Analysis of Statistical Models with Linked Data

Partha Lahiri and Judith Law

University of Maryland, College Park

plahiri@umd.edu

August 24, 2015
4th Baltic-Nordic Conference on Survey Statistics

Record Linkage

A goal of record linkage is to join together two files that contain information on the same individuals using available information (matching fields), which typically does not include unique, error-free personal codes.

Examples of matching fields

- Some matching fields may carry a lot of information for identifying individuals (e.g., surname, first name, age or date of birth)
- Some others may contain very little (e.g., race or sex).

A layout of the two files to be linked

*

Table 1: A layout of the two files to be linked

File A			File B		
	matching fields			matching fields	
	$v_1, v_2, \ldots, v_k$	x		$w_1, w_2, \ldots, w_k$	y
a_1			b_1		
a_2			b_2		
$\vdots$			$\vdots$		
a_n			b_n		

Comparison Vectors

$\gamma \equiv \gamma(a,b)^T = \{\gamma_1(a,b), \gamma_2(a,b), \ldots, \gamma_K(a,b)\}$, where:

- K is the number of fields used for comparison
- (a,b) is a pair of records, one from File A and one from File B
  - for a binary field

$$\gamma_k(a,b) = \begin{cases} 1 & \text{if } v_k(a) = w_k(b) \\ 0 & \text{if } v_k(a) \neq w_k(b) \end{cases}$$

  - for a non-binary field, $\gamma_k(a,b) \in [0,1]$

    (e.g. string comparator metrics of Jaro (1989) and Winkler (1990))

Example: Possible comparison vectors for $K = 3$ binary matching fields

$$(0,0,0), (0,0,1), (0,1,0),$$
$$(0,1,1), (1,0,0), (1,0,1),$$
$$(1,1,0), (1,1,1)$$

Comparison Vectors

*

Table 2: Comparison vectors - excerpt from a large dataset

		Middle		Year	Month	Day	Zip	Hosp	Medical
FN	LN	Initial	Gender	Birth	Birth	Birth	Code	ID	ID
0.448	0.437	0	0	0	0	0	0	1	1
0.464	0.483	0	0	0	0	0	0	1	1
0.429	0.500	0	1	1	1	0	0	0	0
0.550	0.633	0	1	1	1	0	1	0	0
0.000	0.578	0	1	1	1	1	0	0	0
0.625	0.730	0	1	1	1	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
0.167	0.178	1	1	1	1	1	1	1	0

Optimal Linkage Rule

Fellegi and Sunter, (1969) suggested an optimal linkage rule based on the following likelihood ratio:

$$R = \frac{\mathrm{P}(\gamma \mid r \in M)}{\mathrm{P}(\gamma \mid r \in U)}$$

At a prespecified error levels for false links (μ) and false nonlinks (λ), the optimal cutoffs are as follows:

- if $R \geq upper$, then designate the pair as a *link*
- if $upper > R > lower$, then postpone the decision pending clerical review
- if $R \leq lower$, then designate the pair as a *nonlink*

The record linkage rule is optimal insofar as it sends the minimum number of record pairs to clerical review at prespecified error levels.

Three Issues

- Not all possible pairs of records are compared. Instead, pairs are compared within blocks of records that are similar in terms of basic characteristics, such as geography or first letter of last name.

- $P(\gamma|M)$ and $P(\gamma|U)$ are unknown; they must be estimated under a model using certain assumptions. The performance of the procedure in terms of actual versus specified error rates is sensitive to estimates of probabilities and choice of upper and lower (Belin 1993; Belin and Rubin 1995).

- For a record in file A there might be several candidate links within a particular block in file B. We assume in this work that only one of the records in file B is a true link for the record in file A. Given estimated probabilities, in practice, single links for individual records are chosen according to some procedure.

Mixture Models

A $G \geq 2$ class mixture model for γ:

$$P(\gamma) = \sum_{g=1}^{G} \pi_g P(\gamma|\text{class } g),$$

where

- π_g: probability that a record pair belongs to the mixture class g
- $P(\gamma|\text{class } g)$: pmf of γ in class g

Comment:

The probability

$$P(\text{class } g|\gamma) = \frac{\pi_g P(\gamma|\text{class } g)}{\sum_{h=1}^{G} \pi_h P(\gamma|\text{class } h)}$$

can be used to partition the record pairs into designated links and nonlinks and to estimate error rates (Larsen and Rubin 2001).

Estimation of Mixture Model Parameters

- For our application, we have chosen $G = 2$ as in Lahiri and Larsen (2005). The two classes of the mixture model correspond to the links or matches (M) and nonlinks or non-matches (U).

- For $G = 2$, $P(\text{class } g|\gamma)$ is a monotonic function of R.

