Introduction to data linkage

Juan Claramunt

Leiden University

January 2020

Software

- R package: "RecordLinkage"
- Github: jclaramunt

Introduction

- Combine data sources
- Improve quality

Introduction

- Since 2000's:
  - Big data
  - Digitized registers (administrative data)
  - Bio banks
- Huge amount of data
- Different levels of quality
- Different goals

Introduction

- Types:
  - Deterministic
  - Probabilistic

Deterministic linkage

- Linkage key
  - Unique (BSN)
  - Not unique (Gender+Postal code+Date of birth)
- Linkage rules

Deterministic linkage

Decision rule for pair (i,j):

$$x_{ij} = \begin{cases} 1 & \text{if } f_{ij} \geq \beta \\ 0 & \text{otherwise} \end{cases}$$

where:

- β is the number of key variables that must agree.
- $f_{ij} = \sum_k \gamma_{kij}$
- γ_{kij} (linkage rule) is the indicator function of the agreement of the record i and j on the variable k.

Deterministic linkage

- Depends on data quality
- Higher probability of false negatives (missed matches)
- Not unique keys might lead to false positives (false matches).

Probabilistic linkage

- Generate all possible pairs
- Large datasets: Blocking. Ideally using high quality variables.
- Weights are computed for each pair. Determine a treshold depending on the linkage quality
- Link pairs with a weight above the treshold.

Probabilistic linkage

- Two important concepts:
  - m_i is the probability that variable i has the same value given that both records are from the same unit.
  - u_i is the probability that variable i has the same value given that both records are **NOT** from the same unit.
- m and u need to be estimated (e.g. using EM algorithms).

Probabilistic linkage

- How do we define the weights for the linkage?
- Option 1:

$$w_i = \begin{cases} ln(\frac{m_i}{u_i}) & \text{if value of } i \text{ is equal/similar} \\ ln(\frac{1-m_i}{1-u_i}) & \text{if value of } i \text{ is different} \end{cases}$$

$$w = \sum_i w_i$$

Probabilistic linkage

- Option 2:

$$w_i^a = log_2(\frac{m_i}{u_i})$$

$$w_i^d = log_2(\frac{1-m_i}{1-u_i})$$

$$w = \sum_i((w_i^a - w_i^d)\delta_i + w_i^d) = \sum_i(\delta_i w_i^a + (1-\delta_i)w_i^d)$$

where $0 \leq \delta_i \leq 1$.

Probabilistic linkage

- Choose a treshold:
  - Optimize false negatives-false positives trade-off.
  - Divive the pairs in three groups: Correct pairs, incorrect pairs and a group for manual inspection.
  - Minimize manual inspection.
- Pairs above the treshold are linked.

Determine quality

- Biggest issue.
- Lack of data.
- Legal (privacy) issues.
- Property issues.

Determine quality

- Gold standard data.
  - Complete dataset.
  - Complete subset.
- Precision (TP/(TP+FP)).
- Sensitivity (TP/(TP+FN)).
- Match rate ((TP+TN)/TR).
- Identify where the method fails.

Determine quality

- Selection bias?
- Compare characteristics of linked and unlinked data.
  - Standardized differences.

Determine quality

- Sensitivity analysis.
  - Linkage rules.
  - Linkage algorithm.
  - Linkage weights.
  - Different tresholds.

Determine quality

- Manual check.
  - Check manually linked records.
  - Identify random errors.
  - Time consuming.

Alternative

- Missing data problem.
  - k-NN
  - PMM
  - MMSM

SPAM

- Multivariate Mixed Method for Statistical Matching (MMSM)
  - Bayesian regression using additional dataset
  - k-NN
  - Hard constraints
  - Soft constraints.

Simulation study

- Compare both methods when there is not a unique linkage key.
- Factors:
  - Size (1600, 16000, 160000).
  - Overlap (10, 60 and 90%).
  - Quality of linkage variables (10, 20 and 30% of errors in linkage variables).
- Add random and systematic errors.

Simulation study

- Evaluation criteria:
  - Precision.
  - Sensitivity.
- Find optimal balance between precision and sensitivity.

Simulation study. Results and conclusions.

- Deterministic:
  - Higher precision.
  - Lower sentitivity.
  - Use it if the overlap is small, the % of errors is small and a high quality linkage key is available.
- Probabilistic:
  - Higher precision.
  - Lower sentitivity.
  - Use it in the remaining cases.