مجله‌ی پژوهش‌های آماری ایران

پژوهشکده‌ی آمار

J. Statist. Res. Iran **9** (2012): 43–60

دوره‌ی ۹، شماره‌ی ۱، بهار و تابستان ۱۳۹۱، صص ۴۳–۶۰

Bayesian and Iterative Maximum Likelihood Estimation of the Coefficients in Logistic Regression Analysis with Linked Data

M. Mohammadzadeh[†,*] and A. Fallah[‡]

[†] Tarbiat Modares University
[‡] Imam Khomeini International University

Abstract. This paper considers logistic regression analysis with linked data. It is shown that, in logistic regression analysis with linked data, a finite mixture of Bernoulli distributions can be used for modeling the response variables. We proposed an iterative maximum likelihood estimator for the regression coefficients that takes the matching probabilities into account. Next, the Bayesian counterpart of the frequentist model is developed. Then, a simulation study is performed to check the applicability and performance of the proposed frequentist and Bayesian methodologies encountering mismatch errors.

Keywords. Record linkage; logistic regression; mixture distributions; EM algorithm.

MSC 2010: 62J05, 14M99.

1 Introduction

Record linkage is a widely used technique in applied statistics and computer sciences that brings information from different databases together, which are related to the same identities such as individuals, places, events and etc. It

[*] Corresponding author

is also a technique for recognition of duplications in a data set. Record linkage, for example, can be used to link between hospital data records and the mortality file included socio-demographic data and clinical diagnosis to discover the main causes of mortality. As another example, linking hospital and police records of traffic can be used to extend the knowledge about causes and outcomes of road traffic and injuries and for preparing more comprehensive picture on traffic injuries. Usually, each pair of records are compared based on their common fields; named as identifier variables, in the absence of any unique and free from error identifier. Newcombe et al. (1959) and Newcombe and Kennedy (1962) considered the record linkage as a statistical problem. While Fellegi and Sunter (1967) formulized this idea as a hypothesis testing problem and derived an optimal decision rule in the sense of minimum indeterminate record pairs for a pre-specified rate of linkage errors. On the other hand, Suchindran et al. (2001) proposed to use logistic regression as an alternative method for record linkage. And also, Verykios et al. (2003) considered the cost of different decisions in record linkage. Many authors notably Winkler (1993, 1994, 1995), Lahiri and Larsen (2000) and Larsen and Rubin (2001) discussed the use of mixture models in record linkage to classify record pairs. By definition, the data set obtained from a record linkage process needs to be referenced to as linked data. As a relatively new field in applied statistics, making various statistical analysis with linked data is one of the secondary aims of record linkage. Due to occurrence of linkage errors, it is necessary to account for these errors during any statistical analysis or inference with linked data. Most of the research in this area has been done in statistical modeling context. Scheuren and Winkler (1993, 1997) studied the effect of linkage errors on the ordinary least squares estimators of linear regression model parameters and proposed a least squares type estimator with adjusted bias. Considering linkage errors, Lahiri and Larsen (2005) proposed an unbiased least squares type estimator for regression coefficients. Additionally, Chambers (2009) used random permutation matrices for estimation of coefficients in linear and logistic regression with linked data. And also, Fallah and Mohammadzadeh (2010) considered the Bayesian regression analysis with linked data using mixture normal distributions. Tancredi and Liseo (2011) propose a hierarchical Bayesian approach for matching statistical records observed on different occasions that can be profitably adopted both in record linkage problems and in capture-recapture applications. In the present paper, we consider the logistic regression analysis with linked data. It is illustrated that, due to mixture structure of the response variable distribution for linked data, the estimation of the regression coefficients can be done based on the theory of finite mixture analysis for Bernoulli distributions. By using this idea, we propose an iterative maximum likelihood and also a Bayesian estimator for the coefficients of logistic regression model that completely considers the matching probabilities and leads to considerably good estimates.

In Section 2, the main idea of record linkage and its probabilistic model are illustrated. Regression analysis with linked data and least squares type estimators proposed in the literature are discussed in Section 3. Initially, in Section 4, a method for logistic regression analysis with linked data based on finite mixture analysis of Bernoulli distributions is proposed. The parameter estimation for proposed method is demonstrated by using the Expectation-Maximization (EM) algorithm. The Bayesian approach of the proposed methodology is developed in Section 5. Section 6 consists of a simulation study which is performed to show the performance of the proposed method, in both frequentist and Bayesian framework followed by the discussion in Section 7.

2 Probabilistic Record Linkage

Consider two files A and B that contain n_A and n_B records, respectively. The interest is in the identification of possible identical records in these two files. The set of all pair records $A \times B = \{(a,b); a \in A, b \in B\}$ can be partitioned to Matched $M = \{(a,b) \in A \times B; a = b\}$ and Non-matched $U = \{(a,b) \in A \times B; a \neq b\}$ sets, where a and b denote the records. Each pair of records (a,b) are compared based on their corresponding comparison vector $\boldsymbol{\gamma}_{ab} = (\gamma_{ab}^1, \dots, \gamma_{ab}^k)$, where k is the number of common fields in two records and γ_{ab}^i is an indicator variable with $\gamma_{ab}^i = 1$ if $x_{ai} = x_{bi}$ and 0 otherwise and x_{ai} and x_{bi} are the values of the ith field in records a and b, respectively. The conditional probability of comparison vector for each pair of records given M and U are compared with each other using the ratio $w(\boldsymbol{\gamma}_{ab}) = \log \frac{P(\boldsymbol{\gamma}_{ab}|M)}{P(\boldsymbol{\gamma}_{ab}|U)}$, which is called the total matching weight (Fellegi and Sunter, 1967). Under conditional independence assumption, i.e. independence of $\gamma_{ab}^i, i = 1, \dots, k$, conditional on M and U, the total matching weight can be written as the sum of individual matching weigh of fields, $w(\boldsymbol{\gamma}_{ab}) = \sum_{i=1}^k \log \frac{P(\gamma_{ab}^i|M)}{P(\gamma_{ab}^i|U)}$. Then, on the basis of the total matching weight, the following linkage rule,

