ALTERNATIVE METHODS FOR RECORD LINKAGE: APPLICATION TO LINKING VITAL RECORDS

Chirayath M. Suchindran, Jack K. Leiss, Ibrahim Salama
Chirayath M. Suchindran, Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7420

Key Words: Multi-level model, Regression Trees, Logistic Regression

INTRODUCTION

There is widespread and growing usage of linking multiple data sources in Public Health Studies. The objective of the linking process is to determine whether two or more records refer to the same person, object or event. Modern computer power has enhanced our capacity to conduct computer linkage of large public health data files. A key step in conducting this linkage is the development of an efficient record-matching algorithm. Formal development of a theory of record linkage started with the pioneering work of Fellegi and Sunter (1969). Several people have worked on extending or modifying their procedure (Jaro 1989; Winkler 1994). Many of these existing procedures have a number of limitations. Most procedures use specialized software to implement the algorithm. In this paper we plan to develop alternative models for record linkage that can be used with widely available statistical software.

In the next section of the paper we briefly review the Fellegi and Sunter method to set up the problem and to understand the limitations of the method. In the subsequent sections we describe modifications of the Fellegi and Sunter method that can overcome some of the limitations of their method. Illustrative examples linking birth and infant death files will be presented next. A brief discussion and conclusion section discusses the strengths and limitations of the proposed methods and discusses some future directions.

FELLEGI AND SUNTER MODEL

The record linkage problem can be formally stated as follows: Suppose that we have two files A and B and file A contains k_a records and file B contains k_b records. The product space $A \times B$ contains $k_a \times k_b$ possible pairs of linked records. Among them there exist (possibly) only $k = \min imum(k_a, k_b)$ true links. The problem is to identify the true links. The linking process usually is done by creating a linkage weight w_j indicating the degree to which the pair j is likely to be a true link. A decision rule is then implemented to declare a pair as a link or non-link or to be a possible link to be determined by further examination. For this purpose Fellegi and Sunter formulated a decision rule as follows: Define

$$m_{i(j)} \sim Probability\{i^{th}\ field\ (\ component\ variable\ )\ of\ j^{th}\ pair\ agrees \mid true\ link\}$$

$$u_{i(j)} \sim Probability\{i^{th}\ field\ (\ component\ variable\ )\ of\ j^{th}\ pair\ agrees \mid true\ non\text{-}link\}$$

and

$$w_{i(j)} = \log \frac{m_{i(j)}}{u_{i(j)}}\ \ if\ i^{th}\ field\ of\ j^{th}\ record\ agrees$$

$$= \log \frac{1-m_{i(j)}}{1-u_{i(j)}}\ \ if\ i^{th}\ field\ of\ j^{th}\ record\ disagrees$$

Then a composite weight for the j[th] record pair is obtained as: $w_j = \sum_i w_{i(j)}$. Fellegi and Sunter proposed a decision rule to classify the record pair as true link or true non-link by defining an upper threshold and a lower threshold. The lower and upper thresholds are determined by a priori error bounds based on rates of false links and false non-links. Usually clerical decisions are recommended when the composite weight falls between the lower and upper bounds.

The justification for the weights derived by Fellegi and Sunter can be formally made as follows:

Let Y denote a random variable (not observed) that takes the value 1 if the record pair is a true link and 0 otherwise. Let X denote a set of predictor variables available in the linked record. In the early applications these variables are taken as the indicator agreements of the common fields (variables) in the files A and B. When Y is known the data vector (Y, X) is considered as complete. In addition if the *record pairs are considered as*

independent observations the likelihood for the data can be written as:

$$L = \prod f(y,x) = \prod f(x \mid y,\theta) g(y \mid \theta)$$

The marginal distribution of Y is assumed to be Bernoulli random variable with parameter θ. Fellegi and Sunter (1969) and Jaro (1989) specified the conditional distributions as follows:
For all record pair j define

$m_i \sim \Pr ob\{i^{th} \ field \ agrees | Y = 1\}$

$u_i \sim \Pr ob\{i^{th} \ field \ agrees | Y = 0\}$

Under the assumption that the *field agreements are independent*

$f(x \mid y = 1) = \prod m_i^{x_{ij}} (1 - m_i)^{1 - x_{ij}}$,

where $x_{ij} = 1$ *if* i^{th} *field of* j^{th} *record matches*

and $f(x \mid y = 0) = \prod u_i^{x_{ij}} (1 - u_i)^{1 - x_{ij}}$

where m_i *and* u_i are unknown parameters and $x = \{X_{ij}\}$. Because Y is unknown the parameter estimates are obtained through the EM algorithm.

