Cleanup and Statistical Analysis of Sets of National Files

William.e.winkler@census.gov

FCSM Conference, November 6, 2013

Outline

1. Background on record linkage
2. Background on edit/imputation
3. Current research on adjusting statistical analyses for linkage error
4. Example
5. Limitations of current models

Goal: Combine sets of files to create larger, cleaner sets of data for policy analyses.

Economics- Companies

```
   Agency A                Agency B

  fuel           ------>   outputs
  feedstocks     ------>   produced
```

Health- Individuals

```
   Receiving                Agencies
    Social Benefits          B1, B2, B3

   Incomes                  Agency I

   Use of Health            Agencies
    Services                 H1, H2
```

File A	Common	File B
A_{11} , ... A_{1n}	Name1, Addr1	B_{11},...B_{1m}
A_{21} , ... A_{2n}	Name2, Addr2	B_{21},...B_{2m}
.		.
.		.
.		.
A_{N1} , ... A_{Nn}	NameN, AddrN	B_{N1},...B_{Nm}

Issues:

1. Clean-up original source files
    a. Modeling/edit/imputation
    b. Data linkage (duplication)
2. Create merged file (data linkage)
3. Adjust statistical analysis for linkage error
    a. Enhancements to current elementary models
    b. Extensions using modeling/edit/imputation and statistical matching

Modeling/edit/imputation

Goals:
1. Fill-in (impute) missing data in a manner that preserves joint distributions in a principled manner
2. Remove contradictory data while preserving joint distributions

Fellegi and Holt (JASA 1976) – Theorem 1 states that implicit edits are necessary for filling in for missing/contradictory data.

Means of implementation: integer programming, set covering algorithms, logic programming, satisfiability

Winkler (1997): Set covering algorithms for enumerating all implicit edits that are 100 times as fast as those developed by IBM for the Italian Labour Force Survey (discrete data).

Winkler (2003): Theory connecting edit with modern imputation as in Little and Rubin (2002).

Winkler (2008, 2010): Theory/fast algorithms for DISCRETE system. Main modeling (EM fitting algorithm) fits a 0.5 billion contingency table in 1000 minutes with epsilon 10^-12 in 200 iterations. Synthetic data generated from models of high quality but allows as much as 80% re-identification risk with small cells. Extensions using convex constraints (Winkler *Ann. Prob.* 1990, 1993) better scale microdata using external constraints.

Methods suitable for junior analysts/programmers, allow clean-up of large national files.

Fellegi-Sunter model of record linkage (*JASA* 1969)

Main Likelihood Ratio and Associated Decision Rule

Two files **A** and **B**

Classify pairs in a product space **A** × **B** from two files A and B into M, the set of true matches, and U, the set of true nonmatches.

Fellegi and Sunter (1969), making rigorous concepts introduced by Newcombe (1959), considered ratios of probabilities of the form:

$$R = P(\gamma \in \Gamma \mid M) / P(\gamma \in \Gamma \mid U) \qquad (1)$$

where γ is an arbitrary agreement pattern in a comparison space Γ. For instance, Γ might consist of eight patterns representing simple agreement or not on the largest name component, street name, and street number.

Alternatively, each $\gamma \in \Gamma$ might additionally account for the relative frequency with which specific values of name components such as "Smith", "Zabrinsky", "AAA", and "Capitol" occur.

The ratio R or any monotonely increasing function of it such as the natural log is referred to as a matching weight (or score).The decision rule is given by:

If $R > UPPER$, then designate pair as a link. (2a)

If $LOWER \leq R \leq UPPER$, then designate pair as a possible link and hold for clerical review. (2b)

If $R < LOWER$, then designate pair as a nonlink. (2c)

The cutoff thresholds *UPPER* and *LOWER* are determined by a priori error bounds on false matches and false nonmatches. Rule (2) agrees with intuition. If $\gamma \in \Gamma$ consists primarily of agreements, then it is intuitive that $\gamma \in \Gamma$ would be more likely to occur among matches than nonmatches and ratio (1) would be large. On the other hand, if $\gamma \in \Gamma$ consists primarily of disagreements, then ratio (1) would be small.

