2010

Regression analysis for longitudinally linked data

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

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

Centre for Statistical and Survey Methodology

The University of Wollongong

Working Paper

2210

Regression analysis for longitudinally linked data

Gunky Kim and Raymond Chambers

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

Regression analysis for longitudinally linked data

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

Abstract

Most probability-based methods used to link records from two distinct data sets corresponding to the same target population do not lead to perfect linkage, i.e. there are linkage errors in the merged data. Chambers (2008) describes modifications to standard methods of regression analysis that can be used with such imperfectly linked data. However, these methods assume that the linkage process is complete, i.e. all records on the two data sets are linked. This paper extends theses ideas to accommodate the situation when the number of data sets are more than two.

key words: Record matching; linkage errors; linear regression; logistic regression; estimating equations.

1 Introduction

In recent years, because of its advantage of creating new information from already existing files by linking them, the linkage process becomes an important research tool in many areas such as health, business, economics and sociology. One important linkage application is where different data sets relating to the same individuals at different points in time are linked to provide a longitudinal data record for each individual, thus permitting longitudinal analysis for these individuals. To illustrate, the Census Data Enhancement project of the

Australian Bureau of Statistics aims to develop a Statistical Longitudinal Census Dataset by linking data from the same individuals over a number of censuses. It is expected that this linked data set will provide a powerful tool for future research into the longitudinal dynamics of the Australian population. However, without access to the same unique identifier in each of the linked data sets, there is always the possibility that linkage errors in the merged data could lead to a longitudinal record ostensibly relating to a single individual being actually made up of a composite of data items from different individuals. This in turn could lead to bias and loss of efficiency for the longitudinal modelling process. Further, as the number of censuses to be linked increase, the structure of linkage error will be more complicated as it will increase more bias and inefficiency for the modelling process.

The work of Neter *et al.* (1965) shows that small mismatching could cause significant response error. Their work has become a foundation of the analysis on the linkage error. Some authors, such as Scheuren and Winkler (1993), Scheuren and Winkler (1997) and Lahiri and Larsen (2005), have tried to extend the work of Neter *et al.* (1965) on regression setting. However, the volume of works on the analysis of the linkage error is not rich. In Chambers (2008), Chambers has developed new methods to adjust the bias in the linear regression parameters for the linkage process when two data sets were merged. In this study, we extended the ideas of Chambers (2008) to accommodate longitudinally linked data sets where the number of merged data sets are more than two.

In general, most of works for linkage error correction has been done when two data sets are merged. However, the linkage error structure of longitudinally linked data sets, when the number of data sets are more than two, are more complicated compared to the linkage error structure of two data sets. As far as our knowledge, this is the first attempt to correct the linkage errors in the merged data sets when the number of data sets are more than two. We will use three data set case as an illustration of our regression analysis, but it is trivial to see that it can be easily extended to deal with any number of data sets.

1.1 Backgrounds and assumptions

Suppose that we are interested in fitting a regression model of the form

$$E_{\boldsymbol{X}}(\boldsymbol{Y}) = f(\boldsymbol{X}; \theta),$$

where f is a known function, but the parameter θ is to be estimated, and $\boldsymbol{X}$ has more than one data sets. For example, consider a linear regression model of the form

$$\boldsymbol{Y} = \beta_0 + \boldsymbol{X}_1\beta_1 + \boldsymbol{X}_2\beta_2 + \epsilon = \boldsymbol{X}\boldsymbol{\beta} + \epsilon,$$

where we have three different files, one for $\boldsymbol{Y}$, one for $\boldsymbol{X}_1$ and one for $\boldsymbol{X}_2$. When these three data sets were created separately, and if there is no unique identifier among them to match each other, matching y_i with the correct values of x_{1i} from one file and x_{2i} from another file could be a difficult task and there could be a strong chance of mismatching. If there exist mismatches, the estimation of $\boldsymbol{\beta}$ could be biased if we ignore them in the estimation process. The purpose of our study is to develop some methodological frames to adjust the bias of $\boldsymbol{\beta}$ estimations when the mismatches are expected.

For the assumptions we made in this papers are:

1. For the case of register-register, there exist a population of N units for all $\boldsymbol{Y}$, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ such that each one of y_i should be linked with one of x_{1i} from one file and x_{2i} from another file.

2. $\boldsymbol{X}$ can be partitioned into Q different blocks[1]. Let us call this block as "m-block".

3. The linkage errors occur only within the m-block, in the sense that records in distinct m-blocks can never be linked. The records from $\boldsymbol{X}$ that make up the q^{th} m-block is denoted $\boldsymbol{X}_q$.

4. In case of sample-register, suppose that we only have sample[2] s from a bench mark register, for example, $\boldsymbol{X}_1$ with possible relation $E(\boldsymbol{Y}|\boldsymbol{X}_1, \boldsymbol{X}_2) = f(\boldsymbol{X}_1, \boldsymbol{X}_2; \boldsymbol{\theta})$ when they are correctly linked. f could be either linear of logistic function.

5. Denote $\boldsymbol{X}_{1s}$ the sample records $\boldsymbol{X}_1$ of the sample size s and some of records in $\boldsymbol{X}_{1s}$ may not be linked to the records in $\boldsymbol{Y}$ or $\boldsymbol{X}_2$.

6. Even though some of records are not linked, we assume that the regression model of linked records would be valid for the non-linked records if the links are found.

[1]See Chambers (2008) for more detailed discussion about the block.

[2]The sample set can be drawn from any data set. To explain our assumptions in more details, here we assume that the samples are drawn from $\boldsymbol{X}_1$, while $\boldsymbol{Y}$ and $\boldsymbol{X}_2$ are registers.

2 Register-register case

When there are three data sets, usually one of them is regraded as a bench mark data set and mismatches happens when someone try to link this bench mark data set with other data sets. Thus, when there are three data sets, we expect that at most two kinds of possible mismatches can happen. For example, if we set $\boldsymbol{X}_1$ as the bench mark data set, possible mismatches happen when we link $\boldsymbol{Y}$ with $\boldsymbol{X}_1$ or link $\boldsymbol{X}_1$ with $\boldsymbol{X}_2$. We will consider the case of one mismatch situations and the case of two possible mismatches case separately. For the two mismatches case, we assume that mismatches from the linkage process between $\boldsymbol{Y}$ and $\boldsymbol{X}_1$ are independent of the mismatches from the linkage process between $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$.

2.1 Three data sets and one mismatch cases: A ratio-type estimator

Note that our model is of the form

$$\boldsymbol{Y} = \beta_0 + \boldsymbol{X}_1\beta_1 + \boldsymbol{X}_2\beta_2 + \epsilon = \boldsymbol{X}\boldsymbol{\beta} + \epsilon,$$

where $\boldsymbol{X} = (\boldsymbol{1}, \boldsymbol{X}_1, \boldsymbol{X}_2)$. Suppose that $\boldsymbol{X}_1$ is the bench mark data set. Then, possible mismatch can happen either when one links records from $\boldsymbol{X}_1$ with $\boldsymbol{Y}$ or when one links records from $\boldsymbol{X}_1$ with $\boldsymbol{X}_2$. However, if the mismatch happens only when one links records from $\boldsymbol{X}_1$ with $\boldsymbol{Y}$ and $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ can be linked perfectly, one can regards $\boldsymbol{X} = (\boldsymbol{1}, \boldsymbol{X}_1, \boldsymbol{X}_2)$ as a one data set, and this case has been dealt extensively in Chambers (2008). Hence, we will only consider the case where mismatch happens when one links records from $\boldsymbol{X}_1$ with $\boldsymbol{X}_2$. Let us call this situation as **Case 0**.

Case 0: When each x_{1i} is correctly linked with corresponding y_i, but some of x_{2i} are not correctly linked with x_{1i}, one has

$$\boldsymbol{Y}_q = \beta_0 + \boldsymbol{X}_{1q}\beta_1 + \boldsymbol{X}^*_{2q}\beta^*_2 + \epsilon_q = \boldsymbol{X}^*_q\boldsymbol{\beta}^* + \epsilon_q,$$

where

$$\boldsymbol{X}^*_q = (\boldsymbol{1}_q, \boldsymbol{X}_{1q}, \boldsymbol{X}^*_{2q}) \ , \ \boldsymbol{X}^*_{2q} = B_{2q}\boldsymbol{X}_{2q}$$

and B_{2q} is a permutation matrix. Note that $\boldsymbol{X}_{2q}$ is not observable, and we only observe $\boldsymbol{X}^*_{2q}$. However, if the matrix B_{2q} is known, one has

$$\boldsymbol{X}_{2q} = B^T_{2q}\boldsymbol{X}^*_{2q}.$$

Thus,

$$\boldsymbol{X}_q = (\mathbf{1}_q, \boldsymbol{X}_{1q}, \boldsymbol{X}_{2q}) = (\mathbf{1}_q, \boldsymbol{X}_{1q}, B_{2q}^T \boldsymbol{X}_{2q}^*).$$

Let

$$\boldsymbol{X}_q^{B_2} = (\mathbf{1}_q, \boldsymbol{X}_{1q}, B_{2q}^T \boldsymbol{X}_{2q}^*). \tag{1}$$

Note that $\boldsymbol{X}_q^{B_2}$ is only observable if B_{2q} is known. But, generally, B_{2q} is unknown and in this case we adapt the *non-informative linkage assumption*[3], that is,

$$E_{\boldsymbol{X}^*}(\boldsymbol{X}_{2q}) = E_{\boldsymbol{X}^*}(B_{2q}^T)\boldsymbol{X}_{2q}^* = E_{B_{2q}}\boldsymbol{X}_{2q}^*,$$

where $E_{B_{2q}}$ satisfies the *exchangeable linkage error model*. It means

$$E_{B_{2q}} = (\lambda_{B_{2q}} - \gamma_{B_{2q}})\boldsymbol{I}_q + \gamma_{B_{2q}}\mathbf{1}_q\mathbf{1}_q^T,$$

where

$$\lambda_{B_{2q}} = \text{pr(correct linkage between } \boldsymbol{X}_{1q} \text{ and } \boldsymbol{X}_{2q}^*)$$

and

$$\gamma_{B_{2q}} = \text{pr(incorrect linkage between } \boldsymbol{X}_{1q} \text{ and } \boldsymbol{X}_{2q}^*).$$

Let

$$\boldsymbol{X}_q^E = E_{\boldsymbol{X}^*}(\boldsymbol{X}_q) = E_{\boldsymbol{X}^*}\big[(\mathbf{1}_q, \boldsymbol{X}_{1q}, \boldsymbol{X}_{2q})\big] = (\mathbf{1}_q, \boldsymbol{X}_{1q}, E_{B_{2q}}\boldsymbol{X}_{2q}^*). \tag{2}$$

Then, by OLS, one has

$$\hat{\boldsymbol{\beta}}^* = \Big[\sum_q (\boldsymbol{X}_q^*)^T \boldsymbol{X}_q^*\Big]^{-1}\Big[\sum_q (\boldsymbol{X}_q^*)^T \boldsymbol{Y}_q\Big],$$

where

$$E_{\boldsymbol{X}^*}(\hat{\boldsymbol{\beta}}^*) = \Big[\sum_q (\boldsymbol{X}_q^*)^T \boldsymbol{X}_q^*\Big]^{-1}\Big[\sum_q (\boldsymbol{X}_q^*)^T \boldsymbol{X}_q^E\Big]\boldsymbol{\beta} = \boldsymbol{D}_1\boldsymbol{\beta}.$$

Hence, if the matrix $E_{B_{2q}}$ is known and the inverse of $\boldsymbol{D}_1$ exists, a ratio form of an unbiased estimator of $\boldsymbol{\beta}$ is of the form

$$\hat{\boldsymbol{\beta}}_{R1} = \boldsymbol{D}_1^{-1}\hat{\boldsymbol{\beta}}^*.$$

Let

$$\begin{aligned} \boldsymbol{f}_q &= \boldsymbol{X}_q\boldsymbol{\beta}, \\ \boldsymbol{f}_q^* &= \boldsymbol{X}_q^*\boldsymbol{\beta}, \\ \boldsymbol{f}_q^E &= \boldsymbol{X}_q^E\boldsymbol{\beta}. \end{aligned} \tag{3}$$

[3]We assume that the distribution of B_{2q} is independent of $\boldsymbol{X}_{2q}^*$ given $\boldsymbol{X}^*$.

Proposition 1. *An asymptotic variance estimator of* $\hat{\boldsymbol{\beta}}_{R1}$ *can be defined by*

$$\hat{\boldsymbol{V}}_1(\hat{\boldsymbol{\beta}}_{R1}) = \Big[\sum_q (\boldsymbol{X}_q^*)^T \boldsymbol{X}_q^E\Big]^{-1}\Big[\sum_q (\boldsymbol{X}_q^*)^T \hat{\boldsymbol{V}}_1(\boldsymbol{Y}_q)\boldsymbol{X}_q^*\Big]\Big(\Big[\sum_q (\boldsymbol{X}_q^*)^T \boldsymbol{X}_q^E\Big]^{-1}\Big)^T,$$

where

$$\hat{\boldsymbol{V}}_1(\boldsymbol{Y}_q) = \hat{\sigma}^2 \boldsymbol{I}_q + \hat{\boldsymbol{V}}_{B_{2q}}.$$

Here, $\hat{\boldsymbol{V}}_1(\boldsymbol{Y}_q)$ *can be estimated by*

$$\hat{\sigma}^2 = N^{-1}\sum_q (\boldsymbol{Y}_q - \boldsymbol{f}_q^E)^T(\boldsymbol{Y}_q - \boldsymbol{f}_q^E)$$

and, given $\boldsymbol{f}_{B_{2q}}^* := \boldsymbol{X}_{2q}^*\beta_2$,

$$\boldsymbol{V}_{B_{2q}} = \text{diag}\Big[(1-\lambda_{B_{2q}})\big\{\lambda_{B_{2q}}(f_{B_{2q},i}^* - \bar{f}_{B_{2q}}^*)^2 + \bar{f}_{B_{2q}}^{*(2)} - (\bar{f}_{B_{2q}}^*)^2\big\}\Big],$$

where $\boldsymbol{f}_{B_{2q}}^* = (f_{B_{2q},i}^*)$ *and* $\bar{f}_{B_{2q}}^*$, $\bar{f}_{B_{2q}}^{*(2)}$ *are the averages of* $f_{B_{2q},i}^*$ *and their squares respectively in* $\boldsymbol{f}_{B_{2q}}^*$.

2.2 Three data sets and two mismatches cases: A ratio-type estimator

When there are three data sets and and two mismatches in the data linkage processes, there are two possible scenarios.

- **Case 1:** $\boldsymbol{Y}$ is the bench mark data set and the linkages between $\boldsymbol{Y}$ and $\boldsymbol{X}_1$ and the linkages between $\boldsymbol{Y}$ and $\boldsymbol{X}_2$ are done with some errors.
- **Case 2:** Either $\boldsymbol{X}_1$ or $\boldsymbol{X}_2$ is the bench mark data set[4] and the linkage between the bench mark data and other $\boldsymbol{X}$ data set and the linkage between the bench mark data set and $\boldsymbol{Y}$ are done with some errors.

Let us consider the **Case 1** first. So, we assume that the data set for y_i is correctly recorded, but there are mismatches between y_i and x_{1i} as well as between y_i and x_{2i}. Also, we assume that mismatches between y_i and x_{1i} are independent of the mismatches between y_i and x_{2i}. In this case, our regression model is of the form

$$\boldsymbol{Y}_q = \beta_0 + \boldsymbol{X}_{1q}^*\beta_1^* + \boldsymbol{X}_{2q}^*\beta_2^* + \epsilon_q = \boldsymbol{X}_q^*\boldsymbol{\beta}^* + \epsilon_q,$$

[4]In this paper, we assume that $\boldsymbol{X}_1$ is the bench mark.

where

$$\boldsymbol{X}_q^* = (\mathbf{1}_q, \boldsymbol{X}_{1q}^*, \boldsymbol{X}_{2q}^*)\ ,\ \boldsymbol{X}_{1q}^* = B_{1q}\boldsymbol{X}_{1q}\ \text{ and }\ \boldsymbol{X}_{2q}^* = B_{2q}\boldsymbol{X}_{2q},$$

and B_{1q} and B_{2q} are permutation matrices. If B_{1q} and B_{2q} are known, one has

$$\boldsymbol{X}_q = (\mathbf{1}_q, \boldsymbol{X}_{1q}, \boldsymbol{X}_{2q}) = (\mathbf{1}_q, B_{1q}^T\boldsymbol{X}_{1q}^*, B_{2q}^T\boldsymbol{X}_{2q}^*).$$

Since, B_{1q} and B_{2q} are unknown in general, we apply the non-informative linkage assumption so that

$$X_q^{E2} = E_{\boldsymbol{X}^*}(\boldsymbol{X}_q) = (\mathbf{1}_q, E_{B_{1q}}\boldsymbol{X}_{1q}^*, E_{B_{2q}}\boldsymbol{X}_{2q}^*), \tag{4}$$

where,

$$E_{B_{iq}} = (\lambda_{B_{iq}} - \gamma_{B_{iq}})\boldsymbol{I}_q + \gamma_{B_{iq}}\mathbf{1}_q\mathbf{1}_q^T$$

and

$$\lambda_{B_{iq}} = \text{pr(correct linkage between } \boldsymbol{Y}_q \text{ and } \boldsymbol{X}_{iq}^*)$$

$$\gamma_{B_{iq}} = \text{pr(incorrect linkage between } \boldsymbol{Y}_q \text{ and } \boldsymbol{X}_{iq}^*).$$

Then, by OLS,

$$\hat{\boldsymbol{\beta}}^* = \Big[\sum_q (\boldsymbol{X}_q^*)^T\boldsymbol{X}_q^*\Big]^{-1}\Big[\sum_q (\boldsymbol{X}_q^*)^T\boldsymbol{Y}_q\Big],$$

where

$$E_{\boldsymbol{X}^*}(\hat{\boldsymbol{\beta}}^*) = \Big[\sum_q (\boldsymbol{X}_q^*)^T\boldsymbol{X}_q^*\Big]^{-1}\Big[\sum_q (\boldsymbol{X}_q^*)^T\boldsymbol{X}_q^{E2}\Big]\boldsymbol{\beta} = \boldsymbol{D}_2\boldsymbol{\beta}.$$

Hence, if the matrices $E_{B_{1q}}$ and $E_{B_{2q}}$ are known and the inverse of $\boldsymbol{D}_2$ exists, a ratio form of an unbiased estimator of $\boldsymbol{\beta}$ is of the form

$$\hat{\boldsymbol{\beta}}_{R2} = \boldsymbol{D}_2^{-1}\hat{\boldsymbol{\beta}}^*.$$

Let

$$\boldsymbol{f}_q^{E2} = \boldsymbol{X}_q^{E2}\boldsymbol{\beta}.$$

Proposition 2. *An asymptotic variance estimator of $\hat{\boldsymbol{\beta}}_{R2}$ can be defined by*

$$\hat{\boldsymbol{V}}_2(\hat{\boldsymbol{\beta}}_{R2}) = \Big[\sum_q (\boldsymbol{X}_q^*)^T\boldsymbol{X}_q^{E2}\Big]^{-1}\Big[\sum_q (\boldsymbol{X}_q^*)^T\hat{\boldsymbol{V}}_2(\boldsymbol{Y}_q)\boldsymbol{X}_q^*\Big]\Big(\Big[\sum_q (\boldsymbol{X}_q^*)^T\boldsymbol{X}_q^{E2}\Big]^{-1}\Big)^T,$$

where

$$\hat{\boldsymbol{V}}_2(\boldsymbol{Y}_q) = \hat{\sigma}^2\boldsymbol{I}_q + \hat{\boldsymbol{V}}_{B_{1q}} + \hat{\boldsymbol{V}}_{B_{2q}}.$$

Here, $\hat{\boldsymbol{V}}_2(\boldsymbol{Y}_q)$ can be estimated by

$$\hat{\sigma}^2 = N^{-1}\sum_q (\boldsymbol{Y}_q - \boldsymbol{f}_q^{E2})^T(\boldsymbol{Y}_q - \boldsymbol{f}_q^{E2})$$

and, given $\boldsymbol{f}^*_{B_{jq}} := \boldsymbol{X}^*_{jq}\beta_j$ *for* $j = 1$ *or* 2 *,*

$$\boldsymbol{V}_{B_{jq}} = \text{diag}\left[(1-\lambda_{B_{jq}})\{\lambda_{B_{jq}}(f^*_{B_{jq},i} - \bar{f}^*_{B_{jq}})^2 + \bar{f}^{*(2)}_{B_{jq}} - (\bar{f}^*_{B_{jq}})^2\}\right],$$

where $\boldsymbol{f}^*_{B_{jq}} = (f^*_{B_{jq},i})$ *and* $\bar{f}^*_{B_{jq}}$, $\bar{f}^{*(2)}_{B_{jq}}$ *are the averages of* $f^*_{B_{jq},i}$ *and their squares respectively in* $\boldsymbol{f}^*_{B_{jq}}$.

Now, we are considering the **Case 2**. When some of x_{1i} are incorrectly linked with corresponding y_i or with x_{2i}, our regression model is of the form

$$\boldsymbol{Y}^*_q = \beta_0 + \boldsymbol{X}_{1q}\beta_1 + \boldsymbol{X}^*_{2q}\beta^*_2 + \epsilon_q = \boldsymbol{X}^*_q\boldsymbol{\beta}^* + \epsilon_q,$$

where

$$\boldsymbol{Y}^*_q = A_q\boldsymbol{Y}_q,\ \boldsymbol{X}^*_{2q} = B_{2q}\boldsymbol{X}_{2q}$$

and A_q and B_{2q} are permutation matrices. By non-informative linkage assumption[5] on A_q, one has

$$E_{\boldsymbol{X}^*}(\boldsymbol{Y}^*_q) = E_{\boldsymbol{X}^*}(A_q\boldsymbol{Y}_q) = E_{\boldsymbol{X}^*}(A_q)E_{\boldsymbol{X}^*}(\boldsymbol{Y}_q) = E_{A_q}E_{\boldsymbol{X}^*}(\boldsymbol{Y}_q) = E_{A_q}\boldsymbol{X}^E_q\boldsymbol{\beta}, \tag{5}$$

where

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

with

$$\lambda_{A_q} = \text{pr(correct linkage between } \boldsymbol{X}_{1q} \text{ and } \boldsymbol{Y}^*_q)$$

and

$$\gamma_{A_q} = \text{pr(incorrect linkage between } \boldsymbol{X}_{1q} \text{ and } \boldsymbol{Y}^*_q).$$

Further, we assume that the mismatch between x_{1i} and y_i is uncorrelated[6] with the mismatch between x_{1i} and x_{2i}. With these assumption, by OLS, one has

$$\begin{aligned}\hat{\boldsymbol{\beta}}^* &= \Big[\sum_q (\boldsymbol{X}^*_q)^T\boldsymbol{X}^*_q\Big]^{-1}\Big[\sum_q (\boldsymbol{X}^*_q)^T\boldsymbol{Y}^*_q\Big] \\ &= \Big[\sum_q (\boldsymbol{X}^*_q)^T\boldsymbol{X}^*_q\Big]^{-1}\Big[\sum_q (\boldsymbol{X}^*_q)^T A_q\boldsymbol{Y}_q\Big]\end{aligned}$$

and

$$E_{\boldsymbol{X}^*}(\hat{\boldsymbol{\beta}}^*) = \Big[\sum_q (\boldsymbol{X}^*_q)^T\boldsymbol{X}^*_q\Big]^{-1}\Big[\sum_q (\boldsymbol{X}^*_q)^T E_{A_q}\boldsymbol{X}^E_q\Big]\boldsymbol{\beta} = \boldsymbol{D}_3\boldsymbol{\beta}.$$

[5]Here we assume the randomness of the linkage error between $\boldsymbol{Y}^*_q$ and $\boldsymbol{X}^*_q$. See Chambers (2008) for a more detailed discussion.

[6]We will try to relax this assumption soon.