$$d(\boldsymbol{\gamma}_{ab}) = \begin{cases} (a,b) & \text{non-matched} & w(\boldsymbol{\gamma}_{ab}) < C_1 \\ (a,b) & \text{possibly matched} & C_1 < w(\boldsymbol{\gamma}_{ab}) < C_2 \\ (a,b) & \text{matched} & w(\boldsymbol{\gamma}_{ab}) > C_2, \end{cases} \tag{1}$$

refers each pair records to one of three sets of matched, non-matched and possibly matched records. It is important to note that the records which fall in the possibly matched region must be checked manually. The linkage rule (1) is optimal in the sense of smallest possibly matched area for pre-specified linkage errors rates. To use this linkage rule we shall specify the threshold values and estimate the model parameters. The threshold values C_1 and C_2 are selected based on a priori determined admissible error rates (Fellegi and Sunter, 1967). The maximum likelihood estimation of the parameters of Fellegi-Sunter model, $P(\gamma^i_{ab}|M)$, $P(\gamma^i_{ab}|U), (a,b) \in A \times B$ and $P(M)$ can be obtained by using the EM algorithm (Jaro, 1989). Note that, the efficiency of this linkage rule solely depends on the accuracy of parameter estimations and selection of threshold values (Bellin and Rubin, 1995). Following many authors such as Burkard and Derigs (1980), Jaro (1989) and Lahiri and Larsen (2005), we assumed the one-to-one matching procedure. i.e. for each record in a file there may be only one possibly matched pair in the other file.

3 Linear Regression

Suppose two files A and B are linked and $n(\leqslant \min(n_A, n_B))$ pair of records are declared as matched. In this paper, our interest is in the regression analysis of a response variable y and a set of covariates $\boldsymbol{x}$, that are correspondent with two linked files A and B, respectively. Consider the linear regression model

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \dots, n, \tag{2}$$

where $\boldsymbol{x}_i = (x_{i1}, \dots, x_{ip})'$ is the vector of covariates, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)$ denotes the regression coefficients and $\varepsilon_i, i = 1, \dots, n$ are zero mean, fixed variance and uncorrelated errors. Due to statistical uncertainty of the probabilistic record linkage procedure, some pairs of $(y_i, \boldsymbol{x}_i)$ might not correspond; that is, $\boldsymbol{x}_i$ may correspond to another response value such as $y_{j(\neq i)}$. Suppose that $\boldsymbol{x}_i$ corresponds to y_i with probability $q_{ii} \in (0,1)$ and to y_j with probability $q_{ij} \in (0,1)$. According to Lahiri and Larsen (2005) we can define a new variable

$$z_i = \begin{cases} y_i & \text{with probability } \; q_{ii} \\ y_j & \text{with probability } \; q_{ij} \qquad (j \neq i) = 1, \ldots, n, \end{cases} \tag{3}$$

where $\sum_{j=1}^{n} q_{ij} = 1, i = 1, \ldots, n$ then the least squares estimator for $\boldsymbol{\beta}$ that is given by $\hat{\boldsymbol{\beta}}_{LS} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{z}$, where $\boldsymbol{X}_{n\times p} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$, is a biased estimator. Scheuren and Winkler (1993, 1997) studied the effect of mismatch errors on this estimator and proposed an adjusted biased estimator as $\hat{\boldsymbol{\beta}}_{SW} = \hat{\boldsymbol{\beta}}_{LS} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\hat{\boldsymbol{B}}$, where $\boldsymbol{B} = (B_1, \ldots, B_n)'$, $B_i = (q_{i1} - 1)y_i + \sum_{(j\neq i)=1}^{n} q_{ij}y_i, i = 1, \ldots, n$ and $\hat{\boldsymbol{B}}$ denotes the estimator of $\boldsymbol{B}$. Lahiri and Larsen (2000, 5005) proposed an unbiased estimator for $\boldsymbol{\beta}$ as $\hat{\boldsymbol{\beta}}_U = (\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{W}'\boldsymbol{z}$, where $\boldsymbol{W} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n)'$, $\boldsymbol{q}_i = (q_{i1}, \ldots, q_{in})$ and $\boldsymbol{w}_i = \boldsymbol{q}_i'\boldsymbol{X} = \sum_{j=1}^{n} q_{ij}\boldsymbol{x}_i'$, $i = 1, \ldots, n$ is a $1 \times p$ vector. Since B and $\boldsymbol{z}$ are explicit functions of linkage probabilities $q_{_{ij}}, i, j(\neq i) = 1, \ldots, n$, the estimators $\hat{\boldsymbol{\beta}}_{SW}$ and $\hat{\boldsymbol{\beta}}_U$ obviously are dependent on linkage probabilities. We note that, the linkage probabilities henceforth would be known since their corresponding estimates are provided by Fellegi-Sunter linkage procedure applied to the linked data sets before starting the logistic regression analysis on them.

