Regression Analysis Using Longitudinally Linked Data

Kim, Gunky
University of Wollongong, Centre for Statistical and Survey Methodology
Northfields Av.
Wollongong 2522, Australia
E-mail: gkim@uow.edu.au

Chambers, Raymond
University of Wollongong, Centre for Statistical and Survey Methodology
Northfields Av.
Wollongong 2522, Australia
E-mail: ray@uow.edu.au

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. As an example, Brook *et al.* (2008) claims that the Western Australia Data Linkage Unit has been able to produce 708 research outputs, comprising journal articles, reports, presentations, conference proceedings and thesis, during the period of 1995-2003 though the comprehensive system of linked health records in Australia.

In Australia, 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. As a preparation of this project, a quality test for the the linkage process between a sample data and a census data has been done and the results are reported in Bishop and Khoo (2007). They found that their linkage procedure provides 87% correct linkage rate when names and address are used. These figures are quite common in many studies done in Australia. For example, Holman *et al.* (1998) showed that a linkage procedure done in 1996-1997 in Western Australia provides 87%, while the hospital morbidity data in Victoria in 1993-1994 showed 78-86% of accuracy. These rates will be lower when the actual names and addresses are not provided for linkage procedure, which is the most possible scenario for the Statistical Longitudinal Census Dataset of the Australian Bureau of Statistics due to the strict confidentiality regulations. 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 increases, the structure of linkage error will be more complicated and it will increase more bias and inefficiency for the modelling process.

The work of Neter *et al.* (1965) shows that a small amount of 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, 1997) and Lahiri and Larsen (2005), have tried to extend the work of Neter *et al.* (1965) on regression setting. However, their works only work for the situation where two data sets are to be merged. But, the linkage error structure of linked data sets, when the number of data sets to be merged 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. Furthermore, we also considered the case where one data set is a sample and others are registers, and some of sample cannot be linked to other registers.

Methodological development

The aim of this section is to develop an empirical Best Linear Unbiased Estimator of a regression model from the merged data sets when there exist some linkage errors among them. Especially, we are interested the case where a sample data set has be merged with other registers to form a new data set. To do that, we started with the case where all the data sets are registers. The main reason is that most of theoretical developments can be done under this situation and the sample to registers case can be extended from it easily. Further, to explain some of main ideas, we start with a ratio-type estimator.

1 A ratio-type estimator: when all data sets are registers

Note that our model is of the form

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

Suppose that $\boldsymbol{X}_1$ is the bench mark data set. When some of x_{1i} are incorrectly linked with corresponding y_i or with x_{2i}, our regression model becomes 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 $\boldsymbol{X}^*_q = (\mathbf{1}_q, \boldsymbol{X}_{1q}, \boldsymbol{X}^*_{2q})$. Note that $\boldsymbol{Y}^*_q$ and $\boldsymbol{X}^*_{2q}$ are the values that are linked, with some linkage errors, to the values of $\boldsymbol{X}_{1q}$ and, theoretically, one has

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

where A_q and B_{2q} are permutation matrices. In reality, $\boldsymbol{X}_{2q}$ 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,

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

Let

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

Because B_{2q} is unknown in general, we adapt the *non-informative linkage assumption*, that is,

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

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

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

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

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

Then, by non-informative linkage assumption on A_q, one has

$$(3) \quad E_{\boldsymbol{X}^*}(\boldsymbol{Y}^*_q) = E_{\boldsymbol{X}^*}(A_q\boldsymbol{Y}_q) = E_{\boldsymbol{X}^*}(A_q)E_{\boldsymbol{X}^*}(\boldsymbol{Y}_q) = E_{A_q}E_{\boldsymbol{X}^*}(\boldsymbol{Y}_q) = E_{A_q}\boldsymbol{X}^E_q\boldsymbol{\beta},$$

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 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

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

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

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

Let $\boldsymbol{f}_q = \boldsymbol{X}_q\boldsymbol{\beta}$, $\boldsymbol{f}^*_q = \boldsymbol{X}^*_q\boldsymbol{\beta}$ and $\boldsymbol{f}^E_q = \boldsymbol{X}^E_q\boldsymbol{\beta}$.

