NIASRA

NATIONAL INSTITUTE FOR APPLIED STATISTICS RESEARCH AUSTRALIA

National Institute for Applied Statistics Research Australia

The University of Wollongong

Working Paper

17-13

Maximum Likelihood Logistic Regression with Auxiliary Information for Probabilistically Linked Data

Gunky Kim and Raymond Chambers

National Institute for Applied Statistics Research Australia, University of Wollongong, Wollongong NSW 2522. Phone +61 2 4221 5435, Fax +61 2 4221 4845. Email: anica@uow.edu.au

Maximum Likelihood Logistic Regression with Auxiliary Information for Probabilistically Linked Data

Gunky Kim and Ray Chambers
University of Wollongong, NSW 2522, AUSTRALIA

Abstracts

Despite the huge potential benefits, any analysis of probabilistically linked data cannot avoid the problem of linkage errors. These errors occur when probability-based methods are used to link or match records from two or more distinct data sets corresponding to the same target population, and they can lead to biased analytical decisions when they are ignored. Previous studies aimed at resolving this problem have assumed that the analyst has access to all the information used in the data linkage process. In practice, however, most analysts are secondary analysts, with only partial access to information about the linkage error structure. As a consequence, our previous research has focused on using an estimating equations approach to develop bias correction methods for secondary analysis of probabilistically linked data. In this paper we extend this approach to maximum likelihood estimation, using the missing information principle to accommodate the more realistic scenario of dependent linkage errors in both linear and logistic regression settings. We also develop the maximum likelihood solution when population auxiliary information in the form of population summary statistics is available. We also show that the main advantage from inclusion of population summary information is to correct small sample bias.

Keywords: Probabilistic record matching; Linkage errors; Regression modeling; Auxiliary data.

1. Introduction

Record linkage has been a very popular research tool in many areas such as health, economics and sociology. One of important issues in record linkage process is to deal with the record linkage errors. When there is no unique identifier, the probabilistic record linkage process would produce some unwanted record linkage errors. Most of recent research works have been focused on the reduction of record linkage error rates in the probabilistic record linkage process. However, this approach does not provide error free record linkage data and, as Neter et al. (1965) indicated, a small amount of linkage error in a linked data set could cause significant error when we ignore the linkage errors in the linked data. Inspired by Neter et al. (1965), Scheuren and Winkler (1993,1997) and Lahiri and Larsen (2005) have tried to adjust the bias due to the linkage errors in the linear regression setting using the weights used in the probabilistic record matching process. Their unbiased estimators are very useful when all the data sets and the weights used in the record matching process are available to the analyst. However, due to strict confidential policies, the second analyst cannot access to all the information used in the data linkage process. With only partial access to information about the linkage error structure, the second analyst cannot use the methods of Scheuren and Winkler (1993,1997) or Lahiri and Larsen (2005). To overcome this problem, our previous researches have focused on using an estimating equations approach to develop bias correction methods for secondary analysis of probabilistically linked data. In this paper we extend this approach to maximum likelihood estimation, using the missing information principle to accommodate the more realistic scenario of dependent linkage errors in both linear and logistic regression settings. A problem in sample data analysis is that small sample may not represent whole population and it may lead a small sample bias in its analysis.

Using the missing information principle can be used to avoid the unwanted small sample bias naturally.

2. Methodology developments

In this section, we explain how the missing information principle provides a natural mechanism for incorporating the auxiliary information, such as summary information, into likelihood analysis of the probabilistic linked sample data. The general assumptions for the analysis of the probabilistic linked sample data can be found in Kim and Chambers (2012), but to incorporate the auxiliary information, we further assume that the sum (or mean) of each data sets to be linked are known to the second analyst. Details are:

1. All registers have complete coverage of the target population and are of size *N*. In particular, for each distinct population unit there exist unique records in each of $\mathbf{Y}$ and $\mathbf{X}$ that correspond to this population unit.
2. $\mathbf{Y}$ and $\mathbf{X}$ can each be partitioned into *Q* 'match blocks' or '*m*-blocks' such that linkage errors occur only within them. That is, records in distinct *m*-blocks can never be linked. We denote quantities associated with the q^{th} *m*-block by a subscript of *q*. Thus the M_q records making up the q^{th} *m*-block within $\mathbf{X}$ are denoted $\mathbf{X}_q$, etc.
3. Linkage errors within an *m*-block are independent of any regression errors associated with observations from that *m*-block.