4 Logistic Regression

Logistic regression is a standard approach for the analysis of binary or categorical data (see e.g. Hosmer and Lemeshow, 2000). In this Section we consider the logistic regression analysis with linked data. Consider the logistic regression model

$$y_i|\boldsymbol{x}_i \sim \text{Bernoulli}(\pi_{y_i}), \qquad i = 1, \ldots, n,$$

where $y_i \in \{0, 1\}$ is a binary response variable in file A and π_{y_i} denotes the success probability in a Bernoulli trail that is a nonlinear function of covariates $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})'$ in file B, given by

$$\pi_{y_i} = P(y_i = 1|\boldsymbol{x}_i, \boldsymbol{\beta}) = \frac{1}{1 + \exp\left(-\boldsymbol{x}_i'\boldsymbol{\beta}\right)}, \qquad i = 1, \ldots, n, \tag{4}$$

where $\boldsymbol{\beta}$ represents the unknown regression coefficients vector. According to the definition of the random variable z_i in (3), one can write

$$\begin{aligned} p(z_i) &= \sum_{j=1}^{n} P(z_i = y_j) P(z_i | z_i = y_j) \\ &= \sum_{j=1}^{n} q_{ij} \pi_{y_j}^{z_i} (1 - \pi_{y_j})^{1 - z_i}, \quad i = 1, \ldots, n. \end{aligned} \tag{5}$$

Hence, each random variable z_i has a mixture distribution of n Bernoulli distributions with mixing coefficients satisfying $0 < q_{ij} < 1$ and $\sum_{j=1}^{n} q_{ij} = 1$. We assumed a fixed number of components $n(\leqslant \min(n_A, n_B))$, due to known number of matched record pairs in an one-to-one matching procedure. Thus the log-likelihood function can be written as

$$\ell(\boldsymbol{\pi}_y | \boldsymbol{X}, \boldsymbol{z}) = \sum_{i=1}^{n} \log \left\{ q_{ii} \pi_{y_i}^{z_i} (1 - \pi_{y_i})^{1 - z_i} + \sum_{\substack{(j \neq i)=1 \\ i=1,\ldots,n}}^{n} q_{ij} \pi_{y_j}^{z_i} (1 - \pi_{y_j})^{1 - z_i} \right\},$$

where $\boldsymbol{\pi}_y = (\pi_{y_1}, \ldots, \pi_{y_n})$ are unknown success probabilities defined in equation (4), $\boldsymbol{X}_{n \times p} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$ denotes the matrix of covariates observations and $\boldsymbol{z}$ shows the vector of response variables. As mentioned in Section 3, the mixing coefficients in the mixture distribution are exactly the linkage probabilities in Fellegi-Sunter model and are assumed to be known in this stage. It should be noted that, there are some difficulties related to computing these probabilities, for more details see Bellin and Rubin (1995) and also Lahiri and Larsen (2005). We demonstrate here the application of the EM algorithm for driving the maximum likelihood estimate of model parameters. In the finite mixture analysis, the observable data $\boldsymbol{z} = (\boldsymbol{z}_1', \ldots, \boldsymbol{z}_n')'$ is usually considered as incomplete data, in the sense that the associated component-labels, are not available. This incomplete version of data can be completed by using an indicator vector $\boldsymbol{\delta}_i = (\delta_{i1}, \ldots, \delta_{in})'$ in the component generating $\boldsymbol{z}_i$, where $\delta_{ij} = 1$ if $z_i = y_i$ and $\delta_{ij} = 0$, if $z_i = y_j$, when $j \neq i$. Here, δ_{ij}, $j = 1, \ldots, n$, for a given i, is a random vector with Multinomial distribution $\mathcal{MN}(1, q_{i1}, \ldots, q_{in})$. The complete log-likelihood function, that is usually more suitable for application of EM algorithm, is therefore declared to be

$$\begin{aligned}\ell_C(\boldsymbol{\pi}_y|\boldsymbol{X},\boldsymbol{z},\boldsymbol{\Delta}) &= \sum_{i=1}^{n}\sum_{j=1}^{n}\delta_{ij}\left\{\log q_{ij} + \log\left(\pi_{y_j}^{z_i}(1-\pi_{y_j})^{1-z_i}\right)\right\}\\ &= \sum_{i=1}^{n}\sum_{j=1}^{n}\delta_{ij}\left\{\log q_{ij} + z_i\log\left(\frac{\pi_{y_j}}{1-\pi_{y_j}}\right) + \log(1-\pi_{y_j})\right\}.\end{aligned} \tag{6}$$

Here the matrix $\boldsymbol{\Delta} = (\delta_{ij})_{n\times n}$ is treated as unobservable or missing data, usually. The EM algorithm replaces each missing data δ_{ij} by its expected value, in Expectation step (E-step) and then on the basis of the estimated missing data in the E-step, it finds the parameter values that maximizes the completed log-likelihood function, in the Maximization step (M-step) (Dempster et al., 1977). It is easy to show that for complete log-likelihood (6), the E-step replaces each δ_{ij} by

$$\frac{q_{ij}\pi_{y_j}^{z_i}(1-\pi_{y_j})^{1-z_i}}{\sum_{k=1}^{n} q_{ik}\pi_{y_k}^{z_i}(1-\pi_{y_k})^{1-z_i}},$$

and the M-step drives the maximum likelihood estimates of the mixing coefficients,

$$\hat{\pi}_{y_j}(\boldsymbol{z}) = \frac{\sum_{i=1}^{n}\delta_{ij}z_i}{\sum_{i=1}^{n}\delta_{ij}}, \qquad j=1,\ldots,n.$$