Figure 1. Log Frequency vs Weight Links

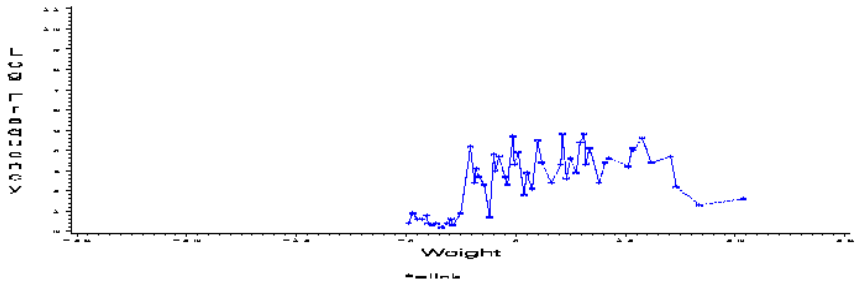


Figure 2. Log Frequency vs Weight Nonlinks

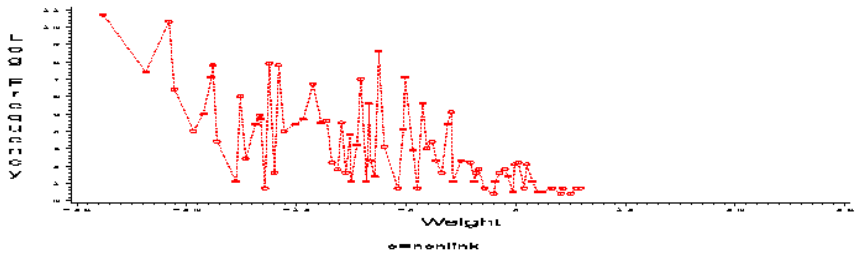


Figure 3. Log Frequency vs Weight Links and Nonlinks Combined

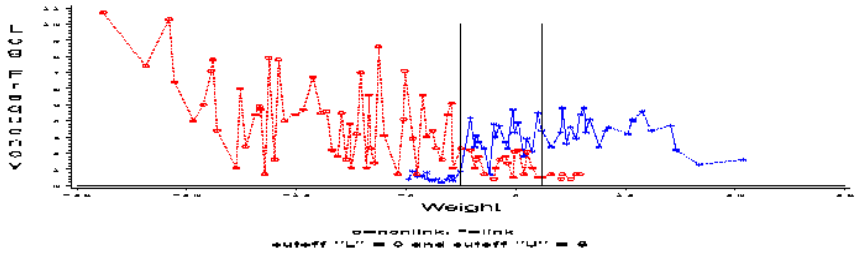


Figure 1a. Good Matching Scenario

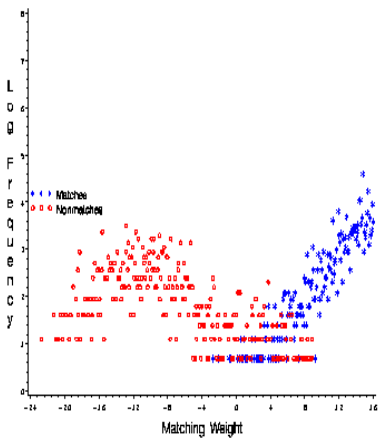


Figure 1b. Mediocre Matching Scenario

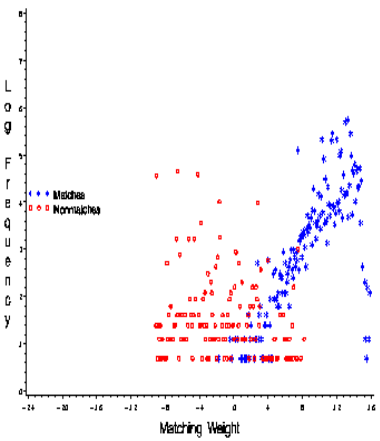


Figure 1c. 1st Poor Matching Scenario

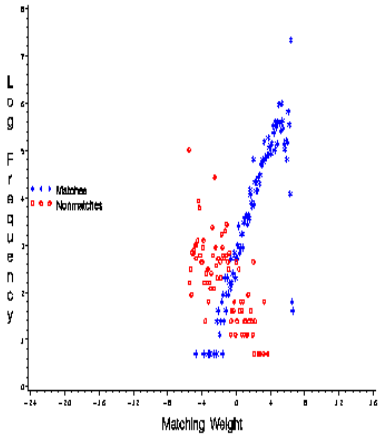
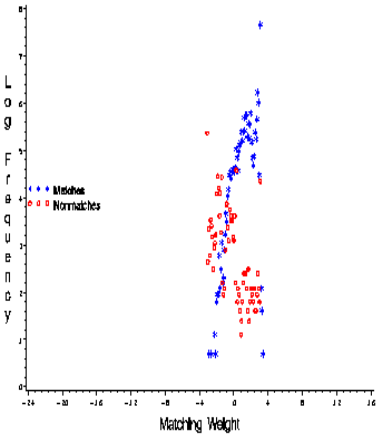


Figure 1d. 2nd Poor Matching Scenario



Adjusting statistical analyses for linkage error

Winkler (1991), Winkler and Scheuren (1991), Scheuren and Winkler (1993, 1997), Lahiri and Larsen (2005 *JASA*), Chambers et al. (2009, 2012), Chipperfield et al. (*SM* 2011), Tancredi and Liseo (*AoAP* 2011, *JOS* 2011), Goldstein et. (*Stat. Meth.* 2009, *Stat. in Medicine* 2011)

Research in infancy. Main starting need is estimating the 'true' probability of match for most pairs. Methods break down significantly with 'realistic' data.

Scheuren-Winkler (1993) **(**also Lahiri-Larsen 2000, 2005**)**

Files A and B are matched.

$Y = X\beta + \varepsilon.$

$$Z_i = \begin{cases} Y_i \text{ with probability } p_i \\ Y_j \text{ with probability } q_{ij} \text{ for } j \neq i, \end{cases}$$

$p_i + \sum_j q_{ij} = 1.$

$E(Z) = (1/n) \sum_i \mathrm{E}(Z|i) =$

$(1/n) \sum_i (Y_i\, p_i + \sum_j Y_j\, q_{ij}) =$

$(1/n) \sum_i Y_i + (1/n) \sum_i [Y_i\, (-h_i) + Y_{\varphi(i)}\, h_i] = \mu_Y + \mathrm{B},$

