Data Integration, Data Linkage, and Linked Data Analysis

Tiziana Tuoto

Italian National Institute for Statistics, Istat

tuoto@istat.it

IASS Webinar 38: March 27, 2024

Outline

Data Linkage

Record Linkage aka entity resolution, de-duplication

the process of merging together potentially noisy data sets in the absence of a unique identifier, both to remove duplicated information and to increase the informative content of each single file.

Data Linkage: Motivations

Linking data sets is useful for different and complementary purposes:

1. to obtain a larger reference data set or frame, with a more complete unit-level picture, to be used repeatedly
   - Cover several tasks
   - Improve statistics quality
2. to enable more comprehensive population studies, model-based statistical analyses via the additional information, one-off studies
   - Linear and non-linear regressions
   - Capture-recapture models for population size estimation
   - Small area estimation
   - Survivor analysis

Data Linkage: Identifiers

- When unique identifiers are known exactly, the linkage process can be accomplished without errors, no specific consequences on the downstream statistical procedures.
- When unique identifiers are not available, the linkage is based on common matching variables and some decision rules. We must deal with the uncertainty related to the linkage step.
  - Basic logic: if records agree (disagree) on most of the matching variables, it is likely (unlikely) that they might correspond to the same entity.
  - Intermediate situations: when only some matching variables agree, we need to make a decision whether the records are a true match or a non-match.

Challenges in Matching Variables

Matching variables can show discrepancy and false agreement.

Discrepancy

Disparity between two data sources concerning the same information

It can be due to errors, variations, and missing data.

- Differences in data collection and maintenance.
- Data changes over time.
- Strings: Misspelling, transpositions, fused or split words, persons with the same name, nicknames.
- Numerical variables: Rounding, transposed numbers, insertions, deletions.

Challenges in Matching Variables

Robbie and Robin Williams

Anne Hathaway and Shakespeare's wife

Micheal Douglas and (aka) Micheal Keaton

Marilyn Monroe and Norma Jeane Mortenson Baker Monroe

Two approaches to the linkage

Deterministic Linkage

A fixed set of rules determines which records are matches and which are not.

Probabilistic Linkage

A statistical model deals with agreements and discrepancies in matching variables to estimate some linkage probability.

In this talk, we focus on probabilistic linkage

The Fellegi and Sunter model

Fellegi, I. and Sunter, A. (1969) A Theory for Record Linkage. Journal of the American Statistical Association, 64, 1183-1210.

- Fellegi and Sunter (1969), formalizing Newcombe's ideas, is the first statistical framework for record linkage.
- Jaro (1989) proposed an EM-based latent class algorithms.
- Since then, several improvements have been proposed.

Fellegi and Sunter model in a nutshell

Main steps:

- Pair-wise comparison: records from the different data sources are compared each other
- Agreement/disagreement: a metric is used to measure the "similarity" between pairs
- A decision (either based on a test or a posterior probability) is taken, to classify all the pairs between **links** and **non links**
- Output: among all the compared pairs, few links and a huge number of non links. Downstream analyses usually only involve links.

Linkage Errors

The classification procedure aims at minimising two kinds of errors:

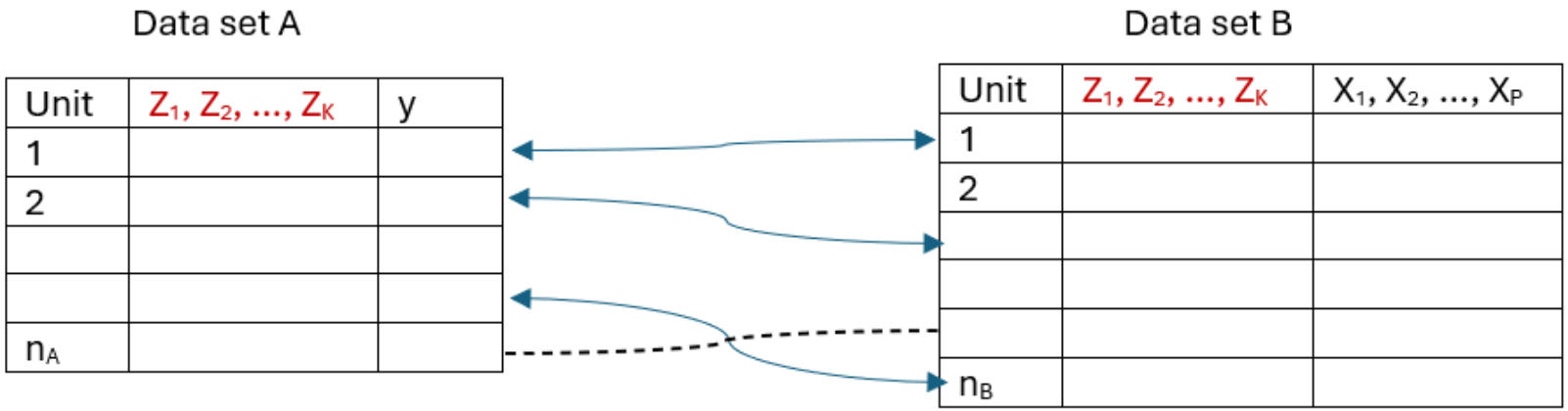
Linkage Errors

- false links or false positive, when a pair is classified as a link but actually the two records refer to different units,
- missing match or false negative, when the pair is classified as a non-link but actually the two records belong to the same unit.

The Fellegi and Sunter model: formally

Simplest case of 2 data sets A and B.

- Data set A, n_A records, and data set B, n_B records
- $Z_1, Z_2, ..., Z_k$ are common matching variables



The pairs space

Ω is the Cartesian product of records in the two data sets, $|\Omega| = n_A \times n_B$

$$(a, b) \in A \times B = \Omega$$

The set Ω composed by all the pairs of records can be split in two disjoint sets of pairs:

- Matches: $M = (a, b) : a = b$
- Non-matches: $U = (a, b) : a \neq b$

$$M \cup U = \Omega \text{ and } M \cap U = \varnothing$$

The objective of a record linkage method is to determine what pairs are in M

Configuration matrix

All the pairs in $\Omega = A \times B$ can be represented as a matrix C, where every row refers to a unit in A, and every column to a unit in B. The elements in the matrix, c_{ab}, $a \in A$, $b \in B$, can only assume two values:

- $c_{ab} = 1$ if $(a, b) \in M$
- $c_{ab} = 0$ if $(a, b) \in U$

The record linkage problem can be formulated as the estimation of the matrix $C = \{c_{ab}\}$

Comparison Function

For each pair (a, b), the matching variables $Z_1, Z_2, ..., Z_k$ are compared according to some comparison function:

$$\gamma_{ab,k} = d_k(z_{a,k}, z_{b,k})$$

d_k is a comparison (similarity) function: how similar are the values of the matching variable Z_k on the units a and b.
The comparison function d_k facilitates the classification of the pairs:

- high level of similarity for the pairs $(a, b) \in M$
- low level of similarity for the pairs $(a, b) \in U$

The Mixture

We can **observe** the overall frequency distribution for the comparison vector $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_k)$ in Ω

$$f(\gamma_k) = \frac{n(\gamma_k)}{n_a \times n_b}$$

where $n(\gamma_k)$ is number of comparisons in Ω with values γ_k
The observed distribution is a mixture of two unobserved distributions:

- the frequency distribution of the comparison γ_k in M: $m(\gamma_k) = \frac{n_m(\gamma_k)}{|M|}$ where $n_m(\gamma_k)$ is number of comparisons in M
- the frequency distribution of the comparison γ_k in U: $u(\gamma_k) = \frac{n_u(\gamma_k)}{|U|}$ where $n_u(\gamma_k)$= number of comparisons in U

weighted by the proportion of matches in Ω: $p = \frac{|M|}{n_a \times n_b}$

Estimation

- The observed distribution for the comparison vector $P(\gamma)$ is a mixture

$$P(\gamma) = P(\gamma|M)p + P(\gamma|U)(1-p) = m(\gamma)p + u(\gamma)(1-p)$$

- EM algorithm estimates $(m(\gamma), u(\gamma), p)$ under some simplifying assumptions, e.g. conditional independence of the matching variables
- The (log-)likelihood ratio