In many practical situations, a sample *s* of records from the register $\mathbf{X}$ is selected, and this sample is then independently linked to the two separate population registers $\mathbf{Y}$ and $\mathbf{X}$. In this second situation we make the following additional assumptions:

4. Not all records in $\mathbf{X}$ can be linked. However, this 'non-linkage' is at random, so the same regression model holds for the linked and non-linked records.
5. For each *m*-block, the means of $\mathbf{Y}_q$ and $\mathbf{X}_q$, $\bar{\mathbf{y}}_q$ and $\bar{\mathbf{X}}_q$, are known.

We consider a case where population values of two variables y and $\mathbf{X}$ are stored on two separate databases. A sample of record from $\mathbf{X}$ database is matched to the records on y database with some possible record linkage errors. Suppose that population values of two variables y and $\mathbf{X}$ are stored in *q* different *m*-blocks. Because of record linkage errors, the linked y-values in qth *m*-block, $\mathbf{y}_q^*$ is not the same as the true $\mathbf{y}_q$ that are not observable. However, theoretically, the relation between the linked y-values and the true y-values can be defined by

$$\mathbf{y}_q^* = \mathbf{A}_q \mathbf{y}_q,$$

where $\mathbf{A}_q$ is an unobservable random permutation matrix. One important issue in dealing with record linkage errors is to define the expectation of $\mathbf{A}_q$. The previous studies including Lahiri and Larsen (2005) used the weights used in the probabilistic record matching process to define the expectation of $\mathbf{A}_q$. However, this information is not available to the secondary analyst in most cases. One way to overcome this problem is to use the correct matching rate between y and $\mathbf{X}$ databases to define the expectation of $\mathbf{A}_q$. Let λ_q be the correct matching rate between y and $\mathbf{X}$ in qth *m*-block and let $E(\mathbf{A}_q) = \mathbf{T}_q = [t_{ij}^q]$. Define $t_{ii}^q = \lambda_q$ and $t_{ij}^q = (1-\lambda_q)/(M_q - 1)$ if $i \neq j$. This correct matching rate

λ_q can be obtained publically or can be estimated using a small audit samples. See Kim and Chambers (2012) for the details.

2.1 Linear regression case

Suppose that the values in y and $\mathbf{X}$ databases has the following relation

$$\mathbf{y}_q = \mathbf{X}_q\beta+\varepsilon_q,$$

where $var(\varepsilon_q)=\sigma^2$. Here, we are interested in a case where some of sample records from $\mathbf{X}$-database, $\mathbf{X}_{sq}$, has been linked to the records in y-database, $\mathbf{y}^*_{sq}$, with possible linkage errors. We assume that the selection of linked sample is noninformative. Initially, we start with the case where $\mathbf{X}_q$ is known. However, this assumption will be dropped later naturally because, in the end, the knowledge of $\mathbf{X}_q$ can be replaced by the values of $\mathbf{X}_{sq}$ and $\bar{\mathbf{X}}_q$ in our equations.

Note that, by assuming that the selection of linked sample is noninformative, we can partition y-values into the linked sample y-values, $\mathbf{y}^*_{sq}$, and the rest of y-values, $\mathbf{y}^*_{rq}$. Then, the permutation matrix and its expectation matrix can be partitioned accordingly as

$$\mathbf{A}_q = \begin{pmatrix} \mathbf{A}_{sq} \\ \mathbf{A}_{rq} \end{pmatrix} \text{ and } \mathbf{T}_{Aq} = \begin{pmatrix} \mathbf{T}_{Asq} \\ \mathbf{T}_{Arq} \end{pmatrix}.$$

Let $\mathbf{f}_q = E(\mathbf{y}_q \mid \mathbf{X}_q) = \mathbf{X}_q\beta$ and $\mathrm{T}_{Asq} = E(\mathbf{A}_{sq})$. Then, with the results in Kim and Chambers (2012),

$$\begin{aligned} E(\mathbf{y}^*_{sq} \mid \mathbf{X}_q) &= \mathbf{T}_{Asq}\mathbf{f}_q, \\ Var(\mathbf{y}^*_{sq} \mid \mathbf{X}_q) &= \sigma^2\mathbf{I}_{sq} + Var(\mathbf{A}_{sq}\mathbf{f}_q) \\ &= \sigma^2 \sum_{sq}, \end{aligned}$$

where details of $\sum_{sq}$ can be found in Chambers (2009). Further, with other variances and covariances, it follows that