where $h_i = 1 - p_i$.

Under an assumption of 1-1 matching, for each $i = 1, \ldots, n$, there exists at most one j such that $q_{ij} > 0$. We let φ be defined by $\varphi(i) = j$.

Common thread of later research (after SW 93, 97)

1. All research assumes that given true probability of a match p_{ij} for every pair $(a_i , b_j) \in A \times B$.
Presently only Belin-Rubin (JASA 1995) and Winkler (2006, also 2002) have unsupervised learning models (somewhat inaccurate) for estimating these probabilities.

2. Most assume linear regression analysis is goal $Y = B X + \varepsilon$.
Issue: Masks (somewhat hides) some difficulties. Also, most make additional simplifying assumptions. All use simulated data that drastically reduces (usually eliminates) the messy-data situations of real data.

3. Chipperfield, Bishop & Campbell (SM 2011) provides a nice model for discrete data that serves to illustrate many difficulties that do not show up in models with continuous data.

Chipperfield et al. model

$(a_i , b_j) \in A \times B$. Only look at $x \in A$ and $y \in B$.
x is a univariate representation of multivariate; y is univariate representation
each $x_i \in A$ may be associated with many r_{Ai} ; similarly each y_j .

Build a loglinear model M on the pairs (x, y) obtained after record linkage.

Want to get better estimates (closer to original data) than the estimates from crude tabulations from observed data.

Underlying truth representation

$w^*_{ic|x} = 1$ if $y^*_i = c$, and $x_i = x$; else $w^*_{ic|x} = 0$.