$$w_{ab} = P(\gamma|M)/P(\gamma|U) = m(\gamma)/u(\gamma)$$

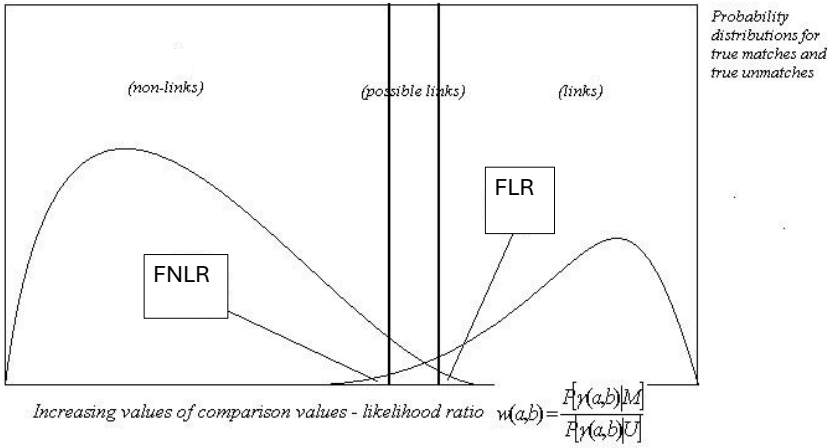
is then used to classify the pair (a, b)

Decision

Classification of the pairs:

- $w_{ab} > T_M \Rightarrow$(a,b) $\in$ Links
- $T_M \geq w_{ab} \geq T_U \Rightarrow$(a,b) $\in$ Possible Links
- $w_{ab} < T_U \Rightarrow$(a,b) $\in$ No-Links

T_M and T_U are chosen to minimize the size of Possible Links as well as linkage errors



- False links FLR $= P(w_{ab} \geq T_M | ab \in U)$
- False non-links FNLR $= P(w_{ab} < T_U | ab \in M)$

FS model in action

It is a competitive unsupervised probabilistic approach for record linkage, robust wrt the classification of the pairs.

Scalable, fast and efficient open-source implementation of the FS model based on the EM algorithm in the R package `fastLink`

> Enamorado T, Fifield B, Imai K. Using a Probabilistic Model to Assist Merging of Large-Scale Administrative Records. American Political Science Review. 2019;113(2):353-371.

Main challenges with FS model and some research directions

- The pairs are not independent: NO i.i.d.
- Scalability
  - Naive comparison of all record pairs is quadratic: n_A=100, n_B=100, $|M|$=10, $\Omega = 10^4$, $|U| = 9990$
  - for large files remove all likely no-matches is essential: blocking procedures
- 1-1 matching, i.e. a given record should be matched at most once, requires ad hoc solutions
- Supervised machine learning for alternative estimation, based on labeled samples
- Extensions to deal with the frequency distribution of the matching variables
- Extensions to three and more files, preserving transitivity
- Privacy preserving RL, with encrypted identifiable data

The Bayesian approach

- Bayesian version of the classical FS approach: the likelihood function provided by the set of multiple comparison vectors is used to estimate the matching configuration matrix C via MCMC methods, e.g. Fortini et al (2001), Larsen (2005), Sadinle (2017)
- As in FS approach, they work with a small set of highly informative variables
- Scalability

A Bayesian application: RL using diagnosis codes

> Hejblum, B., Weber, G., Liao, K. et al. (2019) Probabilistic record linkage of de-identified research datasets with discrepancies using diagnosis codes. Nature Scientific Data.

De-identified datasets:

- No personal identifiers are available
- Dates are considered protected health information
- Only diagnoses codes, in a standard vocabulary (ICD-x)

A Bayesian application: RL using diagnosis codes

- Diagnoses are informative characteristics, they uniquely identify most units within a single datasets
- Peculiarity: high dimensionality of diagnosis codes and their sparse information content
- Presence of a diagnosis is usually more informative than its absence, but with common code the absence is quite informative