$$\begin{pmatrix} \mathbf{y}_q \\ \mathbf{y}^*_{sq} \\ \overline{y_q} \end{pmatrix} \Bigg| \mathbf{X}_q \sim N\left\{ \begin{pmatrix} \mathbf{f}_q \\ \mathbf{T}_{Asq}\mathbf{f}_q \\ \overline{f_q} \end{pmatrix}, \sigma^2 \begin{pmatrix} \mathbf{I}_q & \mathbf{T}^T_{Asq} & \mathbf{1}_q / M_q \\ . & \sum_{sq} & \mathbf{T}_{Asq}\mathbf{1}_q / M_q \\ . & . & M_q^{-1} \end{pmatrix} \right\}$$

and it leads to the estimator of the form

$$\beta = \left[\sum_q \mathbf{X}^T_q \mathbf{R}_{sq} \begin{pmatrix} \mathbf{T}_{Asq}\mathbf{X}_q \\ \bar{X}^T_q \end{pmatrix} \right]^{-1} \sum_q \mathbf{X}^T_q \mathbf{R}_{sq} \begin{pmatrix} \mathbf{y}^*_q \\ \bar{y}_q \end{pmatrix},$$

where

$$\mathbf{R}_{sq} = \begin{pmatrix} \mathbf{T}_{Asq}^T & \mathbf{1}_q / M_q \end{pmatrix} \begin{pmatrix} \Sigma_{sq} & \mathbf{T}_{Asq}\mathbf{1}_q / M_q \\ \left[\mathbf{T}_{Asq}\mathbf{1}_q / M_q\right]^T & M_q^{-1} \end{pmatrix}^{-1}.$$

However, by the definition of $\mathbf{T}_{Asq}$,

$$\mathbf{X}_q^T \mathbf{R}_{sq} = (\mathbf{C}_{sq}^T \ \bar{\mathrm{X}}_q^T) \begin{pmatrix} \Sigma_{sq} & \mathbf{T}_{Asq}\mathbf{1}_q / M_q \\ (\mathbf{T}_{Asq}\mathbf{1}_q / M_q)^T & M_q^{-1} \end{pmatrix}^{-1},$$

where $\mathbf{C}_{sq} = (\lambda_q - \gamma_q)\mathbf{X}_{sq} + M_q \gamma_q \bar{X}_q$. Then, the estimation of β depends on $\mathbf{X}_q$ only through $\mathbf{X}_{sq}$ and $\bar{\mathrm{X}}_q$. Therefore, the estimation of β can be possible with linked sample data set and the summary statistics provided.

2.2 logistic regression case

Many survey data can be binary or categorical. In this section, we will explain how the missing information principle can be applied in logistic regression model when the summary statistics are available where record matching between two samples is not perfect. The logistic model we are using is of the form

$$y_i \mid X_i \sim \text{independent Bernoulli}\{\pi(x_i)\},$$
$$\pi(x_i) = \Pr(y_i = 1 \mid x_i) = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}.$$

Considering the case of linked sample, the usual score function for β can be separated with linked sample set and non-linked set such that

$$sc(\beta_0) = \sum_U y_i - \sum_U \pi(x_i),$$
$$sc(\beta_1) = \sum_s x_i \{y_i - \pi(x_i)\} + E_s\left(\sum_r x_i y_i\right) - \sum_r x_i \pi(x_i),$$

where U denotes the population set, s (r) denotes the set of linked sample (nonlinked) population units and E_s denotes the conditional expectation on linked sample. However, when there exist linkage errors,

$$sc(\beta_1) = \sum_q \mathbf{X}_{sq}^T \{\mathbf{y}_{sq} - \pi_{sq}(x)\} + E_s\left(\sum_r x_i y_i\right) - \sum_r x_i \pi(x_i)$$
$$= \sum_q \mathbf{X}_{sq}^T \{\mathbf{A}_{sq}\mathbf{y}_q^* - \pi_{sq}(x)\} + E_s\left(\sum_r x_i y_i\right) - \sum_r x_i \pi(x_i),$$

where $\pi_{sq}(x) = \left(\pi(x_1), \ldots, \pi(x_{m_q})\right)^T$. By considering the expectation of the unobservable $\mathbf{A}_{sq}$, the modified score function can be of the form

$$sc^*(\beta_1) = \sum_q \mathbf{X}_{sq}^T \{(\lambda_q - \gamma_q)\mathbf{y}_{sq}^* + M_q \gamma_q \bar{y}_q - \pi_{sq}(x)\} + E_s\left(\sum_r x_i y_i - \sum_r x_i \pi(x_i)\right),$$

where the first part of the score function depends only on the linked sample units. For the second part that is the expectation of nonlinked part, we adapt the methods explained in Chambers et al. (2012).