Thus, if the matrices $E_{X^*}(B_{2q}) = E_{B_{2q}}$ and $E_{X^*}(A_q) = E_{A_q}$ are known and the inverse of $\boldsymbol{D}_3$ exists, a ratio form of an unbiased estimator of $\boldsymbol{\beta}$ for this case is of the form

$$\hat{\boldsymbol{\beta}}_{R3} = \boldsymbol{D}_3^{-1}\hat{\boldsymbol{\beta}}^*.$$

Proposition 3. *An asymptotic variance estimator of $\hat{\boldsymbol{\beta}}_{R3}$ can be defined by*

$$\hat{\boldsymbol{V}}_3(\hat{\boldsymbol{\beta}}_{R3}) = \Big[\sum_q (\boldsymbol{X}_q^*)^T \boldsymbol{X}_q^E\Big]^{-1}\Big[\sum_q (\boldsymbol{X}_q^*)^T \hat{\boldsymbol{V}}_3(\boldsymbol{Y}_q^*)\boldsymbol{X}_q^*\Big]\Big(\Big[\sum_q (\boldsymbol{X}_q^*)^T \boldsymbol{X}_q^E\Big]^{-1}\Big)^T,$$

where

$$\hat{\boldsymbol{V}}_3(\boldsymbol{Y}_q^*) = \hat{\sigma}^2\boldsymbol{I}_q + \hat{\boldsymbol{V}}_{B_{2q}} + \hat{\boldsymbol{V}}_{C_{2q}}.$$

Here, $\hat{\boldsymbol{V}}_3(\boldsymbol{Y}_q^*)$ can be estimated by

$$\hat{\sigma}^2 = N^{-1}\Big(\sum_q (\boldsymbol{Y}_q^* - \boldsymbol{f}_q^E)^T(\boldsymbol{Y}_q^* - \boldsymbol{f}_q^E) - 2\sum_q (\boldsymbol{f}_q^E)^T\big[\boldsymbol{I}_q - E_{A_q}\big]\boldsymbol{f}_q^E\Big)$$

and, given $\boldsymbol{f}_{B_{2q}}^* := \boldsymbol{X}_{2q}^*\beta_2$,

$$\boldsymbol{V}_{B_{2q}} = \operatorname{diag}\Big[(1-\lambda_{B_{2q}})\big\{\lambda_{B_{2q}}(f_{B_{2q},i}^* - \bar{f}_{B_{2q}}^*)^2 + \bar{f}_{B_{2q}}^{*(2)} - (\bar{f}_{B_{2q}}^*)^2\big\}\Big],$$

where $\boldsymbol{f}_{B_{2q}}^* = (f_{B_{2q},i}^*)$ and $\bar{f}_{B_{2q}}^*$, $\bar{f}_{B_{2q}}^{*(2)}$ are the averages of $f_{B_{2q},i}^*$ and their squares respectively in $\boldsymbol{f}_{B_{2q}}^*$. Further, one has

$$\boldsymbol{V}_{C_{2q}} = A_q\, Var_{\boldsymbol{X}^*}\Big[E_{\boldsymbol{X}^*}(\boldsymbol{Y}_q|B_{2q})\Big]A_q^T,$$

and it can be estimated by

$$\boldsymbol{V}_{C_{2q}} = \operatorname{diag}\Big[(1-\lambda_{C_{2q}})\big\{\lambda_{C_{2q}}(f_{B_{2q},i}^* - \bar{f}_{B_{2q}}^*)^2 + \bar{f}_{B_{2q}}^{*(2)} - (\bar{f}_{B_{2q}}^*)^2\big\}\Big],$$

where $\boldsymbol{f}_{B_{2q}}^* = (f_{B_{2q},i}^*)$ and $\bar{f}_{B_{2q}}^*$, $\bar{f}_{B_{2q}}^{*(2)}$ are the averages of $f_{B_{2q},i}^*$ and their squares respectively in $\boldsymbol{f}_{B_{2q}}^*$. Moreover, $C_{B_{2q}} = A_q B_{2q}^T$ and $\lambda_{C_{2q}}$ is the probability of correct linkages in $C_{B_{2q}}$.

2.3 The estimating function

we will modify the estimating functions used in Chambers (2008) to accommodate the longitudinal linkage case.

Suppose that one has $E(Y|\boldsymbol{X}) = g(\boldsymbol{X};\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ can be estimated by solving

$$\boldsymbol{H}(\boldsymbol{\theta}) = 0,$$

and $\boldsymbol{H}(\boldsymbol{\theta})$ is a function that satisfies $E_X\big[\boldsymbol{H}(\boldsymbol{\theta}_0)\big] = 0$ when $\boldsymbol{\theta}_0$ is the true value of $\boldsymbol{\theta}$. Let ∂_θ be the partial differentiation operator with respect to $\boldsymbol{\theta}$. Suppose that $\hat{\boldsymbol{\theta}}$ satisfies $\boldsymbol{H}(\hat{\boldsymbol{\theta}}) = 0$. Then, under some regularity condition for the smoothness and Taylor expansion,

$$0 = \boldsymbol{H}(\hat{\boldsymbol{\theta}}) \approx \boldsymbol{H}(\boldsymbol{\theta}_0) + (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\partial_\theta \boldsymbol{H}(\boldsymbol{\theta}_0).$$

If $\boldsymbol{H}(\boldsymbol{\theta})$ is an unbiased estimating function and $\partial_\theta \boldsymbol{H}(\boldsymbol{\theta}_0)$ is non-singular, one has

$$E_X\big[\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\big] \approx -\big[\partial_\theta \boldsymbol{H}(\boldsymbol{\theta}_0)\big]^{-1} E_X\big[\boldsymbol{H}(\boldsymbol{\theta}_0)\big] = \mathbf{0}.$$

Then, the variance function for $\hat{\boldsymbol{\theta}}$ can be derived by

$$\mathrm{Var}_X(\hat{\boldsymbol{\theta}}) \approx \big[\partial_\theta \boldsymbol{H}(\boldsymbol{\theta}_0)\big]^{-1} \mathrm{Var}_X\big[\boldsymbol{H}(\boldsymbol{\theta}_0)\big] \Big(\big[\partial_\theta \boldsymbol{H}(\boldsymbol{\theta}_0)\big]^{-1}\Big)^T.$$

In Chambers (2008), the estimating function is of the form

$$\boldsymbol{H}(\boldsymbol{\theta}) = \sum_q \boldsymbol{G}_q(\boldsymbol{\theta})\big\{\boldsymbol{Y}_q - \boldsymbol{f}_q\big\},$$

where $\boldsymbol{f}_q = E_X(\boldsymbol{Y}_q)$ and $\boldsymbol{G}_q(\boldsymbol{\theta})$[7] is a function of $\boldsymbol{\theta}$ and $\boldsymbol{X}_q$ but not of $\boldsymbol{Y}_q$.

In the longitudinal case for the three data set, we have three different cases to consider. Firstly, consider the **Case 0** where Y and $\boldsymbol{X}_1$ are correctly linked, but $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are not correctly linked. Hence, we can observe true $\boldsymbol{Y}_q$, but we cannot observe the true $\boldsymbol{X}$. Instead, we observe $\boldsymbol{X}^*$, which is of the form

$$\boldsymbol{X}^* = (\mathbf{1}, \boldsymbol{X}_1, \boldsymbol{X}_2^*) \ , \ \boldsymbol{X}_2^* = B_2 \boldsymbol{X}_2$$

and B_2 is a permutation matrix that is not observable in general. Then, a naive estimating function can be of the form

$$\boldsymbol{H}^*(\boldsymbol{\theta}) = \sum_q \boldsymbol{G}_q(\boldsymbol{\theta})\big\{\boldsymbol{Y}_q - \boldsymbol{f}_q^*(\boldsymbol{\theta})\big\},$$

where $\boldsymbol{f}_q^*(\boldsymbol{\theta}) = \boldsymbol{X}_q^* \boldsymbol{\beta}$. Then, it is easy to see that the estimator from the naive estimating function is biased, because

$$E_{\boldsymbol{X}^*}(\boldsymbol{Y}_q) = \boldsymbol{f}_q^E(\boldsymbol{\theta}) \neq \boldsymbol{f}_q^*(\boldsymbol{\theta}).$$

Thus, an unbiased estimating function can be of the form

$$\boldsymbol{H}_1^*(\boldsymbol{\theta}) = \sum_q \boldsymbol{G}_q(\boldsymbol{\theta})\big\{\boldsymbol{Y}_q - \boldsymbol{f}_q^E(\boldsymbol{\theta})\big\}, \tag{6}$$

[7]Some examples of $\boldsymbol{G}_q$ for different estimators are given in the simulation section.

where $\boldsymbol{f}_q^E(\boldsymbol{\theta}) = \boldsymbol{X}_q^E\boldsymbol{\beta} = (\mathbf{1}_q, \boldsymbol{X}_{1q}, E_{B_{2q}}\boldsymbol{X}_{2q}^*)\boldsymbol{\beta}$.

Let us consider the **Case 1** where $\boldsymbol{Y}$ is the bench mark data set and the linkages between $\boldsymbol{Y}$ and $\boldsymbol{X}_1$ and the linkages between $\boldsymbol{Y}$ and $\boldsymbol{X}_2$ are done with some errors. In this case, we have similar estimating function

$$\boldsymbol{H}_2^*(\boldsymbol{\theta}) = \sum_q \boldsymbol{G}_q(\boldsymbol{\theta})\big\{\boldsymbol{Y}_q - \boldsymbol{f}_q^{E2}(\boldsymbol{\theta})\big\}, \tag{7}$$

where, by (4), $\boldsymbol{f}_q^{E2}(\boldsymbol{\theta}) = \boldsymbol{X}_q^{E2}\boldsymbol{\beta} = (\mathbf{1}_q, E_{B_{1q}}\boldsymbol{X}_{1q}^*, E_{B_{2q}}\boldsymbol{X}_{2q}^*)\boldsymbol{\beta}$.

Now, consider the **Case 2** where $\boldsymbol{X}_1$ is the bench mark data set and the linkage between $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ and the linkage between $\boldsymbol{X}_1$ and $\boldsymbol{Y}$ are done with some errors. In this case, $\boldsymbol{Y}_q^*$ is observed instead of $\boldsymbol{Y}_q$ and also the true $\boldsymbol{X}$ is not observable. Instead, we observe $\boldsymbol{X}^*$, which is of the form

$$\boldsymbol{X}^* = (\mathbf{1}, \boldsymbol{X}_1, \boldsymbol{X}_2^*)\ ,\ \boldsymbol{X}_2^* = B_2\boldsymbol{X}_2$$

and B_2 is a permutation matrix that is not observable in general. Hence, a naive estimating function can be of the form

$$\boldsymbol{H}^*(\boldsymbol{\theta}) = \sum_q \boldsymbol{G}_q(\boldsymbol{\theta})\big\{\boldsymbol{Y}_q^* - \boldsymbol{f}_q^*(\boldsymbol{\theta})\big\},$$

where $\boldsymbol{f}_q^*(\boldsymbol{\theta}) = \boldsymbol{X}_q^*\boldsymbol{\beta}$. Then, as before, it is easy to see that the estimator from the naive estimation function is biased, because

$$E_{\boldsymbol{X}^*}(\boldsymbol{Y}_q^*) = E_{A_q}\boldsymbol{f}_q^E(\boldsymbol{\theta}) \neq \boldsymbol{f}_q^*(\boldsymbol{\theta}).$$

Hence, by (2), (5) and (28), an unbiased estimator is of the form

$$\boldsymbol{H}_3^*(\boldsymbol{\theta}) = \sum_q \boldsymbol{G}_q(\boldsymbol{\theta})\big\{\boldsymbol{Y}_q^* - E_{A_q}\boldsymbol{f}_q^E(\boldsymbol{\theta})\big\}, \tag{8}$$

and the estimator $\hat{\boldsymbol{\theta}}_3^*$ is defined as the the solution of

$$\boldsymbol{H}_3^*(\hat{\boldsymbol{\theta}}_3^*) = 0.$$

Theorem 4. *Let $\hat{\boldsymbol{\theta}}_1^*$ be the solution of (6). Then, an asymptotic variance estimator is of the form*

$$V_{1|X^*}(\hat{\boldsymbol{\theta}}_1^*) = \Big[\sum_q \boldsymbol{G}_q\partial_\theta\boldsymbol{f}_q^E(\hat{\boldsymbol{\theta}}_1^*)\Big]^{-1}\Big[\sum_q \boldsymbol{G}_q\hat{\Sigma}_q^{*1}\boldsymbol{G}_q^T\Big]\Big(\Big[\sum_q \boldsymbol{G}_q\partial_\theta\boldsymbol{f}_q^E(\hat{\boldsymbol{\theta}}_1^*)\Big]^{-1}\Big)^T$$

where,

$$\hat{\Sigma}_q^{*1} = \hat{\sigma}_q^2 \boldsymbol{I}_q + \hat{\boldsymbol{V}}_{B_{2q}}.$$

Also, let $\hat{\boldsymbol{\theta}}_2^*$ *be the solution of (7). Then, an asymptotic variance estimator is of the form*

$$V_{2|X^*}(\hat{\boldsymbol{\theta}}_2^*) = \Big[\sum_q \boldsymbol{G}_q \partial_\theta \boldsymbol{f}_q^{E2}(\hat{\boldsymbol{\theta}}_2^*)\Big]^{-1}\Big[\sum_q \boldsymbol{G}_q \hat{\Sigma}_q^{*2} \boldsymbol{G}_q^T\Big]\Big(\Big[\sum_q \boldsymbol{G}_q \partial_\theta \boldsymbol{f}_q^{E2}(\hat{\boldsymbol{\theta}}_2^*)\Big]^{-1}\Big)^T,$$

where, by (26), $\hat{\Sigma}_q^{*2} = \hat{\sigma}_q^2 \boldsymbol{I}_q + \hat{\boldsymbol{V}}_{B_{1q}} + \hat{\boldsymbol{V}}_{B_{2q}}$.

Finally, the asymptotic variance estimator for the solution of (8) is of the form

$$V_{3|X^*}(\hat{\boldsymbol{\theta}}_3^*) = \Big[\sum_q \boldsymbol{G}_q E_{A_q} \partial_\theta \boldsymbol{f}_q^{E}(\hat{\boldsymbol{\theta}}_3^*)\Big]^{-1}\Big[\sum_q \boldsymbol{G}_q \hat{\Sigma}_q^{*3} \boldsymbol{G}_q^T\Big]\Big(\Big[\sum_q \boldsymbol{G}_q E_{A_q} \partial_\theta \boldsymbol{f}_q^{E}(\hat{\boldsymbol{\theta}}_3^*)\Big]^{-1}\Big)^T,$$

where,

$$\hat{\Sigma}_q^{*3} = \hat{\sigma}_q^2 \boldsymbol{I}_q + \hat{\boldsymbol{V}}_{C_{2q}} + \hat{\boldsymbol{V}}_{A_q}.$$

2.4 Variance estimation when linkage probabilities are estimated

So far, we assume that we know the correct linkage probabilities which is a very strong assumption. In this subsection, we consider the case where the correct linkage probabilities are estimated by checking a random 'audit' sample of linked records in each m-block. More details of this audit estimates when there are two data sets can be found in Chambers (2008), and we will modify his idea to accommodate the cases when there are more than two data sets.

Let us consider the **Case 2** where x_{1i} is neither correctly linked with corresponding y_i, nor with x_{2i}. In this case, we need to consider two different linkage probabilities:

$$\lambda_{A_q} = \text{pr(correct linkage between } \boldsymbol{X}_{1q} \text{ and } \boldsymbol{Y}_q^*)$$
$$\lambda_{B_{2q}} = \text{pr(correct linkage between } \boldsymbol{X}_{1q} \text{ and } \boldsymbol{X}_{2q}^*),$$

where there is no correlation between them. Thus, the estimating function (8) can be replaced by

$$\boldsymbol{H}_3^*(\boldsymbol{\theta}, \lambda_A, \lambda_{B_2}) = \sum_q \boldsymbol{G}_q\big\{\boldsymbol{Y}_q^* - E_{A_q}(\lambda_{A_q})\boldsymbol{f}_q^E(\boldsymbol{\theta}, \lambda_{B_{2q}})\big\} = \sum_q \boldsymbol{U}_q(\boldsymbol{\theta}_q, \lambda_{A_q}, \lambda_{B_{2q}}),$$

and a first order Taylor series approximation is of the form

$$\begin{aligned}\mathbf{0} &= \boldsymbol{H}_3^*(\hat{\boldsymbol{\theta}}_3^{**}, \hat{\lambda}_A, \hat{\lambda}_{B_2}) \\ &\approx \boldsymbol{H}_3^*(\boldsymbol{\theta}_0, \lambda_A^0, \lambda_{B_2}^0) + \partial_\theta \boldsymbol{H}_3^*(\boldsymbol{\theta}_0, \lambda_A^0, \lambda_{B_2}^0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \\ &\quad + \partial_{\lambda_A} \boldsymbol{H}_3^*(\boldsymbol{\theta}_0, \lambda_A^0, \lambda_{B_2}^0)(\hat{\lambda}_A - \lambda_A^0) + \partial_{\lambda_{B_2}} \boldsymbol{H}_3^*(\boldsymbol{\theta}_0, \lambda_A^0, \lambda_{B_2}^0)(\hat{\lambda}_{B_2} - \lambda_{B_2}^0),\end{aligned}$$

where $\boldsymbol{\theta}_0$, λ_A^0 and $\lambda_{B_2}^0$ denote the *true* values of $\boldsymbol{\theta}$, λ_A and λ_{B_2} respectively. Denote

$$\begin{aligned}\boldsymbol{H}_0^* &= \boldsymbol{H}_3^*(\boldsymbol{\theta}_0, \lambda_A^0, \lambda_{B_2}^0), \\ \partial_{\lambda_1} &= \partial_{\lambda_A} \quad \text{and} \\ \partial_{\lambda_2} &= \partial_{\lambda_{B_2}}.\end{aligned}$$

Then, one has

$$\hat{\boldsymbol{\theta}}_3^{**} = \boldsymbol{\theta}_0 - \left[\partial_\theta \boldsymbol{H}_0^*\right]^{-1}\Big[\boldsymbol{H}_0^* + \partial_{\lambda_1}\boldsymbol{H}_0^*(\hat{\lambda}_A - \lambda_A^0) + \partial_{\lambda_2}\boldsymbol{H}_0^*(\hat{\lambda}_{B_2} - \lambda_{B_2}^0)\Big].$$

If the estimates of the linkage probabilities are obtained by a random audit sample (of the size m_q^A for λ_{A_q} and m_q^B for $\lambda_{B_{2q}}$) of linked records, one has

$$\mathrm{Var}_{\boldsymbol{X}^*}(\lambda_{A_q}) = (m_q^A)^{-1}\lambda_{A_q}(1 - \lambda_{A_q})$$

and

$$\mathrm{Var}_{\boldsymbol{X}^*}(\lambda_{B_{2q}}) = (m_q^B)^{-1}\lambda_{B_{2q}}(1 - \lambda_{B_{2q}}).$$

Theorem 5. *An asymptotic variance estimator of $\hat{\boldsymbol{\theta}}_3^{**}$ is of the form*

$$\begin{aligned}\hat{\boldsymbol{V}}^{\lambda}_{3|\boldsymbol{X}^*}(\hat{\boldsymbol{\theta}}_3^{**}) = \left[\partial_\theta \hat{\boldsymbol{H}}_0^*\right]^{-1}\Big[\hat{V}_{3|\boldsymbol{X}^*}\left(\boldsymbol{\theta}_3^{**}\right) + \left(\partial_{\lambda_1}\hat{\boldsymbol{H}}_0^*\right) Var_{\boldsymbol{X}^*}(\hat{\lambda}_A)\left(\partial_{\lambda_1}\hat{\boldsymbol{H}}_0^*\right)^T \\ + \left(\partial_{\lambda_2}\hat{\boldsymbol{H}}_0^*\right) Var_{\boldsymbol{X}^*}(\hat{\lambda}_{B_2})\left(\partial_{\lambda_2}\hat{\boldsymbol{H}}_0^*\right)^T\Big]\left\{\left[\partial_\theta \hat{\boldsymbol{H}}_0^*\right]^{-1}\right\}^T,\end{aligned}$$

where

$$\begin{aligned}\partial_{\lambda_1}\hat{\boldsymbol{H}}_0^* &= -\sum_q \boldsymbol{G}_q\Big[(M_q - 1)^{-1}(M_q\boldsymbol{I}_q - \mathbf{1}_q\mathbf{1}_q^T)\Big]\hat{\boldsymbol{f}}_q^E(\hat{\boldsymbol{\theta}}, \hat{\lambda}_{B_{2q}}) \quad \textit{and} \\ \partial_{\lambda_2}\hat{\boldsymbol{H}}_0^* &= -\sum_q \boldsymbol{G}_q E_{A_q}\Big[(M_q - 1)^{-1}(M_q\boldsymbol{I}_q - \mathbf{1}_q\mathbf{1}_q^T)\Big]\boldsymbol{X}_{2q}^*\hat{\beta}_2.\end{aligned}$$

and $\hat{V}_{3|\boldsymbol{X}^*}$ is the asymptotic variance estimator for $\hat{\boldsymbol{\theta}}_3^*$ in the Theorem 4.

*Similarly, an asymptotic variance estimator for $\hat{\boldsymbol{\theta}}_2^{**}$, the unbiased estimator for the **Case 1** when the linkage probabilities are unknown, can be represented by*

$$\begin{aligned}\hat{\boldsymbol{V}}^{\lambda}_{2|\boldsymbol{X}^*}(\hat{\boldsymbol{\theta}}_2^{**}) = \left[\partial_\theta \hat{\boldsymbol{H}}_0^*\right]^{-1}\Big[\hat{V}_{2|\boldsymbol{X}^*}\left(\boldsymbol{\theta}_2^{**}\right) + \left(\partial_{\lambda_{B_1}}\hat{\boldsymbol{H}}_0^*\right) Var_{\boldsymbol{X}^*}(\hat{\lambda}_{B_1})\left(\partial_{\lambda_{B_1}}\hat{\boldsymbol{H}}_0^*\right)^T \\ + \left(\partial_{\lambda_{B_2}}\hat{\boldsymbol{H}}_0^*\right) Var_{\boldsymbol{X}^*}(\hat{\lambda}_{B_2})\left(\partial_{\lambda_{B_2}}\hat{\boldsymbol{H}}_0^*\right)^T\Big]\left\{\left[\partial_\theta \hat{\boldsymbol{H}}_0^*\right]^{-1}\right\}^T,\end{aligned}$$

where,

$$\lambda_{B_{1q}} = pr(\textit{correct linkage between } \boldsymbol{Y}_q \textit{ and } \boldsymbol{X}^*_{1q}) \textit{ and}$$
$$\partial_{\lambda_{B_1}} \hat{\boldsymbol{H}}^*_0 = -\sum_q \boldsymbol{G}_q \Big[(M_q - 1)^{-1} (M_q \boldsymbol{I}_q - \mathbf{1}_q \mathbf{1}_q^T) \Big] \boldsymbol{X}^*_{1q} \hat{\beta}_1.$$

Finally, an asymptotic variance estimator for $\hat{\boldsymbol{\theta}}^{**}_1$*, the unbiased estimator for the* ***Case 0*** *when the linkage probabilities are unknown, can be represented by*

$$\hat{\boldsymbol{V}}^{\lambda}_{1|\boldsymbol{X}^*}(\hat{\boldsymbol{\theta}}^{**}_1) = \left[\partial_\theta \hat{\boldsymbol{H}}^*_0\right]^{-1} \Big[\hat{V}_{1|\boldsymbol{X}^*}(\boldsymbol{\theta}^{**}_1) + \left(\partial_{\lambda_{B_2}} \hat{\boldsymbol{H}}^*_0\right) Var_{\boldsymbol{X}^*}(\hat{\lambda}_{B_2}) \left(\partial_{\lambda_{B_2}} \hat{\boldsymbol{H}}^*_0\right)^T \Big] \Big\{ \left[\partial_\theta \hat{\boldsymbol{H}}^*_0\right]^{-1} \Big\}^T.$$

2.5 Simulation

We use simulation to compare the performances of different estimators we considered in this study. The linear model we used in this simulation is of the form

$$\boldsymbol{Y} = 1 + 5\boldsymbol{X}_1 + 8\boldsymbol{X}_2 + \boldsymbol{\epsilon},$$

where $\boldsymbol{X}_1$ were drawn from the standard normal distribution and $\boldsymbol{X}_2$ were drawn from the normal distribution with mean= 2 and the variance of 4. $\boldsymbol{\epsilon}$ were drawn from the standard normal distribution as well.