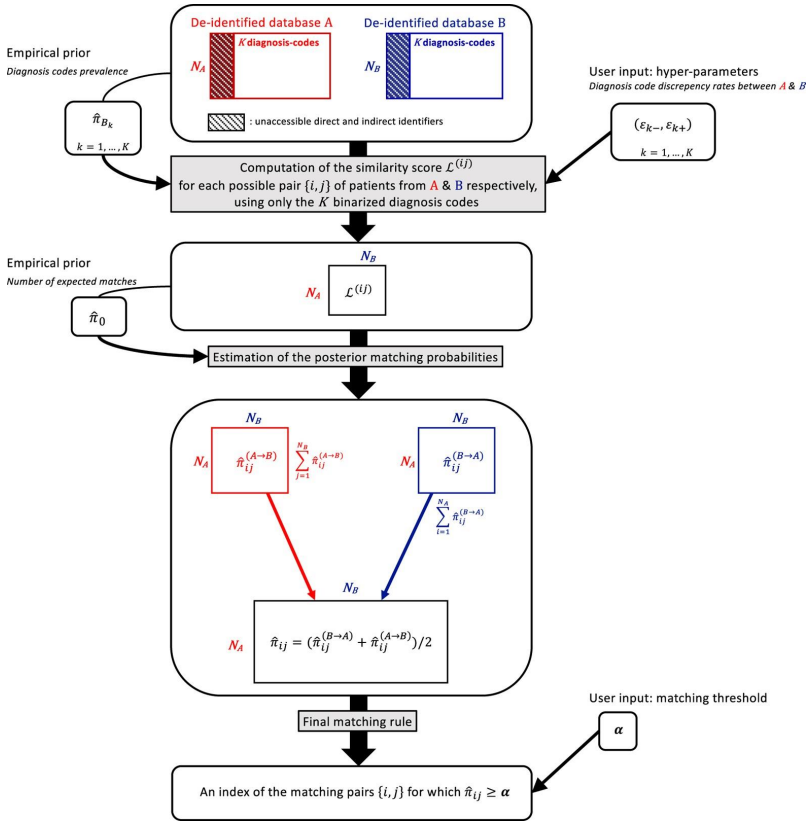
A Bayesian application: RL using diagnosis codes

- Data set A: composed by n_A units
- Data set B: composed by n_B units
- K common variables: binarized diagnosis codes - K are thousands
- $A^{(i)}$ vector of the K variables for unit i from A
- $B^{(j)}$ vector of the K variables for unit j from B
- $M_{(ij)}$ is the binary indicator of whether i and j are a match
- π_{ij} is the posterior probability of being a match for the pair i, j

Bayesian model

- Empirical Bayes priors for the marginal prevalence of codes: $\pi_{A_k} = P(A_k^i = 1)$ distribution of $A^{(i)}$ does not depended on $M_{(ij)}$. Same for π_{B_k} on $B^{(j)}$
- Discrepancy probabilities:
  - $\epsilon_- = P(B^{(j)} = 0 | A_k^i = 1, M_{(ij)} = 1)$
  - $\epsilon_+ = P(B^{(j)} = 1 | A_k^i = 0, M_{(ij)} = 1)$
- $m_k = \pi_{A_k}(1 - \epsilon_-) + (1 - \pi_{A_k})(1 - \epsilon_+)$
- $u_k = \pi_{A_k}\pi_{B_k} + (1 - \pi_{A_k})(1 - \pi_{B_k})$
- Derive the log-likelihood ratio $\mathcal{L}_k(A_k^{(i)}, B_k^{(j)}) = m_k/u_k$
- Derive the posterior probability $\pi_{ij}^{(A \Rightarrow B)}$, $\pi_{ij}^{(B \Rightarrow A)}$, and $\hat{\pi}_{ij}$

A Bayesian application: RL using diagnosis codes



A Bayesian application: RL using diagnosis codes

Good linkage performance when:

- Overlapping information: sufficient diagnosis information shared between the datasets and overlapping units
- Informativeness of variables: enough rare diagnosis

Open-source implementation with diagnosis code data:

package `ludic` available on CRAN

Bayesian Latent Entity Model

The Latent Entity model overcomes the pair-wise comparison among records in different files.
The observed data are perturbed versions of a set of matching variables drawn from a finite population:

- fixed and known size, initially in Steorts et al (2016)
- population size is considered as an unknown parameter in a joint model for RL and population size estimation

Tancredi, A., Steorts, R., Liseo B. (2020) A Unified Framework for De-Duplication and Population Size Estimation (with Discussion) Bayesian Anal. 15(2): 633-682

Latent Entity model

- Naturally works with L lists
- Hierarchical model:
  - N unobserved population units generate the fully observed records, independently collected across the L lists
  - The super-population variables generate the observed ones through a particular measurement error model, the hit-miss model by Copas and Hilton (1990)
  - The matching variables probabilities and the distortion probabilities are unknown quantities, as well as the records partition and the population size

Latent Entity model

- The final likelihood function is available in a closed form for small clusters (size < 3), or via a recursive formula.
- Records belonging to different clusters are independent.
- Records within the same clusters are dependent, since they refer to the same entity.
- Micro-clustering approach on dependence structures. The hit-miss model is a convenient choice, the resulting marginal distribution of the observed variables is the product of within-cluster distributions
- Role of different prior distributions on the population size and the partition space, e.g. Pitman-Yor process facilitates the micro-clustering effect.