Their functional forms depend on the amount of information available. When all the values in the x-database are available instead of $\mathbf{X}_{sq}$ and $\bar{\mathbf{X}}_q$,

$$E_s\left(\sum_r x_i y_i - \sum_r x_i \pi(x_i)\right) = \sum_r x_i \pi(x_i)\left[\frac{1-\pi(x_i)-\{1-\pi(x_i)\}\hat{B}_r^{t_{ry}}}{\pi(x_i)+\{1-\pi(x_i)\}\hat{B}_r^{t_{ry}}}\right],$$

where t_{ry} is the sum of nonlinked y-values and $\hat{B}_r^{t_{ry}} = \exp\left[\frac{\sum_r \pi(x_i) - t_{ry}}{\sum_r \pi(x_i)\{1-\pi(x_i)\}}\right]$.

For the case where all the values in the x-database are not available but only $\mathbf{X}_{sq}$ and $\bar{\mathbf{X}}_q$ are available, the score functions obtained by using smear techniques are of the form

$$sc^*(\beta_0) = \sum_q \left(t_{y_q} - \sum_s \left\{ \pi(x_i) - \frac{M_q - m_q}{m_q}\pi(\tilde{x}_i)\right\}\right),$$

$$sc^*(\beta_1) = \sum_q \left[\begin{array}{l} \mathbf{X}_{sq}\left\{(\lambda_q - \gamma_q)\mathbf{y}_{sq}^* + M_q\gamma_q \bar{y}_q - \pi_{sq}(x)\right\}^T \\ \quad + \frac{M_q - m_q}{m_q}\sum_s \tilde{x}_i \pi(\tilde{x}_i)\left[\frac{1-\pi(\tilde{x}_i)-\{1-\pi(\tilde{x}_i)\}\tilde{B}_r^{t_{ry}}}{\pi(\tilde{x}_i)+\{1-\pi(\tilde{x}_i)\}\tilde{B}_r^{t_{ry}}}\right] \end{array}\right],$$

where $\tilde{x}_i = \bar{x}_r - \bar{x}_s + x_i$, t_{y_q} is the sum of y values in q-block and

$$\tilde{B}_r^{t_{ry}} = \exp\left[\frac{\frac{M_q - m_q}{m_q}\sum_s \pi(\tilde{x}_i) - t_{ry}}{\frac{M_q - m_q}{m_q}\sum_s \pi(\tilde{x}_i)\{1-\pi(\tilde{x}_i)\}}\right].$$

3. Simulation results

We use simulation to investigate the performance of our estimators in terms of relative bias and relative RMSE. We also compare the performance of our estimators that contain the missing information principle idea with other estimators.

3. 1 Simulation results for linear regression

The model used in this simulation is $y = 1 + 5x + \varepsilon$. There are 3 different blocks and the pairs (y_i^*, x_i) were generated according to the correct linkage rate between them. The set of correct linkage rate between y-database and X-database in each block is (1.0, 0.95, 0.75). The total number of simulations is 1000.

The estimators for this simulation are:

1. Full MLE: No linkage errors and all of y and X values are known. This is a benchmark estimator.
2. Naïve: Linkage errors exist, but ignore them in the analysis. Also assume that only linked sample values are presented.
3. Eblue: Adjust linkage errors using the correct linkage error, but do not use the missing

information principle. An empirical Best Linear Unbiased Estimator that was developed in Kim and Chambers (2012).

4. MIP-MLE: Adjust linkage errors and also use the missing information principle.

Table1: *Simulation results for linear regression, in terms of relative bias and relative RMSE where the population size is (1000,1000,1000) and true correct linkage rate is (1.0, 0.95,0.75) for each block.*

Estimator	Relative Bias		Relative RMSE	
	Intercept	slope	Intercept	slope
Linked sample size in each block (300,300,300)				
Full MLE	-0.09	0.02	3.71	2.89
Naïve	24.63	-9.92	25.93	23.08
Eblue	-0.45	0.10	7.88	6.19
MIP-MLE	-0.30	0.10	7.58	6.60
Linked sample size in each block (50,50,50)				
Full MLE	0.10	-0.02	3.68	2.89
Naïve	25.68	-10.18	32.63	27.69
Eblue	1.63	-0.57	20.26	15.97
MIP-MLE	0.81	-0.30	18.50	16.54
Linked sample size in each block (20,20,20)				
Full MLE	0.25	-0.04	3.51	2.74
Naïve	25.93	-10.46	41.30	34.38
Eblue	2.55	-1.21	32.34	25.60
MIP-MLE	0.92	-0.35	28.45	25.43

Our simulation results indicate that the missing information principle is not necessary when the sample size is large (when the linked sample size for each block is 300). However, as the linked sample size gets smaller, the performance of MIP-MLE gets better than that of Eblue. Especially, when the linked sample size is less than 50, we believe that MIP-MLE would provide most reasonable unbiased estimates of parameters.