In this simulation, we consider all three cases we have studied:

- **Case 0**: $\boldsymbol{X}_1$ is the bench mark data set and the mismatch happens only from the linkage between $\boldsymbol{X}_1$ and r $\boldsymbol{X}_2$.
- **Case 1**: $\boldsymbol{Y}$ is the bench mark data set and the linkages between $\boldsymbol{Y}$ and $\boldsymbol{X}_1$ and the linkages between $\boldsymbol{Y}$ and $\boldsymbol{X}_2$ are done with some errors.
- **Case 2**: $\boldsymbol{X}_1$ is the bench mark data set and the linkage between $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ data set and the linkage between $\boldsymbol{X}_1$ and $\boldsymbol{Y}$ are done with some errors.

Here, we will only explain how we generate the data sets for **Case 2**. Generating the data sets for other cases are quite trivial.

There are three m-blocks and in each m block, the pairs (x_{1i}, x^*_{2i}) were generated according to an independent exchangeable linkage error model. Further, given $\boldsymbol{X}^*_i = (1, x_{1i}, x^*_{2i})$, the pairs $(y^*_i, \boldsymbol{X}^*_i)$ were generated according to another independent exchangeable linkage error model. In this simulation, we use three m-blocks of sizes 500 for each m-block. Also we

assume that the probability of correct linkage between $\boldsymbol{Y}_q^*$ and $\boldsymbol{X}_q^*$ and probability of correct linkage between $\boldsymbol{X}_{1q}$ and $\boldsymbol{X}_{2q}^*$ are known.

The estimators for the simulations are

1. the naive OLS estimator (ST),
2. the ratio-type estimator (R),
3. the Lahiri-Larsen estimator (A) and
4. the empirical Best Linear Unbiased Estimator, EBLUE, (C).

Note that different estimating functions have different form of $\boldsymbol{G}_q$. In our case,

1. the naive estimator: $\boldsymbol{G}_q = (\boldsymbol{X}_q^*)^T$,
2. the Lahiri-Larsen estimator: $\boldsymbol{G}_q = (\hat{E}_{A_q}\boldsymbol{X}_q^E)^T$ and
3. the EBLUE: $\boldsymbol{G}_q = (\hat{E}_{A_q}\boldsymbol{X}_q^E)^T(\hat{\sigma}_q^2\boldsymbol{I}_q + \hat{\boldsymbol{V}}_{C_{2q}} + \hat{\boldsymbol{V}}_{A_q})^{-1}$.

The assumptions on the probability of correct linkage on each m-block are

- the probability of correct linkage between $\boldsymbol{Y}_q^*$ and $\boldsymbol{X}_{1q}$: $\lambda_{A_1} = 1$, $\lambda_{A_2} = 0.95$ and $\lambda_{A_3} = 0.75$,
- the probability of correct linkage between $\boldsymbol{Y}_q$ and$\boldsymbol{X}_{1q}^*$: $\lambda_{B_{11}} = 1$, $\lambda_{B_{12}} = 0.95$ and $\lambda_{B_{13}} = 0.75$ and
- the probability of correct linkage between $\boldsymbol{X}_{1q}$ and $\boldsymbol{X}_{2q}^*$: $\lambda_{B_{21}} = 1$, $\lambda_{B_{22}} = 0.85$ and $\lambda_{B_{23}} = 0.8$.

Under the above scenario, the estimators were independently simulated 1000 times. The regression parameters were estimated using the four estimators. The following plot box represent the overall performance of the estimators.

Clearly, the ration-type estimator, the Lahiri-Larsen estimator and the EBLUE correct the bias due to incorrect linkage, and the EBLUE outperforms other estimators, that was also noted in Chambers (2008) where two registers were merged. These observations are consistent for all three cases. It is worth to note that the EBLUE(C) outperforms all other estimators in general. The figures 1-3 clearly show that EBLUE is the best one. However, our simulation shows that the relative biases of EBLUE, when λs are unknown, are

larger than the Lahiri-Larsen estimator and the ratio-type estimator. But the overall relative RMSE are smaller than other estimators.

[Table 1 here.]

[Table 2 here.]

[Table 3 here.]

3 Sample-register case

In this section, we consider the case where we only observe a sample s of records from the bench mark data set. Suppose that $\boldsymbol{X}_1$ is the bench mark data set. When all the records in $\boldsymbol{X}_1$-register are linked to the records in$\boldsymbol{X}_2$-register and $\boldsymbol{Y}$-register, all of the sample records s from $\boldsymbol{X}_1$-register are perfectly linked with some records in $\boldsymbol{X}_2$-register and $\boldsymbol{Y}$-register. However, in reality, some records in the sample s cannot be linked to a record in $\boldsymbol{X}_2$-register or $\boldsymbol{Y}$-register. We will consider these two cases separately.

3.1 Sample-register case: When sample records are perfectly linked

As before, we will consider three different cases, **Case 0**, **Case 1** and **Case 2**.

Let us start with **Case 2**. If all records in the sample s are linked to the records in$\boldsymbol{X}_2$-register and $\boldsymbol{Y}$-register, We can assume that the sample s is a part of $\boldsymbol{X}_1$-register that is complete register-register linkage. Hence we can use a weighted estimating function. In this subsection, we will modify the estimating function approach to accommodate this sample-register linkage.

When the sample records s from $\boldsymbol{X}_1$-register are linked to $\boldsymbol{X}_2$-register and $\boldsymbol{Y}$-register, we observe a subset s_q of M_q records from $\boldsymbol{Y}_q^*$, which we denote by $\boldsymbol{Y}_{sq}^*$. More precisely, let M_q be the population number in the q^{th} m-block, and let m_{sq} be the sample size in the q^{th} m-block. We use a subscript of sq to denote quantities that depend on the sample records in the q^{th} m-block. Similarly, the subscript of rq is used to indicate quantities that depend on the non-sample records in the q^{th} m-block.

Under perfect linkage of the sample data, when there is no linkage error, the true parameter θ_0 can be estimated by solving the estimating equation

$$\boldsymbol{H}_s(\boldsymbol{\theta}) = \sum_q \boldsymbol{G}_{sq}\big\{\boldsymbol{Y}_{sq} - \boldsymbol{f}_{sq}(\boldsymbol{\theta})\big\},$$

where $\boldsymbol{G}_{sq}$ is modified by the sample weights w_{sq} that depend on the ratio of the sample size from the population. When there exist linkage errors and we ignore the errors, the estimating equation is then of the form

$$\boldsymbol{H}^*_s(\boldsymbol{\theta}) = \sum_q \boldsymbol{G}_{sq}\big\{\boldsymbol{Y}^*_{sq} - \boldsymbol{f}^*_{sq}(\boldsymbol{\theta})\big\},$$

where

$$\boldsymbol{Y}^*_{sq} = A_{sq}\boldsymbol{Y}^*_q$$

and

$$A_q = \begin{pmatrix} A_{sq} \\ A_{rq} \end{pmatrix} = \begin{pmatrix} A_{ssq} & A_{srq} \\ A_{rsq} & A_{rrq} \end{pmatrix}$$

is the sample/non-sample decomposition of the complete linkage process in the q^{th} m-block. This estimating equation leads to a bias because $E_{\boldsymbol{X}^*}(\boldsymbol{Y}^*_{sq}) \neq \boldsymbol{f}_{sq}$. To correct the bias, by using the fact that

$$E_{\boldsymbol{X}^*}(\boldsymbol{Y}^*_{sq}) = E_{\boldsymbol{X}^*}(A_{sq}\boldsymbol{Y}^*_q) = E_{A_{sq}}\boldsymbol{f}^E_q(\boldsymbol{\theta}),$$

we modify this estimating equation

$$\begin{aligned} \boldsymbol{H}^{adj}_s(\boldsymbol{\theta}) &= \sum_q \boldsymbol{G}_{sq}\big\{\boldsymbol{Y}^*_{sq} - E_{A_{sq}}\boldsymbol{f}^E_q(\boldsymbol{\theta})\big\} \\ &= \sum_q \boldsymbol{G}_{sq}\big\{\boldsymbol{Y}^*_{sq} - E_{A_{ssq}}\boldsymbol{f}^E_{sq}(\boldsymbol{\theta}) - E_{A_{srq}}\boldsymbol{f}^E_{rq}(\boldsymbol{\theta})\big\}, \end{aligned} \tag{9}$$

where

$$E_{A_q} = \begin{pmatrix} E_{A_{sq}} \\ E_{A_{rq}} \end{pmatrix} = \begin{pmatrix} E_{A_{ssq}} & E_{A_{srq}} \\ E_{A_{rsq}} & E_{A_{rrq}} \end{pmatrix} \tag{10}$$

is the corresponding sample/non-sample decomposition of the expected value E_{A_q} of A_q. Now, by the definition of E_{A_q}, one has

$$E_{A_{ssq}} = \Big(\frac{\lambda_{A_q}M_q - 1}{M_q - 1}\Big)\boldsymbol{I}_{sq} + \Big(\frac{1 - \lambda_{A_q}}{M_q - 1}\Big)\mathbf{1}_{sq}\mathbf{1}^T_{sq}$$

and

$$E_{A_{srq}} = \Big(\frac{1 - \lambda_{A_q}}{M_q - 1}\Big)\mathbf{1}_{sq}\mathbf{1}^T_{rq}$$

so that (9) becomes

$$\boldsymbol{H}_{s}^{adj}(\boldsymbol{\theta}) = \sum_{q} \boldsymbol{G}_{sq}\big\{\boldsymbol{Y}_{sq}^{*} - \Big(\frac{\lambda_{A_q} M_q - 1}{M_q - 1}\Big)\boldsymbol{I}_{sq}\boldsymbol{f}_{sq}^{E}(\boldsymbol{\theta}) - \Big(\frac{1 - \lambda_{A_q}}{M_q - 1}\Big)\mathbf{1}_{sq}\mathbf{1}_{q}^{T}\boldsymbol{f}_{q}^{E}(\boldsymbol{\theta})\big\}.$$

Using a weighting approach[8], the unknown value $\mathbf{1}_q^T \boldsymbol{f}_q^E(\boldsymbol{\theta})$ can be replaced by $\boldsymbol{w}_{sq}^T \boldsymbol{f}_{sq}^E(\boldsymbol{\theta})$ under the assumption that the samples are chosen randomly from the population. This leads us to the estimating function of the form

$$\boldsymbol{H}_{ws3}^{adj}(\boldsymbol{\theta}) = \sum_{q} \boldsymbol{G}_{sq}\big\{\boldsymbol{Y}_{sq}^{*} - \tilde{E}_{A_{sq}}\boldsymbol{f}_{sq}^{E}(\boldsymbol{\theta})\big\}, \tag{11}$$

where

$$\tilde{E}_{A_{sq}} = \Big(\frac{\lambda_{A_q} M_q - 1}{M_q - 1}\Big)\boldsymbol{I}_{sq} + \Big(\frac{1 - \lambda_{A_q}}{M_q - 1}\Big)\mathbf{1}_{sq}\boldsymbol{w}_{sq}^{T}.$$

For the **Case 1**, formulae are similar, but simpler than those in **Case 2**. Note that, in this case, we observe true $\boldsymbol{Y}_{sq}$. Hence, the estimating function is of the form

$$\boldsymbol{H}_{ws2}^{adj}(\boldsymbol{\theta}) = \sum_{q} \boldsymbol{G}_{sq}\big\{\boldsymbol{Y}_{sq} - \boldsymbol{f}_{sq}^{E2}(\boldsymbol{\theta})\big\},$$

where

$$\begin{aligned}
\boldsymbol{f}_{sq}^{E2} &= \boldsymbol{X}_{sq}^{E2}\boldsymbol{\beta} = (\mathbf{1}_{sq}, \tilde{E}_{B_{1sq}}\boldsymbol{X}_{1sq}^{*}, \tilde{E}_{B_{2sq}}\boldsymbol{X}_{2sq}^{*})\boldsymbol{\beta}, \\
\tilde{E}_{B_{1sq}} &= \Big(\frac{\lambda_{B_{1q}} M_q - 1}{M_q - 1}\Big)\boldsymbol{I}_{sq} + \Big(\frac{1 - \lambda_{B_{1q}}}{M_q - 1}\Big)\mathbf{1}_{sq}\boldsymbol{w}_{sq}^{T} \quad \text{and} \\
\tilde{E}_{B_{2sq}} &= \Big(\frac{\lambda_{B_{2q}} M_q - 1}{M_q - 1}\Big)\boldsymbol{I}_{sq} + \Big(\frac{1 - \lambda_{B_{2q}}}{M_q - 1}\Big)\mathbf{1}_{sq}\boldsymbol{w}_{sq}^{T}.
\end{aligned}$$

Finally, for the **Case 0**, it has simplest forms for their formulae since there is only one mismatch. The estimating function is of the form

$$\boldsymbol{H}_{ws1}^{adj}(\boldsymbol{\theta}) = \sum_{q} \boldsymbol{G}_{sq}\big\{\boldsymbol{Y}_{sq} - \boldsymbol{f}_{sq}^{E}(\boldsymbol{\theta})\big\},$$

where

$$\begin{aligned}
\boldsymbol{f}_{sq}^{E} &= \boldsymbol{X}_{sq}^{E}\boldsymbol{\beta} = (\mathbf{1}_{sq}, \boldsymbol{X}_{1sq}, \tilde{E}_{B_{2sq}}\boldsymbol{X}_{2sq}^{*})\boldsymbol{\beta} \quad \text{and} \\
\tilde{E}_{B_{2sq}} &= \Big(\frac{\lambda_{B_{2q}} M_q - 1}{M_q - 1}\Big)\boldsymbol{I}_{sq} + \Big(\frac{1 - \lambda_{B_{2q}}}{M_q - 1}\Big)\mathbf{1}_{sq}\boldsymbol{w}_{sq}^{T}.
\end{aligned}$$

Theorem 6. *Let $\hat{\boldsymbol{\theta}}_3^{s*}$ be the solution of the estimating equation (11). Then, under the assumption that we know true λ_{A_q} and $\lambda_{B_{2q}}$, an asymptotic variance estimator is of the form*

$$\boldsymbol{V}_{3|\boldsymbol{X}^*}^{ws}(\hat{\boldsymbol{\theta}}_3^{s*}) = \Big[\sum_{q} \boldsymbol{G}_{sq}\tilde{E}_{A_{sq}}\partial_{\theta}\boldsymbol{f}_{sq}^{E}(\hat{\boldsymbol{\theta}})\Big]^{-1}\Big[\sum_{q} \boldsymbol{G}_{sq}\hat{\Sigma}_{sq}\boldsymbol{G}_{sq}^{T}\Big]\Big(\Big[\sum_{q} \boldsymbol{G}_{sq}\tilde{E}_{A_{sq}}\partial_{\theta}\boldsymbol{f}_{sq}^{E}(\hat{\boldsymbol{\theta}})\Big]^{-1}\Big)^{T},$$

[8]In this article, we simply use weight $\boldsymbol{w}_{sq} = (\frac{M_q}{m_{sq}})\mathbf{1}_q$.

where

$$\hat{\Sigma}_{sq} \approx \mathrm{diag}\Big(\frac{(\lambda_{A_q}M_q-1)d_i+M_q(1-\lambda_{A_q})\bar{d}_{sq}}{M_q-1}+(1-\lambda_{A_q})\big[\lambda_{A_q}(f_i^E-\bar{f}_{sq}^E)^2+\bar{f}_{sq}^{E(2)}-(\bar{f}_{sq}^E)^2\big]; i\in s_q\Big)$$

with $D_{sq}=\mathrm{diag}\{d_i; i\in s_q\}\approx Var_{\boldsymbol{X}^*}(\boldsymbol{Y}_{sq})$ *and* $\bar{d}_{sq}$ *is the mean of* $\{d_i; i\in s_q\}$.

Let $\hat{\boldsymbol{\theta}}_2^{s*}$ *be the solution of the estimating equation for the* ***Case 1****. Then, under the assumption that we know true* $\lambda_{B_{1q}}$ *and* $\lambda_{B_{2q}}$*, an asymptotic variance estimator is of the form*

$$\boldsymbol{V}_{2|\boldsymbol{X}^*}^{ws}(\hat{\boldsymbol{\theta}}_2^{s*})=\Big[\sum_q \boldsymbol{G}_{sq}\partial_\theta \boldsymbol{f}_{sq}^{E2}(\hat{\boldsymbol{\theta}}_2^{s*})\Big]^{-1}\Big[\sum_q \boldsymbol{G}_{sq}\hat{\Sigma}_{sq}^{(2)}\boldsymbol{G}_{sq}^T\Big]\Big(\Big[\sum_q \boldsymbol{G}_{sq}\partial_\theta \boldsymbol{f}_{sq}^{E2}(\hat{\boldsymbol{\theta}}_2^{s*})\Big]^{-1}\Big)^T,$$

where

$$\hat{\Sigma}_{sq}^{(2)}=\hat{\sigma}_{sq}^2\boldsymbol{I}_{sq}+\hat{\boldsymbol{V}}_{B_{1_{sq}}}+\hat{\boldsymbol{V}}_{B_{2_{sq}}}.$$

Finally, let $\hat{\boldsymbol{\theta}}_1^{s*}$ *be the solution of the estimating equation for the* ***Case 0****. Then, an asymptotic variance estimator is of the form*

$$\boldsymbol{V}_{1|\boldsymbol{X}^*}^{ws}(\hat{\boldsymbol{\theta}}_1^{s*})=\Big[\sum_q \boldsymbol{G}_{sq}\partial_\theta \boldsymbol{f}_{sq}^{E}(\hat{\boldsymbol{\theta}}_1^{s*})\Big]^{-1}\Big[\sum_q \boldsymbol{G}_{sq}\hat{\Sigma}_{sq}^{(1)}\boldsymbol{G}_{sq}^T\Big]\Big(\Big[\sum_q \boldsymbol{G}_{sq}\partial_\theta \boldsymbol{f}_{sq}^{E}(\hat{\boldsymbol{\theta}}_1^{s*})\Big]^{-1}\Big)^T,$$

where,

$$\hat{\Sigma}_{sq}^{(1)}=\hat{\sigma}_{sq}^2\boldsymbol{I}_{sq}+\hat{\boldsymbol{V}}_{B_{2_{sq}}}.$$

Note that the above asymptotic variance estimator assumes that the λ_{A_q}, $\lambda_{B_{1q}}$ and $\lambda_{B_{2q}}$ are known. If we need to estimate these probabilities, the asymptotic variance estimator should have more terms that count the estimations of λ_{A_q}, $\lambda_{B_{1q}}$ and $\lambda_{B_{2q}}$. To see this, note that, when λ_{A_q} and $\lambda_{B_{2q}}$ are estimated, (11) becomes

$$\boldsymbol{H}_{ws3,\lambda}^{adj}(\boldsymbol{\theta},\lambda_A,\lambda_{B_2})=\sum_q \boldsymbol{G}_{sq}\big\{\boldsymbol{Y}_{sq}^*-\tilde{E}_{A_{sq}}(\lambda_A)\boldsymbol{f}_{sq}^E(\boldsymbol{\theta},\lambda_{B_2})\big\}. \tag{12}$$

In this case, the asymptotic variance estimator is of the form

$$\begin{aligned}\mathrm{Var}_{\boldsymbol{X}^*}(\hat{\boldsymbol{\theta}})\approx\big[\partial_\theta \boldsymbol{H}_{ws3,0}^{adj}\big]^{-1}\Big[&\mathrm{Var}_{\boldsymbol{X}^*}\big(\boldsymbol{H}_{ws3,0}^{adj}\big)+\big(\partial_{\lambda_A}\boldsymbol{H}_{ws3,0}^{adj}\big)\mathrm{Var}_{\boldsymbol{X}^*}(\lambda_A)\big(\partial_{\lambda_A}\boldsymbol{H}_{ws3,0}^{adj}\big)^T\\&+\big(\partial_{\lambda_{B_2}}\boldsymbol{H}_{ws3,0}^{adj}\big)\mathrm{Var}_{\boldsymbol{X}^*}(\lambda_{B_2})\big(\partial_{\lambda_{B_2}}\boldsymbol{H}_{ws3,0}^{adj}\big)^T\Big]\Big\{\big[\partial_\theta \boldsymbol{H}_{ws3,0}^{adj}\big]^{-1}\Big\}^T,\end{aligned} \tag{13}$$

where

$$\begin{aligned}\boldsymbol{H}_{ws3,0}^{adj}&=\boldsymbol{H}_{ws3,\lambda}^{adj}(\boldsymbol{\theta}_0,\lambda_A^0,\lambda_{B_2}^0),\\ \partial_{\lambda_A}\boldsymbol{H}_{ws3,0}^{adj}&=-\sum_q \boldsymbol{G}_{sq}\Big[(M_q-1)^{-1}(M_q\boldsymbol{I}_{sq}-\boldsymbol{1}_{sq}\boldsymbol{w}_{sq}^T)\Big]\boldsymbol{f}_{sq}^E(\boldsymbol{\theta}_0,\lambda_{B_{2q}}^0),\\ \partial_{\lambda_{B_2}}\boldsymbol{H}_{ws3,0}^{adj}&=-\sum_q \boldsymbol{G}_{sq}\tilde{E}_{A_{sq}}\Big[(M_q-1)^{-1}(M_q\boldsymbol{I}_{sq}-\boldsymbol{1}_{sq}\boldsymbol{w}_{sq}^T)\Big]\boldsymbol{X}_{2q}^*\beta_2.\end{aligned} \tag{14}$$

Corollary 7. *Let $\hat{\boldsymbol{\theta}}_3^{s**}$ be the solution of the estimating equation (12). Then, an asymptotic variance estimator is of the form*

$$\boldsymbol{V}_{3|\boldsymbol{X}^*}^{ws,\lambda}(\hat{\boldsymbol{\theta}}_3^{s**}) = \big[\partial_\theta \hat{\boldsymbol{H}}_{ws3,0}^{adj}\big]^{-1}\Big[\boldsymbol{V}_{3|\boldsymbol{X}^*}^{ws}(\hat{\boldsymbol{\theta}}_3^{s**}) + \big(\partial_{\lambda_1}\hat{\boldsymbol{H}}_{ws3,0}^{adj}\big)\, Var_{\boldsymbol{X}^*}(\hat{\lambda}_A)\big(\partial_{\lambda_1}\hat{\boldsymbol{H}}_{ws3,0}^{adj}\big)^T + \big(\partial_{\lambda_2}\hat{\boldsymbol{H}}_{ws3,0}^{adj}\big)\, Var_{\boldsymbol{X}^*}(\hat{\lambda}_{B_2})\big(\partial_{\lambda_2}\hat{\boldsymbol{H}}_{ws3,0}^{adj}\big)^T\Big]\Big\{\big[\partial_\theta \hat{\boldsymbol{H}}_{ws3,0}^{adj}\big]^{-1}\Big\}^T.$$

*Also, Let $\hat{\boldsymbol{\theta}}_2^{s**}$ be the solution of the estimating equation for the **Case 1**. Then, an asymptotic variance estimator is of the form*

$$\boldsymbol{V}_{2|\boldsymbol{X}^*}^{ws,\lambda}(\hat{\boldsymbol{\theta}}_2^{s**}) = \big[\partial_\theta \hat{\boldsymbol{H}}_{ws2,0}^{adj}\big]^{-1}\Big[\boldsymbol{V}_{2|\boldsymbol{X}^*}^{ws}(\hat{\boldsymbol{\theta}}_2^{s**}) + \big(\partial_{\lambda_{B_1}}\hat{\boldsymbol{H}}_{ws2,0}^{adj}\big)\, Var_{\boldsymbol{X}^*}(\hat{\lambda}_{B_1})\big(\partial_{\lambda_{B_1}}\hat{\boldsymbol{H}}_{ws2,0}^{adj}\big)^T + \big(\partial_{\lambda_{B_2}}\hat{\boldsymbol{H}}_{ws2,0}^{adj}\big)\, Var_{\boldsymbol{X}^*}(\hat{\lambda}_{B_2})\big(\partial_{\lambda_{B_2}}\hat{\boldsymbol{H}}_{ws2,0}^{adj}\big)^T\Big]\Big\{\big[\partial_\theta \hat{\boldsymbol{H}}_{ws2,0}^{adj}\big]^{-1}\Big\}^T,$$

where

$$\begin{aligned}
\boldsymbol{H}_{ws2,0}^{adj} &= \boldsymbol{H}_{ws2}^{adj}(\boldsymbol{\theta}_0, \lambda_{B_1}, \lambda_{B_2}),\\
Var_{\boldsymbol{X}^*}(\lambda_{B_{1q}}) &= (m_q^{B_1})^{-1}\lambda_{B_{1q}}(1-\lambda_{B_{1q}}),\\
\partial_{\lambda_{B_1}}\boldsymbol{H}_{ws2,0}^{adj} &= -\sum_q \boldsymbol{G}_{sq}\Big[(M_q-1)^{-1}(M_q\boldsymbol{I}_{sq} - \mathbf{1}_{sq}\boldsymbol{w}_{sq}^T)\Big]\boldsymbol{X}_{1q}^*\beta_1.
\end{aligned}$$

*Further, let $\hat{\boldsymbol{\theta}}_1^{s**}$ be the solution of the estimating equation for the **Case 0**. Then, an asymptotic variance estimator is of the form*

$$\boldsymbol{V}_{1|\boldsymbol{X}^*}^{ws,\lambda}(\hat{\boldsymbol{\theta}}_1^{s**}) = \big[\partial_\theta \hat{\boldsymbol{H}}_{ws1,0}^{adj}\big]^{-1}\Big[\boldsymbol{V}_{1|\boldsymbol{X}^*}^{ws}(\hat{\boldsymbol{\theta}}_1^{s**}) + \big(\partial_{\lambda_{B_2}}\hat{\boldsymbol{H}}_{ws1,0}^{adj}\big)\, Var_{\boldsymbol{X}^*}(\hat{\lambda}_{B_2})\big(\partial_{\lambda_{B_2}}\hat{\boldsymbol{H}}_{ws1,0}^{adj}\big)^T\Big]\Big\{\big[\partial_\theta \hat{\boldsymbol{H}}_{ws1,0}^{adj}\big]^{-1}\Big\}^T.$$

3.2 Sample-register case: When sample records are not perfectly linked

When some records are not linked, A_q or B_{2q} cannot be a permutation matrix, because the entries of some rows are all zero due to non-linkage. However, we can still use similar ideas introduced in the previous subsection.