Latent Entity model

- Data are informative on how many distinct population entities have been observed at the sample level to estimate the population size N and which sample records gather around each one of the entities to perform the linkage
- Joint model for record linkage and linear regression when response Y and covariates $(X_1, \ldots, X_p)$ come from different files (Steorts et al, 2018)
- Linkage uncertainty is introduced in the inferential process and the information coming from the regression analysis can modify the posterior on the cluster structure.
- This feed-back effect may or not be welcome. When the regression model is not a good choice, the feed-back effect may increase the noise and worsen the linkage process.

Linkage Errors in downstream analysis

Linkage Errors

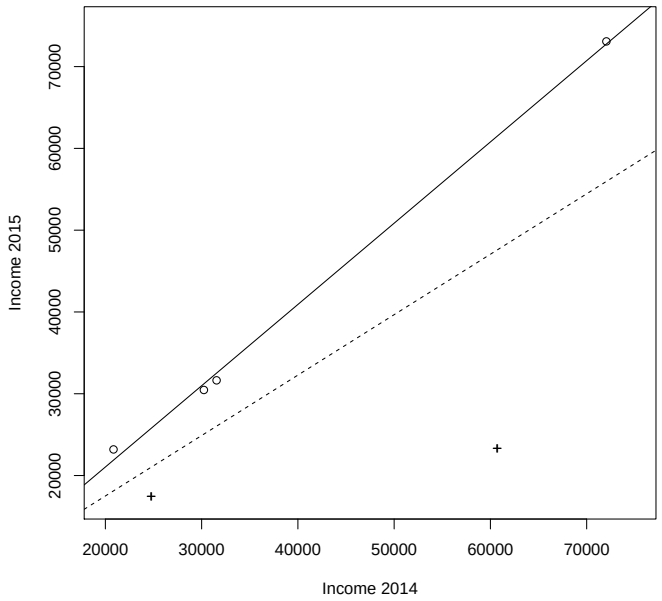
The classification procedure might produce two kinds of errors:

- the false links or false positive
- the missing match or false negative

Standard statistical analyses can be misleading in the presence of linkage errors. E.g. false links made up of data values from different population units

Illustration for linear regression

Figure from Zhang & Tuoto (2021) *JRSS-A*



- X income from admin register
- Y income in the following year
- solid line based on true matches (o)
- dash line based on linked data (+)

False links attenuate the statistical relationship in the actual (correctly linked) population: e.g. bias effect towards zero when estimating the slope of the regression line

Primary vs. Secondary analysis

- Primary: linker *and* analyst: maximal data setting
- Secondary: only linked dataset, neither matching variables nor unlinked records
  - Scheuren & Winkler (1993). Regression analysis of data files that are computer matched. bias correction to the Ordinary Least Squares (OLS) estimates
  - Lahiri & Larsen (2005). Regression analysis with linked data *JASA* bias correction for linear regression
  - Han & Lahiri (2019) Statistical analysis with linked data *ISR*, linear and logistic regression
  - Tancredi & Liseo (2016) Regression Analysis with linked data: Problems and possible solutions *Statistica*, feedback effect between linkage and linear regression in a Bayesian approach

Secondary Analysis

Chambers (2009) Regression analysis of probability-linked data. *Statisphere*

- Propose an Exchangeable model for Linkage Error - ELE
- Revise previous proposals, e.g. Lahiri&Larsen in the new setting
- Propose a Best Linear Unbiased Estimator (BLUE) and its empirical (EBLUE) version

Exchangeable Linkage Error Model

- Exchangeable linkage errors model $\rightarrow$ The probability of correct linkage is the same for all records within some blocks
  - Within a block, correct links are more likely than incorrect ones and all correct (incorrect) links are equally likely
- The linkage is complete and 1-to-1: the two files have the same size, refer to the same population, and have no duplicates
  - Ignorable missing links (false negative), all records have a link, only false links are treated
- As a result, linkage errors introduce a permutation of the true values within the blocks
- Linkage errors within a block are independent of regression errors for that block, non-informative linkage