Hence, the maximum likelihood estimate of the vector parameter $\boldsymbol{\pi}_y = (\pi_{y_1},\ldots,\pi_{y_n})'$ can be written as

$$\begin{aligned}\hat{\boldsymbol{\pi}}_y(\boldsymbol{z}) &= (\hat{\pi}_{y_1}(\boldsymbol{z}),\ldots,\hat{\pi}_{y_n}(\boldsymbol{z}))'\\ &= \left(\frac{\sum_{i=1}^{n}\delta_{i1}z_i}{\sum_{i=1}^{n}\delta_{i1}},\ldots,\frac{\sum_{i=1}^{n}\delta_{in}z_i}{\sum_{i=1}^{n}\delta_{in}}\right)'\\ &= \begin{pmatrix} \frac{\delta_{11}}{\sum_{i=1}^{n}\delta_{i1}} & \cdots & \frac{\delta_{n1}}{\sum_{i=1}^{n}\delta_{i1}}\\ \vdots & \ddots & \vdots\\ \frac{\delta_{1n}}{\sum_{i=1}^{n}\delta_{in}} & \cdots & \frac{\delta_{nn}}{\sum_{i=1}^{n}\delta_{in}}\end{pmatrix}\begin{pmatrix} z_1\\ \vdots\\ z_n\end{pmatrix}\\ &= \boldsymbol{V}'\boldsymbol{z},\end{aligned}$$

where the $n \times n$ matrix $\boldsymbol{V}$ depends on the missing data $\boldsymbol{\Delta}$. Using equation (4) it would be concluded that $\boldsymbol{x}_i'\boldsymbol{\beta} = \text{logit}(\pi_{y_i})$, $i=1,\ldots,n$. Hence, one can write $\boldsymbol{X\beta} = \boldsymbol{\Lambda}_y$ or $\boldsymbol{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Lambda}_y$, where $\boldsymbol{\Lambda}_y = (\text{logit}(\pi_{y_1}), \ldots, \text{logit}(\pi_{y_n}))$. Therefore given the estimate $\hat{\boldsymbol{\beta}}^{(t)}$ (or equivalently $\hat{\boldsymbol{\Lambda}}_y^{(t)}(\boldsymbol{z})$) in t-th step, we obtain the following estimate in $(t+1)$-th step

$$\hat{\boldsymbol{\beta}}^{(t+1)} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\hat{\boldsymbol{\Lambda}}_y^{(t)}(\boldsymbol{z}),$$

where $\hat{\boldsymbol{\Lambda}}_y^{(t)}(\boldsymbol{z})$ denotes the estimate of $\boldsymbol{\Lambda}_y$ in the t-th step. The algorithm starts with some initial guess for the maximum likelihood estimation of parameters and proceeds iteratively, considering a convergence criterion (Dempster et al., 1977). Under fairly general conditions, the EM algorithm is numerically stable and robust against the choice of initial values and has reliable global convergence. Of course, the choice of initial values affects the convergence rate of algorithm and it will converge slowly for poor choice of initial values. Dempster et al. (1977) show that the rate of convergence is linear for this algorithm and it depends on the proportion of missing data. Then it is recommended, usually, to try many initial values to find the best convergence rate.

5 Bayesian Approach

For the purpose of developing the Bayesian counterpart of the previously discussed logistic regression model, it is important to note that according to (5), for given probabilities (q_{ii}, q_{ij}) and values (y_i, y_j), z_i is a random variable with conditional probability density function as follow

$$z_i|\boldsymbol{x}_i, \boldsymbol{\beta} \sim p(z_i|\boldsymbol{x}_i, \boldsymbol{\beta}) = \sum_{j=1}^{n} q_{ij} \left(\frac{1}{1+\exp(-\boldsymbol{x}_i'\boldsymbol{\beta})}\right)^{z_i} \left(\frac{\exp(-\boldsymbol{x}_i'\boldsymbol{\beta})}{1+\exp(-\boldsymbol{x}_i'\boldsymbol{\beta})}\right)^{1-z_i}.$$

If we consider that the vector of regression coefficients $\boldsymbol{\beta}$ has a multivariate normal prior distribution with mean vector $\boldsymbol{\theta}$ and covariance matrix $\boldsymbol{\Sigma}$, then, the posterior distribution is given by

$$\begin{aligned} p(\boldsymbol{\beta}|\boldsymbol{z}, \boldsymbol{X}) &\propto p(\boldsymbol{z}|\boldsymbol{X}, \boldsymbol{\beta})p(\boldsymbol{\beta}) \\ &= \left(\prod_{i=1}^{n}\left\{\sum_{j=1}^{n} q_{ij}\, \frac{(\exp(-\boldsymbol{x}_i'\boldsymbol{\beta}))^{1-z_i}}{1+\exp(-\boldsymbol{x}_i'\boldsymbol{\beta})}\right\}\right) \times \phi_p(\boldsymbol{\beta}; \boldsymbol{\theta}, \boldsymbol{\Sigma}), \end{aligned}$$

where $\phi_p(\cdot;\cdot,\cdot)$ denotes the p-dimensional normal density function. The complexity of the posterior distribution for finite mixture models preclude analytic solutions for parameter estimations (Marin et al., 2005). So, in order to sample from this posterior distribution, we use a Metropolis-Hastings algorithm with a multivariate students-t proposal density of the form