Firstly, we consider the **Case 2**. Let $\boldsymbol{X}_{1sq}$ be the set of the sample records from $\boldsymbol{X}_{1q}$. Also let $\boldsymbol{X}_{1slq}$ be the set of sample records in $\boldsymbol{X}_{1sq}$ that are linked both to $\boldsymbol{X}_2$-register and to $\boldsymbol{Y}$-register. Further, let $\boldsymbol{X}_{1suq} := \boldsymbol{X}_{1sq} - \boldsymbol{X}_{1slq}$. Then it represents the set of sampled records in $\boldsymbol{X}_{1sq}$ that cannot be linked either to $\boldsymbol{X}_2$-register or to $\boldsymbol{Y}$-register. Also, let $\boldsymbol{X}_{1rq} := \boldsymbol{X}_{1q} - \boldsymbol{X}_{1sq}$, the set of non-sample records in $\boldsymbol{X}_{1q}$. We assume that, theoretically,

there exists $\boldsymbol{X}_{1rlq}$ that represents the set of non-sample records that can be linked both to $\boldsymbol{X}_2$-register and $\boldsymbol{Y}$-register. Similarly, $\boldsymbol{X}_{1ruq} := \boldsymbol{X}_{1rq} - \boldsymbol{X}_{1rlq}$.

Similarly, under the one to one linkage assumption, $\boldsymbol{Y}^*_q$ can be partitioned into four groups, namely $\boldsymbol{Y}^*_{slq}$, $\boldsymbol{Y}^*_{suq}$, $\boldsymbol{Y}^*_{rlq}$ and $\boldsymbol{Y}^*_{ruq}$. Thus, one has

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

and

$$E(A_q|\boldsymbol{X}^*_q) = E_{A_q} = \begin{pmatrix} E_{slsl,A_q} & E_{slsu,A_q} & E_{slrl,A_q} & E_{slru,A_q} \\ E_{susl,A_q} & E_{susu,A_q} & E_{surl,A_q} & E_{suru,A_q} \\ E_{rlsl,A_q} & E_{rlsu,A_q} & E_{rlrl,A_q} & E_{rlru,A_q} \\ E_{rusl,A_q} & E_{rusu,A_q} & E_{rurl,A_q} & E_{ruru,A_q} \end{pmatrix}.$$

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

$$E(B_{2q}|\boldsymbol{X}^*_q) = E_{B_{2q}} = \begin{pmatrix} E_{slsl,B_{2q}} & E_{slsu,B_{2q}} & E_{slrl,B_{2q}} & E_{slru,B_{2q}} \\ E_{susl,B_{2q}} & E_{susu,B_{2q}} & E_{surl,B_{2q}} & E_{suru,B_{2q}} \\ E_{rlsl,B_{2q}} & E_{rlsu,B_{2q}} & E_{rlrl,B_{2q}} & E_{rlru,B_{2q}} \\ E_{rusl,B_{2q}} & E_{rusu,B_{2q}} & E_{rurl,B_{2q}} & E_{ruru,B_{2q}} \end{pmatrix}.$$

This leads to the estimating equation of the form

$$\begin{aligned} \boldsymbol{H}^{adj}_{sl}(\boldsymbol{\theta}) &= \sum_q \boldsymbol{G}_{slq}\big\{\boldsymbol{Y}^*_{slq} - E_{A_{slq}}\boldsymbol{f}^E_q(\boldsymbol{\theta})\big\} \\ &= \sum_q \boldsymbol{G}_{slq}\big\{\boldsymbol{Y}^*_{slq} - E_{slsl,A_q}\boldsymbol{f}^E_{slq}(\boldsymbol{\theta}) - E_{slsu,A_q}\boldsymbol{f}^E_{suq}(\boldsymbol{\theta}) \\ &\qquad - E_{slrl,A_q}\boldsymbol{f}^E_{rlq}(\boldsymbol{\theta}) - E_{slru,A_q}\boldsymbol{f}^E_{ruq}(\boldsymbol{\theta})\big\}. \end{aligned} \tag{15}$$

Under the exchangeable linkage error model, one has

$$\begin{aligned} E_{slsl,A_q} &= \Big[\frac{\lambda_{A_q}M_q - 1}{M_q - 1}\Big]\boldsymbol{I}_{slq} + \Big[\frac{1-\lambda_{A_q}}{M_q-1}\Big]\mathbf{1}_{slq}\mathbf{1}^T_{slq}, \\ E_{slsu,A_q} &= \Big[\frac{1-\lambda_{A_q}}{M_q-1}\Big]\mathbf{1}_{slq}\mathbf{1}^T_{suq}, \\ E_{slrl,A_q} &= \Big[\frac{1-\lambda_{A_q}}{M_q-1}\Big]\mathbf{1}_{slq}\mathbf{1}^T_{rlq}, \\ E_{slru,A_q} &= \Big[\frac{1-\lambda_{A_q}}{M_q-1}\Big]\mathbf{1}_{slq}\mathbf{1}^T_{ruq}. \end{aligned}$$

It leads (15) to the form of

$$\boldsymbol{H}^{adj}_{sl}(\boldsymbol{\theta}) = \sum_q \boldsymbol{G}_{slq}\big\{\boldsymbol{Y}^*_{slq} - \Big[\frac{\lambda_{A_q}M_q - 1}{M_q - 1}\Big]\boldsymbol{I}_{slq}\boldsymbol{f}^E_{slq}(\boldsymbol{\theta}) - \Big[\frac{1-\lambda_{A_q}}{M_q-1}\Big]\mathbf{1}_{slq}\mathbf{1}^T_q\boldsymbol{f}^E_q\big\}.$$

If we assume that the distribution of $\boldsymbol{Y}^*_{slq}$ is the same as that of $\boldsymbol{Y}$ in the population, the observable population value $\mathbf{1}_q^T \boldsymbol{f}_q^E(\theta)$ can be replaced by weighted sample estimate by $\boldsymbol{w}_{slq}^T \boldsymbol{f}_{slq}^E(\theta)$[9] so that one has

$$\boldsymbol{H}_{wsl}^{adj}(\boldsymbol{\theta}) = \sum_q \boldsymbol{G}_{slq}\big\{\boldsymbol{Y}^*_{slq} - \tilde{E}_{A_{slq}} \boldsymbol{f}_{slq}^E(\boldsymbol{\theta})\big\},$$

where

$$\tilde{E}_{A_{slq}} = \Big[\frac{\lambda_{A_q} M_q - 1}{M_q - 1}\Big]\boldsymbol{I}_{slq} + \Big[\frac{1 - \lambda_{A_q}}{M_q - 1}\Big]\mathbf{1}_{slq}\boldsymbol{w}_{slq}^T.$$

For $\boldsymbol{f}_{slq}^E(\theta)$, note that by (2)

$$\boldsymbol{f}_{slq}^E(\theta) = (\mathbf{1}_{slq}, \boldsymbol{X}_{1slq}, E_{B_{sl,2q}}\boldsymbol{X}^*_{2q})(\beta_0, \beta 1, \beta_2)^T,$$

where

$$E_{B_{sl,2q}}\boldsymbol{X}^*_{2q} = E_{slsl,B_{2q}}\boldsymbol{X}^*_{2slq} + E_{slsu,B_{2q}}\boldsymbol{X}^*_{2suq} + E_{slrl,B_{2q}}\boldsymbol{X}^*_{2rlq} + E_{slru,B_{2q}}\boldsymbol{X}^*_{2ruq}.$$

The exchangeable linkage error model provides that

$$\begin{aligned}
E_{slsl,B_{2q}} &= \Big[\frac{\lambda_{B_{2q}} M_q - 1}{M_q - 1}\Big]\boldsymbol{I}_{slq} + \Big[\frac{1 - \lambda_{B_{2q}}}{M_q - 1}\Big]\mathbf{1}_{slq}\mathbf{1}_{slq}^T,\\
E_{slsu,B_{2q}} &= \Big[\frac{1 - \lambda_{B_{2q}}}{M_q - 1}\Big]\mathbf{1}_{slq}\mathbf{1}_{suq}^T,\\
E_{slrl,B_{2q}} &= \Big[\frac{1 - \lambda_{B_{2q}}}{M_q - 1}\Big]\mathbf{1}_{slq}\mathbf{1}_{rlq}^T,\\
E_{slru,B_{2q}} &= \Big[\frac{1 - \lambda_{B_{2q}}}{M_q - 1}\Big]\mathbf{1}_{slq}\mathbf{1}_{ruq}^T.
\end{aligned}$$

If we also assume that the distribution of $\boldsymbol{X}^*_{2slq}$ is the same as that of $\boldsymbol{X}^*_{2q}$ in the population, then $E_{B_{sl,2q}}\boldsymbol{X}^*_{2q}$ can be replaced by $\tilde{E}_{B_{sl,2q}}\boldsymbol{X}^*_{2slq}$ where

$$\tilde{E}_{B_{sl,2q}} = \Big[\frac{\lambda_{B_{2q}} M_q - 1}{M_q - 1}\Big]\boldsymbol{I}_{slq} + \Big[\frac{1 - \lambda_{B_{2q}}}{M_q - 1}\Big]\mathbf{1}_{slq}\boldsymbol{w}_{slq}^T.$$

Then, $\boldsymbol{f}_{slq}^E(\theta)$ can be evaluated by

$$\boldsymbol{f}_{slq}^E(\theta) = (\mathbf{1}_{slq}, \boldsymbol{X}_{1slq}, \tilde{E}_{B_{sl,2q}}\boldsymbol{X}^*_{2slq})(\beta_0, \beta 1, \beta_2)^T.$$

[9]We will use $\boldsymbol{w}_{slq} = (\frac{M_q}{m_{slq}})\mathbf{1}_{slq}$, where m_{slq} is the number of linked sample records, while M_q is the total population number in $q^{th}m$-block.

Suppose that we know λ_{A_q} and $\lambda_{B_{2q}}$, and let $\hat{\boldsymbol{\theta}}$ be the solution of the estimating equation. To derive the asymptotic variance estimator for $\hat{\theta}$, note that (47) becomes now

$$\mathrm{Var}_{X^*}(\hat{\boldsymbol{\theta}}) \approx \big[\partial_\theta \boldsymbol{H}^{adj}_{wsl}(\boldsymbol{\theta}_0)\big]^{-1}\mathrm{Var}_{X^*}\big[\boldsymbol{H}^{adj}_{wsl}(\boldsymbol{\theta}_0)\big]\Big(\big[\partial_\theta \boldsymbol{H}^{adj}_{wsl}(\boldsymbol{\theta}_0)\big]^{-1}\Big)^T$$

with corresponding estimator of the form

$$\begin{aligned} V_{X^*}(\hat{\boldsymbol{\theta}}) &= \big[\partial_\theta \boldsymbol{H}^{adj}_{wsl}(\boldsymbol{\theta}_0)\big]^{-1} V_{X^*}\big[\boldsymbol{H}^{adj}_{wsl}(\boldsymbol{\theta}_0)\big]\Big(\big[\partial_\theta \boldsymbol{H}^{adj}_{wsl}(\boldsymbol{\theta}_0)\big]^{-1}\Big)^T \\ &\approx \Big[\sum_q \boldsymbol{G}_{slq}\tilde{E}_{A_{slq}}\partial_\theta \boldsymbol{f}^E_{slq}(\hat{\boldsymbol{\theta}})\Big]^{-1}\Big[\sum_q \boldsymbol{G}_{slq}\hat{\Sigma}_{slq}\boldsymbol{G}^T_{slq}\Big]\Big(\Big[\sum_q \boldsymbol{G}_{slq}\tilde{E}_{A_{slq}}\partial_\theta \boldsymbol{f}^E_{slq}(\hat{\boldsymbol{\theta}})\Big]^{-1}\Big)^T, \end{aligned}$$

under the assumption that $\boldsymbol{G}_{slq}$ is independent of $\boldsymbol{\theta}$. By the similar arguments in (48)-(49),

$$\begin{aligned} \Sigma_{slq} &= \mathrm{Var}_{X^*}(\boldsymbol{Y}^*_{slq}) \\ &\approx \mathrm{Var}_{X^*}(A_{slsl,q}\boldsymbol{Y}_{slq}) + \mathrm{Var}_{X^*}(A_{slsu,q}\boldsymbol{Y}_{suq}) + \mathrm{Var}_{X^*}(A_{slrl,q}\boldsymbol{Y}_{rlq}) + \mathrm{Var}_{X^*}(A_{slru,q}\boldsymbol{Y}_{ruq}) \end{aligned}$$

that can be approximated by

$$\begin{aligned} \hat{\Sigma}_{slq} \approx \mathrm{diag}\Big(&\frac{(\lambda_{A_q}M_q - 1)d_i + M_q(1-\lambda_{A_q})\bar{d}_{slq}}{M_q - 1} \\ &+ (1-\lambda_{A_q})\big[\lambda_{A_q}(f^E_i - \bar{f}^E_{slq})^2 + \bar{f}^{E(2)}_{slq} - (\bar{f}^E_{slq})^2\big]; i \in \{1,\ldots,m_{slq}\}\Big). \end{aligned}$$

If we need to estimate λ_{A_q} and $\lambda_{B_{2q}}$, we still can use the asymptotic variance estimator defined by (13)-(14), except that the subscripts sp and ws need to be replaced by slp and wsl. That is, the asymptotic variance estimator is of the form

$$\begin{aligned} \mathrm{Var}_{\boldsymbol{X}^*}(\hat{\boldsymbol{\theta}}) \approx \big[\partial_\theta \boldsymbol{H}^{adj}_{wsl,0}\big]^{-1}\Big[&\mathrm{Var}_{\boldsymbol{X}^*}\big(\boldsymbol{H}^{adj}_{wsl,0}\big) + \big(\partial_{\lambda_1}\boldsymbol{H}^{adj}_{wsl,0}\big)\mathrm{Var}_{\boldsymbol{X}^*}(\hat{\lambda}_A)\big(\partial_{\lambda_1}\boldsymbol{H}^{adj}_{wsl,0}\big)^T \\ &+ \big(\partial_{\lambda_2}\boldsymbol{H}^{adj}_{wsl,0}\big)\mathrm{Var}_{\boldsymbol{X}^*}(\hat{\lambda}_{B_2})\big(\partial_{\lambda_2}\boldsymbol{H}^{adj}_{wsl,0}\big)^T\Big]\Big\{\big[\partial_\theta \boldsymbol{H}^{adj}_{wsl,0}\big]^{-1}\Big\}^T. \end{aligned}$$

Using the above arguments, it is clear that, to deal with **Case 0** and **Case 1** in this case, we can use the formulae in the previous subsection by replacing sp and ws with slp and wsl.

3.3 Simulation

We use simulation to compare the performances of different estimators we considered in this study for the sample to register linkage case. The linear model we used in this simulation is the same as before,

$$Y = 1 + 5X_1 + 8X_2 + \epsilon.$$

Most of assumptions and scenarios we made for the register to register case are the same except that we use the sample here instead of whole population. In this simulation, we considered the case of complete linkage and incomplete linkage separately. For the case of complete linkage, we assume that the sample records s from the bench mark data sets are linked to the records in other registers. The extra assumption we made in this simulation is that the population size of all registers the same and each m-block has 2000 records, and 500 samples are chosen randomly for each m-block. Further, in case of incomplete linkage, we assume that, among 2000 records, half of them cannot be linked. In this incomplete linkage case, we chose 1000 samples. The reason is that because half of them cannot be linked, we might have around 500 samples that are linked to other registers. This assumption will provide another consistent comparisons of the same estimators between the complete linkage case and incomplete linkage case. The results for the complete linkage case can be found in the Table 4–Table 6, while the results for the incomplete case are in Table 7–Table 9

The result shows very similar pattern in the register to register case. Clearly, while the ratio-type estimator, the Lahiri-Larsen estimator and the EBLUE correct the bias due to linkage errors, the EBLUE outperforms all other estimators. Here are the results for the complete linkage case:

[Table 4 here.]

[Table 5 here.]

[Table 6 here.]

Here are the results for the incomplete linkage case:

[Table 7 here.]

[Table 8 here.]

[Table 9 here.]

The results for the sample-register cases are very similar to the register-register cases as long as the sample sizes are similar. One thing to note is that the coverage rates are all higher than 95%. This is not the case when the number of merged data sets are two. One possible explanation is that the variance terms in these cases are more complicated and, as the number of merged data sets increase, the variances increase as well so that the confidence intervals are becoming wider.

4 Conclusion and further research direction

In this paper we extend the linkage error adjusting technique in regression analysis developed in Chambers (2008) to accommodated the situation where the number of merged data sets are more than two. We developed a ration-type estimator for the regression analysis and then it has been extended to more general adjusted estimating function approach. These methods can deal with the case where all the data sets are registers, as well as the case where the bench mark data sets are sample and the others are registers. Even though it hasn't been dealt here, it is easy to see that these methods can naturally accommodate the case where all the data sets are sample. These methods also extended to deal with the situation where some of sample data are failed to be linked to other registers. However, all of these bias correction methods have to pay the price of large variance. Furthermore, in the case of sample-registers case with non-linkage situation, the number of linked sample data, if the the number of merged data sets are increasing, will be decreasing. Thus, we expect some sort of loss of information by merging more data sets. We expect to overcome this limitation by adapting other approaches.

Another limitation of these methods is that we assume that the linkage errors among the data sets occurs randomly. However, there might be some correlation among the linkage errors. To deal with this situation, our model should include more complicated covariance measures in the formulae and it will be dealt in our next research paper.

A Proofs of the Propositions and Theorems

A.1 Proof of Proposition 1

For the variance of the estimator, note that

$$\mathrm{Var}_{\boldsymbol{X}^*}(\hat{\boldsymbol{\beta}}_R) = \boldsymbol{D}_1^{-1}\mathrm{Var}_{\boldsymbol{X}^*}(\hat{\boldsymbol{\beta}}^*)\big(\boldsymbol{D}_1^{-1}\big)^T,$$

where

$$\mathrm{Var}_{\boldsymbol{X}^*}(\hat{\boldsymbol{\beta}}^*) = \Big[\sum_q (\boldsymbol{X}_q^*)^T\boldsymbol{X}_q^*\Big]^{-1}\Big[\sum_q (\boldsymbol{X}_q^*)^T\mathrm{Var}_{\boldsymbol{X}^*}(\boldsymbol{Y}_q)\boldsymbol{X}_q^*\Big]\Big[\sum_q (\boldsymbol{X}_q^*)^T\boldsymbol{X}_q^*\Big]^{-1}.$$

Further, one has

$$\mathrm{Var}_{\boldsymbol{X}^*}(\boldsymbol{Y}_q) = E_{\boldsymbol{X}^*}\Big[\mathrm{Var}_{\boldsymbol{X}^*}(\boldsymbol{Y}_q|B_{2_q})\Big] + \mathrm{Var}_{\boldsymbol{X}^*}\Big[E_{\boldsymbol{X}^*}(\boldsymbol{Y}_q|B_{2_q})\Big].$$

Note that, by (1),

$$E_{\boldsymbol{X}^*}(\boldsymbol{Y}_q|B_{2_q}) = \boldsymbol{X}_q^{B_2}\boldsymbol{\beta}.$$

Thus,

$$\begin{aligned} V_{B_{2q}} = \mathrm{Var}_{\boldsymbol{X}^*}\Big[E_{\boldsymbol{X}^*}(\boldsymbol{Y}_q|B_{2_q})\Big] &= E_{\boldsymbol{X}^*}\Big[\boldsymbol{X}_q^{B_2}\boldsymbol{\beta} - \boldsymbol{X}_q^{E}\boldsymbol{\beta}\Big]^2 \\ &= E_{\boldsymbol{X}^*}\Big[B_{2_q}^T\boldsymbol{X}_{2q}^*\beta_2 - E_{B_{2q}}\boldsymbol{X}_{2q}^*\beta_2\Big]^2. \end{aligned}$$

Denote that

$$\boldsymbol{f}_{B_{2_q}}^* = \boldsymbol{X}_{2q}^*\beta_2.$$

Then, by (16) from Chambers (2008),

$$\boldsymbol{V}_{B_{2q}} = \mathrm{diag}\left[(1-\lambda_{B_{2_q}})\big\{\lambda_{B_{2_q}}(f_{B_{2_q},i}^* - \bar{f}_{B_{2_q}}^*)^2 + \bar{f}_{B_{2_q}}^{*(2)} - (\bar{f}_{B_{2_q}}^*)^2\big\}\right], \tag{16}$$

where $\boldsymbol{f}_{B_{2_q}}^* = (f_{B_{2_q},i}^*)$ and $\bar{f}_{B_{2_q}}^*$, $\bar{f}_{B_{2_q}}^{*(2)}$ are the averages of $f_{B_{2_q},i}^*$ and their squares respectively in $\boldsymbol{f}_{B_{2_q}}^*$. Furthermore, one has

$$\mathrm{Var}_{\boldsymbol{X}^*}(\boldsymbol{Y}_q|B_{2_q}) = E_{\boldsymbol{X}^*}(\boldsymbol{Y}_q - \boldsymbol{X}_q^{B_2}\boldsymbol{\beta})^2 = E_{\boldsymbol{X}^*}(\epsilon_q)^2 = \sigma_q^2\boldsymbol{I}_q.$$

Therefore, one has

$$\mathrm{Var}_{\boldsymbol{X}^*}(\boldsymbol{Y}_q) = \sigma_q^2\boldsymbol{I}_q + \boldsymbol{V}_{B_{2q}} \tag{17}$$

which implies that

$$\mathrm{Var}_{\boldsymbol{X}^*}(\hat{\boldsymbol{\beta}}_R) = \Big[\sum_q (\boldsymbol{X}_q^*)^T\boldsymbol{X}_q^E\Big]^{-1}\Big[\sum_q (\boldsymbol{X}_q^*)^T\Big(\sigma_q^2\boldsymbol{I}_q + \boldsymbol{V}_{B_{2q}}\Big)\boldsymbol{X}_q^*\Big]\Big[\sum_q (\boldsymbol{X}_q^*)^T\boldsymbol{X}_q^E\Big]^{-1}. \tag{18}$$