- The parameters of the mixture model can be estimated using the expectation-maximization (EM) (Dempster, Laird, and Rubin 1977) and expectation-conditional maximization (ECM) (Meng and Rubin 1993) algorithms.

- Several authors, including Larsen and Rubin (2001) and Lahiri and Larsen (2005), have implemented these algorithms for the purposes of record linkage.

Conditional Independence (Fellegi and Sunter, 1969)

$$P(\gamma|\text{class } g) = \prod_{k=1}^{K} P(\gamma_k|\text{class } g),$$

where $P(\gamma_k|\text{class } g)$: probability of γ_k on comparison k in class g.

- Other modeling assumptions are possible and, in some cases, correspond better to the observed data (Larsen and Rubin 2001; Armstrong and Mayda 1993; Thibaudeau 1993).
- A few authors in other contexts have used mixture models applied to discrete data with modeling assumptions other than conditional independence (see, e.g., Becker and Yang 1998; Larsen and Rubin 2001).
- Studies by Winkler (1993, 1994) showed that with certain data, good decision rules are possible under the assumption of conditional independence, even when there are substantial departures from conditional independence.

EM Algorithm to Estimate Matching Weights

We will assume conditional independence, a binary agreement pattern, and the following notation:

- Let r_j be the jth record pair in the cross-product space $A \times B$.
- Let N be the number of record pairs in the cross-product space $A \times B$.
- Let K be the number of fields used for comparison.
- Let γ_{kj} be the agreement indicator of the kth field in the jth pair, where $k = 1, \ldots, K$, and $j = 1, \ldots, N$.

$$\gamma_{kj} = \begin{cases} 1 & \text{if } \textit{match} \\ 0 & \text{if } \textit{nonmatch} \end{cases}$$

- Let $\gamma_j = \{\gamma_{1j}, \ldots, \gamma_{Kj}\}$
- Let $\gamma = \{\gamma_1, \ldots, \gamma_N\}$

EM Algorithm, continued

- Let $m = \{m_1, \ldots, m_K\}$ and $u = \{u_1, \ldots, u_K\}$, where, for a randomly selected pair, r_j:

$$m_k = \mathrm{P}(\gamma_{kj} = 1 \mid r_j \in M)$$
$$u_k = \mathrm{P}(\gamma_{kj} = 1 \mid r_j \in U)$$

- Let $\pi = \frac{\text{Number of record pairs in set } M}{N}$
- Let $g = \{g_1, g_2, \ldots, g_N\}$ be the complete data vector of indicator functions with

$$g_j = \begin{cases} 1 & \text{if } r_j \in M \\ 0 & \text{if } r_j \in U \end{cases}$$

Then the complete data likelihood function is:

$$\mathcal{L}(g, \gamma \mid m, u, \pi) = \prod_{j=1}^{N} \left[\, \pi \, \mathrm{P}(\gamma_j \mid r_j \in M)\right]^{g_j} \left[(1-\pi)\, \mathrm{P}(\gamma_j \mid r_j \in U)\right]^{1-g_j}$$

EM Algorithm, continued

We assume conditional independence, and therefore

$$\mathrm{P}(\gamma_j \mid r_j \in M) = \prod_{k=1}^{K} m_k^{\gamma_{kj}} (1 - m_k)^{1-\gamma_{kj}}$$
$$\mathrm{P}(\gamma_j \mid r_j \in U) = \prod_{k=1}^{K} u_k^{\gamma_{kj}} (1 - u_k)^{1-\gamma_{kj}}$$

Implementation of the EM Algorithm is then carried out with the following steps:

- Set the Initial Values
- The E Step
- The M Step
- Repeat the E Step and M Step until the desired level of precision is attained.

EM Algorithm, continued

Set the Initial Values

The initial values can be based on previous record linkage projects with similar comparison fields. Or, one might use a rough estimate based on analyzing a subset of the current files. The algorithm is not particularly sensitive to starting values and the initial estimates can be guesses, per Herzog (2007).

The E Step

$$\hat{g}_j = \frac{\hat{\pi} \prod_{k=1}^{K} \hat{m}_k^{\gamma_{kj}} (1 - \hat{m}_k)^{1-\gamma_{kj}}}{\hat{\pi} \prod_{k=1}^{K} \hat{m}_k^{\gamma_{kj}} (1 - \hat{m}_k)^{1-\gamma_{kj}} + (1 - \hat{\pi}) \prod_{k=1}^{K} \hat{u}_k^{\gamma_{kj}} (1 - \hat{u}_k)^{1-\gamma_{kj}}}$$

EM Algorithm, continued

The M Step

Partition the M Step into three distinct maximization problems.