$$q(\boldsymbol{\beta}|\boldsymbol{\beta}_0) \propto \left(1 + \frac{(\boldsymbol{\beta} - \boldsymbol{\beta}_0)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)}{k}\right)^{-\frac{k+p}{2}},$$

where k is the degrees of freedom which must be less than 2 so that the jump probabilities vary between 0.25 and 0.5. Gelmann et al. (1996) discussion about the most efficient jumping rules in Metropolis-Hastings algorithm describes this constraint on k. In general finding values of hyperparameters that represent actual prior information can be difficult. Following Kass and Wasserman (1995) we replaced $\boldsymbol{\theta}$ by $\hat{\boldsymbol{\beta}}_{OLS}$ in order to centering the prior distribution of $\boldsymbol{\beta}$ around the OLS estimator which is a naive estimator for $\boldsymbol{\beta}$ and does not take into account the linkage errors. We also used the idea of Zellner (1986) to set $\boldsymbol{\Sigma} = g\hat{\boldsymbol{\Sigma}}_{OLS}$ where g is any positive value. This choice results in a version of prior distribution that is widely used for regression analysis and called g-prior. The fixed value g as a blow-up factor is set to be 100 in order to construct a reasonable noninformative prior distribution.

6 Simulation

In this Section, we performed a simulation study to illustrate the performance of proposed methodology to deal with mismatch errors. We note here that, the effects of linkage errors on estimates of regression coefficients in logistic regression analysis is affected by frequencies of 0 and 1 in the response variable observations. This is due to the fact that, when the frequency of 1 is considerably more than the frequency of 0 or vice versa; the linkage errors may link some equal values that correspond to different observations. In other words, when response value $y_i = 1$ (or 0) is misplaced by $y_{(j \neq i)} = 1$ (or 0), it would not affect the ensuing inferences. Whereas, in regression analysis with linked data for count variables the probability of observing equal response observations is small. Also, for continuous response variables this probability is zero. Since, the most and the least variations of the Bernoulli variables are respectively related to middle and boundary values of the success probability, we choose two values of 0.2 and 0.5 for the success probability of Bernoulli trials π_y. We follow the format of George and McCulloch (1993) and Raftery et al. (1997) in our simulation. We simulated n independent values for each covariate x_1, x_2 and x_3 from normal distribution with mean $\mu = \frac{\log(\pi_y^{-1}-1)}{\beta_1+\beta_2+\beta_3}$ and variance 1, where the μ is determined in such a way that the success probability π_y would be approximately equal to 0.2 or 0.5. The response was generated using the model

$$y|\boldsymbol{x}, \boldsymbol{\beta} \sim \text{Bernoulli}([1+\exp(-(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)]^{-1}),$$

taking the regression coefficients $\beta_1 = 3$, $\beta_2 = 1$ and $\beta_3 = 2$. Then, using equation (3) the vector $\boldsymbol{y}$ is transformed to $\boldsymbol{z}$. In order to assess the effects of linkage errors on proposed methods, we also considered four different range of variations for Matching Rate. We generated the true matching probabilities, $q_{ii}, i = 1, \ldots, n$, uniformly on $\mathcal{A}_1 = (0.90, 0.95)$, $\mathcal{A}_2 = (0.85, 0.90)$, $\mathcal{A}_3 = (0.80, 0.85)$ and $\mathcal{A}_4 = (0.75, 0.80)$ and then for each i=1,...,n, the false matching probabilities, $q_{ij(\neq i)}, j = 1, \ldots, n$, are generated uniformly on $(0, 1 - q_{ii})$. To consider the condition $\sum_{j=1}^{n} q_{ij} = 1$, all matching probabilities were standardized via dividing each q_{ij} by $\sum_{j=1}^{n} q_{ij}$. A computation of the Standard Deviation (SD), Bias, Root of Mean Square Error (RMSE) of the maximum likelihood estimator and Bayes estimator along with Iteratively Re-weighted Least Squares (IRLS) estimator of logistic regression coefficients has taken place for different sample sizes n=50, 100, 200, 500, and the mean values of these criteria are presented in Tables 1 to 3. Given a current estimate $\hat{\boldsymbol{\beta}}$ of the regression coefficients, the IRLS estimator is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{R}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{R}\boldsymbol{u},$$

where $\boldsymbol{R}$ is diagonal matrix of weight with components $r_{ii} = \hat{\pi}_{y_i}(1-\hat{\pi}_{y_i}), i = 1, \ldots, n$ and $\boldsymbol{u} = (u_1, \ldots, u_n)$ denotes the working dependent data in which $u_i = \boldsymbol{x}_i'\hat{\boldsymbol{\beta}} + \frac{y_i - \pi_{y_i}}{\pi_{y_i}(1-\pi_{y_i})}, i = 1, \ldots, n$. The resulting estimate of $\boldsymbol{\beta}$ is used to obtain improved fitted values and the procedure is iterated to convergence (McCullagh and Nelder, 1985). The estimates of regression coefficients are obtained by using the glm function of R package. Figure 1 shows the convergence of a typical Metropolis-Hastings algorithm constructed for sampling from posterior distributions of regression coefficients $\boldsymbol{\beta}$ for 4 different rates of matching probabilities.

The number of iteration for computing standard deviations in all cases was fixed to be 500. For each sample size n in each table, the rows correspond to success probability 0.5 and the column $\mathcal{A}_4$ show the cases that the logistic regression analysis with linked data encounters the maximum matching error rate. While the rows correspond to success probability 0.2 and column $\mathcal{A}_1$ show the cases with minimum matching error rate.