To evaluate $\mathrm{Var}_{\boldsymbol{X}^*}(\hat{\boldsymbol{\beta}}_R)$, Then, one has

$$\begin{aligned} (\boldsymbol{Y}_q - \boldsymbol{f}_q^E)^T(\boldsymbol{Y}_q - \boldsymbol{f}_q^E) &= \Big[(\boldsymbol{Y}_q - \boldsymbol{f}_q) - (\boldsymbol{f}_q^E - \boldsymbol{f}_q)\Big]^T\Big[(\boldsymbol{Y}_q - \boldsymbol{f}_q) - (\boldsymbol{f}_q^E - \boldsymbol{f}_q)\Big] \\ &= (\boldsymbol{Y}_q - \boldsymbol{f}_q)^T(\boldsymbol{Y}_q - \boldsymbol{f}_q) && (19) \\ &\quad - (\boldsymbol{Y}_q - \boldsymbol{f}_q)^T(\boldsymbol{f}_q^E - \boldsymbol{f}_q) - (\boldsymbol{f}_q^E - \boldsymbol{f}_q)^T(\boldsymbol{Y}_q - \boldsymbol{f}_q) && (20) \\ &\quad + (\boldsymbol{f}_q^E - \boldsymbol{f}_q)^T(\boldsymbol{f}_q^E - \boldsymbol{f}_q). && (21) \end{aligned}$$

Note that, by the definition,

$$\sum_q (\boldsymbol{Y}_q - \boldsymbol{f}_q)^T(\boldsymbol{Y}_q - \boldsymbol{f}_q) = N\hat{\sigma}^2. \tag{22}$$

Further,

$$E_{\boldsymbol{X}^*}\Big[(\boldsymbol{Y}_q - \boldsymbol{f}_q)^T(\boldsymbol{f}_q^E - \boldsymbol{f}_q) + (\boldsymbol{f}_q^E - \boldsymbol{f}_q)^T(\boldsymbol{Y}_q - \boldsymbol{f}_q)\Big] = 0 \tag{23}$$

because $\boldsymbol{Y}_q - \boldsymbol{f}_q = \epsilon_q$ and $\mathrm{cov}(\epsilon_q, \boldsymbol{X}_q) = 0$. Moreover, one has

$$\begin{aligned} E_{\boldsymbol{X}^*}\Big[(\boldsymbol{f}_q^E - \boldsymbol{f}_q)^T(\boldsymbol{f}_q^E - \boldsymbol{f}_q)\Big] &= E_{\boldsymbol{X}^*}\Big[(\boldsymbol{f}_q^E)^T(\boldsymbol{f}_q^E - \boldsymbol{Y}_q) + (\boldsymbol{f}_q^E)^T(\boldsymbol{Y}_q - \boldsymbol{f}_q) \\ &\quad + (\boldsymbol{f}_q)^T(\boldsymbol{f}_q - \boldsymbol{Y}_q) + (\boldsymbol{f}_q)^T(\boldsymbol{Y}_q - \boldsymbol{f}_q^E)\Big] \\ &= 0 \end{aligned} \tag{24}$$

because $E_{\boldsymbol{X}^*}\big(\boldsymbol{Y}_q - \boldsymbol{f}_q^E\big) = 0$. Thus, by (19)-(24),

$$\hat{\sigma}^2 = N^{-1}\sum_q (\boldsymbol{Y}_q - \boldsymbol{f}_q^E)^T(\boldsymbol{Y}_q - \boldsymbol{f}_q^E). \tag{25}$$

Consequently, $\mathrm{Var}_{\boldsymbol{X}^*}(\hat{\boldsymbol{\beta}}_R)$ can be evaluated by using (25), (16) and (18).

A.2 Proof of Proposition 2

For the variance of the estimator, one has

$$\mathrm{Var}_{\boldsymbol{X}^*}(\hat{\boldsymbol{\beta}}_R) = \boldsymbol{D}_2^{-1}\mathrm{Var}_{\boldsymbol{X}^*}(\hat{\boldsymbol{\beta}}^*)\big(\boldsymbol{D}_2^{-1}\big)^T,$$

where

$$\mathrm{Var}_{\boldsymbol{X}^*}(\hat{\boldsymbol{\beta}}^*) = \Big[\sum_q (\boldsymbol{X}_q^*)^T\boldsymbol{X}_q^*\Big]^{-1}\Big[\sum_q (\boldsymbol{X}_q^*)^T\mathrm{Var}_{\boldsymbol{X}^*}(\boldsymbol{Y}_q)\boldsymbol{X}_q^*\Big]\Big[\sum_q (\boldsymbol{X}_q^*)^T\boldsymbol{X}_q^*\Big]^{-1}.$$

Note that one has

$$\mathrm{Var}_{\boldsymbol{X}^*}(\boldsymbol{Y}_q) = E_{\boldsymbol{X}^*}\Big[\mathrm{Var}_{\boldsymbol{X}^*}(\boldsymbol{Y}_q|B_{1_q}, B_{2_q})\Big] + \mathrm{Var}_{\boldsymbol{X}^*}\Big[E_{\boldsymbol{X}^*}(\boldsymbol{Y}_q|B_{1_q}, B_{2_q})\Big].$$

Then, by the assumption that the mismatches found in $\boldsymbol{X}_{1q}^*$ are not correlated with the mismatches found in $\boldsymbol{X}_{2q}^*$,

$$\begin{aligned} V_{B_q} &= \mathrm{Var}_{\boldsymbol{X}^*}\Big[E_{\boldsymbol{X}^*}(\boldsymbol{Y}_q|B_{1_q}, B_{2_q})\Big] = E_{\boldsymbol{X}^*}\Big[\boldsymbol{X}_q^{B_2}\boldsymbol{\beta} - \boldsymbol{X}_q^E\boldsymbol{\beta}\Big]^2 \\ &= E_{\boldsymbol{X}^*}\Big[(B_{1_q}^T\boldsymbol{X}_{1q}^*\beta_1 - E_{B_{1q}}\boldsymbol{X}_{1q}^*\beta_1) + (B_{2_q}^T\boldsymbol{X}_{2q}^*\beta_2 - E_{B_{2q}}\boldsymbol{X}_{2q}^*\beta_2)\Big]^2 \\ &= E_{\boldsymbol{X}^*}\Big[B_{1_q}^T\boldsymbol{X}_{1q}^*\beta_1 - E_{B_{1q}}\boldsymbol{X}_{1q}^*\beta_1\Big]^2 + E_{\boldsymbol{X}^*}\Big[B_{2_q}^T\boldsymbol{X}_{2q}^*\beta_2 - E_{B_{2q}}\boldsymbol{X}_{2q}^*\beta_2\Big]^2 \\ &= V_{B_{1q}} + V_{B_{2q}}, \end{aligned} \tag{26}$$

where $V_{B_{2q}}$ is defined in (16) and $V_{B_{1q}}$ also can be defined similarly. Then, by the similar arguments to (17)-(25), one has

$$\mathrm{Var}_{\boldsymbol{X}^*}(\hat{\boldsymbol{\beta}}_R) = \Big[\sum_q (\boldsymbol{X}_q^*)^T \boldsymbol{X}_q^E\Big]^{-1}\Big[\sum_q (\boldsymbol{X}_q^*)^T\Big(\sigma_q^2 \boldsymbol{I}_q + \boldsymbol{V}_{B_q}\Big)\boldsymbol{X}_q^*\Big]\Big[\sum_q (\boldsymbol{X}_q^*)^T \boldsymbol{X}_q^E\Big]^{-1},$$

where $\hat{\sigma}^2$ can be evaluated by (25).

A.3 Proof of Proposition 3

To derive the variance of $\hat{\boldsymbol{\beta}}_R$, note that

$$\mathrm{Var}_{\boldsymbol{X}^*}(\hat{\boldsymbol{\beta}}_R) = \boldsymbol{D}_3^{-1}\mathrm{Var}_{\boldsymbol{X}^*}(\hat{\boldsymbol{\beta}}^*)\big(\boldsymbol{D}_3^{-1}\big)^T,$$

where

$$\mathrm{Var}_{\boldsymbol{X}^*}(\hat{\boldsymbol{\beta}}^*) = \Big[\sum_q (\boldsymbol{X}_q^*)^T \boldsymbol{X}_q^*\Big]^{-1}\Big[\sum_q (\boldsymbol{X}_q^*)^T \mathrm{Var}_{\boldsymbol{X}^*}(\boldsymbol{Y}_q^*)\boldsymbol{X}_q^*\Big]\Big[\sum_q (\boldsymbol{X}_q^*)^T \boldsymbol{X}_q^*\Big]^{-1}.$$

Hence, we need to calculate $\mathrm{Var}_{\boldsymbol{X}^*}(\boldsymbol{Y}_q^*)$ first in order to derive the variance of $\hat{\boldsymbol{\beta}}_R$. Note that

$$\mathrm{Var}_{\boldsymbol{X}^*}(\boldsymbol{Y}_q^*) \neq \mathrm{Var}_{\boldsymbol{X}}(\boldsymbol{Y}_q).$$

To see this, one has

$$\mathrm{Var}_{\boldsymbol{X}^*}(\boldsymbol{Y}_q^*) = E_{\boldsymbol{X}^*}\Big[\mathrm{Var}_{\boldsymbol{X}^*}(\boldsymbol{Y}_q^*|A_q)\Big] + \mathrm{Var}_{\boldsymbol{X}^*}\Big[E_{\boldsymbol{X}^*}(\boldsymbol{Y}_q^*|A_q)\Big]. \tag{27}$$

Then, by (2) and (3)

$$E_{\boldsymbol{X}^*}(\boldsymbol{Y}_q^*|A_q) = A_q E_{\boldsymbol{X}^*}(\boldsymbol{Y}_q) = A_q \boldsymbol{X}_q^E \boldsymbol{\beta} = A_q \boldsymbol{f}_q^E.$$

Note that $\boldsymbol{f}_q$ is not observable, because it is the expectation of $\boldsymbol{Y}_q$, that is also not observable, under completely correct linkage. $\boldsymbol{f}_q^*$ is observable ,but it contains incorrect linkage between $\boldsymbol{X}_{1q}$ and $\boldsymbol{X}_{2q}^*$. $\boldsymbol{f}_q^E$ is a adjusted version of $\boldsymbol{f}_q^*$ to eliminate the bias due to incorrect linkage between $\boldsymbol{X}_{1q}$ and $\boldsymbol{X}_{2q}^*$. Also let $\boldsymbol{V}_{A_q} = \mathrm{Var}_{\boldsymbol{X}^*}\Big[E_{\boldsymbol{X}^*}(\boldsymbol{Y}_q^*|A_q)\Big]$. Then, one has[10]

$$\boldsymbol{V}_{A_q} = \mathrm{Var}_{\boldsymbol{X}^*}(A_q \boldsymbol{f}_q^E).$$

[10]One way to estimate $\boldsymbol{V}_{A_q}$ is using (16) from Chambers (2008). Then,

$$\boldsymbol{V}_{A_q} = \mathrm{diag}\Big[(1-\lambda_{A_q})\{\lambda_{A_q}(f_{q,i}^E - \bar{f}_q^E)^2 + \bar{f}_q^{E(2)} - (\bar{f}_q^E)^2\}\Big], \tag{28}$$

where $\boldsymbol{f}_q^E = (f_{q,i}^E)$ and $\bar{f}_q^E$, $\bar{f}_q^{E(2)}$ are the averages of $f_{q,i}^E$ and their squares respectively in $\boldsymbol{f}_q^E$.

Further,

$$\begin{aligned}\mathrm{Var}_{\boldsymbol{X}^*}(\boldsymbol{Y}_q^*|A_q) &= \mathrm{Var}_{\boldsymbol{X}^*}(A_q\boldsymbol{Y}_q) = A_q\mathrm{Var}_{\boldsymbol{X}^*}(\boldsymbol{Y}_q)A_q^T \\ &= A_q\Big(E_{\boldsymbol{X}^*}\Big[\mathrm{Var}_{\boldsymbol{X}^*}(\boldsymbol{Y}_q|B_{2_q})\Big]\Big)A_q^T + A_q\Big(\mathrm{Var}_{\boldsymbol{X}^*}\Big[E_{\boldsymbol{X}^*}(\boldsymbol{Y}_q|B_{2_q})\Big]\Big)A_q^T,\end{aligned} \tag{29}$$

because one has

$$\mathrm{Var}_{\boldsymbol{X}^*}(\boldsymbol{Y}_q) = E_{\boldsymbol{X}^*}\Big[\mathrm{Var}_{\boldsymbol{X}^*}(\boldsymbol{Y}_q|B_{2_q})\Big] + \mathrm{Var}_{\boldsymbol{X}^*}\Big[E_{\boldsymbol{X}^*}(\boldsymbol{Y}_q|B_{2_q})\Big]. \tag{30}$$

Note that, by (1),

$$E_{\boldsymbol{X}^*}(\boldsymbol{Y}_q|B_{2_q}) = \boldsymbol{X}_q^{B_2}\boldsymbol{\beta}.$$

Thus,

$$\begin{aligned}\mathrm{Var}_{\boldsymbol{X}^*}\Big[E_{\boldsymbol{X}^*}(\boldsymbol{Y}_q|B_{2_q})\Big] &= E_{\boldsymbol{X}^*}\Big[\boldsymbol{X}_q^{B_2}\boldsymbol{\beta} - \boldsymbol{X}_q^{E}\boldsymbol{\beta}\Big]^2 \\ &= E_{\boldsymbol{X}^*}\Big[B_{2_q}^T\boldsymbol{X}_{2q}^*\beta_2 - E_{B_{2q}}\boldsymbol{X}_{2q}^*\beta_2\Big]^2.\end{aligned} \tag{31}$$

Denote that

$$\boldsymbol{f}_{B_{2_q}}^* = \boldsymbol{X}_{2q}^*\beta_2.$$

Also, let

$$C_{B_{2q}} = A_qB_{2_q}^T,$$

which is another permutation matrix, and let

$$E_{C_{2q}} = E_{\boldsymbol{X}^*}\big(A_qB_{2_q}^T\big).$$

Further, let

$$\boldsymbol{V}_{C_{2q}} = A_q\mathrm{Var}_{\boldsymbol{X}^*}\Big[E_{\boldsymbol{X}^*}(\boldsymbol{Y}_q|B_{2_q})\Big]A_q^T. \tag{32}$$

Then, one has[11]

$$\boldsymbol{V}_{C_{2q}} = E_{\boldsymbol{X}^*}\Big[C_{2_q}\boldsymbol{f}_{B_{2_q}}^*(\boldsymbol{f}_{B_{2_q}}^*)^TC_{2_q}^T\Big] - E_{C_{2_q}}\boldsymbol{f}_{B_{2_q}}^*(\boldsymbol{f}_{B_{2_q}}^*)^TE_{C_{2_q}}^T.$$

Furthermore, one has

$$\mathrm{Var}_{\boldsymbol{X}^*}(\boldsymbol{Y}_q|B_{2_q}) = E_{\boldsymbol{X}^*}(\boldsymbol{Y}_q - \boldsymbol{X}_q^{B_2}\boldsymbol{\beta})^2 = E_{\boldsymbol{X}^*}(\epsilon_q)^2 = \sigma_q^2\boldsymbol{I}_q.$$

[11]By (16) from Chambers (2008),

$$\boldsymbol{V}_{C_{2_q}} = \mathrm{diag}\Big[(1-\lambda_{C_{2_q}})\big\{\lambda_{C_{2_q}}(f_{B_{2_q},i}^* - \bar{f}_{B_{2_q}}^*)^2 + \bar{f}_{B_{2_q}}^{*(2)} - (\bar{f}_{B_{2_q}}^*)^2\big\}\Big],$$

where $\boldsymbol{f}_{B_{2_q}}^* = (f_{B_{2_q},i}^*)$ and $\bar{f}_{B_{2_q}}^*$, $\bar{f}_{B_{2_q}}^{*(2)}$ are the averages of $f_{B_{2_q},i}^*$ and their squares respectively in $\boldsymbol{f}_{B_{2_q}}^*$.

Hence,

$$A_q\Big(E_{\boldsymbol{X}^*}\Big[\mathrm{Var}_{\boldsymbol{X}^*}(\boldsymbol{Y}_q|B_{2_q})\Big]\Big)A_q^T = A_q\sigma_q^2\boldsymbol{I}_q A_q^T = \sigma_q^2 A_q A_q^T = \sigma_q^2\boldsymbol{I}_q. \tag{33}$$

Thus, by (29), (30), (32) and (33)

$$\begin{aligned} E_{\boldsymbol{X}^*}\Big[\mathrm{Var}_{\boldsymbol{X}^*}(\boldsymbol{Y}_q^*|A_q)\Big] &= E_{\boldsymbol{X}^*}\Big[A_q\mathrm{Var}_{\boldsymbol{X}^*}(\boldsymbol{Y}_q)A_q^T\Big] \\ &= E_{\boldsymbol{X}^*}\Big\{\sigma_q^2\boldsymbol{I}_q + \boldsymbol{V}_{C_{2q}}\Big\} \\ &= \sigma_q^2\boldsymbol{I}_q + \boldsymbol{V}_{C_{2q}}. \end{aligned} \tag{34}$$

Then, by (27), (32) and (34),

$$\mathrm{Var}_{\boldsymbol{X}^*}(\boldsymbol{Y}_q^*) = \sigma_q^2\boldsymbol{I}_q + \boldsymbol{V}_{C_{2q}} + \boldsymbol{V}_{A_q} = \Sigma_q^*. \tag{35}$$

Consequently, one has

$$\mathrm{Var}_{\boldsymbol{X}^*}(\hat{\boldsymbol{\beta}}^*) = \Big[\sum_q (\boldsymbol{X}_q^*)^T\boldsymbol{X}_q^*\Big]^{-1}\Big[\sum_q (\boldsymbol{X}_q^*)^T\Big(\sigma_q^2\boldsymbol{I}_q + \boldsymbol{V}_{C_{2q}} + \boldsymbol{V}_{A_q}\Big)\boldsymbol{X}_q^*\Big]\Big[\sum_q (\boldsymbol{X}_q^*)^T\boldsymbol{X}_q^*\Big]^{-1},$$

and

$$\begin{aligned} V_R &= \mathrm{Var}_{\boldsymbol{X}^*}(\hat{\boldsymbol{\beta}}_R) \\ &= \Big[\sum_q (\boldsymbol{X}_q^*)^T E_{Aq}\boldsymbol{X}_q^E\Big]^{-1}\Big[\sum_q (\boldsymbol{X}_q^*)^T\Big(\sigma_q^2\boldsymbol{I}_q + \boldsymbol{V}_{C_{2q}} + \boldsymbol{V}_{A_q}\Big)\boldsymbol{X}_q^*\Big]\Big[\sum_q (\boldsymbol{X}_q^*)^T E_{Aq}\boldsymbol{X}_q^E\Big]^{-1}. \end{aligned}$$

To define $\hat{\boldsymbol{V}}_R$, the estimator of $\boldsymbol{V}_R$, let

$$\hat{\boldsymbol{f}}_{B_{2_q}}^* = \boldsymbol{X}_{2q}^*\hat{\beta}_2$$

and

$$\hat{\boldsymbol{f}}_q^E = \boldsymbol{X}_q^E\hat{\boldsymbol{\beta}},$$

where $\hat{\beta}_2$ and $\hat{\boldsymbol{\beta}}$ are the estimates of β_2 and $\boldsymbol{\beta}$ respectively. Then, $\hat{\boldsymbol{V}}_{A_q}$ and $\hat{\boldsymbol{V}}_{C_{2q}}$ can be estimated by replacing $\boldsymbol{f}_q^E$ and $\boldsymbol{f}_{B_{2_q}}^*$, in $\boldsymbol{V}_{A_q}$ and $\boldsymbol{V}_{C_{2q}}$, with $\hat{\boldsymbol{f}}_q^E$ and $\hat{\boldsymbol{f}}_{B_{2_q}}^*$ respectively.

Now, to estimate σ^2, one has

$$\begin{aligned} (\boldsymbol{Y}_q^* - \boldsymbol{f}_q^E)^T(\boldsymbol{Y}_q^* - \boldsymbol{f}_q^E) &= (\boldsymbol{Y}_q^*)^T\boldsymbol{Y}_q^* - (\boldsymbol{Y}_q^*)^T\boldsymbol{f}_q^E - (\boldsymbol{f}_q^E)^T\boldsymbol{Y}_q^* + (\boldsymbol{f}_q^E)^T\boldsymbol{f}_q^E \\ &= \boldsymbol{Y}_q^T A_q^T A_q\boldsymbol{Y}_q - \boldsymbol{Y}_q^T\boldsymbol{f}_q - \boldsymbol{f}_q^T\boldsymbol{Y}_q + \boldsymbol{f}_q^T\boldsymbol{f}_q \\ &\quad + \boldsymbol{Y}_q^T\boldsymbol{f}_q + \boldsymbol{f}_q^T\boldsymbol{Y}_q - \boldsymbol{f}_q^T\boldsymbol{f}_q \\ &\quad - (\boldsymbol{Y}_q^*)^T\boldsymbol{f}_q^E - (\boldsymbol{f}_q^E)^T\boldsymbol{Y}_q^* + (\boldsymbol{f}_q^E)^T\boldsymbol{f}_q^E, \end{aligned} \tag{36}$$

where

$$E_{\boldsymbol{X}^*}\sum_q\Big(\boldsymbol{Y}_q^TA_q^TA_q\boldsymbol{Y}_q-\boldsymbol{Y}_q^T\boldsymbol{f}_q-\boldsymbol{f}_q^T\boldsymbol{Y}_q+\boldsymbol{f}_q^T\boldsymbol{f}_q\Big)=E_{\boldsymbol{X}^*}\sum_q\Big[(\boldsymbol{Y}_q-\boldsymbol{f}_q)^T(\boldsymbol{Y}_q-\boldsymbol{f}_q)\Big]=N\sigma^2. \tag{37}$$

Also, one has

$$E_{\boldsymbol{X}^*}\big(\boldsymbol{f}_q^T\boldsymbol{Y}_q-\boldsymbol{f}_q^T\boldsymbol{f}_q\big)=E_{\boldsymbol{X}^*}\big(\boldsymbol{f}_q^T[\boldsymbol{Y}_q-\boldsymbol{f}_q]\big)=E_{\boldsymbol{X}^*}\big(\boldsymbol{f}_q^T\epsilon_q\big)=0. \tag{38}$$

Further,

$$\begin{aligned}\boldsymbol{Y}_q^T\boldsymbol{f}_q-(\boldsymbol{Y}_q^*)^T\boldsymbol{f}_q^E-(\boldsymbol{f}_q^E)^T\boldsymbol{Y}_q^*+(\boldsymbol{f}_q^E)^T\boldsymbol{f}_q^E&=\boldsymbol{Y}_q^T\boldsymbol{f}_q-(\boldsymbol{Y}_q^*)^T\boldsymbol{f}_q^E-(\boldsymbol{f}_q^E)^T\boldsymbol{Y}_q^*+(\boldsymbol{f}_q^E)^T\boldsymbol{f}_q^E\\&\quad-\boldsymbol{Y}_q^T\boldsymbol{f}_q^E+\boldsymbol{Y}_q^T\boldsymbol{f}_q^E-(\boldsymbol{f}_q^E)^T\boldsymbol{f}_q+(\boldsymbol{f}_q^E)^T\boldsymbol{f}_q-(\boldsymbol{f}_q^E)^T\boldsymbol{f}_q^E+(\boldsymbol{f}_q^E)^T\boldsymbol{f}_q^E\\&=\big[\boldsymbol{Y}_q^T\boldsymbol{f}_q^E-(\boldsymbol{Y}_q^*)^T\boldsymbol{f}_q^E\big]+\big[(\boldsymbol{f}_q^E)^T\boldsymbol{f}_q^E-(\boldsymbol{f}_q^E)^T\boldsymbol{Y}_q^*\big] \qquad (39)\\&\quad+\big[\boldsymbol{Y}_q^T-(\boldsymbol{f}_q^E)^T\big]\boldsymbol{f}_q+\big[(\boldsymbol{f}_q^E)^T-\boldsymbol{Y}_q^T\big]\boldsymbol{f}_q^E+(\boldsymbol{f}_q^E)^T\big[\boldsymbol{f}_q-\boldsymbol{f}_q^E\big]. \qquad (40)\end{aligned}$$