The expectation step for implementing this algorithm can be expressed in terms of the probability that the j^{th} record is a true link as:

$$\log \frac{E(Y_j \mid X)}{1 - E(Y_j \mid X)} = \sum_i X_{ij} \log \frac{m_i}{u_i} + \sum_i (1 - X_{ij}) \log \frac{1 - m_i}{1 - u_i} + \log \frac{\theta}{1 - \theta} \qquad (1)$$

The equation (1) shows that the log odds for a match is given by the composite weight proposed by the Fellegi and Sunter with an additional constant term for each record. These results suggest that an estimated prediction equation of the log odds avoiding the assumption of independence of records as well as independence of fields matching may provide an improvement in the classification decision. In the following sections we will follow these guidelines to modify Fellegi and Sunter procedure.

REFORMULATION OF FELLEGI AND SUNTER MODEL

As before assume that there are A×B records that are either true links or false links.
Let us assume that Y_i is a binary random variable taking the value of 1 if the record is a true link and 0 otherwise. For the moment assume that all linked pairs in the A×B are independent observations (we will deal with this problem later). Note that we do not observe Y_i. Let X_i denote the set of all known predictors. In most linkage models (including Fellegi and Sunter) the predictor variables X_i are indicators of agreement of common fields from A and B. In general this requirement is not necessary. One can keep the field characteristics themselves as predictor variables to improve the prediction. Let us assume that it is possible to make an initial assignment of the records into true link or not using a surrogate measure. This initial assignment is considered as an imperfect assignment. (We will discuss later how one will choose this surrogate measure to make this initial assignment). Let Z denote a binary random variable indicating the true link status as assigned by the initial assignment.

Notations

$P[Y = y] = P_y(x, \beta)$ where β is a set of unknown parameters.

$P[Z = z \mid \gamma, Y = 1]$ *and* $P[Z = z \mid \gamma, Y = 0]$ stand for the conditional distributions of Z given Y. Therefore the marginal distribution of Z is given by:

$$P[Z = z \mid x, \gamma, \beta] = P[Z = z \mid \gamma, Y = 1] P[Y = 1 \mid X, \beta] + P[Z = z \mid \gamma, Y = 0] P[Y = 0 \mid X, \beta]$$

Also note that

$$E(Y \mid Z = z) = P(Y = 1 \mid Z = z) = \frac{P(Y = 1 Z = z)}{P(Z = z)} = \frac{P(Z = z \mid Y = 1) P(Y = 1)}{P(Z = z)}$$

Denote

$P[z = 1 \mid y = 1] = \gamma_1$ (Sensitivity)

$P[z = 0 \mid y = 0] = \gamma_2$ (Specificity)

With these notations we can express the probability of true match conditional on the surrogate value in terms of the sensitivity and specificity parameters.

Data Structure and Likelihood

The observed data {Z, X} now form an incomplete data vector and {Y, Z, X) the complete data vector. Assuming that all N= A×B are independent

observations, the likelihood function for β *and* γ given Z, Y and X is

$$L = \prod_{i=1}^{N} P[y_i, z_i \mid X_i, \beta, \gamma] = \prod_{i=1}^{N} P[z_i \mid Y, \gamma] P[Y \mid X, \beta]$$

Both parameters β *and* γ are estimated simultaneously by maximizing this likelihood. However Y is not known. Hence other procedures such as EM algorithm are used to estimate the parameters. The log-likelihood can be written in two parts as:

$$\log L = \log l(\gamma \mid Z, Y) + \log l(\beta \mid Y, X)$$

where $\log l(\gamma \mid Z, Y) =$

$$\sum_{i=1}^{N} Y_i \log P[Z_i = z_i \mid \gamma, Y_i = 1]$$

$$+ \sum_{i=1}^{N} (1 - Y_i) \log P[Z_i = z_i \mid \gamma, Y_i = 0]\}$$

and $\log l(\beta \mid Y, X)$

$$= \sum_{i=1}^{N} \{Y_i \log P_y(X, \beta) + (1 - Y_i) \log(1 - P_y(X, \beta)\}$$

Comments on the likelihood.