1. $\hat{m}_k = \frac{\sum_{j=1}^{N} \hat{g}_j \gamma_{kj}}{\sum_{j=1}^{N} \hat{g}_j}, \quad k = 1, \ldots, K$
2. $\hat{u}_k = \frac{\sum_{j=1}^{N} (1-\hat{g}_j) \gamma_{kj}}{\sum_{j=1}^{N} (1-\hat{g}_j)}, \quad k = 1, \ldots, K$
3. $\hat{\pi} = \frac{\sum_{j=1}^{N} \hat{g}_j}{N}$

Repeat the E Step and M Step until the desired level of precision is attained.

An Illustration Using the EM Algorithm

We illustrate the approach on a set of manufactured data. We assume that a current site has 100 pairs of which 20 are true matches. This implies that blocking or other methods have been employed to reduce the number of pairs, otherwise the total matches could be at most 10. We manufacture the data with 10 comparison fields. The agreement patterns for the 100 pairs are binary, and are generated from a Binomial distribution with a probability of a match shown in the next table. We refer to these 100 pairs as the data at the current site.

The True Values of Parameters

Parameter	True Value	Parameter	True Value
π	0.200		
m_1	0.800	u_1	0.350
m_2	0.750	u_2	0.340
m_3	0.720	u_3	0.320
m_4	0.820	u_4	0.310
m_5	0.710	u_5	0.330
m_6	0.840	u_6	0.250
m_7	0.830	u_7	0.300
m_8	0.900	u_8	0.270
m_9	0.820	u_9	0.200
m_{10}	0.850	u_{10}	0.150

An Illustration Using the EM Algorithm, continued

After generating the data for the current site, we wish to use probabilistic record linkage to classify each pair as either a match or a non-match.

We assume that the marginal probabilities of each field are independent and use the EM Algorithm on the current site to estimate the m and u probabilities. We repeat this algorithm using the new estimate as the current estimate for each iteration until there is convergence on all parameters.

The results are extremely close to the actual probabilities used to generate the data, as shown on the next slide.

Results of the EM Algorithm

Parameter	True Value	EM Initial Values	EM Estimate	Error
π	0.200	0.180	0.198	0.002
m_1	0.750	0.800	0.747	0.003
m_2	0.700	0.800	0.695	0.005
m_3	0.650	0.800	0.660	-0.010
m_4	0.700	0.800	0.691	0.009
m_5	0.600	0.800	0.599	0.001
m_6	0.850	0.800	0.847	0.003
m_7	0.750	0.800	0.750	0.000
m_8	0.850	0.800	0.851	-0.001
m_9	1.000	0.800	1.000	0.000
m_{10}	0.950	0.800	0.944	0.006

Results of the EM Algorithm, continued

Parameter	True Value	EM Initial Values	EM Estimate	Error
u_1	0.300	0.280	0.302	-0.002
u_2	0.325	0.280	0.327	-0.002
u_3	0.313	0.280	0.311	0.001
u_4	0.263	0.280	0.266	-0.003
u_5	0.325	0.280	0.326	-0.001
u_6	0.238	0.280	0.240	-0.002
u_7	0.263	0.280	0.264	-0.001
u_8	0.325	0.280	0.326	-0.001
u_9	0.175	0.280	0.177	-0.002
u_{10}	0.163	0.280	0.166	-0.003

An Illustration Using the EM Algorithm, continued

Then we calculate the ratio by the Fellegi-Sunter method, and take the logarithm with base 2 of the ratio to get the “matching weight”. Since the EM Algorithm gave us the estimated proportion of true matches is .2 and we have 100 record pairs, we conclude that the 20 with the largest match weight values are links. The next table shows that we have no false matches and no false non-matches.

Results from the EM Algorithm, continued

Sorted	$\mathrm{P}(\gamma \mid r \in M)$	$\mathrm{P}(\gamma \mid r \in U)$	Ratio (R)	Matching Weight	Concl.
[9,]	0.0724587	0.0000016	44962.39	15.46	link
[4,]	0.0485488	0.0000033	14571.15	13.83	link
[5,]	0.0485488	0.0000033	14571.15	13.83	link
[3,]	0.0317674	0.0000033	9584.08	13.23	link
[6,]	0.0323388	0.0000045	7262.38	12.83	link
[18,]	0.0126731	0.0000033	3804.69	11.89	link
[14,]	0.0166929	0.0000099	1692.15	10.72	link
[2,]	0.0107494	0.0000077	1402.14	10.45	link
[13,]	0.0081648	0.0000104	784.50	9.62	link
[11,]	0.0109870	0.0000152	723.69	9.50	link
[20,]	0.0094995	0.0000189	501.68	8.97	link
[8,]	0.0084799	0.0000171	496.70	8.96	link
[15,]	0.0058285	0.0000141	412.74	8.69	link