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

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

where

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

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

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

and,

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

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

$$\boldsymbol{V}_{C_{2q}} = (1-\lambda_{B_{2_q}})\operatorname{diag}\Big[\big(M_q - 1\big)^{-1}\big[(\lambda_{A_q}M_q - 1)d_i + M_q(1-\lambda_{A_q})\bar{d}_q\big]; i \in q\Big],$$

where $d_i = \lambda_{B_{2q}}(f_{B_{2q},i}^* - \bar{f}_{B_{2q}}^*)^2 + \bar{f}_{B_{2q}}^{*(2)} - (\bar{f}_{B_{2q}}^*)^2$ *and* $\bar{d}_q$ *is the mean of* $\{d_i; i \in q\}$.

2 The estimating function: when all data sets are registers

Godambe (1960) developed the estimating function method where MLE is a special case of this method. Since then there have been many researches in this direction. In this subsection, we briefly review the estimating function approach. Our main purpose is to develop an optimal estimating function approach called the *empirical Best Linear Unbiased Estimator.*

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 $\boldsymbol{f}_q^*(\boldsymbol{\theta}) = \boldsymbol{X}_q^*\boldsymbol{\beta}$. In this case, a *naive estimator* can be found by solving $\boldsymbol{H}^*(\theta) = 0$ where $\boldsymbol{G}_q = \big(\partial_\theta f_q\big)^T = \boldsymbol{X}^T$. 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) and (3), an unbiased estimator is of the form

$$\boldsymbol{H}_3^*(\boldsymbol{\theta}) = \sum_q \boldsymbol{G}_q(\boldsymbol{\theta})\big\{\boldsymbol{Y}_q^* - E_{A_q}\boldsymbol{f}_q^E(\boldsymbol{\theta})\big\}, \tag{4}$$

and an unbiased estimator $\hat{\boldsymbol{\theta}}^*$ can be defined as the the solution of

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

Theorem 2. *The asymptotic variance estimator for the solution of (4) is of the form*

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

where,

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

can be estimated by the same methods in the propositions 1.

3 Sample-registers case: When sample records are not perfectly linked

In this section, we consider the case where we only observe a sample $\boldsymbol{s}$ of records from the bench mark data set. Suppose that $\boldsymbol{X}_1$ is the bench mark data set. In this section we consider the case where some records in the sample $\boldsymbol{s}$ cannot be linked to a record in $\boldsymbol{X}_2$-register or $\boldsymbol{Y}$-register.

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 $\boldsymbol{w}_{slq}^T \boldsymbol{f}_{slq}^E(\theta)$[1] so that one has

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

where

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

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

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

where

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

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

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

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

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

Suppose that we know λ_{A_q} and $\lambda_{B_{2q}}$, and let $\hat{\boldsymbol{\theta}}^{s*}$ be the solution of the estimating equation. Then it is clear that the asymptotic variance is of the form

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

Theorem 3. *Under the assumption that $\boldsymbol{G}_{slq}$ is independent of $\boldsymbol{\theta}$, an estimator of the asymptotic variance of $\hat{\boldsymbol{\theta}}$ is of the form*

$$\hat{V}^{sl}(\hat{\boldsymbol{\theta}}^{s*}) = \Big[\sum_q \hat{\boldsymbol{G}}_{slq}\tilde{E}_{A_{slq}}\partial_\theta \boldsymbol{f}^E_{slq}(\hat{\boldsymbol{\theta}})\Big]^{-1}\Big[\sum_q \hat{\boldsymbol{G}}_{slq}\hat{\Sigma}_{slq}\hat{\boldsymbol{G}}^T_{slq}\Big]\Big(\Big[\sum_q \hat{\boldsymbol{G}}_{slq}\tilde{E}_{A_{slq}}\partial_\theta \boldsymbol{f}^E_{slq}(\hat{\boldsymbol{\theta}})\Big]^{-1}\Big)^T,$$

where

$$\hat{\Sigma}_{slq} = \hat{\sigma}^2\boldsymbol{I}_{slq} + \hat{\boldsymbol{V}}_{A_{slq}} + \hat{\boldsymbol{V}}_{C_{2slq}}.$$

[1] 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.