In this work we assume that $P[Y = 1] = \frac{1}{1 + e^{X\beta}}$ a logistic function. When Y is binary taking values 1 and 0 the maximum likelihood estimates of β can be found using routines that fit logistic regression. When the model is specified correctly and when there are large numbers of observations available the regression parameters can be consistently estimated even when the assumption of independence of observations is violated. However the standard error of the estimates will not be correctly estimated. The estimation of correct standard errors may not be crucial for classification purposes.

EM Algorithm

Because of the unknown nature of the true link variable Y, EM algorithm is used to get the regression parameters and subsequent classification schemes. The algorithm is implemented as follows: First obtain initial values of β *and* γ . Use this to update values Y using its expected values . Use the updated values of Y to get new estimates of β *and* γ . Continue the process until the parameter estimates converge.

Implementation of the classification procedure

In order to reduce the computational burden a number of steps can be taken for the implementation of the algorithm. The goal is to use readily available software to implement the algorithm. Therefore we adapt the following steps:

1. An initial assignment of the true link status is made on the basis of agreement on a single field. This field was chosen in such a way that the proportion of true links be close to the actual (or approximate) proportion of the true links.
2. Using this initial assignment fit a logistic regression model with other field characteristics (which includes indicators of other field agreements and other characteristics of the record). Obtain the regression coefficients, say b, and the estimated variance covariance matrix of b, say $v_{b.}$
3. It is well known that in classification, if the same observation that is used to fit the model is also used to estimate the classification error, the resulting error count estimate is biased. In order to reduce this bias, one will remove the binary observation to be classified from the data, re-estimate the parameters of the model, and then classify the observation based on the new parameter estimates. The logistic procedure in SAS uses the following simple procedure to get the one observation removed parameter estimates. The one step estimate b_j after removing the j[th] observation is given by:

$$b_j^{'} = b - \frac{(y_j - \hat{p}_j)}{1 - h_{jj}} \hat{V}_b (1 \quad x_j)'$$

where $\hat{p}_j$ is the predicted probability based on b (full sample) h_{jj} is the diagonal element of the hat matrix and in the logistic regression model, it is given by

$$h_{jj} = \hat{p}_j \hat{q}_j (1, x'_j) \hat{V}_b (1, x_j)'$$

4. Let $\hat{p}_j^*$ denote the predicted probability based on b_j . For a given cut off point z, the j^{th} record is classified as true link if

$\hat{p}_j^* \geq z$. We choose the cut off point z such that the sum of sensitivity and sensitivity is the maximum.

5. Usually continued iteration under the EM algorithm is necessary to get the final estimates and classification. When steps 1-4 are implemented, in practice, the estimates converge after a single iteration.

A MULTI-LEVEL MODEL FOR RECORD LINKAGE

In the proposed model above we assumed that all the record pairs are independent observations. This assumption may not be valid in many situations. Also when the linking files are large (as in birth and death linking where the birth record file contains a large number of observations) the product space $A \times B$ of all possible pairs is also very large to manipulate. To reduce this computational burden a stratification scheme based on the common variables in the two files is introduced (the idea is very similar to the blocking scheme introduced by Jaro 1989). It is assumed that the true links come within a stratum. A multi-level model can be used for the purpose of classifying true links. The model estimation will take into account the uncertainties caused by stratification as well as the dependence of observations within the strata. The model is formulated as follows:

Suppose the files come from K strata and each stratum contains

$$N_i = A_i \times B_i \text{ records and } N = \sum_{i=1}^{K} N_i$$

Let $Y_{j(i)}$ denote a Bernoulli random variable taking a value of 1 if the j[th] record in i[th] stratum is a true link and 0 otherwise (This random variable is not observable).