Results from the EM Algorithm, continued

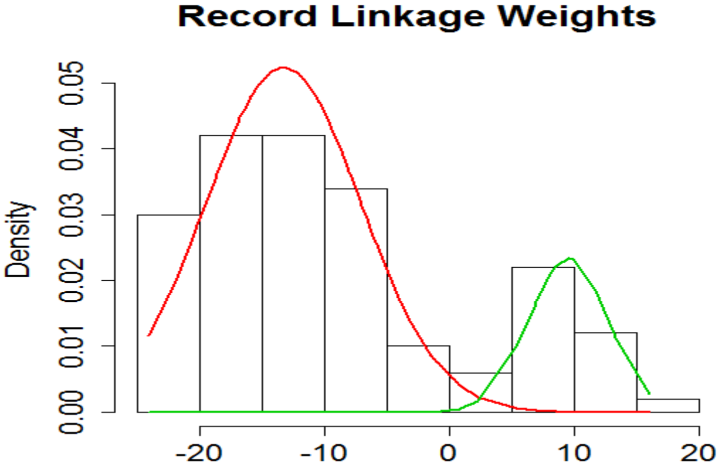
Sorted	$\mathrm{P}(\gamma \mid r \in M)$	$\mathrm{P}(\gamma \mid r \in U)$	Ratio	Matching Weight	Concl.
[1,]	0.0083452	0.0000206	404.91	8.66	link
[17,]	0.0070880	0.0000191	370.43	8.53	link
[19,]	0.0022136	0.0000171	129.69	7.02	link
[12,]	0.0019066	0.0000294	64.96	6.02	link
[10,]	0.0019409	0.0000394	49.22	5.62	link
[7,]	0.0019562	0.0000422	46.40	5.54	link
[16,]	0.0007471	0.0000415	18.00	4.17	link
[58,]	0.0001164	0.0001642	0.71	-0.50	
[40,]	0.0000943	0.0002973	0.32	-1.66	
[91,]	0.0000506	0.0004358	0.12	-3.11	
[68,]	0.0000385	0.0005917	0.06	-3.94	
[73,]	0.0000177	0.0006239	0.03	-5.14	

An Illustration Using the EM Algorithm, continued

To estimate the false-match rate, the method proposed by Fellegi-Sunter (1969), requires the summation of the probabilites of oberving gamma given that the record pair is not a match, for those record pairs in the group deemed a match. In our example, the false-match rate is the sum of the first 20 rows in the third column in the previous table which is approximately .00032. Furthermore, the estimated false-nonmatch rate is the sum of the last 80 rows in the second column which totals .00042. Note, that we assumed conditional independence to calculate the marginal probabilities, and if conditional independence truly held, this method would provide a reliable estimate. However, as mentioned in Section 2.3, while we can get good results by relying on the assumption of conditional independence to calculate the matching weights and determine which pairs are a link, even when the assumption does not hold, we cannot reliably estimate the false-match rate unless we make additional assumptions. One method which we illustrate next, is to make additional assumptions regarding the distribution of the matching weights.

Illustration Using a Bayesian Approach

The graphs of the probability density functions of the two normal distributions. The curve on the left represents the density function for the non-matches. The curve on the right represents the density function for the matches.



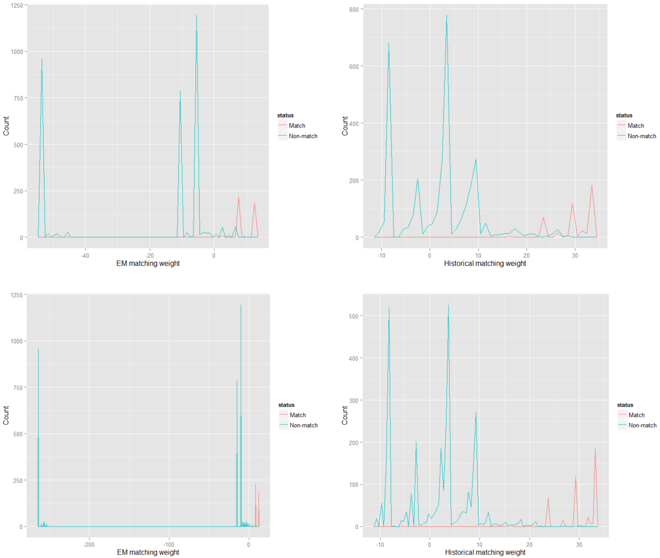


Figure 1: Frequency of EM and historical matching weights by true status of the 3,711 cases in Set 1 (top two) and the 3,615 cases in Set 2 (bottom two.)