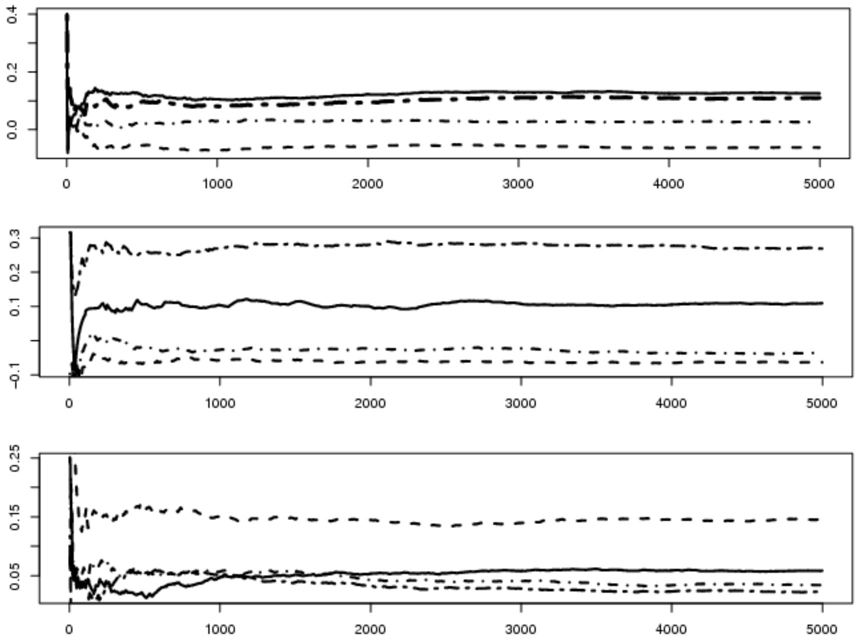

Figure 1. Convergence plots of the generated samples from posterior distributions of β_1 (top), β_2 (middle) and β_3 (bottom) for different rates of matching probabilities $\mathcal{A}_1$ to $\mathcal{A}_4$.

As one can see, in Table 1, the values of standard deviations for the ML and Bayes estimators are both less than the corresponding values for the IRLS estimator. Although, the standard deviation of Bayes estimator is somehow larger than ML estimator, but they have approximately similar efficiency in terms of standard deviation. When Matching errors increases from $\mathcal{A}_1$ to $\mathcal{A}_4$, the SD of IRLS estimator increases where as it remains without considerable changes for ML and Bayes estimators. The results presented in Table 2 show that all of the three estimators have negative biases value. Despite of the results of Table 1, the IRLS and ML estimator look more reliable than Bayes estimator. In terms of bias criterion it seems that it works very good for negligible matching errors and the ML estimator. Although for large matching errors the Bayes estimator dominates IRLS estimator based on biasedness. In general and in terms of RMSE of estimators presented in Table 3, one can say that the Bayes and especially the ML estimator are more efficient than IRLS estimator for large values of matching errors and can be preferred compared to this estimator. This is absolutely due to this fact that, despite of the IRLS estimator, the ML and Bayes estimators truly take

Table 1. The mean of SD of regression coefficients estimators for various values of sample size, success probability and matching error rate.

			Matching Rates			
n	π_y	Estimator	$\mathcal{A}_1$	$\mathcal{A}_2$	$\mathcal{A}_3$	$\mathcal{A}_4$
50	0.2	IRLS	0.2704	0.4039	0.4414	0.4631
		ML	0.1785	0.2104	0.2127	0.2212
		Bayes	0.1653	0.2490	0.2752	0.2831
	0.5	IRLS	0.3221	0.3840	0.3457	0.3458
		ML	0.1571	0.1438	0.1517	0.1508
		Bayes	0.2090	0.2271	0.2188	0.2151
100	0.2	IRLS	0.4719	0.5421	0.5279	0.5335
		ML	0.1348	0.1618	0.1636	0.1640
		Bayes	0.2863	0.3336	0.3271	0.2739
	0.5	IRLS	0.4824	0.4922	0.4332	0.3818
		ML	0.1346	0.1503	0.1698	0.1562
		Bayes	0.3107	0.3178	0.2772	0.2409
200	0.2	IRLS	0.3539	0.3813	0.3238	0.3676
		ML	0.1181	0.1163	0.1155	0.1123
		Bayes	0.2362	0.2511	0.2145	0.2009
	0.5	IRLS	0.1586	0.1866	0.2001	0.1943
		ML	0.0937	0.1103	0.1193	0.1151
		Bayes	0.1102	0.1340	0.1342	0.1291
500	0.2	IRLS	0.1304	0.1664	0.1643	0.1693
		ML	0.0671	0.0733	0.0733	0.0710
		Bayes	0.1006	0.1136	0.1101	0.0911
	0.5	IRLS	0.1253	0.1357	0.1253	0.1245
		ML	0.0567	0.0685	0.0706	0.0701
		Bayes	0.1213	0.1171	0.1014	0.0945

the probability of true and false matching in record linkage procedure into account. At this point, it should be noted that, in this simulation study, we considered a non-informative prior distribution in Bayesian paradigm. The efficiency of Bayes estimator can be improved by considering a more suitable prior for regression coefficients. Both IRLS and ML estimators encounter the convergence problem in some cases, such as small sample sizes and when

Table 2. The mean of bias of regression coefficients estimators for various values of sample size, success probability and matching error rate.