Denote $\theta_{ij} = P\{Y_{j(i)} = 1\}$. We formulate the model using the logit link function as:

$$\log it(\theta_{ij}) = \alpha + u_i + X'_{ij}\beta .$$

The u_i s are stratum specific parameters and are initially assumed to be distributed as

N $(0, \sigma_u^2)$. Because $Y_{j(i)}$ are unknown, as before, we first make an imperfect allocation using a field matching variable. This matching variable will be chosen as the one that gives a marginal distribution of the true link as close as the true one.

The random effects model is estimated first using the imperfect assignment.

Because the beta-normal distribution has no closed form some approximate methods need to be used to estimate the model parameters. A simple procedure is to use a predictive quasi-likelihood (PQL) or a second order Taylor series approximation (PQL2) to obtain the parameter estimates. A Bayesian approach using MCMC via Metropolis-Hastings Algorithm can also be used in the model estimation. The NLMIX procedure in SAS uses the PQL procedure to get the parameter estimates. The Bayesian procedure can be implemented through software such Mlwin or WINBUG.

After the parameter estimates are obtained on the basis of the initial assignment, a prediction is made as to the true link status of a record on the basis of the estimated random effects model. Ideally some iteration is necessary until the parameters converge to improve the parameter estimates obtained through the initial imperfect classification. However, in practice, with a careful choice of initial assignment the algorithm usually stops after a single iteration.

Once the parameter estimates (including the random effects u_i) are obtained the model will be used to make conditional prediction for observations within each stratum. As before we will choose a cut off point that maximizes the sum of sensitivity and sensitivity (It is possible to choose stratum specific cut off points by maximizing the sum of sensitivity and specificity within a stratum).

ILLUSTRATIVE EXAMPLES

We present here two illustrative examples of the proposed methods. The first example illustrates the use of logistic regression with a surrogate indicator assignment. The second illustrates the use of multilevel model.

Example 1

This illustrative example performs a logistic regression model assignment for linking birth and death records. For this purpose a data set containing 10600 record pairs were created with 50% of the pairs known to be true links (In this case the true link status is known because the original data set was manually examined). All pairs of observations formed by linking birth and death files were initially compared to see whether there

was agreement on 18 variables. In this file all variables were coded 1 if the birth and death file agree and 0 otherwise. The working file for this analysis is a one-to-one matched file in the sense that one birth record is linked to one death record. With 50% true link, the file of 10600 records contained 5300 hundred true links. The first step in the analysis is to find a surrogate variable that matches the marginal proportion of true links. Among the 18 variables in the list the variable *child's day of birth* was chosen for the initial assignment. This assignment put 5385 records as true link and 5215 as true non-link. Because in this case the true status is known (which is not the case normally) we can calculate the error rate of this assignment. In this case 69 (1.32%) of the true links were assigned as non-links and 154 (2.86%) of the true non-links were classified as true links. Thus the total number of misclassified records for this initial assignment was 223 (2.1% of all records). The initial logistic regression model with the initial assignment was done. For this illustration we arbitrarily selected a set of independent variables. The included variables are child's sex, child's month of birth, state of birth, soundex code, child's first name and father's last name. (The choice is completely arbitrary. A formal analysis will be conducted later). The SAS procedure PROC LOGISTIC was used to get the parameter estimates dropping one observation at a time. With the parameter estimates predicted probabilities were obtained for each observation. The sum of sensitivity and specificity was found to be maximum when probability of match was set at greater than 0.460. The resulting classification compared with the true link status (which happened to be known in this example) is shown in Table 1. The table shows that overall 99.47% of the records were classified correctly. Fifty one (0.96%) of the 5300 true links were incorrectly classified as non-links whereas five (0.09%) of the 5300 true non-links were classified as links. An examination of the regression diagnostics failed to pick up the 56 misclassified records. However, it picked up all the records in the initial classification as well as in the final classification. As expected, a single iteration produced the final result

Table 1: The predicted and true link status

Status	Predicted link (Row %)	Predicted Non-link (Row %)	TOTAL
True link	5249 (99.04)	5l (0.96)	5300
True non-link	5 (0.09)	5295 (99.91)	5300
TOTAL	5254	5346	10600

Example 2

In this example a subset of 3315 observations was selected arbitrarily from the 10600 observations used in example 1. These observations were also placed in 18 strata in an arbitrary manner. Unlike in the example 1, in this data set only 38.46 percent of the observations were true links. We use this smaller data set to illustrate the multi-level model. A logistic regression model as in Example 1 was run first that resulted in misclassification of 2.65% (88 out of 3315 records). A multilevel logistic regression model was then run using the Mlwin software. The Metropolis Hastings algorithm was used in estimating the parameter estimates. The criterion sum of sensitivity and specificity was used to assign the link status. The results are shown in table 3.