Evaluation of error levels for false links

*

Table 3: A comparison of error levels for false links for two data sets when R is estimated using historical data (EM algorithm)

Set 1				Set 2			
μ	0.05	0.025	0.005	μ	0.05	0.025	0.005
Implied number of links	1172	920	713	Implied number of links	1077	825	619
	(488)	(359)	(226)		(429)	(347)	(223)
True status				True status			
links	441	441	441	links	441	441	441
	(426)	(358)	(225)		(427)	(347)	(223)
nonlinks	731	479	272	nonlinks	636	384	178
	(62)	(1)	(1)		(2)	(0)	(0)
True error level for false links	0.624	0.521	0.381	True error level for false links	0.591	0.465	0.288
	(0.127)	(0.003)	(0.004)		(.005)	(0)	(0)

Model

Notation:

- $y = (y_1, \cdots, y_n)'$, a $n \times 1$ vector of responses
- $X = col_{1 \leq i \leq n} x_i'$, a $n \times p$ design matrix, where x_i is a $p \times 1$ vector of known covariates

Model:

$$E(y) = g(X, \beta), \;\; V(y) = V(X, \beta),$$

where

- β is a $p \times 1$ vector of unknown coefficients
- $g(X, \beta) = [g(x_1, \beta), \ldots, g(x_n, \beta)]'$, a $n \times 1$ vector, where the function form of $g(\cdot)$ is known
- $V(X, \beta) = ((v_{ij}(X, \beta)))$, a $n \times n$ matrix, where the functional forms of $v_{ij}(\cdot)$, are known

Our goal is to estimate β when the true data pairs (x_i, y_i) $i = 1, \ldots, n$ are not observable. Instead, the record linkage procedure produces pairs $(x_i, z_i),\; i = 1, \ldots, n$ in which z_i may or may not correspond to y_i.

The Scheuren-Winkler (SW) Model

Scheuren and Winkler (1993) considered the following model for $z = (z_1, \ldots, z_n)'$ given y:

$$z_i = \begin{cases} y_i & \text{with probability } q_{ii} \\ y_j & \text{with probability } q_{ij} \text{ for } j \neq i, j = 1, \ldots, n, \end{cases}$$

where $\sum_{j=1}^{n} q_{ij} = 1,\ i = 1, \ldots, n.$

In our application, $q_{ij} \equiv q_{ij}(\psi)$ represents the matching probability of the pair (i, j) obtained from the mixture model.

Define

- $q_i \equiv q_i(\psi) = (q_{i1}, \ldots, q_{in})'$
- $Q \equiv Q(\psi) = col_{1 \leq i \leq n} q_i'$, a $n \times n$ matrix of matching probabilities depending on ψ

Marginal model on z

Note that under the SW model,

$$\begin{aligned} E(z|y) &= Qy \\ V(z|y) &= ((y' A_{ij} y)), \end{aligned}$$

where $A_{ij} \equiv A_{ij}(\psi)$, a $n \times n$ matrix depending on ψ.

Thus the marginal mean vector and variance covariance matrix of z are given by

$$\begin{aligned} E(z) &= Q(\psi) g(X;\beta) \equiv \eta(X;\beta,\psi), \text{ (say)} \\ V(z) &= Q(\psi) V(X;\beta) Q'(\psi) + B(X;\beta,\psi) \equiv \Sigma(X;\beta,\psi), \text{ (say)} \end{aligned}$$

where $B(X;\beta,\psi) = ((b_{ij}(X;\beta,\psi)))$, with

$$b_{ij}(X;\beta,\psi) = \text{tr}\left\{ A_{ij}(\psi) \left[V(X;\beta) + g(X;\beta) g'(X;\beta) \right] \right\}.$$

Estimation of β with the linked data

We propose to estimate β using the following optimal estimating equation:

$$\left[\frac{\partial \eta}{\partial \beta'}\right]' \Sigma^{-1}(z-\eta)=0,$$

where 0 is a $p \times 1$ vector of zeroes; $\frac{\partial g}{\partial \beta'}$ is a $n \times p$ matrix of partial derivatives. We denote this estimator $\hat{\beta}(\psi)$.

- For the special case of linear regression model, the estimator is identical to the unbiased estimator of β proposed by Lahiri and Larsen (2005). For recent development in this area, see Chambers (2009), Kim and Chambers (2012a,b).
- When ψ is unknown, we estimate β by $\hat{\beta}(\hat{\psi})$, where $\hat{\psi}$ is the estimator of ψ from the EM algorithm.
- The properties of $\hat{\beta}(\hat{\psi})$ are expected to be similar to those of $\hat{\beta}(\psi)$ because the distribution of the matching variables (e.g., last name, phone number), which determines the distribution of ψ, is usually independent of the response variable y (e.g., income) and hence of z.

Estimator of variance-covariance matrix of the proposed estimator

Note that

$$V[\hat{\beta}(\hat{\psi})] = E\left\{V[\hat{\beta}(\hat{\psi})|\hat{\psi}]\right\} + V\left\{E[\hat{\beta}(\hat{\psi})|\hat{\psi}]\right\}$$

- We generally expect the second term to be negligible. Then we can apply the general jackknife method proposed by Jiang, Lahiri and Wan (2002) to estimate $V[\hat{\beta}(\hat{\psi})|\hat{\psi}]$ and hence to estimate $V[\hat{\beta}(\hat{\psi})]$
- To incorporate the additional uncertainty due to the estimation of ψ (second term), we can apply parametric bootstrap method using the mixture model (see Lahiri 2003; Larsen and Lahiri 2005).