Take a (likely very large) sample s_c to get (possibly only somewhat) good estimates of $w^*_{ic|x}$. Use EM model to get estimates for all $\hat{p}_{xy}$. The sample s_c gives

$$\hat{p}_{xy^*} = (\Sigma_{sc}\ w^*_{ic|x}\ \delta_i\ )/\ (\Sigma_{sc}\ w^*_{ic|x}). \qquad (6)$$

$$\tilde{\pi}_{c|x} = \tilde{n}_{c|x}/(\Sigma_c\ \tilde{n}_{c|x})^{-1}, \qquad (7)$$

where

$$\tilde{n}_{c|x} = \Sigma_i\ \tilde{w}_{ic|x}, \qquad (8)$$

$$\tilde{w}_{ic|x} = w^*_{ic|x}\ \hat{p}_{xy^*} + (1- \hat{p}_{xy^*})\ \tilde{\pi}_{c|x} \text{ if } i \notin s_c,$$

$$= w^*_{ic|x} \text{ if } i \in s_c \qquad (9)$$

$$= \tilde{\pi}_{c|x} \text{ if } i \in s_c \quad \text{and } \delta_i = 0 \ (\delta_i \text{ is indicator that true match}),$$

The (semi-supervised) EM procedure is

1. Calculate $\hat{p}_{xy*}$ from (6),
2. Initialize $\tilde{\pi}_{c|x}^{(0)}$ and then calculate $\tilde{w}_{c|x}^{(0)}$ from (9) and then $\tilde{n}_{c|x}^{(0)}$ from (8),
3. Calculate $\tilde{\pi}_{c|x}^{(t)}$ from (7) using $\tilde{n}_{c|x}^{(t-1)}$,
4. Calculate $\tilde{w}_{c|x}^{(t)}$ from (9) using $\tilde{\pi}_{c|x}^{(t)}$ and then calculate $\tilde{n}_{c|x}^{(t)}$ from (8) using $\tilde{w}_{c|x}^{(t)}$,
5. Iterate between 3 and 4 until convergence.

Empirical example (Chipperfield et al.):
Three values (employed, unemployed, not in labor force) against same values in another file for a later time period. Sample size 1000. ~100 for each combination of cells across time periods. Very little variation.

Data (2000 PUMS data for one State)

-------- A data ----------					-------- B data ------		
Sex	age	race	marit	educ	occup	house	income
2	16	2	5	16	3	5	40

2560 data patterns 600 data patterns

Split records – induce matching error

1. 8000 0.01 error
2. 8000 0.02 error
3. 8000 0.05 error
4 8000 0.08 error
5. 8000 0.12 error
6. 8000 0.15 error
7. 7926 0.20 error

1	015	1	5	05	10	04	000	1	2
1	015	1	5	05	10	04	002	.	3
1	015	1	5	05	10	04	004	.	1
1	015	1	5	05	10	04	005	.	1
1	015	1	5	05	10	06	027	.	1
1	015	1	5	06	00	00	000	165	151
1	015	1	5	06	00	00	001	12	18
1	015	1	5	06	00	00	002	.	2
1	015	1	5	06	00	00	005	.	1
1	015	1	5	06	00	02	000	4	4
1	015	1	5	06	00	03	001	1	1
1	015	1	5	06	00	04	000	2	2
1	015	1	5	06	00	04	001	1	1
1	015	1	5	06	00	04	003	.	1
1	015	1	5	06	00	04	004	.	1
1	015	1	5	06	00	04	006	.	2
1	015	1	5	06	00	05	003	.	2
1	015	1	5	06	00	06	003	.	1

Discrete data (contingency table). A few large clumps (higher counts in a few cells) and very many small cells.

After matching: Counts in large cells have lower counts. Many sampling zeros now have small counts (less than 5).

Issue 1. A sample of size 0.25 of the number of pairs will not yield accurate estimates of $\hat{p}_{xy*}$ for most pairs. Indeed a substantial number of estimates will be zero.

Issue 2. $\tilde{\pi}_{c|x}^{(0)}$must be dispersed over 600 values (all but one false).

The two loglinear models.
There is no way for the Chipperfield et al. procedure to get reasonable estimates for $\tilde{p}_{xy*}$ for most pairs (in most data situations).
Later model after matching has lost many interaction components.
Many coefficients in the original model are lost.

How to improve.

1. Extension of correction model with more detail of the matching process. Allows *far smaller* sample size and better targeting of cells in sample.

2. Large count cells in merged data must be present in original data (i.e., must be true matches). Allows further better targeting of sample to cells.

3. Edit restraints using subject matter expertise and auxiliary files. Can eliminate many false matches that can then be rematched.