Then it is easy to see that the expectation of (40) is zero. Also,

$$E_{\boldsymbol{X}^*}\Big(\big[\boldsymbol{Y}_q^T\boldsymbol{f}_q^E-(\boldsymbol{Y}_q^*)^T\boldsymbol{f}_q^E\big]+\big[(\boldsymbol{f}_q^E)^T\boldsymbol{f}_q^E-(\boldsymbol{f}_q^E)^T\boldsymbol{Y}_q^*\big]\Big)=2(\boldsymbol{f}_q^E)^T\big[\boldsymbol{I}_q-E_{A_q}\big]\boldsymbol{f}_q^E \tag{41}$$

Therefore,by (36)–(41),

$$\hat{\sigma}^2=N^{-1}\Big(\sum_q(\boldsymbol{Y}_q^*-\boldsymbol{f}_q^E)^T(\boldsymbol{Y}_q^*-\boldsymbol{f}_q^E)-2\sum_q(\boldsymbol{f}_q^E)^T\big[\boldsymbol{I}_q-E_{A_q}\big]\boldsymbol{f}_q^E\Big).$$

A.4 Proof of Theorem 4

Let $\hat{\boldsymbol{\theta}}_1^*$ be the solution of (6). Then, the asymptotic variance of $\hat{\boldsymbol{\theta}}_1^*$ is of the form

$$\mathrm{Var}_{X^*}(\hat{\boldsymbol{\theta}}_1^*)\approx\big[\partial_\theta\boldsymbol{H}_1^*(\boldsymbol{\theta}_0)\big]^{-1}\mathrm{Var}_{X^*}\big[\boldsymbol{H}_1^*(\boldsymbol{\theta}_0)\big]\Big(\big[\partial_\theta\boldsymbol{H}_1^*(\boldsymbol{\theta}_0)\big]^{-1}\Big)^T.$$

Note that, in general $\boldsymbol{G}_q(\boldsymbol{\theta})$ is a function of both $\boldsymbol{\theta}$ and $\mathbf{X}$, but, in our case, we only consider the case where $\boldsymbol{G}_q$ is a function of $\mathbf{X}$. Thus,

$$\partial_\theta\boldsymbol{H}_1^*(\boldsymbol{\theta})=\sum_q\boldsymbol{G}_q\partial_\theta\boldsymbol{f}_q^E(\boldsymbol{\theta}).$$

Further, by (17), one has

$$\begin{aligned}\mathrm{Var}_{X^*}\big[\boldsymbol{H}_1^*(\boldsymbol{\theta})\big]&=\sum_q\boldsymbol{G}_q\mathrm{Var}_{X^*}\big(\boldsymbol{Y}_q\big)\boldsymbol{G}_q^T\\&=\sum_q\boldsymbol{G}_q\Big[\sigma_q^2\boldsymbol{I}_q+\boldsymbol{V}_{B_{2_q}}\Big]\boldsymbol{G}_q^T\\&=\sum_q\boldsymbol{G}_q\Sigma_q^{*1}\boldsymbol{G}_q^T.\end{aligned}$$

Therefore, the asymptotic variance estimator is of the form

$$\boldsymbol{V}_{1|\boldsymbol{X}^*}(\hat{\boldsymbol{\theta}}_1^*) = \Big[\sum_q \boldsymbol{G}_q \partial_\theta \boldsymbol{f}_q^E(\hat{\boldsymbol{\theta}}_1^*)\Big]^{-1}\Big[\sum_q \boldsymbol{G}_q \hat{\Sigma}_q^{*1}\boldsymbol{G}_q^T\Big]\Big(\Big[\sum_q \boldsymbol{G}_q \partial_\theta \boldsymbol{f}_q^E(\hat{\boldsymbol{\theta}}_1^*)\Big]^{-1}\Big)^T.$$

Let us consider the **Case 1** where $\boldsymbol{Y}$ is the bench mark data set and the linkages between $\boldsymbol{Y}$ and $\boldsymbol{X}_1$ and the linkages between $\boldsymbol{Y}$ and $\boldsymbol{X}_2$ are done with some errors. In this case, we have similar estimating function

$$\boldsymbol{H}_2^*(\boldsymbol{\theta}) = \sum_q \boldsymbol{G}_q(\boldsymbol{\theta})\big\{\boldsymbol{Y}_q - \boldsymbol{f}_q^{E2}(\boldsymbol{\theta})\big\},$$

but, by (4), $\boldsymbol{f}_q^{E2}(\boldsymbol{\theta}) = \boldsymbol{X}_q^{E2}\boldsymbol{\beta} = (\mathbf{1}_q, E_{B_{1q}}\boldsymbol{X}_{1q}^*, E_{B_{2q}}\boldsymbol{X}_{2q}^*)\boldsymbol{\beta}$. This leads the asymptotic variance estimator of the form

$$\boldsymbol{V}_{2|\boldsymbol{X}^*}(\hat{\boldsymbol{\theta}}_2^*) = \Big[\sum_q \boldsymbol{G}_q \partial_\theta \boldsymbol{f}_q^{E2}(\hat{\boldsymbol{\theta}}_2^*)\Big]^{-1}\Big[\sum_q \boldsymbol{G}_q \hat{\Sigma}_q^{*2}\boldsymbol{G}_q^T\Big]\Big(\Big[\sum_q \boldsymbol{G}_q \partial_\theta \boldsymbol{f}_q^{E2}(\hat{\boldsymbol{\theta}}_2^*)\Big]^{-1}\Big)^T,$$

where, by (26), $\hat{\Sigma}_q^{*2} = \hat{\sigma}_q^2\boldsymbol{I}_q + \hat{\boldsymbol{V}}_{B_{1q}} + \hat{\boldsymbol{V}}_{B_{2q}}$.

Finally, the asymptotic variance of $\hat{\boldsymbol{\theta}}_3^*$ is of the form

$$\mathrm{Var}_{X^*}(\hat{\boldsymbol{\theta}}_3^*) \approx \big[\partial_\theta \boldsymbol{H}_3^*(\boldsymbol{\theta}_0)\big]^{-1}\mathrm{Var}_{X^*}\big[\boldsymbol{H}_3^*(\boldsymbol{\theta}_0)\big]\Big(\big[\partial_\theta \boldsymbol{H}_3^*(\boldsymbol{\theta}_0)\big]^{-1}\Big)^T, \tag{42}$$

where,

$$\partial_\theta \boldsymbol{H}_3^*(\boldsymbol{\theta}) = \sum_q \boldsymbol{G}_q E_{A_q}\partial_\theta \boldsymbol{f}_q^E(\boldsymbol{\theta}). \tag{43}$$

Further, by (35), one has

$$\begin{aligned}\mathrm{Var}_{X^*}\big[\boldsymbol{H}_3^*(\boldsymbol{\theta})\big] &= \sum_q \boldsymbol{G}_q \mathrm{Var}_{X^*}\big(\boldsymbol{Y}_q^*\big)\boldsymbol{G}_q^T \\ &= \sum_q \boldsymbol{G}_q\Big[\sigma_q^2\boldsymbol{I}_q + \boldsymbol{V}_{C_{2q}} + \boldsymbol{V}_{A_q}\Big]\boldsymbol{G}_q^T \\ &= \sum_q \boldsymbol{G}_q\Sigma_q^{*3}\boldsymbol{G}_q^T.\end{aligned}$$

Therefore, the asymptotic variance estimator is of the form

$$\boldsymbol{V}_{3|\boldsymbol{X}^*}(\hat{\boldsymbol{\theta}}_3^*) = \Big[\sum_q \boldsymbol{G}_q E_{A_q}\partial_\theta \boldsymbol{f}_q^E(\hat{\boldsymbol{\theta}}_3^*)\Big]^{-1}\Big[\sum_q \boldsymbol{G}_q \hat{\Sigma}_q^{*3}\boldsymbol{G}_q^T\Big]\Big(\Big[\sum_q \boldsymbol{G}_q E_{A_q}\partial_\theta \boldsymbol{f}_q^E(\hat{\boldsymbol{\theta}}_3^*)\Big]^{-1}\Big)^T,$$

as required.

A.5 Proof of the Theorem 5

Let $\lambda_1 = \lambda_A$ and $\lambda_2 = \lambda_{B_2}$. Then the variance of $\hat{\boldsymbol{\theta}}_3^{**}$ can be approximated by

$$\begin{aligned}\mathrm{Var}_{\boldsymbol{X}^*}(\hat{\boldsymbol{\theta}}_3^{**}) &\approx \big[\partial_\theta \boldsymbol{H}_0^*\big]^{-1}\mathrm{Var}_{\boldsymbol{X}^*}\Big[\boldsymbol{H}_0^* + \partial_{\lambda_1}\boldsymbol{H}_0^*(\hat{\lambda}_A - \lambda_A) + \partial_{\lambda_2}\boldsymbol{H}_0^*(\hat{\lambda}_{B_2} - \lambda_{B_2})\Big]\Big\{\big[\partial_\theta \boldsymbol{H}_0^*\big]^{-1}\Big\}^T \\ &= \big[\partial_\theta \boldsymbol{H}_0^*\big]^{-1}\Big[\mathrm{Var}_{\boldsymbol{X}^*}\big(\boldsymbol{H}_0^*\big) + \big(\partial_{\lambda_1}\boldsymbol{H}_0^*\big)\mathrm{Var}_{\boldsymbol{X}^*}(\lambda_A)\big(\partial_{\lambda_1}\boldsymbol{H}_0^*\big)^T \\ &\qquad + \big(\partial_{\lambda_2}\boldsymbol{H}_0^*\big)\mathrm{Var}_{\boldsymbol{X}^*}(\lambda_{B_2})\big(\partial_{\lambda_2}\boldsymbol{H}_0^*\big)^T\Big]\Big\{\big[\partial_\theta \boldsymbol{H}_0^*\big]^{-1}\Big\}^T.\end{aligned} \tag{44}$$

To derive $\partial_{\lambda_i}\boldsymbol{H}_0^*$, we assume that the distribution of λ_i is independent[12] of the distribution of $\boldsymbol{H}_0^*$. Then, by the similar arguments in Chambers (2008),

$$\begin{aligned}\partial_{\lambda_1}\boldsymbol{H}_0^* &= \partial_{\lambda_1}\sum_q \boldsymbol{G}_q\big\{\boldsymbol{Y}_q^* - E_{A_q}(\lambda_{A_q})\boldsymbol{f}_q^E(\boldsymbol{\theta}, \lambda_{B_{2q}})\big\} \\ &= -\sum_q \boldsymbol{G}_q\Big[\partial_{\lambda_1}E_{A_q}(\lambda_{A_q})\Big]\boldsymbol{f}_q^E(\boldsymbol{\theta}, \lambda_{B_{2q}}) \\ &= -\sum_q \boldsymbol{G}_q\Big[(M_q - 1)^{-1}(M_q\boldsymbol{I}_q - \mathbf{1}_q\mathbf{1}_q^T)\Big]\boldsymbol{f}_q^E(\boldsymbol{\theta}, \lambda_{B_{2q}})\end{aligned} \tag{45}$$

and

$$\begin{aligned}\partial_{\lambda_2}\boldsymbol{H}_0^* &= \partial_{\lambda_2}\sum_q \boldsymbol{G}_q\big\{\boldsymbol{Y}_q^* - E_{A_q}(\lambda_{A_q})\boldsymbol{f}_q^E(\boldsymbol{\theta}, \lambda_{B_{2q}})\big\} \\ &= -\sum_q \boldsymbol{G}_q E_{A_q}(\lambda_{A_q})\Big[\partial_{\lambda_2}\boldsymbol{f}_q^E(\boldsymbol{\theta}, \lambda_{B_{2q}})\Big] \\ &= -\sum_q \boldsymbol{G}_q E_{A_q}(\lambda_{A_q})\Big[\partial_{\lambda_2}(\beta_0 + \boldsymbol{X}_{1q}\beta_1 + E_{B_{2q}}\boldsymbol{X}_{2q}^*\beta_2)\Big] \\ &= -\sum_q \boldsymbol{G}_q E_{A_q}(\lambda_{A_q})\Big[\partial_{\lambda_2}(E_{B_{2q}})\boldsymbol{X}_{2q}^*\beta_2\Big] \\ &= -\sum_q \boldsymbol{G}_q E_{A_q}\Big[(M_q - 1)^{-1}(M_q\boldsymbol{I}_q - \mathbf{1}_q\mathbf{1}_q^T)\Big]\boldsymbol{X}_{2q}^*\beta_2.\end{aligned} \tag{46}$$

Therefore, the variance $\mathrm{Var}_{\boldsymbol{X}^*}(\hat{\boldsymbol{\theta}}_3^*)$ can be evaluated by substituting the estimated values of (43), (45) and (46) into (44).

For the **Case 1** where Y is the bench mark data set and the linkages between Y and $\boldsymbol{X}_1$ and the linkages between Y and $\boldsymbol{X}_2$ are done with some errors, the variance of $\mathrm{Var}_{\boldsymbol{X}^*}(\hat{\boldsymbol{\theta}}_2^*)$ is

[12]This assumption was originally introduced in Chambers (2008).

of the form

$$\begin{aligned}\mathrm{Var}_{\boldsymbol{X}^*}(\hat{\boldsymbol{\theta}}_2^*) &\approx \left[\partial_\theta \boldsymbol{H}_0^*\right]^{-1}\mathrm{Var}_{\boldsymbol{X}^*}\Big[\boldsymbol{H}_0^* + \partial_{\lambda_{B_1}}\boldsymbol{H}_0^*(\hat{\lambda}_{B_1}-\lambda_{B_1}) + \partial_{\lambda_{B_2}}\boldsymbol{H}_0^*(\hat{\lambda}_{B_2}-\lambda_{B_2})\Big]\Big\{\left[\partial_\theta \boldsymbol{H}_0^*\right]^{-1}\Big\}^T \\ &= \left[\partial_\theta \boldsymbol{H}_0^*\right]^{-1}\Big[\mathrm{Var}_{\boldsymbol{X}^*}\big(\boldsymbol{H}_0^*\big) + \big(\partial_{\lambda_{B_1}}\boldsymbol{H}_0^*\big)\mathrm{Var}_{\boldsymbol{X}^*}(\lambda_{B_1})\big(\partial_{\lambda_{B_1}}\boldsymbol{H}_0^*\big)^T \\ &\qquad + \big(\partial_{\lambda_{B_2}}\boldsymbol{H}_0^*\big)\mathrm{Var}_{\boldsymbol{X}^*}(\lambda_{B_2})\big(\partial_{\lambda_{B_2}}\boldsymbol{H}_0^*\big)^T\Big]\Big\{\left[\partial_\theta \boldsymbol{H}_0^*\right]^{-1}\Big\}^T,\end{aligned}$$

where,

$$\begin{aligned}\lambda_{B_{1q}} &= \mathrm{pr}(\text{correct linkage between } Y \text{ and } \boldsymbol{X}_{1q}^*), \\ \boldsymbol{H}_0^* &= \boldsymbol{H}_2^*(\boldsymbol{\theta}_0, \lambda_{B_1}, \lambda_{B_2}).\end{aligned}$$

Further, it is easy to see that

$$\partial_{\lambda_{B_1}}\boldsymbol{H}_0^* = -\sum_q \boldsymbol{G}_q\Big[(M_q-1)^{-1}(M_q\boldsymbol{I}_q - \mathbf{1}_q\mathbf{1}_q^T)\Big]\boldsymbol{X}_{1q}^*\beta_1$$

and

$$\partial_{\lambda_{B_2}}\boldsymbol{H}_0^* = -\sum_q \boldsymbol{G}_q\Big[(M_q-1)^{-1}(M_q\boldsymbol{I}_q - \mathbf{1}_q\mathbf{1}_q^T)\Big]\boldsymbol{X}_{2q}^*\beta_2.$$

Finally, for the **Case 0**, one has

$$\begin{aligned}\mathrm{Var}_{\boldsymbol{X}^*}(\hat{\boldsymbol{\theta}}_1^*) &\approx \left[\partial_\theta \boldsymbol{H}_0^*\right]^{-1}\mathrm{Var}_{\boldsymbol{X}^*}\Big[\boldsymbol{H}_0^* + \partial_{\lambda_{B_2}}\boldsymbol{H}_0^*(\hat{\lambda}_{B_2}-\lambda_{B_2})\Big]\Big\{\left[\partial_\theta \boldsymbol{H}_0^*\right]^{-1}\Big\}^T \\ &= \left[\partial_\theta \boldsymbol{H}_0^*\right]^{-1}\Big[\mathrm{Var}_{\boldsymbol{X}^*}\big(\boldsymbol{H}_0^*\big) + \big(\partial_{\lambda_{B_2}}\boldsymbol{H}_0^*\big)\mathrm{Var}_{\boldsymbol{X}^*}(\lambda_{B_2})\big(\partial_{\lambda_{B_2}}\boldsymbol{H}_0^*\big)^T\Big]\Big\{\left[\partial_\theta \boldsymbol{H}_0^*\right]^{-1}\Big\}^T,\end{aligned}$$

where,

$$\begin{aligned}\lambda_{B_{2q}} &= \mathrm{pr}(\text{correct linkage between } \boldsymbol{X}_{1q} \text{ and } \boldsymbol{X}_{2q}^*), \\ \boldsymbol{H}_0^* &= \boldsymbol{H}_1^*(\boldsymbol{\theta}_0, \lambda_{B_2})\end{aligned}$$

with

$$\partial_{\lambda_{B_2}}\boldsymbol{H}_0^* = -\sum_q \boldsymbol{G}_q\Big[(M_q-1)^{-1}(M_q\boldsymbol{I}_q - \mathbf{1}_q\mathbf{1}_q^T)\Big]\boldsymbol{X}_{2q}^*\beta_2.$$

A.6 Proof of the Theorem 6

Let $\hat{\boldsymbol{\theta}}_3^{s*}$ be the solution of the estimating equation (11). To derive the asymptotic variance estimator for $\hat{\boldsymbol{\theta}}_3^{s*}$, note that by (42),

$$\mathrm{Var}_{X^*}(\hat{\boldsymbol{\theta}}_3^{s*}) \approx \left[\partial_\theta \boldsymbol{H}_{ws}^{adj}(\boldsymbol{\theta}_0)\right]^{-1}\mathrm{Var}_{X^*}\left[\boldsymbol{H}_{ws}^{adj}(\boldsymbol{\theta}_0)\right]\Big(\left[\partial_\theta \boldsymbol{H}_{ws}^{adj}(\boldsymbol{\theta}_0)\right]^{-1}\Big)^T \tag{47}$$

with corresponding estimator of the form

$$\begin{aligned}\boldsymbol{V}^{ws}_{3|\boldsymbol{X}^*}(\hat{\boldsymbol{\theta}}^{s*}_3) &= \big[\partial_\theta \boldsymbol{H}^{adj}_{ws}(\boldsymbol{\theta}_0)\big]^{-1}\boldsymbol{V}^{ws}_{3|\boldsymbol{X}^*}\big[\boldsymbol{H}^{adj}_{ws}(\boldsymbol{\theta}_0)\big]\Big(\big[\partial_\theta \boldsymbol{H}^{adj}_{ws}(\boldsymbol{\theta}_0)\big]^{-1}\Big)^T \\ &\approx \Big[\sum_q \boldsymbol{G}_{sq}\tilde{E}_{A_{sq}}\partial_\theta \boldsymbol{f}^E_{sq}(\hat{\boldsymbol{\theta}}^{s*}_3)\Big]^{-1}\Big[\sum_q \boldsymbol{G}_{sq}\hat{\Sigma}_{sq}\boldsymbol{G}^T_{sq}\Big]\Big(\Big[\sum_q \boldsymbol{G}_{sq}\tilde{E}_{A_{sq}}\partial_\theta \boldsymbol{f}^E_{sq}(\hat{\boldsymbol{\theta}}^{s*}_3)\Big]^{-1}\Big)^T,\end{aligned}$$

under the assumption that $\boldsymbol{G}_{sq}$ is independent of $\boldsymbol{\theta}$. Next step is to define Σ_{sq}. Note that

$$\begin{aligned}\Sigma_{sq} &= \mathrm{Var}_{X^*}(\boldsymbol{Y}^*_{sq}) \\ &= \mathrm{Var}_{X^*}(A_{ssq}\boldsymbol{Y}_{sq} + A_{srq}\boldsymbol{Y}_{rq}) \\ &= \mathrm{Var}_{X^*}(A_{ssq}\boldsymbol{Y}_{sq}) + 2\mathrm{cov}_{X^*}(A_{ssq}\boldsymbol{Y}_{sq}, A_{srq}\boldsymbol{Y}_{rq}) + \mathrm{Var}_{X^*}(A_{srq}\boldsymbol{Y}_{rq}).\end{aligned} \tag{48}$$

Further, by (30) and with similar arguments in (31)-(34), one has

$$\begin{aligned}\mathrm{Var}_{X^*}(\boldsymbol{Y}_q) &= E_{\boldsymbol{X}^*}\Big[\mathrm{Var}_{\boldsymbol{X}^*}(\boldsymbol{Y}_q|B_{2_q})\Big] + \mathrm{Var}_{\boldsymbol{X}^*}\Big[E_{\boldsymbol{X}^*}(\boldsymbol{Y}_q|B_{2_q})\Big] \\ &= \sigma^2_q\boldsymbol{I}_q + \boldsymbol{V}_{B_{2q}},\end{aligned}$$

where

$$\boldsymbol{V}_{B_{2q}} = \mathrm{Var}_{\boldsymbol{X}^*}\Big[E_{\boldsymbol{X}^*}(\boldsymbol{Y}_q|B_{2_q})\Big]$$

that can be approximated with a diagonal matrix[13] by the same argument in (16) from Chambers (2008). Thus, $\mathrm{Var}_{X^*}(\boldsymbol{Y}_q)$ can be approximately regarded as a diagonal matrix and set $\mathrm{Var}_{X^*}(\boldsymbol{Y}_q) \approx D_q = \mathrm{diag}\{d_i; i \in q\}$. In this case, one has

$$\mathrm{cov}_{X^*}(A_{ssq}\boldsymbol{Y}_{sq}, A_{srq}\boldsymbol{Y}_{rq}) \approx 0.$$

Also, (48) becomes

$$\begin{aligned}\Sigma_{sq} &\approx \mathrm{Var}_{X^*}(A_{ssq}\boldsymbol{Y}_{sq}) + \mathrm{Var}_{X^*}(A_{srq}\boldsymbol{Y}_{rq}) \\ &= E_{\boldsymbol{X}^*}\Big[\mathrm{Var}_{\boldsymbol{X}^*}(A_{ssq}\boldsymbol{Y}_{sq}|A_q)\Big] + \mathrm{Var}_{\boldsymbol{X}^*}\Big[E_{\boldsymbol{X}^*}(A_{ssq}\boldsymbol{Y}_{sq}|A_q)\Big] \\ &\quad + E_{\boldsymbol{X}^*}\Big[\mathrm{Var}_{\boldsymbol{X}^*}(A_{srq}\boldsymbol{Y}_{rq}|A_q)\Big] + \mathrm{Var}_{\boldsymbol{X}^*}\Big[E_{\boldsymbol{X}^*}(A_{srq}\boldsymbol{Y}_{rq}|A_q)\Big] \\ &= E_{\boldsymbol{X}^*}\Big(A_{ssq}\mathrm{Var}_{\boldsymbol{X}^*}(\boldsymbol{Y}_{sq})A^T_{ssq}\Big) + E_{\boldsymbol{X}^*}\Big(A_{srq}\mathrm{Var}_{\boldsymbol{X}^*}(\boldsymbol{Y}_{rq})A^T_{srq}\Big) \\ &\quad + \mathrm{Var}_{\boldsymbol{X}^*}\Big(A_{ssq}\boldsymbol{f}^E_{sq} + A_{srq}\boldsymbol{f}^E_{rq}\Big) \\ &\approx E_{\boldsymbol{X}^*}\Big(A_{ssq}D_{sq}A^T_{ssq}\Big) + E_{\boldsymbol{X}^*}\Big(A_{srq}D_{rq}A^T_{srq}\Big) + \mathrm{Var}_{\boldsymbol{X}^*}\Big(A_{ssq}\boldsymbol{f}^E_{sq} + A_{srq}\boldsymbol{f}^E_{rq}\Big)\end{aligned}$$

[13] By (16) from Chambers (2008),

$$\boldsymbol{V}_{B_{2q}} \approx \mathrm{diag}\Big[(1-\lambda_{B_{2q}})\{\lambda_{B_{2q}}(f^*_{B_{2q},i} - \bar{f}^*_{B_{2q}})^2 + \bar{f}^{*(2)}_{B_{2q}} - (\bar{f}^*_{B_{2q}})^2\}\Big],$$

where $\boldsymbol{f}^*_{B_{2q}} = (f^*_{B_{2q},i})$ and $\bar{f}^*_{B_{2q}}$, $\bar{f}^{*(2)}_{B_{2q}}$ are the averages of $f^*_{B_{2q},i}$ and their squares respectively in $\boldsymbol{f}^*_{B_{2q}}$.

where $D_{sq} = \text{diag}\{d_i; i \in s_q\} \approx \text{Var}_{\boldsymbol{X}^*}(\boldsymbol{Y}_{sq})$ and $D_{rq} = \text{diag}\{d_i; i \in r_q\} \approx \text{Var}_{\boldsymbol{X}^*}(\boldsymbol{Y}_{rq})$. Let $\bar{d}_{sq}$ be the mean of $\{d_i; i \in s_q\}$. This approximation approach and the same arguments in (66)-(68) from Chambers (2008) lead to the estimate

$$\hat{\Sigma}_{sq} \approx \text{diag}\Big(\frac{(\lambda_{A_q}M_q - 1)d_i + M_q(1-\lambda_{A_q})\bar{d}_{sq}}{M_q - 1} + (1-\lambda_{A_q})\big[\lambda_{A_q}(f_i^E - \bar{f}_{sq}^E)^2 + \bar{f}_{sq}^{E(2)} - (\bar{f}_{sq}^E)^2\big]; i \in s_q\Big) \tag{49}$$

under the assumption that we know $\boldsymbol{f}_{sq}^E$. However, since we only have sample records s, we do not have B_{2q}. We only have B_{2sq} theoretically. Then by the similar arguments in (10)-(11), we can estimate $\boldsymbol{f}_{sq}^E$ using

$$\tilde{E}_{B_{2sq}} = \Big(\frac{\lambda_{B_{2q}}M_q - 1}{M_q - 1}\Big)\boldsymbol{I}_{sq} + \Big(\frac{1-\lambda_{B_{2q}}}{M_q - 1}\Big)\mathbf{1}_{sq}\boldsymbol{w}_{sq}^T.$$

The proofs for the **Case 1** and the **Case 0** are trivial.