			Matching Rates			
n	π_y	Estimator	$\mathcal{A}_1$	$\mathcal{A}_2$	$\mathcal{A}_3$	$\mathcal{A}_4$
50	0.2	IRLS	-0.4906	-0.5464	-0.6887	-0.7101
		ML	-0.1180	-0.1218	-0.3011	-0.3785
		Bayes	-0.8856	-0.8126	-0.8985	-0.9808
	0.5	IRLS	-0.3884	-0.5038	-0.7976	-0.9545
		ML	-0.2673	-0.0978	-0.3388	-0.4829
		Bayes	-0.9453	-0.9493	-0.9398	-1.0000
100	0.2	IRLS	-0.2020	-0.1764	-0.4626	-0.7018
		ML	-0.2255	-0.0911	-0.2982	-0.4261
		Bayes	-0.6772	-0.9126	-0.1912	-0.5435
	0.5	IRLS	-0.2628	-0.1915	-0.5975	-0.8153
		ML	-0.3583	-0.0200	-0.2782	-0.4338
		Bayes	-0.6647	-0.9568	-0.2484	-0.6616
200	0.2	IRLS	-0.2800	-0.1141	0.0553	-0.2930
		ML	-0.4070	-0.0554	-0.1685	-0.2930
		Bayes	-0.5819	-0.8421	-0.2601	-0.3201
	0.5	IRLS	-0.4108	-0.6323	-0.8358	-0.9707
		ML	-0.4310	-0.0916	-0.2564	-0.4147
		Bayes	-0.4719	-0.2759	-0.3514	-0.4417
500	0.2	IRLS	-0.2060	-0.4721	-0.6552	-0.8120
		ML	-0.2316	-0.1114	-0.2948	-0.4162
		Bayes	-0.5992	-0.2745	-0.2901	-0.6901
	0.5	IRLS	-0.3174	-0.5846	-0.8171	-0.9650
		ML	-0.3857	-0.2829	-0.2829	-0.4363
		Bayes	-0.7131	-0.4798	-0.3346	-0.5900

success probability is closed to 0 or 1, in which their values are often unavailable in these situations, e.g. $\pi_y = 0.1$ and $n = 50$. The problems related to the ML estimator for small sample sizes and boundary values of success probability, may return to the usual major problem of non-identifiability in finite mixture analysis. But, Perpinan and Renals (2000) gave empirical support to the fact that the estimation of Bernoulli mixture distributions produces

Table 3. The mean of RMSE of regression coefficients estimators for various values of sample size, success probability and matching error rate.

			Matching Rates			
n	π_y	Estimator	$\mathcal{A}_1$	$\mathcal{A}_2$	$\mathcal{A}_3$	$\mathcal{A}_4$
50	0.2	IRLS	0.5602	0.6795	0.8180	0.8478
		ML	0.2140	0.2431	0.3686	0.4384
		Bayes	0.9009	0.8499	0.9397	1.0208
	0.5	IRLS	0.5046	0.6335	0.8693	1.0152
		ML	0.3100	0.1739	0.3712	0.5059
		Bayes	0.9681	0.9761	0.9649	1.0229
100	0.2	IRLS	0.5133	0.5701	0.7019	0.8816
		ML	0.2627	0.1857	0.3401	0.4566
		Bayes	0.7352	0.9717	0.3789	0.6086
	0.5	IRLS	0.5493	0.5281	0.7380	0.9003
		ML	0.3827	0.1516	0.3259	0.4611
		Bayes	0.7337	1.0082	0.3722	0.7041
200	0.2	IRLS	0.4513	0.3980	0.3285	0.4701
		ML	0.4238	0.1288	0.2043	0.3138
		Bayes	0.6280	0.8787	0.3371	0.3779
	0.5	IRLS	0.4404	0.6593	0.8594	0.9900
		ML	0.4411	0.1434	0.2828	0.4304
		Bayes	0.4846	0.3067	0.3762	0.4602
500	0.2	IRLS	0.2438	0.5006	0.6755	0.8295
		ML	0.2411	0.1334	0.3038	0.4222
		Bayes	0.6076	0.2971	0.3103	0.6961
	0.5	IRLS	0.3413	0.6001	0.8267	0.9730
		ML	0.3898	0.2911	0.2916	0.4419
		Bayes	0.7233	0.4939	0.3496	0.5975

meaningful results.

Therefore, even when the number of mixture components is large, non-identifiability is not a serious problem for the mixture of the Bernoulli distributions. Although, the small sample size does not make an important problem in Bayesian paradigm, but the Metropolis-Hastings algorithm converges slow when the success probability, π_y, is closed to 0 or 1.

7 Discussion

When two files are linked, due to some usual unavoidable mismatch errors, some pair records might link incorrectly. Hence, when we are interested in the logistic regression analysis with linked data, some response and covariate values may not correspond to each other and the usual logistic regression analysis methods can not be used due to their poor efficiency. Specially, our simulation showed that these methods were seriously affected by large values of matching errors. Hence it is necessary to develop new methods that can truly take into account the probabilistic mismatched errors. We showed that, in this situation, the distribution of the response variable is a finite mixture of Bernoulli distributions with known mixing coefficients. From a frequentis point of view, the maximum likelihood estimation of logistic regression coefficients can be driven using EM algorithm through a iterative procedure as we explained in this paper. The Bayesian estimations also can be deriven by considering a prior distribution for logistic regression model parameters and sampling from their corresponding posterior distribution. Of course, the main problem in both frequentis and Bayesian paradigm is related to large number of components in mixture distribution of response variable. As a solution, one can consider a blocking scheme for record linkage procedure that partitions the full cross product of record comparisons into mutually exclusive blocks and reduces he number of matched records (Evangelista et al., 2010). This leads to considerably smaller number of component in mixture distribution and simpler estimation procedure for parameters of regression model.