4 Simulation

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

$$\boldsymbol{Y}_t = \beta_0 + \beta_1 \boldsymbol{X}_t + \alpha_2 \boldsymbol{Y}_{t-1} + \boldsymbol{\epsilon}_t.$$

One scenario of this type model is that $\boldsymbol{Y}_t$ represents a health system costs of individuals for a given year that can be collected from an administrative data while $\boldsymbol{X}_t$ represents some health risk factor of individuals collected from hospital patients data. Because $\boldsymbol{Y}_t$ and $\boldsymbol{Y}_{t-1}$ are measured in different time, there are no autocorrelations between them so that the linear regression can be applied in this model.

In this simulation, we set $\boldsymbol{\theta} = (\beta_0, \beta_1, \alpha_2) = (1, 3, 0.7)$. $\boldsymbol{X}_t$ were drawn from the normal distribution with mean of 2 and the variance of 4. $\boldsymbol{\epsilon}_t$ were drawn from the standard normal distribution. $\boldsymbol{Y}_{t-1}$ has been generated from $\boldsymbol{Y}_{t-1} = 1 + 2\boldsymbol{X}_{t-1} + \boldsymbol{\epsilon}_{t-1}$ where the distribution of $\boldsymbol{X}_{t-1}$ is the same as the distribution of $\boldsymbol{X}_t$ and $\boldsymbol{\epsilon}_{t-1}$ were drawn from the standard normal distribution as well.

Here, we assume that $\boldsymbol{X}_t$ is the bench mark set and there exist linkage errors between $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t$ as well as between $\boldsymbol{X}_t$ and $\boldsymbol{Y}_{t-1}$ when all three data sets are linked together. To be consistent with the notations from the previous sections, we drop t, the notation for time. Further, we use $\boldsymbol{Y}$ for $\boldsymbol{Y}_t$, $\boldsymbol{X}_1$ for $\boldsymbol{X}_t$ and $\boldsymbol{X}_2$ for $\boldsymbol{Y}_{t-1}$.

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.

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).

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

In this simulation, we consider the case where all the data sets are registers as well as the case where a sample data sets are merged with other registers:

- For the case of all registers, we use three m-blocks of size 500 for each m-block.
- For the case of sample to registers case, the population size of all registers are the same and each m-block has 2000 records. Further, 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.

Under the above scenario, the estimators were independently simulated 1000 times. The regression parameters were estimated using the four estimators. The following plot boxes 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 cases. It is worth to note that the EBLUE(C) outperforms all other estimators in general. The figures 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.

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.

5 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.

Simulation results for the linear regression

Table 1: Simulation results linear regression : 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
Register to register: Simulation results for the intercept estimator						
ST	128.81	128.81	129.97	129.97	0	0
R	-0.64	0.23	19.62	38.98	97.2	100
A	-0.51	3.73	16.52	32.24	98.8	100
C	0.43	7.13	8.02	17.86	99.0	100
Register to register: Simulation results for the first slope estimator						
ST	-9.94	-9.94	17.51	17.51	0	0
R	0.07	-0.06	3.43	6.67	95.9	100
A	0.05	-0.35	3.03	5.69	95.6	100
C	-0.07	-0.74	1.38	3.03	97.7	100
Register to register: Simulation results for the second slope estimator						
ST	-19.78	-19.78	17.76	17.76	0	0
R	0.04	0.03	3.22	5.01	95.3	100
A	0.04	-0.48	2.62	4.07	97.3	100
C	-0.03	-0.81	1.36	2.30	97.8	100
Sample to register: Simulation results for the intercept estimator						
ST	129.75	129.75	130.98	130.98	0	0
R	0.74	0.99	20.32	39.87	95.5	100
A	0.61	4.25	17.69	33.48	96.0	100
C	0.71	7.13	8.70	18.16	97.9	100
Sample to register: Simulation results for the first slope estimator						
ST	-10.08	-10.08	17.71	17.71	0	0
R	-0.09	-0.10	3.24	6.89	96.3	100
A	-0.08	-0.37	2.86	5.91	96.8	100
C	-0.09	-0.72	1.39	3.13	97.8	100
Sample to register: Simulation results for the second slope estimator						
ST	-19.90	-19.90	16.88	16.88	0	0
R	-0.14	-0.20	3.33	5.08	95.0	100
A	-0.12	-0.67	2.71	4.18	96.6	100
C	-0.07	-0.84	1.38	2.36	98.8	100

