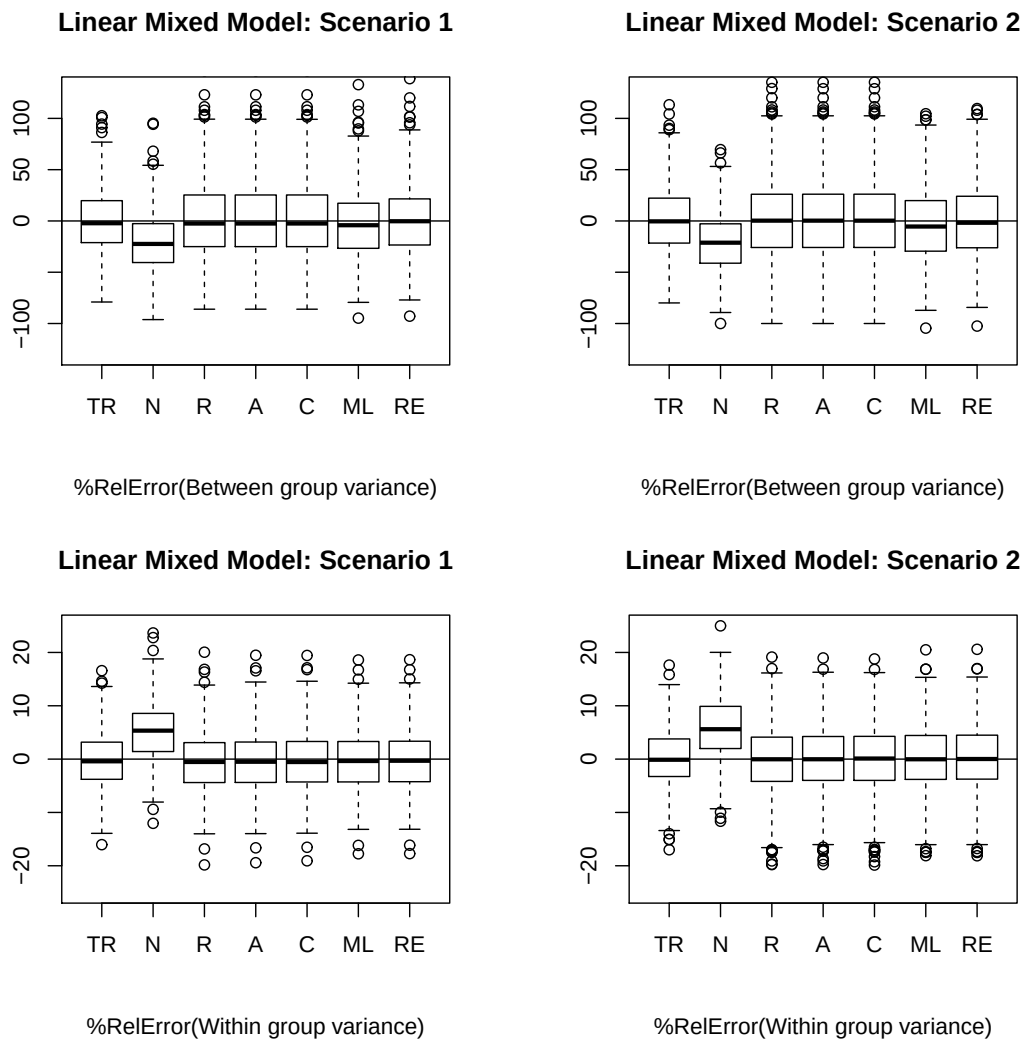
Linear Mixed Model: Scenario 1
Linear Mixed Model: Scenario 2
−100
0
50
100
TR
N
R
A
C
ML
RE
%RelError(Between group variance)
%RelError(Between group variance)
Linear Mixed Model: Scenario 1
Linear Mixed Model: Scenario 2
−20
0
10
20
TR
N
R
A
C
ML
RE
%RelError(Within group variance)
%RelError(Within group variance)

Figure 2