Efficient modelling of record linked data

A missing data perspective

Harvey Goldstein

Record Linkage Methodology Research Group
Institute of Child Health
University College London
and
Centre for Multilevel Modelling
University of Bristol

University of BRISTOL

UCL

Record Linkage

- Consider 2 data files: File of interest (FOI) and the linking data file (LDF) and assume for simplicity all cases within FOI also exist in LDF
- Readily extended to multiple LDFs and cases missing from LDF as in deaths file
- We want variables of interest (VOI) from LDF to add to FOI records and we have a set of *matching variables* (MV) that enable us to link records in each file.
- Deterministic matching relies on a unique (and error free) combination of MV values having a one-to-one relationship from the FOI to the LDF.
- *Probabilistic record matching* arises in the common case when this cannot be assumed, e.g. misspelling of names or transcription errors, resulting in many *possible* matches. These are traditionally assigned 'matching weights'

Think of it as a missing data problem

Extended FOI contains 2 sets of variables: Set A where all are missing (i.e. in LDF) and set B which are available – with possible holes

Set A variables		Set B variables		
0	0	X	X	X
0	0	X	0	X
0	0	X	X	0
0	0	0	0	X

Research problem is to change the 0s to Xs. This is a particular case of missing data and our approach is to use an extension of Multiple Imputation (MI) techniques.

The focus is on data analysis.

Applying MI to extended FOI

- We cannot directly use MI for set A since all of them are missing.
- So: consider filling in some of them *with certainty* from a LDF – the deterministic matching stage.
- We now have something like this – where first record has no definite match

Set A		Set B		
0	0	X	X	X
0	X	X	0	X
X	X	X	X	0
X	X	0	0	X

Note e.g. that some of the imported values may be missing.

At this point we might choose simply to use *multiple imputation* for the remaining missing data since we have information to do this.

This can often produce acceptable estimates – e.g. if data MAR.

Can we do better by using *probabilistic* importation of data values?

Probabilistic record matching as it exists

- First – the deterministic stage - we ascertain and link the 'certain' records
- For unmatched residue probabilistic matching method produces a weight for each unlinked 'candidate' LDF record, for each remaining FOI record.
- If the maximum of these over the LDF records is greater than a chosen threshold a match is accepted for the LDF individual corresponding to this maximum value:
  - Consider say 3 matching variables sex, DOB, name, we may observe a pattern g=$\{d_1, d_2, d_3\}$ where the d_i are distance or similarity measures
  - We compute the probability of observing that pattern of values
  - A) Given that it is a match P(g|M)
  - B) Given that it is not a match P(g|NM)
  - Then compute R=P(g|M)/P(g|NM) and W=log(R)
  - R can be obtained via e.g. training data or otherwise estimated using existing data where matching status is known.
  - The cut-off threshold for W to accept a match has to be chosen, for example, to minimise the percentage of 'false positives'.
  - In practice W is computed for each MV and then summed and cut-off based on sum. Essentially an 'independence' assumption.

Probabilistic matching - problems

- A threshold has to be chosen: some possible matches therefore rejected
- Even if threshold is high some chosen matches will be wrong and these ‘measurement errors’ should be carried through to the analysis, but typically are not.
- (Jaro, M. (1995). "Probabilistic linkage of large public health data files." Statistics in Medicine **14**: 491-498.)
- What we really want, for data analysis purposes, is not to carry the record, but the LDF data values – the VOI.
- So consider the following:

Extending the probabilistic matching model

- If we can assign, for each ‘candidate’ record in the LDF a probability that it is the correct match, then we can adapt our imputation by treating these probabilities as constituting a ‘prior’ distribution.
- Formally we combine the imputation likelihood for the missing set A variables with the prior for each candidate record to form a posterior distribution for these records from which we choose the largest.
- We also can choose a lower threshold so that if none exceeds then standard MI is used.
- To obtain MAR we can condition on the matching variables as well as all other variables in the model of interest (MOI) in obtaining the imputation likelihood.
- Especially useful when probability of a correct match depends on values of the LDF variables.

Advantages

- Combining prior and likelihood will tend more often to select the correct record.
- Some bias will still remain but can be minimised since threshold for acceptance can be made very high (e.g. a probability of 0.95)
- At imputation stage we can condition on auxiliary variables to satisfy ignorability assumption (MAR)
- If elimination of bias is priority and a large enough proportion can successfully be unequivocally matched then standard MI can be used.

Implementation and Software

- Multiply imputed datasets produced and model fits combined in usual way (Rubin's rules).
- Matlab routines available and new STATJR software at Bristol will develop these and improve efficiency.
- Currently will handle mixtures of normal and binary variables and also multilevel data.
- If implemented routinely requires ancillary data (allowing matching probabilities to be estimated) from the matching process to be supplied to data analyst.
- PPRL procedures need to recognise that 'matching probabilities' need to be transferred along with encrypted (hashed) MV values.

- Goldstein, H., Harron, K., and Wade, A. (2012). The analysis of record linked data using multiple imputation with data value priors. Statistics in medicine, DOI: 10.1002/sim.5508