Maximum Likelihood Logistic Regression with Auxiliary Information for Probabilistically Linked Data

Gunky Kim[1] and Ray Chambers
University of Wollongong, NSW 2522, AUSTRALIA
[1] Corresponding author: Gunky Kim, e-mail: gkim@uow.edu.au

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. Our simulation results show that an incorrect assumption of independent linkage errors can lead to insufficient linkage error bias correction, while an approach that allows for correlated linkage errors appears to fully correct this bias. 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 has 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 protection 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). 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.

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.

To be more specific, 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 blocks. Because of record linkage errors, the linked y-values in q-block, $\mathbf{y}_q^*$ is not the same as the true $\mathbf{y}_q$ that are not observable. 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 $\mathbf{A}_q$ that are not available to the secondary analyst. One way to overcome this problem is to use the correct matching rate between y and x databases to define the expectation of $\mathbf{A}_q$. This correct matching rate 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 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 a sample of records from 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 sample is noninformative. Initially, we assume that $\mathbf{X}_q$ is known. However, this assumption will be dropped later naturally. 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),

$$E(\mathbf{y}_{sq}^* \mid \mathbf{X}_q) = \mathbf{T}_{Asq} \mathbf{f}_q,$$

$$Var(\mathbf{y}_{sq}^* \mid \mathbf{X}_q) = \sigma^2 \mathbf{I}_{sq} + Var(\mathbf{A}_{sq} \mathbf{f}_q)$$

$$= \sigma^2 \Sigma_{sq}.$$

Also with other variances and covariances, it follows that

$$\left(\begin{array}{c}\mathbf{y}_q\\ \mathbf{y}^*_{sq}\\ \overline{y}_q\end{array}\right)\Bigg|\mathbf{X}_q \sim N\left\{\left(\begin{array}{c}\mathbf{f}_q\\ \mathbf{T}_{Asq}\mathbf{f}_q\\ \overline{f}_q\end{array}\right),\sigma^2\left(\begin{array}{ccc}\mathbf{I}_q & \mathbf{T}^T_{Asq} & \mathbf{1}_q/M_q\\ . & \Sigma_{sq} & \mathbf{T}_{Asq}\mathbf{1}_q/M_q\\ . & . & M_q^{-1}\end{array}\right)\right\}$$

and it leads to the estimator of the form

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

where

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

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

$$\mathbf{X}^T_q\mathbf{R}_{sq}=(\mathbf{C}^T_{sq}\ \overline{\mathrm{X}}^T_q)\left(\begin{array}{cc}\Sigma_{sq} & \mathbf{T}_{Asq}\mathbf{1}_q/M_q\\ (\mathbf{T}_{Asq}\mathbf{1}_q/M_q)^T & M_q^{-1}\end{array}\right)^{-1},$$

where $\mathbf{C}_{sq}=(\lambda_q-\gamma_q)\mathbf{X}_{sq}+M_q\gamma_q\overline{X}_q$. Note that λ_q is the correct matching rate in q-block and $\gamma_q=(1-\lambda_q)/M_q$ where M_q is the total number of records in $\mathbf{X}_q$. Then, the estimation of β depends on $\mathbf{X}_q$ only through $\mathbf{X}_{sq}$ and $\overline{\mathrm{X}}_q$. Therefore, the estimation of β can be possible with linked sample data set and their summary statistics.

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 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 sample, the usual score function for β can be separated with sample set and nonsample set of the form

$$sc(\beta_0)=\sum_U y_i-\sum_U \pi(x_i),$$

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

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

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

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

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

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

where the first part of the score function depends only on the linked sample units. For the second part that is the expectation of nonsample 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,

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

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

For the case where all the values in the x-database are not available, the score functions obtained by using smear techniques are of the form

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

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

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 compare the performance of our estimators using the missing information principle (MIP) 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 simulations is 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. MIP-MLE1: Adjust linkage errors and also use the missing information principle. x-values and sum of y-values are known as auxiliary information.
5. 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	0.004	0.000	0.116	0.043
Naïve	1.143	-0.432	1.156	0.435
Sample adjust	-0.034	-0.217	0.440	0.244
MIP-MLE1	-0.084	-0.074	0.416	0.158
MIP-MLE2	-0.084	-0.076	0.416	0.159
Linked sample size in each block (200,200,200)				
Full MLE	-0.005	0.001	0.121	0.044
Naïve	1.142	-0.430	1.163	0.435
Sample adjust	-0.089	-0.201	0.768	0.284
MIP-MLE1	-0.097	-0.039	0.500	0.182
MIP-MLE2	-0.098	-0.041	0.502	0.181
Linked sample size in each block (100,100,100)				
Full MLE	-0.003	-0.002	0.123	0.044
Naïve	1.151	-0.431	1.195	0.442
Sample adjust	-0.310	-0.139	3.315	0.904
MIP-MLE1	-0.155	-0.027	0.791	0.307
MIP-MLE2	-0.150	-0.019	0.748	0.285

Simulation results show that MIP-MLEs perform well in general if the linked sample size is greater than or equal to 100. However, we also notice that the linkage error correction methods do not work if the linked sample size is small, say less than 50.

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. However, we still need 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.