Introduction to Probabilistic Record Linking

John M. Abowd and Lars Vilhuber
March 2005
Overview

- Introduction to probabilistic record linking
  What is record linking, what is it not, what is the theory?
- Probabilistic record linking: Applications and examples
  How do you do it, what do you need, what are the possible complications?
- Examples of record linking
  Do it yourself record linking
From imputing to linking

- "Probabilistic record linkage"
  Merge match with imperfect link variables

- "Massively imputed"
  Common variables/values, but datasets can’t be linked

- "Simulated data"
  No common variables, only moments transferred

- "Classical"
  Merge match by link variable

Availability of linked data vs. Precision of link
Definitions of record linkage

• “a procedure to find pairs of records in two files that represent the same entity”
• “identify duplicate records within a file”
Uses of record linkage

- Merging two files for microdata analysis
  - CPS base survey to a supplement
  - SIPP interviews to each other
  - Merging years of the Business Register
  - Merging two years of CPS
  - Merging financial info to firm survey
- Updating a survey frame
  - Based on business lists
  - Based on tax records
- Disclosure review of potential public use microdata
- Data mining by businesses and law enforcement
Types of record linkage

- Merging two files for microdata analysis
  - CPS base survey to a supplement
  - SIPP interviews to each other
  - Merging years of Business Register
  - Merging two years of CPS
  - Merging financial info to firm survey
  - Deterministic linkage: survey-provided IDs
  - Probabilistic linkage: imperfect or no IDs

- Updating a survey frame
  - Based on business lists
  - Based on tax records
  - Probabilistic linkage: imperfect or no IDs

- Disclosure review of potential public use microdata
  - Probabilistic linkage: no IDs
Basic definitions and Notation

• Set of entities $A$, $B$
• Associated files $\alpha(A)$, $\beta(B)$
• Matches

$$M = \{(\alpha(a), \beta(b)) | a = b\}$$

• Nonmatches

$$U = \{(\alpha(a), \beta(b)) | a \neq b\}$$
Agreement patterns

- Comparison space
  \[ \alpha(a) \times \beta(b) \rightarrow \Gamma \]
- Comparison vector
  \[ \gamma \in \Gamma \]
- Components of comparison vector take on finitely many values, typically \( \{0,1\} \)
Example Agreement pattern

• 3 binary comparisons test whether
  - $\gamma_1$ pair agrees on last name
  - $\gamma_2$ pair agrees on first name
  - $\gamma_3$ pair agrees on street name

• Sample agreement pattern:
  $\gamma=(1,0,1)$
Linkage rule

- A linkage rule defines a record pair’s status based on its agreement pattern
  - Link (L)
  - Undecided (Clerical, C)
  - Non-link (N)

\[ F : \Gamma \rightarrow \{L, C, N\} \]
Conditional probabilities

• Probability that a record pair has agreement pattern $\gamma$ given that it is a match [nonmatch]
  
  \[
P(\gamma|M) \\
P(\gamma|U)
  \]

• Agreement ratio
  
  \[
  R(\gamma) = \frac{P(\gamma|M)}{P(\gamma|U)}
  \]

  This ratio will determine the distinguishing power of the comparison $\gamma$. 

© John M. Abowd and Lars Vilhuber 2005, all rights reserved
Error rates

• **False match**: a linked pair that is not a match (type II error)

• **False match rate**: probability that a designated link (L) is a nonmatch: \( \mu = P(L|U) \)

• **False nonmatch**: a nonlinked pair that is a match (type I error)

• **False nonmatch rate**: probability that a designated nonlink is a match: \( \lambda = P(N|M) \)
Fundamental Theorem

1. Order the comparison vectors \( \{ \gamma^j \} \) by \( R(\gamma) \)

2. Choose upper \( T_u \) and lower \( T_l \) cutoff values for \( R(\gamma) \)

3. Linkage rule:

\[
F : \begin{cases} 
\gamma_j \in L & \text{if } R(\gamma_j) \geq T_u \\
\gamma_j \in N & \text{if } R(\gamma_j) \leq T_l \\
\gamma_j \in C & \text{else}
\end{cases}
\]
Fundamental theorem (cont.)

• Error rates are

\[
\mu = \sum_{\gamma^j \in \Gamma} P(\gamma^j \mid U)P(L \mid \gamma^j) = \sum_{\gamma^j \in L} P(\gamma^j \mid U)
\]

\[
\lambda = \sum_{\gamma^j \in \Gamma} P(\gamma^j \mid M)P(N \mid \gamma^j) = \sum_{\gamma^j \in N} P(\gamma^j \mid M)
\]
Fundamental Theorem (3)

- Fellegi & Sunter (JASA, 1969):
  If the error rates for the elements of the comparison vector are conditionally independent, then given the overall error rates \((\mu, \lambda)\), the the linkage rule \(F\) minimizes the probability associated with an agreement pattern \(\gamma\) being placed in the clerical review set. \((\text{optimal linkage rule})\)
Applying the theory

• The theory holds on any subset of match pairs (blocks)

• Ratio $R$: matching weight or total agreement weight

• Optimality of decision rule heavily dependent on the probabilities $P(\gamma|M)$ and $P(\gamma|U)$
Acknowledgements

• This lecture is based in part on a 2000 and 2004 lecture given by William Winkler, William Yancey and Edward Porter at the U.S. Census Bureau


• Examples are all purely fictitious, but inspired from true cases presented in the above lecture, in Abowd & Vilhuber (2004).