The Six Sigma Method

Goal: To create a linked data so that the probability of false links is minimized.

- n: the number of matches among the N records in the file.
- Assume that n is approximately $N[N\pi, \sigma^2 = N\pi(1-\pi)]$, for large N, where π is the probability of link.
- A conservative approach: Choose the maximum possible value of n, say n_{max} $(0 \le n_{max} \le N)$, such that $P(n < n_{max})$ is the minimum.
- If π is known and N large, we may take $n_{max} = N\hat{\pi} - 3\sigma$. But π is unknown and is estimated. To take care of the extra variability we propose to use $n_{max} = N\hat{\pi} - 6\hat{\sigma}$, where $\hat{\sigma} = \sqrt{N\hat{\pi}(1-\hat{\pi})}$.
- We call a method six sigma record linkage method if the linked file consists of n_{max} records from the top when they are sorted by the matching probability estimates in decreasing order.

Why $6\hat{\sigma}$?

- We borrowed the concept of $6\hat{\sigma}$ method from the statistical quality control literature.

- We also know that $P(n < n_{max})$ would be higher when π is estimated because of the extra variability due to the estimation of π. The variance of n that incorporates the extra variability due to estimation of π is given by

$$V(N\hat{\pi}) = E\left[N\hat{\pi}(1-\hat{\pi})\right] + V(N\hat{\pi}).$$

- The first term can be unbiasedly estimated by $N\hat{\pi}(1-\hat{\pi})$. The estimator $N\hat{\pi}(1-\hat{\pi})$ of $V(N\hat{\pi})$ ignores the second term $V(N\hat{\pi})$, which is of the same order as that of the first term. So doubling to $6\hat{\sigma}$ makes sense, although one can propose a more precise limit by estimating the second term by $[Nse(\hat{\pi})]^2$, where $se(\hat{\pi})$ is the asymptotic variance of $\hat{\pi}$.

Simulations – Set 1

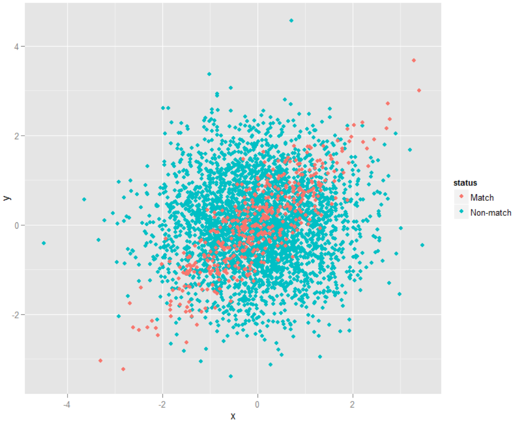


Figure 2: Scatter plot of simulated (x, y) for a typical replication and true status (link or nonlink) of the $N = 3,711$ cases in Set 1.

Simulations – Set 1

*

Table 4: EM estimate of $\pi = \mathrm{P}(\mathrm{M})$ from the mixture model and actual percentages of false match and non-match correct up to two decimal places for Set 1.

	Case 1				Case 2				Case 3			
Cutoff Method	Estimated $\mathrm{P}(M) - 3\sigma$				Estimated $\mathrm{P}(M) - 6\sigma$				True $\mathrm{P}(M)$			
Percent deemed matches		12.58%				10.86%				11.88%		
Linked data size		467				403				441		
	EM Algo		Historical		EM Algo		Historical		EM Algo		Historical	
Errors												
False matches	52	11.13%	56	11.99%	1	0.25%	37	9.18%	26	5.90%	56	12.70%
False nonmatches	26	0.80%	30	0.92%	39	1.18%	75	2.27%	26	0.80%	56	1.71%
Total	78	2.10%	86	2.32%	40	1.08%	112	3.02%	52	1.40%	112	3.02%

Simulations – Set 2

*

Table 5: EM estimate of $\mathrm{P}(M)$ from the mixture model and actual percentages of false match and non-match correct up to two decimal places for Set 2.