Table 2: Classification table using Multi-level model

Status	Predicted true link	Predicted non-link	TOTAL
True link	1993 (97.70)	47 (2.30)	2040
True non-link	2 (0.16)	1273 (99.84)	1275
TOTAL	1995	1320	3315

Table 2 shows that the total error rate with the multilevel model is 1.48% (49/3315) which is smaller than the 2.65% attained by the simple logistic regression model.

Classification and Regression Trees (CART) algorithm for linkage

Tree structured statistical models such as CART have lately become an attractive tool for statistical analysis (Breiman et.al 1984). It is a useful tool for decision making (such as a record is a true link or not) and its implementation is now possible due to the widespread availability of computer software. Tree construction uses the method of recursive

partitioning of a universe (in this case the universe of all possible pairs with identification as true link or not). In a normal situation a learning sample with true status of linkage and predictors of the true status are known to develop a decision rule for subsequent application. However, in the linkage situation the true status is unknown and therefore we start with a surrogate indicator of the true status, as done in the application of the logistic regression models.

.

Illustrative Example

For illustration we use the same data set of 10600 records that was used for illustrating the logistic regression models. In this data set it was known that 5300 records were true links. The child's day of birth was used as surrogate indicator of the true link status. For illustration, as in the case of logistic regression model, we included only six predictor variables: match child's sex, child's month of birth, state of birth, soundex code, child's first name and father's last name. Tree construction based on this sample was performed using classification and trees procedure in S+.2000. The final tree consisted of 13 nodes. For each record belonging to a terminal node a predicted probability of a true link is computed. Based on these predicted probabilities each record was classified as true link or not (to be consistent with the logistic regression model choose a cut off point that maximizes the sum of sensitivity and specificity based on the initial assignment). Because the true link status of the record is known in this case we can compute the error rate. These misclassification rates are given in Table 3. Table 3 shows that the algorithm misclassified 68 of the 10600 records. Of these 68 misclassification errors 64 of them occurred by misclassifying true links as non-links. As seen in Table 1, the logistic regression model misclassified a total of 56 records with 51 true links classified as non-links. The advantage of the tree-based procedure is that nodes with low predicted probability often contain misclassified records. A comparison with the known true link status confirmed this suspicion. This observation gives a working rule to further examine the records (perhaps manually) in the nodes where the probability of a match is far from one or zero.

Table 3: Misclassification rates

Status	Predicted link	Predicted Non-link	Total
True link	5236 (98.79%)	64 (1.21%)	5300
True non-link	4 (0.08)	5296 (99.92)	5300
Total	5240	5360	10600

DISCUSSION AND CONCLUSION

In this paper we present three alternative methods for record linkage. The methods are easy to implement using widely available software and they avoid some of the weakness of some of the existing methods. Illustrative examples of linking birth and death records are given in the paper. All examples show that the proposed method works well in the context they are used here. More sophisticated examples need to be created to fully assess the method. Currently we are working on creating such examples with more dependency in the data. The apparent success of proposed logistic regression as seen in these examples also opens the door for trying other models. The classification and regression trees algorithm is claimed to have better properties than the logistic regression model classification method. Additional work is needed to make recommendations on choice of surrogate measures of link for making the initial classification. Classifications based on criteria other than the sum of sensitivity and specificity also need to be examined further.

REFERENCES

Breiman, L., Friedman, J.H., Olshen, R.A., and C. J. Stone(1984) Classification and Regression Trees. Monterey: Wadsworth and Brooks/Cole.

Fellegi, I. P. and A. B. Sunter (1969), "A Theory for Record Linkage", *Journal of the American Statistical Association*, 64: 1183-1210.

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*, 84: 414-420.

Winkler, W. E. (1994), "Advanced Methods for Record Linkage", *in Proceedings of Survey Research Methods Section, American Statistical Association*: 667-671.