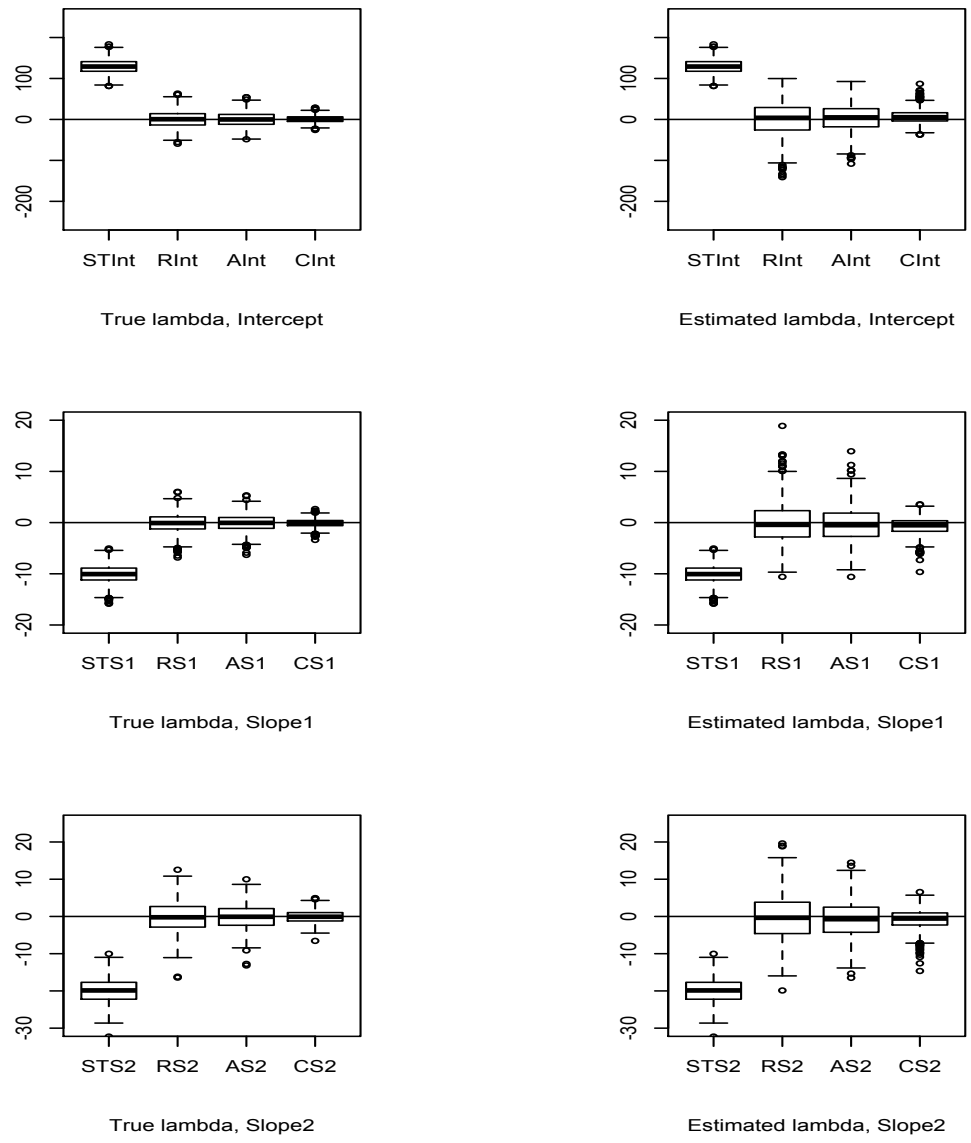

Figure 1: Simulated percentage relative errors for intercept and slope coefficients in linear regression under random linkage errors: Sample - Register with incomplete linkage.

References

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

Brook, E. L., Rosman, D. L., and Holman, C. D. J. (2008). Public good through data linkage: measuring research outputs from the western australian data linkage system. *AUSTRALIAN AND NEW ZEALAND JOURNAL OF PUBLIC HEALTH*, **32**(1), 19–23.

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

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

Holman, C. D. J., Bass, A. J., lan L. Rouse, and Hobbs, M. S. (1998). Population-based linkage of health records in western australia: development of a health services research linked database. *AUSTRALIAN AND NEW ZEALAND JOURNAL OF PUBLIC HEALTH*, **23**(5), 453–459.

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.