	Case 1				Case 2				Case 3			
Cutoff Method	Estimated $\mathrm{P}(M) - 3\sigma$				Estimated $\mathrm{P}(M) - 6\sigma$				True $\mathrm{P}(M)$			
Percent deemed matches		10.24%				8.60%				12.20%		
Linked data size		370				311				441		
	EM Algo		Historical		EM Algo		Historical		EM Algo		Historical	
Errors												
False matches	0	0.00%	2	0.54%	0	0.00%	0	0.00%	14	3.17%	3	0.68%
False nonmatches	71	2.19%	73	2.25%	130	3.93%	130	3.93%	14	0.44%	3	0.09%
Total	71	1.96%	75	2.07%	130	3.60%	130	3.60%	28	0.77%	6	0.17%

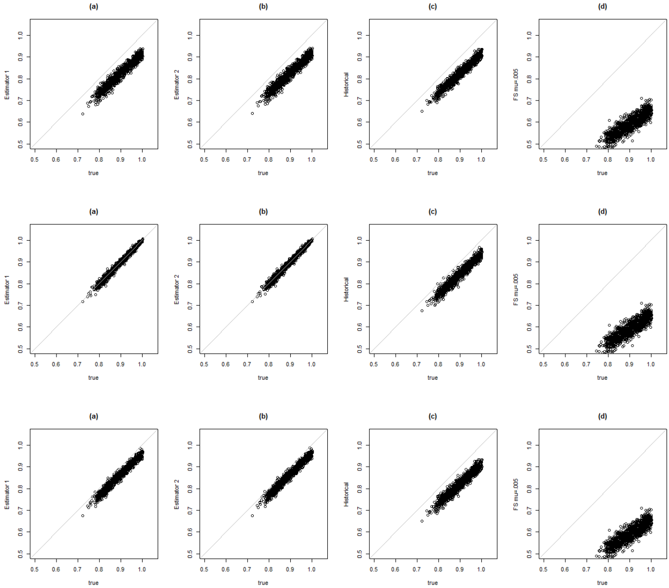


Figure 3: Comparison of four estimates of β for 1000 simulated datasets. Set 1, Cases 1, 2 and 3 of simulation conditions. Plots of (a) Estimator 1, (b) Estimator 2, (c) Historical weights (d) FS with $\mu = .005$ versus the OLS of β without mismatch errors. Diagonal lines have slope 1.

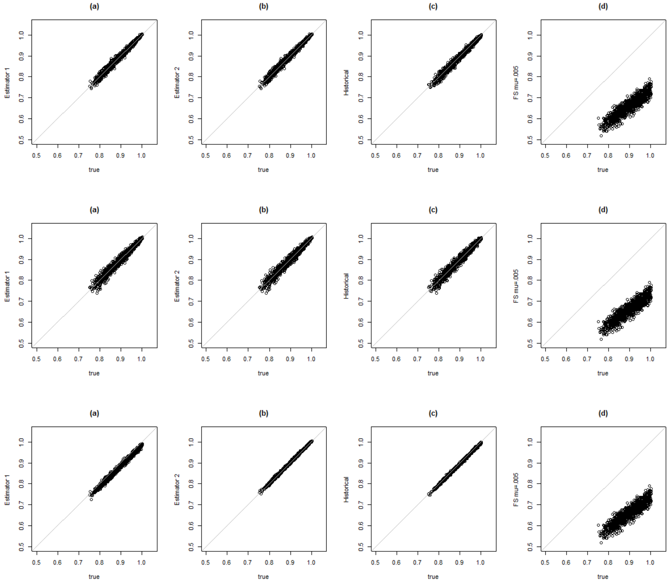


Figure 4: Comparison of four estimates of β for 1000 simulated datasets. Set 2, Cases 1, 2 and 3 of simulation conditions. Plots of (a) Estimator 1, (b) Estimator 2, (c) Historical weights (d) FS with $\mu = .005$ versus the OLS of β without mismatch errors. Diagonal lines have slope 1.

Simulations - Set 1

*

Table 6: Comparison of Average Absolute Deviations (AAD) and percent improvement of the proposed estimators of β (estimators 1-3) over the OLS based on the Fellegi-Sunter cutoff with $\mu = 0.005$

Estimator 3 is "historical"		Estimator 1	Estimator 2	Estimator 3
Case 1	AAD	0.075	0.073	0.075
	Impr. Over Hist	.99	1.02	
Case 2	AAD	0.006	0.006	0.056
	Impr. Over Hist	9.09	9.09	
Case 3	AAD	0.035	0.031	0.078
	Impr. Over Hist	2.26	2.37	

Estimator 3 is "FS method"		Estimator 1	Estimator 2	Estimator 3
Case 1	AAD	0.075	0.073	0.308
	Impr. Over FS	4.08	4.21	
Case 2	AAD	0.006	0.006	0.308
	Impr. Over FS	50.15	50.15	
Case 3	AAD	0.035	0.031	0.308
	Impr. Over FS	8.91	9.34	

Simulations - Set 2

*

Table 7: Comparison of Average Absolute Deviations (AAD) and percent improvement of the proposed estimators of β (estimators 1-3) over the OLS based on the Fellegi-Sunter cutoff with $\mu = 0.005$