A.7 Proof of the Corollary 7

Let $\hat{\boldsymbol{\theta}}_2^*$ be the solution of the estimating equation. When we need to estimate $\lambda_{B_{1q}}$ and $\lambda_{B_{2q}}$, an asymptotic variance estimator is of the form

$$\begin{aligned}\text{Var}_{\boldsymbol{X}^*}(\hat{\boldsymbol{\theta}}_2^*) \approx \big[\partial_\theta \boldsymbol{H}_{ws2,0}^{adj}\big]^{-1}\Big[&\text{Var}_{\boldsymbol{X}^*}\big(\boldsymbol{H}_{ws2,0}^{adj}\big) + \big(\partial_{\lambda_1}\boldsymbol{H}_{ws2,0}^{adj}\big)\text{Var}_{\boldsymbol{X}^*}(\hat{\lambda}_{B_1})\big(\partial_{\lambda_1}\boldsymbol{H}_{ws2,0}^{adj}\big)^T \\ &+ \big(\partial_{\lambda_2}\boldsymbol{H}_{ws2,0}^{adj}\big)\text{Var}_{\boldsymbol{X}^*}(\hat{\lambda}_{B_2})\big(\partial_{\lambda_2}\boldsymbol{H}_{ws2,0}^{adj}\big)^T\Big]\Big\{\big[\partial_\theta \boldsymbol{H}_{ws2,0}^{adj}\big]^{-1}\Big\}^T,\end{aligned}$$

where

$$\begin{aligned}\boldsymbol{H}_{ws2,0}^{adj} &= \boldsymbol{H}_{ws2}^{adj}(\boldsymbol{\theta}_0, \lambda_{B_1}^0, \lambda_{B_2}^0), \\ \partial_{\lambda_1} &= \partial_{\lambda_{B_1}}, \\ \partial_{\lambda_2} &= \partial_{\lambda_{B_2}}, \\ \text{Var}_{\boldsymbol{X}^*}(\lambda_{B_{1q}}) &= (m_q^{B_1})^{-1}\lambda_{B_{1q}}(1-\lambda_{B_{1q}}), \\ \text{Var}_{\boldsymbol{X}^*}(\lambda_{B_{2q}}) &= (m_q^{B_2})^{-1}\lambda_{B_{2q}}(1-\lambda_{B_{2q}}), \\ \partial_{\lambda_1}\boldsymbol{H}_{ws2,0}^{adj} &= -\sum_q \boldsymbol{G}_{sq}\Big[(M_q-1)^{-1}(M_q\boldsymbol{I}_{sq} - \mathbf{1}_{sq}\boldsymbol{w}_{sq}^T)\Big]\boldsymbol{X}_{1q}^*\beta_1 \quad \text{and} \\ \partial_{\lambda_2}\boldsymbol{H}_{ws2,0}^{adj} &= -\sum_q \boldsymbol{G}_{sq}\Big[(M_q-1)^{-1}(M_q\boldsymbol{I}_{sq} - \mathbf{1}_{sq}\boldsymbol{w}_{sq}^T)\Big]\boldsymbol{X}_{2q}^*\beta_2.\end{aligned}$$

Finally, for the **Case 0**, it has simplest forms for their formulae since there is only one mismatch. The estimating function is of the form

$$\boldsymbol{H}_{ws1}^{adj}(\boldsymbol{\theta}) = \sum_q \boldsymbol{G}_{sq}\big\{\boldsymbol{Y}_{sq} - \boldsymbol{f}_{sq}^E(\boldsymbol{\theta})\big\},$$

where

$$\boldsymbol{f}_{sq}^{E} = \boldsymbol{X}_{sq}^{E}\boldsymbol{\beta} = (\mathbf{1}_{sq}, \boldsymbol{X}_{1sq}, \tilde{E}_{B_{2sq}}\boldsymbol{X}_{2sq}^{*})\boldsymbol{\beta} \quad \text{and}$$
$$\tilde{E}_{B_{2sq}} = \Big(\frac{\lambda_{B_{2q}}M_q - 1}{M_q - 1}\Big)\boldsymbol{I}_{sq} + \Big(\frac{1-\lambda_{B_{2q}}}{M_q - 1}\Big)\mathbf{1}_{sq}\boldsymbol{w}_{sq}^{T}.$$

Let $\hat{\boldsymbol{\theta}}_1^*$ be the solution of the estimating equation. When we need to estimate $\lambda_{B_{2q}}$, the asymptotic variance estimator is of the form

$$\begin{aligned}\mathrm{Var}_{\boldsymbol{X}^*}(\hat{\boldsymbol{\theta}}_1^*) \approx \big[\partial_\theta \boldsymbol{H}_{ws1,0}^{adj}\big]^{-1}\Big[&\mathrm{Var}_{\boldsymbol{X}^*}\big(\boldsymbol{H}_{ws1,0}^{adj}\big) \\ &+ \big(\partial_{\lambda_2}\boldsymbol{H}_{ws1,0}^{adj}\big)\mathrm{Var}_{\boldsymbol{X}^*}(\hat{\lambda}_{B_2})\big(\partial_{\lambda_2}\boldsymbol{H}_{ws1,0}^{adj}\big)^{T}\Big]\Big\{\big[\partial_\theta \boldsymbol{H}_{ws1,0}^{adj}\big]^{-1}\Big\}^{T},\end{aligned}$$

where

$$\boldsymbol{H}_{ws1,0}^{adj} = \boldsymbol{H}_{ws1}^{adj}(\boldsymbol{\theta}_0, \lambda_{B_2}^0),$$
$$\partial_{\lambda_2} = \partial_{\lambda_{B_2}}.$$

References

Chambers, R. (2008). Regression analysis of probability-linked data. *Statisphere*, **4**. http://www.statisphere.govt.nz/official-statistics-research/series/vol-4.htm.

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

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

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

Scheuren, F. and Winkler, W. E. (1997). Regression analysis of data files that are computer matched–part ii. *Survey Methodology*, **23**, 157–165.

Tables

Table 1: Simulation results linear regression for register to register of the **Case 0**: in terms of relative bias, RMSE and the actual coverage percentage for nomial 95% confidence intervals

Estimator	Relative Bias		Relative RMSE		Coverage	
	λ known	λ unknown	λ known	λ unknown	λ known	λ unknown
Simulation results for the intercept estimator						
ST	186.38	186.38	188.51	188.51	0	0
R	-0.76	-2.37	31.11	69.23	99.3	99.8
A	-0.68	3.35	28.69	61.54	99.6	99.8
C	0.45	12.94	14.39	38.63	100	100
Simulation results for the first slope estimator						
ST	-0.16	-0.16	9.04	9.04	94.1	94.1
R	-0.14	-0.14	8.94	8.96	98.6	100
A	-0.14	-0.14	8.94	8.96	98.6	100
C	-0.12	-0.13	5.78	6.05	100	100
Simulation results for the second slope estimator						
ST	-11.64	-11.64	33.28	33.28	0	0
R	0.05	0.15	5.48	12.20	97.5	100
A	0.05	-0.21	5.05	10.84	98.2	100
C	-0.03	-0.81	2.34	6.72	100	100

Table 2: Simulation results linear regression for register to register of the **Case 1**: in terms of relative bias, RMSE and the actual coverage percentage for nomial 95% confidence intervals

Estimator	Relative Bias		Relative RMSE		Coverage	
	λ known	λ unknown	λ known	λ unknown	λ known	λ unknown
Simulation results for the intercept estimator						
ST	187.22	187.22	189.39	189.39	0	0
R	0.08	1.28	31.18	71.19	99.4	100
A	0.10	7.03	28.83	63.82	99.8	100
C	1.12	15.33	14.53	40.35	100	100
Simulation results for the first slope estimator						
ST	-9.90	-9.90	24.05	24.05	32.4	32.4
R	0.10	0.34	10.37	13.11	99.1	100
A	0.08	-0.01	9.49	11.56	99.6	100
C	0.03	-0.14	5.70	7.39	100	100
Simulation results for the second slope estimator						
ST	-11.69	-11.69	33.44	33.44	0	0
R	0.00	-0.07	5.49	12.55	97.2	100
A	0.00	-0.43	5.07	11.25	98.0	100
C	-0.07	-0.95	2.41	7.08	100	100

Table 3: Simulation results linear regression for register to register of the **Case 2**: in terms of relative bias, RMSE and the actual coverage percentage for nomial 95% confidence intervals

Estimator	Relative Bias		Relative RMSE		Coverage	
	λ known	λ unknown	λ known	λ unknown	λ known	λ unknown
Simulation results for the intercept estimator						
ST	314.13	314.13	315.87	315.87	0	0
R	-1.09	0.06	38.52	82.16	99.9	100
A	-0.96	7.40	31.43	66.91	99.9	100
C	0.52	10.94	11.43	31.53	100	100
Simulation results for the first slope estimator						
ST	-10.17	-10.17	25.70	25.70	46.9	46.9
R	-0.20	-0.19	12.87	14.76	99.6	100
A	-0.18	-0.45	11.65	13.22	99.7	100
C	-0.12	-0.65	5.42	6.83	100	100
Simulation results for the second slope estimator						
ST	-19.66	-19.66	55.89	55.89	0	0
R	0.07	0.00	6.80	14.57	98.5	100
A	0.06	-0.46	5.54	11.86	99.4	100
C	-0.04	-0.69	1.81	5.52	100	100

Table 4: Simulation results linear regression for sample to register of the **Case 0** wtih complete linkage: in terms of relative bias, RMSE and the actual coverage percentage for nomial 95% confidence intervals

Estimator	Relative Bias		Relative RMSE		Coverage	
	λ known	λ unknown	λ known	λ unknown	λ known	λ unknown
Simulation results for the intercept estimator						
ST	184.71	184.71	187.36	187.36	0	0
R	-2.03	-7.02	34.24	71.10	98.5	99.9
A	-1.85	-0.47	32.30	63.24	99.3	99.9
C	0.09	11.81	17.27	40.35	100	100
Simulation results for the first slope estimator						
ST	-0.25	-0.25	8.48	8.48	95.6	95.6
R	-0.26	-0.29	8.30	8.35	99.7	100
A	-0.26	-0.29	8.30	8.34	99.7	100
C	-0.17	-0.23	5.41	5.74	100	100
Simulation results for the second slope estimator						
ST	-11.56	-11.56	33.04	33.04	0	0
R	0.12	0.43	5.30	12.28	97.2	100
A	0.11	0.02	4.90	10.85	97.9	100
C	-0.01	-0.75	2.24	6.91	100	100

Table 5: Simulation results linear regression for sample to register of the **Case 1** wtih complete linkage: in terms of relative bias, RMSE and the actual coverage percentage for nomial 95% confidence intervals

Estimator	Relative Bias		Relative RMSE		Coverage	
	λ known	λ unknown	λ known	λ unknown	λ known	λ unknown
Simulation results for the intercept estimator						
ST	187.83	187.83	190.87	190.87	0	0
R	1.45	2.22	36.44	72.40	98.5	100
A	1.41	7.81	34.32	65.79	99.4	100
C	1.34	15.59	18.43	42.39	100	100
Simulation results for the first slope estimator						
ST	-10.07	-10.07	24.34	24.34	29.4	29.4
R	-0.11	0.03	9.99	13.17	99.0	100
A	-0.10	-0.29	9.18	11.57	99.6	100
C	-0.09	-0.35	5.60	7.43	100	100
Simulation results for the second slope estimator						
ST	-11.69	-11.69	33.43	33.43	0	0
R	-0.06	-0.11	5.50	12.36	97.1	100
A	-0.06	-0.45	5.07	11.17	98.4	100
C	-0.06	-0.95	2.36	7.15	100	100

Table 6: Simulation results linear regression for sample to register of the **Case 2** wtih complete linkage: in terms of relative bias, RMSE and the actual coverage percentage for nomial 95% confidence intervals

Estimator	Relative Bias		Relative RMSE		Coverage	
	λ known	λ unknown	λ known	λ unknown	λ known	λ unknown
Simulation results for the intercept estimator						
ST	316.83	316.83	319.38	319.38	0	0
R	0.37	-5.17	46.61	89.02	98.8	100
A	0.44	4.46	40.35	72.36	99.5	100
C	1.08	10.12	15.69	32.56	100	100
Simulation results for the first slope estimator						
ST	-10.09	-10.09	25.36	25.36	47.4	47.4
R	-0.11	-0.02	12.36	14.55	99.1	100
A	-0.12	-0.33	11.16	12.86	99.5	100
C	-0.08	-0.60	5.07	6.27	100	100
Simulation results for the second slope estimator						
ST	-19.71	-19.71	56.06	56.06	0	0
R	0.06	0.41	7.21	15.23	98.3	100
A	0.05	-0.20	5.90	12.08	99.0	100
C	-0.03	-0.60	1.92	5.21	100	100

Table 7: Simulation results linear regression for sample to register of the **Case 0** with incomplete linkage: in terms of relative bias, RMSE and the actual coverage percentage for nomial 95% confidence intervals

Estimator	Relative Bias		Relative RMSE		Coverage	
	λ known	λ unknown	λ known	λ unknown	λ known	λ unknown
Simulation results for the intercept estimator						
ST	186.61	186.61	189.33	189.33	0	0
R	0.54	-5.33	34.41	74.38	99.1	100
A	0.56	1.17	32.44	66.57	99.3	100
C	1.07	12.40	16.93	40.94	100	100
Simulation results for the first slope estimator						
ST	0.11	0.11	8.52	8.52	95.6	95.6
R	0.08	0.08	8.49	8.57	99.3	100
A	0.07	0.08	8.48	8.57	99.3	100
C	0.05	0.08	5.56	5.91	100	100
Simulation results for the second slope estimator						
ST	-11.66	-11.66	33.33	33.33	0	0
R	-0.03	0.34	5.41	12.72	97.8	100
A	-0.03	-0.07	5.00	11.29	98.7	100
C	-0.07	-0.78	2.32	6.89	100	100

Table 8: Simulation results linear regression for sample to register of the **Case 1** with incomplete linkage: in terms of relative bias, RMSE and the actual coverage percentage for nomial 95% confidence intervals

Estimator	Relative Bias		Relative RMSE		Coverage	
	λ known	λ unknown	λ known	λ unknown	λ known	λ unknown
Simulation results for the intercept estimator						
ST	187.31	187.31	190.18	190.18	0	0
R	0.77	-1.61	35.55	75.43	94.0	100
A	0.83	4.32	33.48	68.18	95.4	100
C	0.26	1.45	6.08	10.13	75.5	100
Simulation results for the first slope estimator						
ST	-10.15	-10.15	24.68	24.68	30.3	30.3
R	-0.19	0.05	10.67	13.55	91.6	100
A	-0.19	-0.30	9.82	12.11	94.6	100
C	0.00	-0.01	2.00	2.08	73.3	100
Simulation results for the second slope estimator						
ST	-11.69	-11.69	33.43	33.43	0	0
R	-0.02	0.13	5.54	13.01	92.0	100
A	-0.03	-0.24	5.11	11.71	93.9	100
C	0.00	-0.08	0.79	1.63	75.7	100

Table 9: Simulation results linear regression for sample to register of the **Case 2** with incomplete linkage: in terms of relative bias, RMSE and the actual coverage percentage for nomial 95% confidence intervals

Estimator	Relative Bias		Relative RMSE		Coverage	
	λ known	λ unknown	λ known	λ unknown	λ known	λ unknown
Simulation results for the intercept estimator						
ST	318.07	318.07	320.64	320.64	0	0
R	2.91	-4.31	46.65	88.60	98.8	100
A	2.79	5.78	40.63	72.89	99.0	100
C	1.90	11.11	15.98	33.80	100	100
Simulation results for the first slope estimator						
ST	-10.25	-10.25	25.87	25.87	46.1	46.1
R	-0.29	0.10	12.84	15.17	99.4	100
A	-0.28	-0.25	11.61	13.38	99.7	100
C	-0.17	-0.62	5.37	6.47	100	100
Simulation results for the second slope estimator						
ST	-19.79	-19.79	56.30	56.30	0	0
R	-0.06	0.39	7.19	15.16	98.2	100
A	-0.05	-0.24	5.91	12.24	99.2	100
C	-0.05	-0.63	1.94	5.60	100	100

Figures

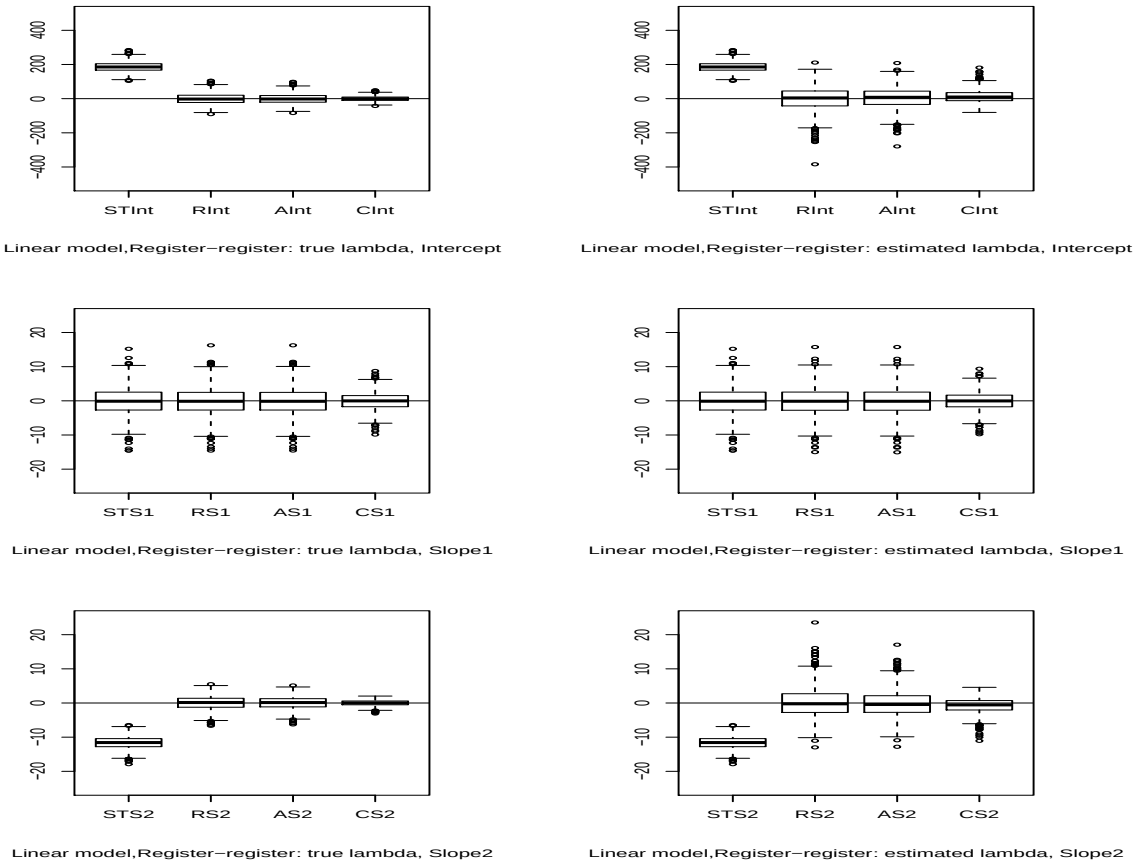


Figure 1: Simulated percentage relative errors for intercept and slope coefficients in linear regression under random linkage errors: Register - Register of the **Case 0**.

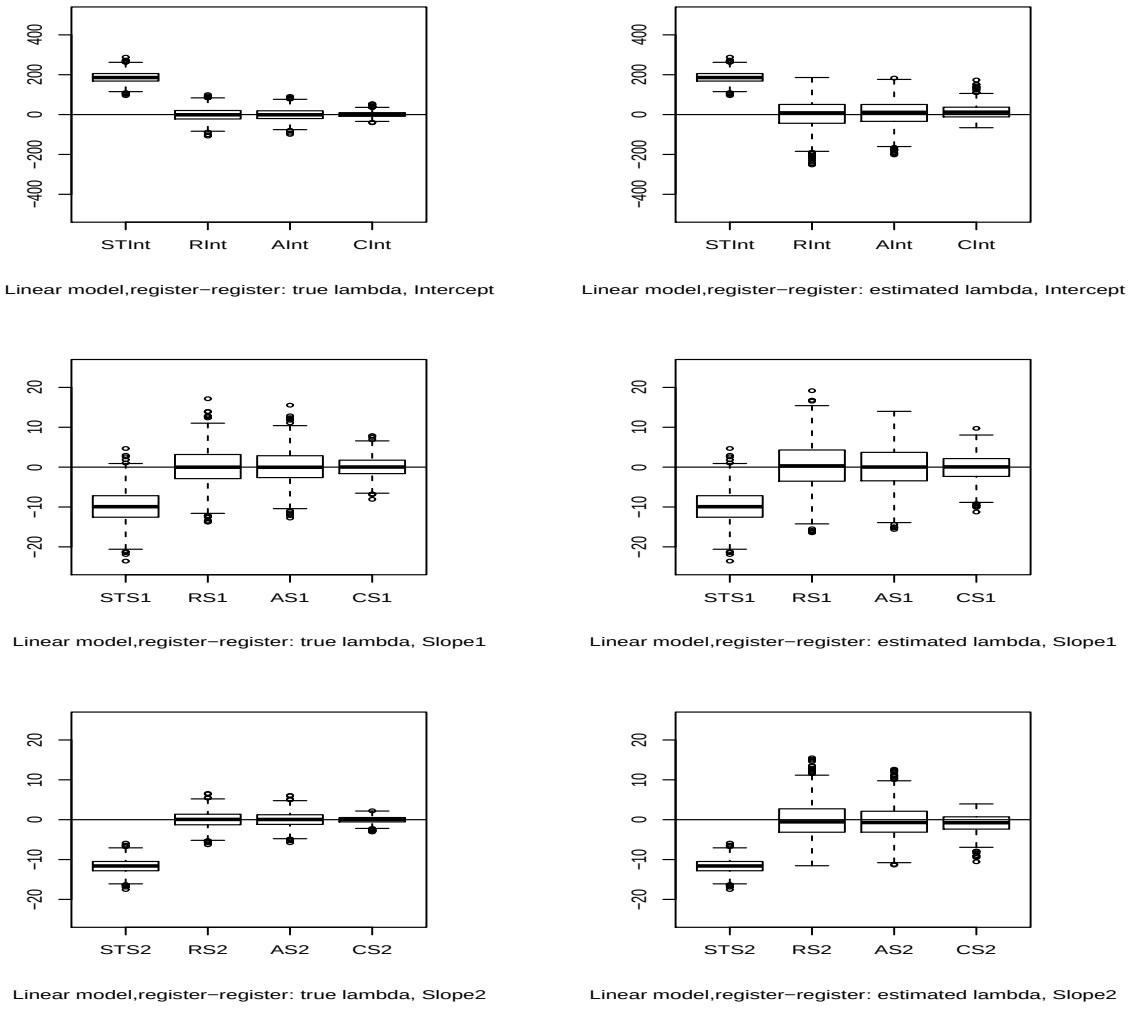


Figure 2: Simulated percentage relative errors for intercept and slope coefficients in linear regression under random linkage errors: Register - Register of the **Case 1**.