Although we only considered logistic regression analysis, but developing other statistical modeling methods for linked data is in the core interest of future researches.

Acknowledgements

We thank the reviewers for their helpful suggestions, which improved our manuscript. Partial support from Ordered and Spatial Data Center of Excellence of Ferdowsi University of Mashhad is acknowledged.

References

Belin, T.R. and Rubin, D.B. (1995). Method for calibrating false-match rates in record linkage. *Journal of the American Statistical Association*, **90**, 694-707.

Burkard, R.E. and Derigs, U. (1980). Assignment and Matching Problems: *Solution Methods with FORTRAN-Programs*, Springer-Verlag, New York.

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

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of Royal Statistical Society*, Ser. B, **39**, 1-38.

Evangelista, L.O., Cortez, E., da Silva, A.S. and Jr., W.M. (2010). Adaptive and flexible blocking for record linkage tasks, *Journal of Information and Data Management*, **1**, 2, 167-181.

Fallah, A. and Mohammadzadeh, M. (2010). Bayesian regression analysis with linked data using mixture normal distributions, *Statistical Papers*, **51**, 421-430.

Fellegi, I.P. and Sunter, A.B. (1967). A theory for record linkage, *Journal of American Statistical Association*, **64**, 694-707.

Gelman, A., Roberts, G. and Gilks, W. (1995). Efficient metropolis jumping rules. *Bayesian Statistics*, **5**, 599-607.

George, E.L. and McCulloch, R.E. (1993). Variable selection via Gibbs sampling, *Journal of American Statistical Association*, **88**, 881-890.

Hosmer, D.W. and Lemeshow, S. (2000). *Applied Logistic Regression*, Wiley, New York.

Jaro, M.A. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida, *Journal of the American Statistical Association*, **89**, 414-420.

Kass, R.E. and Wasserman, L. (1995). A reference bayesian test for nested hypotheses relationship to the schuartz criterion. *Journal of American Statistical Association*, **90**, 928-934.

Lahiri, P. and Larsen, M.D. (2000). Model-Based Analysis of Records Linked Using Mixture Models. American Statistical Association, *Proceedings of the Section on Survey Research Methods*, 11-19.

Lahiri, P. and Larsen, M.D. (2005). Regression analysis with linked data, *Journal of American Statistical Association, Theory and Methods*, **496**, 222-230.

Larsen, M.D. and Rubin, D.B. (2001). Iterative automated record linkage using mixture models, *Journal of the American Statistical Association*, **96**, 32-41.

Marin, J.M., Mengersen, K. and Robert, C. (2005). Bayesian Modelling and Inference on Mixtures of Distributions, In Rao, C. and Dey, D., editors, Handbook of Statistics, **25**, Springer-Verlag, New York.

McCullagh, P. and Nelder, L.A. (1985). *Generalized Linear Models*, Chapman and Hall, London.

Newcomb, H.B., Kennedy, J.M., Axford, S.I. and James, A.P. (1959). Automatic record linkage of vital records, *Science*, **130**, 954-959.

Newcombe, H.B. and Kennedy, J.M. (1962). Record linkage, *Communications of the Association for Computing Machinery*, **5**, 563-566.

Perpinan, M.A.C. and Renals, S.A. (2000). Practical identifiability of finite mixture of multivariate bernoulli distribution, *Neural Computation*, **12**, 141-152.

Quandt, R.E. (1972). A new approach to estimating switching regressions, *Journal of the American Statistical Association*, **67**, 306-310.

Raftery, A.E., Madigan, D. and Hoeting, J. (1997). Bayesian model averaging for linear regression models, *Journal of the American Statistical Association*, **92**, 179-191.

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

Sheuren, F. and Winkler, W.E. (1997). Regression analysis of data files that are computer matched II, *Survey Methodology*, **23**, 157-165.

Suchindran, C.M., Leiss, J.K. and Salama, I. (2001). Alternative methods for record linkage: Application to linking vital records, *Proceedings of the Annual Meeting of the American Statistical Association*, August 5-9.

Tancredi, A. and Liseo, B. (2011). A hierarchical bayesian approach to record linkage and population size problems, *The Annals of Applied Statistics*, **5**, 1553-1585.

Verykios, V., Moustakides, G.V. and Elegy, M.G. (2003). A bayesian decision model for cost optimal record matching, *The VLDB Journal*, **12**, 28-40.

Winkler, W.E. (1993). Improving decision rules in Fellegi-Sunter model of record linkage, American Statistical Association, *Proceeding of Survey Research Methods Section*, 274-279.

Winkler, W.E. (1994). Advanced methods for record linkage, American Statistical Association, *Proceedings of the Section of Survey Research Methods*, 467-472.

Winkler, W.E. (1995). A Matching and Record Linkage, In Cox *et al.*, editor, *Business Survey Methods*, J. Wiley & Sons Inc.

Zellner, A. (1986). On Assessing Prior Distributions and Bayesian Regression Analysis with g-Prior Distributions. In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, North-Holland, Amsterdam, Chapter 5, 233-243.

M. Mohammadzadeh
Department of Statistics,
Tarbiat Modares University,
Tehran, Iran.
email: *mohsen_m@modares.ac.ir*

A. Fallah
Department of Statistics,
Imam Khomeini International University,
Ghazvin, Iran.