Estimator 3 is “historical”		Estimator 1	Estimator 2	Estimator 3
Case 1	AAD	0.006	0.006	0.007
	Impr. Over Hist	1.20	1.20	
Case 2	AAD	0.009	0.009	0.009
	Impr. Over Hist	1.03	1.03	
Case 3	AAD	0.014	0.002	0.004
	Impr. Over Hist	.29	1.17	
Estimator 3 is “FS method”		Estimator 1	Estimator 2	Estimator 3
Case 1	AAD	0.006	0.006	0.238
	Impr. Over FS	38.39	38.39	
Case 2	AAD	0.009	0.009	0.238
	Impr. Over FS	26.91	26.91	
Case 3	AAD	0.014	0.002	0.238
	Impr. Over FS	17.01	68.65	

Comparison of $\hat{\beta}$ Set 1, Case 2

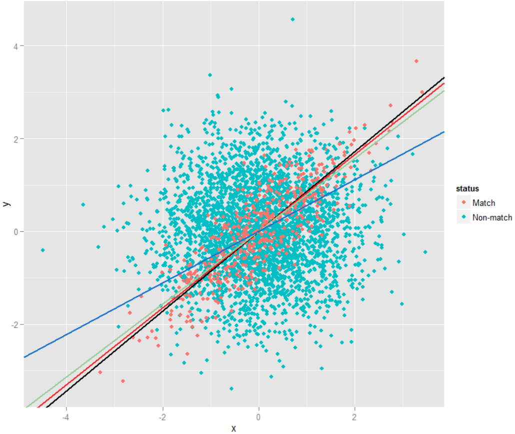


Figure 5: Comparison of three estimates on one replication. Set 1, Case 2 of simulation conditions. Regression lines of (a) the OLS of β without mismatch errors shown as a black line, (b) Estimator 2 shown as a red line, (c) Using historical is shown as a green line, and (d) Using FS method for cutoff with $\mu = .005$ shown as a blue line.

References

- J.B. Armstrong and J.E. Mayda, *Model-based estimaton of record linkage error rates*, Survey Methodology, 19 (1993) 137-147.
- L.E. Baum, T. Petri, G. Soules, and N. Weiss, *A maximization technique occurring in statistical analysis of probabilistic functions in Markov chains*, The Annals of Mathematical Statistics, 41(1) (1970) 164-171.
- M.P. Becker and I. Yang, *Latent class marginal models for cross-classifications of counts*, Sociological Methodology, 28 (1998) 293-325.
- Thomas R. Belin, *Evaluation of sources of variation in record linkage through a factorial experiment*, Survey Methodology, 19 (1993) 13-29.
- Thomas R. Belin and Donald B. Rubin, *A method for calibrating false-match rates in record linkage*, Journal of the American Statistical Association, Vol. 90, No. 430 (June, 1995) 694-707.
- G.E.P. Box and D.R. Cox, *An analysis of transformations (with discussion)*, Journal of the Royal Statistical Society, Ser. B, 26 (1964) 206-252.
- R. Chambers, *Regression analysis of probability-linked data*, Statisphere (2009), 4.

References, continued

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

- I.P. Fellegi and A.B. Sunter, *A theory for record linkage*, Journal of the American Statistical Association, Vol. 64 (1969) 1183-1210.

- T.N. Herzog, F.J. Scheuren and W.E. Winkler, *Data Quality and Record Linkage Techniques*, Springer, New York, NY, 2007.

- M.A. Jaro, *Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida*, Journal of the American Statistical Association, Vol. 84 (1989) 414-420.

- J. Jiang, P. Lahiri, and S-M. Wan, *A unified jackknife theory for empirical best prediction with M-estimation*, Annals of Statistics, 30 (2002) 1782-1810.

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

- G. Kim and R. Chambers, *Regression analysis under probabilistic multi-linkage*, Stat Neerl, 66 (2012) 64-79.

References, continued

- P. Lahiri and M.D. Larsen, *Analysis with linked data*, Journal of the American Statistical Association, Vol. 100, No. 469 (2005) 222-230.

- M.D. Larsen and D. B. Rubin, *Iterative automated record linkage using mixture models*, Journal of the American Statistical Association, Vol. 96 (2001) 32-34.

- X.L. Meng and D. B. Rubin, *Maximum likelihood estimation via the ECM algorithm: a general framework*, Biometrika, Vol. 80 (1993) 267-278.

- D. B. Rubin, *Multiple Imputation for Nonresponse in Surveys*, Wiley, New York, NY, 1987.

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

- Y. Thibaudeau, *The discrimination power of dependency structures in record linkage*, Survey Methodology, 19 (1993) 31-38.

- W.E. Winkler, *String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage*, Proceedings of the Section on Survey Research Methods, American Statistical Association (1990) 354-359.

The End