3. 2 Simulation results for logistic regression

For the logistic regression, we consider the model of the form

$$E(y_i = 1 \mid x_i) = \pi(x_i) = \frac{\exp(-2 + x_i)}{1 + \exp(-2 + x_i)}$$

and the pairs $(y_i^*, \pi(x_i))$ were generated according to the correct linkage rate of (1.0,0.8,0.6). The total number of iterations is 200 for the linked sample size of (60,60,60) otherwise 500.

The estimators for this simulation are

1. Full MLE: No linkage errors and all of y and X values are known. This is a benchmark estimator.
2. Naïve: Linkage errors exist, but ignore them in the analysis. Also assume that only linked sample values are presented.
3. Sample adjust: Adjust linkage errors using the correct linkage error, but do not use the missing information principle. Only linked samples are used.
4. Eblue: Adjust linkage errors using the correct linkage error, but do not use the missing

information principle.

5. MIP-MLE1: Adjust linkage errors and also use the missing information principle. x-values and sum of y-values are known as auxiliary information.
6. MIP-MLE2: Adjust linkage errors and also use the missing information principle. Sum of x-values and sum of y-values are known.

Table2: *Simulation results for logistic regression, in terms of relative bias and relative RMSE: the population size is (1000,1000,1000) and true correct linkage rate is (1.0, 0.8,0.6) for each block.*

Estimator	Relative Bias		Relative RMSE	
	Intercept	slope	Intercept	slope
Linked sample size in each block (300,300,300)				
Full MLE	-0.011	0.005	0.123	0.045
Naïve	1.146	-0.434	1.160	0.438
Sample adjust	-0.045	-0.218	0.302	0.231
Eblue	-0.045	0.020	0.300	0.117
MIP-MLE1	-0.044	-0.090	0.282	0.132
MIP-MLE2	-0.050	-0.090	0.287	0.132
Linked sample size in each block (100,100,100)				
Full MLE	-0.005	0.001	0.119	0.043
Naïve	1.137	-0.427	1.178	0.428
Sample adjust	-0.091	-0.203	0.557	0.243
Eblue	-0.091	0.043	0.546	0.208
MIP-MLE1	-0.094	-0.003	0.496	0.190
MIP-MLE2	-0.107	-0.005	0.504	0.189
Linked sample size in each block (80,80,80)				
Full MLE	-0.004	0.001	0.119	0.042
Naïve	1.098	-0.417	1.156	0.433
Sample adjust	-0.136	-0.192	0.657	0.254
Eblue	-0.137	0.057	0.651	0.255
MIP-MLE1	-0.125	0.019	0.597	0.236
MIP-MLE2	-0.143	0.016	0.616	0.234
Linked sample size in each block (60,60,60): Number of iteration = 200				
Full MLE	0.007	-0.003	0.125	0.044
Naïve	1.145	-0.425	1.206	0.440
Sample adjust	-0.083	-0.194	0.833	0.299
Eblue	-0.074	0.050	0.811	0.338
MIP-MLE1	-0.098	0.021	0.790	0.321
MIP-MLE2	-0.148	0.018	0.853	0.322

Simulation results show that MIP-MLEs perform well even when the sample size is large (300 for each *m*-block). The results also show that the performances of estimators that include the MIP part are getting better as the sample size is getting smaller. However, we also notice that these methods fail to converge quite often if the linked sample size is small, say less than 50. This is the reason why the number of iteration for the last case is 200 rather than 500. Currently, we are searching for an alternative optimization technique to solve this problem.

4. Conclusions

We extend the linkage error correction methods in regression analysis using the missing information principle and we show that these methods performs well under reasonable situation. These methods are especially useful when the sample size is small. However, due to the extra terms in the formulae, these methods are not helpful, especially in linear regression, when the sample size is large (more than 300). Also, in logistic regression, these methods often fail to converge if the sample size is very small and further improvement is required in optimization technique. Despite of these problems, these methods are quite useful to adjust a small sample bias. Currently we are trying to extend these methods to accommodate multi-linked data set where number of linked data set is more than three.

References

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

Chambers, R. L., Steel, D. G., Wang, S. and Welsh, A. (2012). *Maximum Likelihood Estimation for Sample Surveys*. CRC Monographs on Statistics and Applied Probability. Chapman and Hall.

Kim, G. and Chambers, R. (2012). Regression analysis under incomplete linkage. *Computational Statistics and Data Analysis*, 56, 2756–2770.

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.