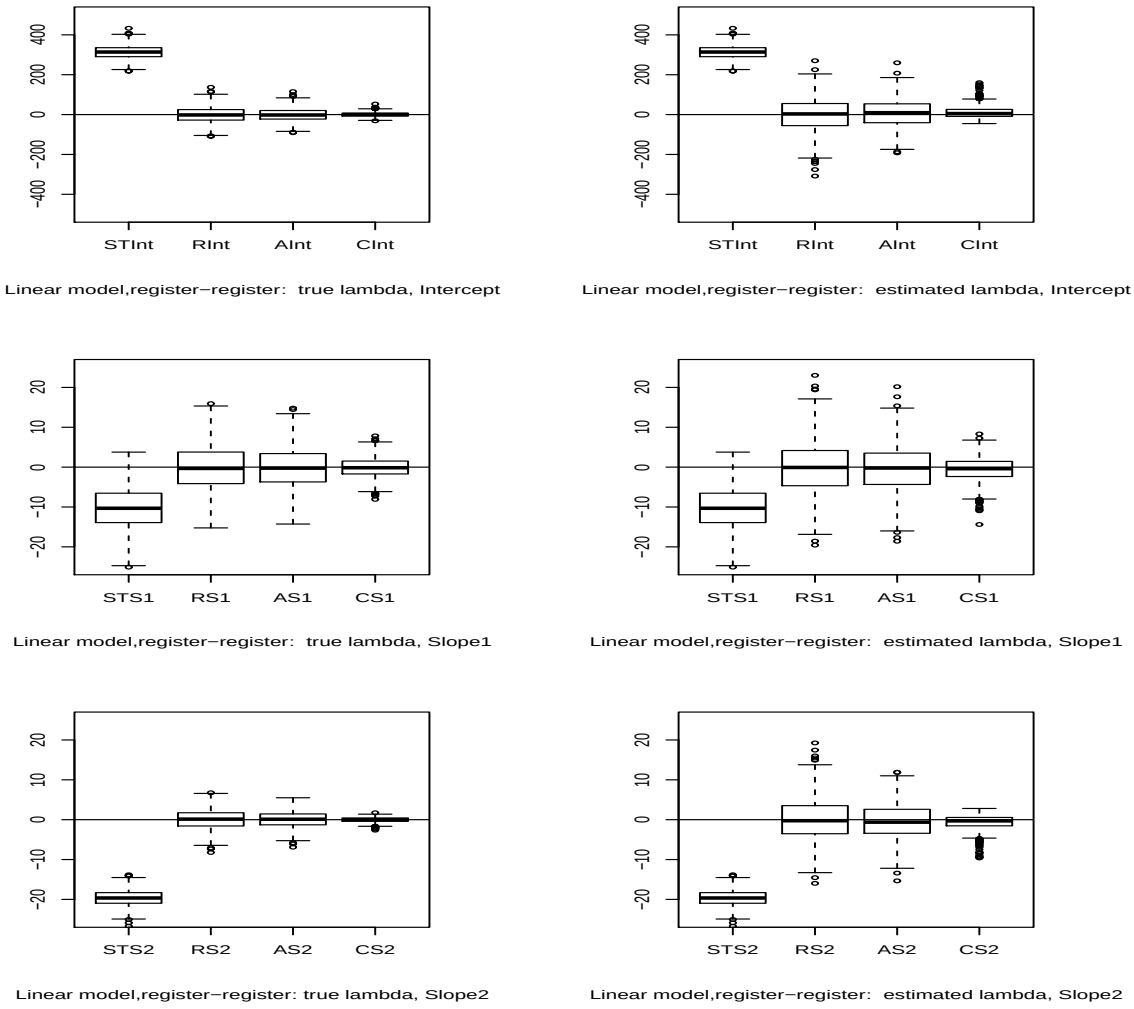


Figure 3: Simulated percentage relative errors for intercept and slope coefficients in linear regression under random linkage errors: Register - Register of the **Case 2**.

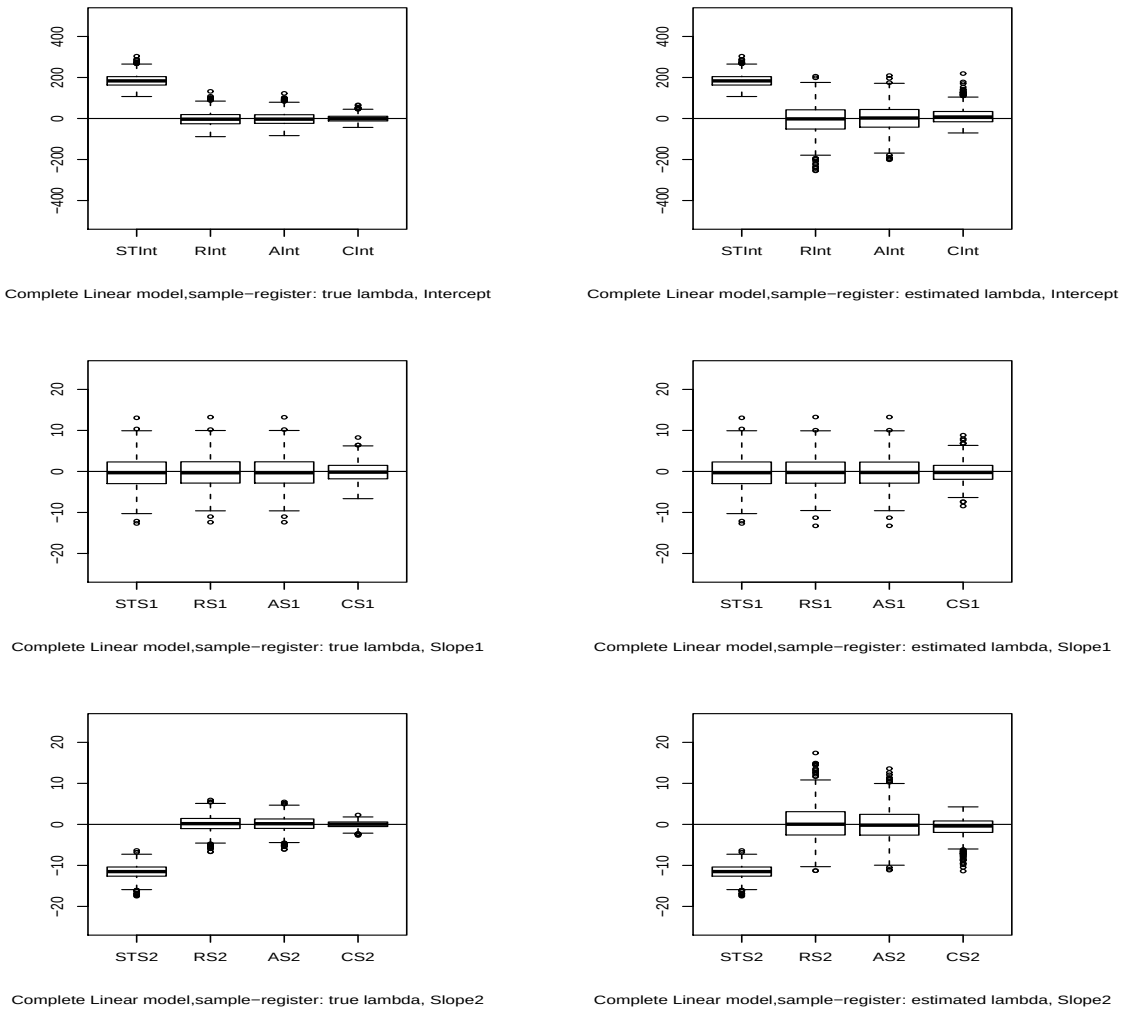
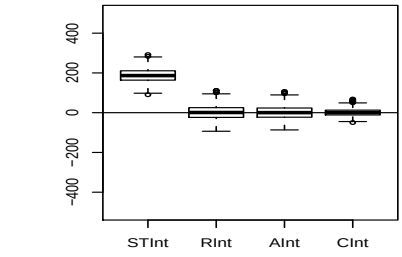
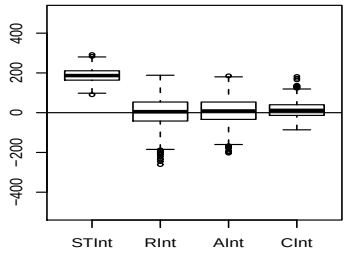


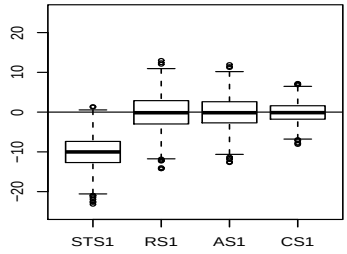
Figure 4: Simulated percentage relative errors for intercept and slope coefficients in linear regression under random linkage errors: Sample - Register of the **Case 0** with complete linkage.



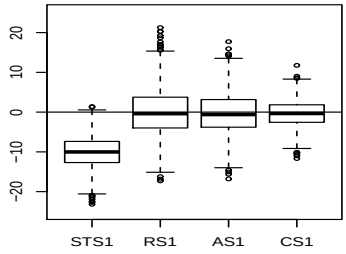
Complete Linear model,sample−register: true lambda, Intercept



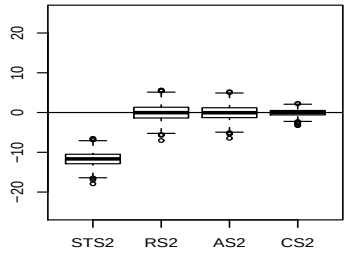
Complete Linear model,sample−register: estimated lambda, Intercept



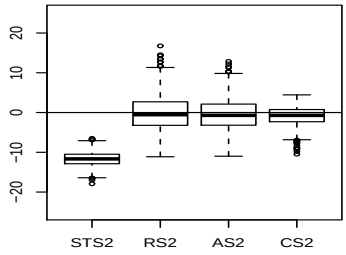
Complete Linear model,sample−register: true lambda, Slope1



Complete Linear model,sample−register: estimated lambda, Slope1



Complete Linear model,sample−register: true lambda, Slope2



Complete Linear model,sample−register: estimated lambda, Slope2

Figure 5: Simulated percentage relative errors for intercept and slope coefficients in linear regression under random linkage errors: Sample - Register of the **Case 1** with complete linkage.

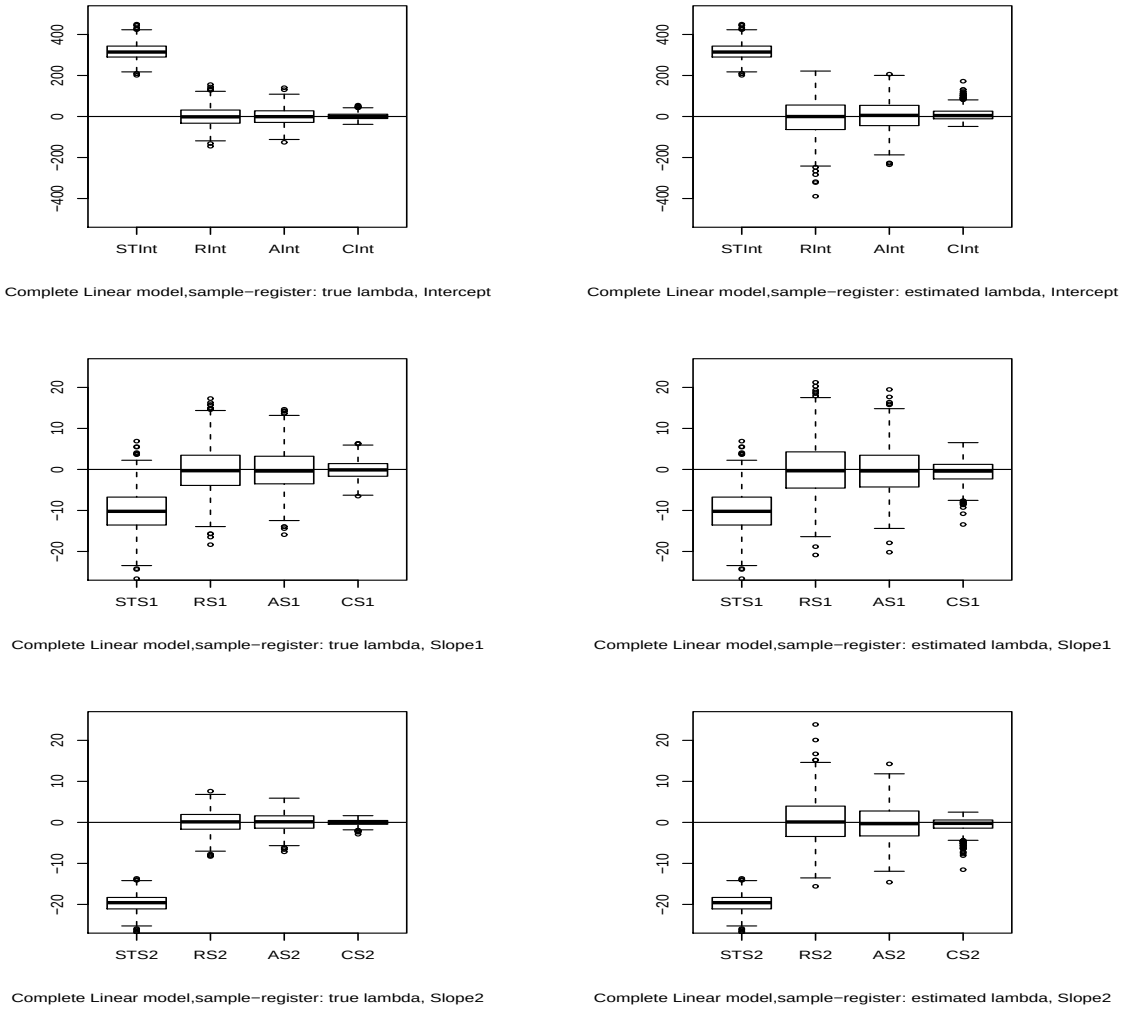
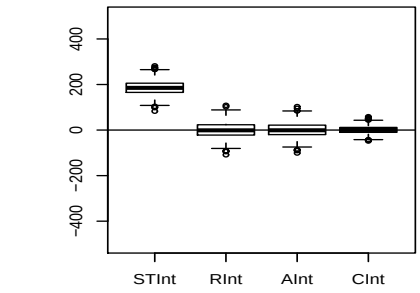
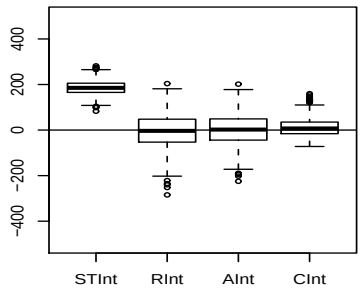


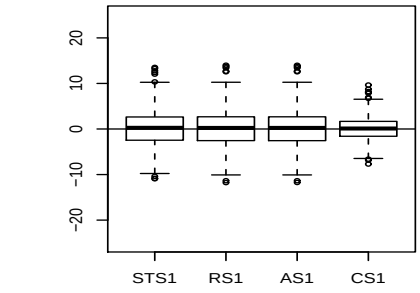
Figure 6: Simulated percentage relative errors for intercept and slope coefficients in linear regression under random linkage errors: Sample - Register of the **Case 2** with complete linkage.



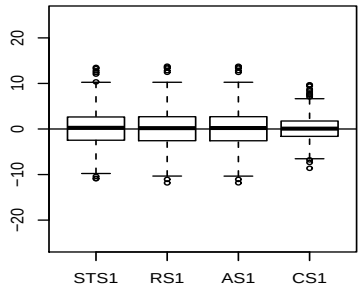
Incomplete Linear model,sample−register: true lambda, Intercept



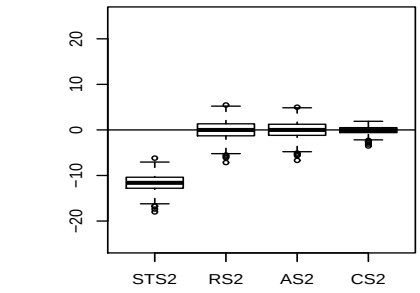
Incomplete Linear model,sample−register: estimated lambda, Intercept



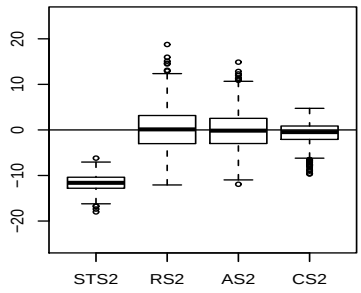
Incomplete Linear model,sample−register: true lambda, Slope1



Incomplete Linear model,sample−register: estimated lambda, Slope1

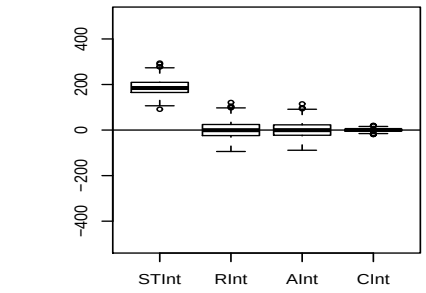


Incomplete Linear model,sample−register: true lambda, Slope2

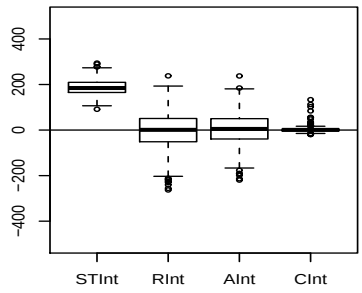


Incomplete Linear model,sample−register: estimated lambda, Slope2

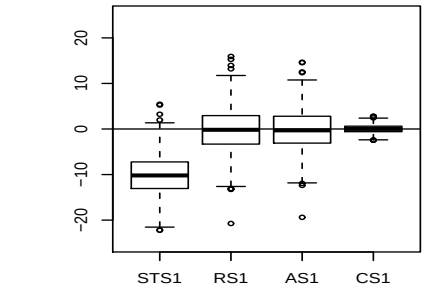
Figure 7: Simulated percentage relative errors for intercept and slope coefficients in linear regression under random linkage errors: Sample - Register of the **Case 0** with incomplete linkage.



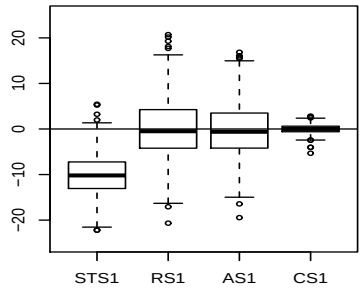
Incomplete Linear model,sample−register: true lambda, Intercept



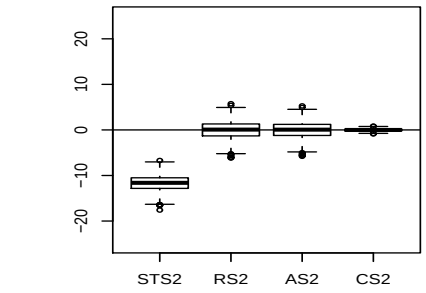
Incomplete Linear model,sample−register: estimated lambda, Intercept



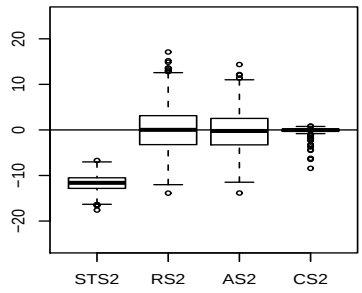
Incomplete Linear model,sample−register: true lambda, Slope1



Incomplete Linear model,sample−register: estimated lambda, Slope1

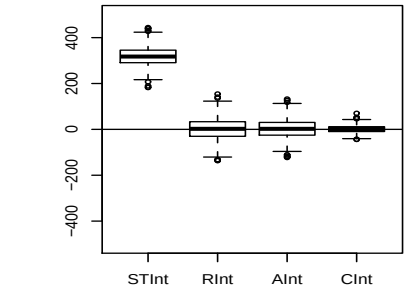


Incomplete Linear model,sample−register: true lambda, Slope2

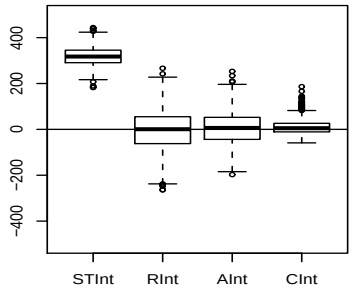


Incomplete Linear model,sample−register: estimated lambda, Slope2

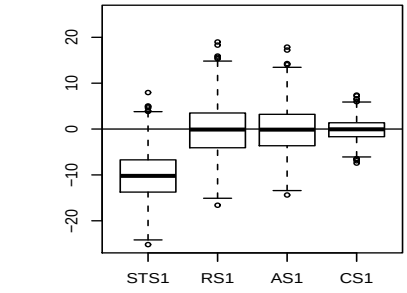
Figure 8: Simulated percentage relative errors for intercept and slope coefficients in linear regression under random linkage errors: Sample - Register of the **Case 1** with incomplete linkage.



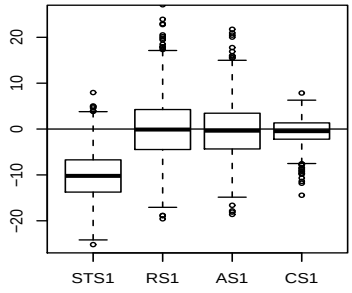
Incomplete Linear model,sample−register: true lambda, Intercept



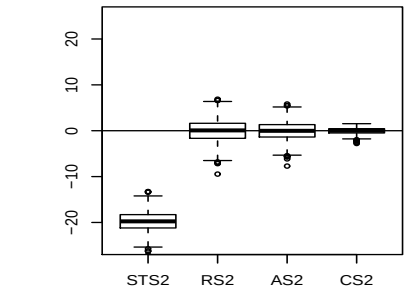
Incomplete Linear model,sample−register: estimated lambda, Intercept



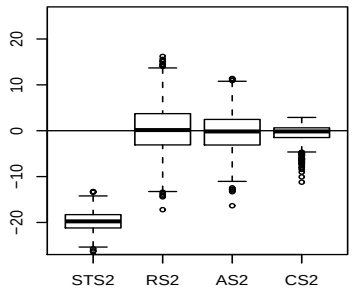
Incomplete Linear model,sample−register: true lambda, Slope1



Incomplete Linear model,sample−register: estimated lambda, Slope1



Incomplete Linear model,sample−register: true lambda, Slope2



Incomplete Linear model,sample−register: estimated lambda, Slope2

Figure 9: Simulated percentage relative errors for intercept and slope coefficients in linear regression under random linkage errors: Sample - Register of the **Case 2** with incomplete linkage.