Exchangeable Linkage Errors model

Y^* is a permutation of the true Y, $Y^* = AY$

$$y_q = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} \text{ and } A_q = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \Rightarrow y_q^* = \begin{pmatrix} y_3 \\ y_2 \\ y_1 \\ y_4 \end{pmatrix}$$

Under ELE, the expected value $E(A|X) = E$ can be written as:

$$E = (\lambda - \psi)I + \psi 11^T$$

where $\lambda = Pr(a_{ii} = 1|X) = Pr(\textit{correct linkage})$ and
$\psi = Pr(a_{ij} = 1|X) = Pr(\textit{incorrect linkage})$.

$$E_q = \begin{pmatrix} \lambda_q & \frac{1-\lambda_q}{M_q-1} & \cdots & \frac{1-\lambda_q}{M_q-1} \\ \frac{1-\lambda_q}{M_q-1} & \lambda_q & \cdots & \frac{1-\lambda_q}{M_q-1} \\ \frac{1-\lambda_q}{M_q-1} & \cdots & \lambda_q & \frac{1-\lambda_q}{M_q-1} \\ \frac{1-\lambda_q}{M_q-1} & \cdots & \frac{1-\lambda_q}{M_q-1} & \lambda_q \end{pmatrix}$$

Estimating Equation Solutions

- G = GEE matrix
- E expected value of the permutation matrix
- Estimating equation for β: $H = G(y^* - EX\beta) = 0$
  - $\hat{\beta} = (GEX)^{-1}Gy^*$
  - sandwich-type approximation to variance
- Standard choices for G
  - $G = X^T \rightarrow$ Least Square
  - $G = X^T E^T \rightarrow$ Lahiri&Larsen
  - $G = X^T E^T \hat{\Sigma}^{-1} \rightarrow$ EBLUE

Relaxing ELE and the complete match space

Zhang, L.-C. and Tuoto, T. (2021). Linkage-data linear regression. Journal of the Royal Statistical Society, Series A, 184, 522-554.

- Linkage data structure:

$$(ab) \in M \not\perp (ab') \in M \qquad M \cup U = A \times B$$

- Incomplete match space $A_M \subset A$ *and* $B_M \subset B$
- Heterogeneous linkage errors $\Pr(a_{ii} = 1 | \ell_i = 1) = \lambda_i$
- Non-informative linkage error

Non-informative linkage error NILE assumption

- Accommodate incomplete match space and heterogeneous linkage errors:

$$\begin{cases} \psi_i = \Pr(\ell_i = 1|X_i) = \Pr(\ell_i = 1) & \text{for } i \in A \\ \psi_j = \Pr(\ell_i = 1|y_j) = \Pr(\ell_j = 1) & \text{for } j \in B \end{cases}$$

$$\lambda_i = \Pr(a_{ii} = 1|\ell_i = 1, X, y) = \begin{cases} \Pr(a_{ii} = 1|\ell_i = 1) & \text{for } i \in A_M \\ 0 & \text{otherwise} \end{cases}$$

- FLR = expected **overall** proportion of false links

$$1 - \lambda = 1 - \sum_{i \in D} \lambda_i / n$$

Pseudo - OLS

- Pseudo-OLS for linear regression coefficient β:

$$\widehat{\beta}_P = \left(\frac{1}{N} X_D^\top X_D\right)^{-1} (\bar{x}\bar{y}^* + \widehat{\lambda}^{-1} S_{xy*})$$
$$= \widehat{\lambda}^{-1}\widehat{\beta}^* - (\widehat{\lambda}^{-1} - 1)\left(\frac{1}{N} X_D^\top X_D\right)^{-1} \bar{x}\bar{y}^*,$$

$S_{xy^*} = \sum_{i \in D}(x_i - \bar{x})(y_i^* - \bar{y}^*)/N$,

$\widehat{\beta}^* = (X_D^\top X_D)^{-1} X_D^\top y^*$ is the *face-value* OLS, and

$\widehat{\lambda}$ is an estimate of the number of correct matches among the links in D.

Variance estimator and Consistency of Pseudo-OLS

$$V(\widehat{\beta}_P) \approx (X_D^\top X_D)^{-1}\sigma^2 + (\frac{1}{N}X_D^\top X_D)^{-1}\Delta(\frac{1}{N}X_D^\top X_D)^{-1}$$

- Under NILE and regularity conditions the Pseudo-OLS is a (asymptotically) consistent estimator
- ELE-based estimators are inconsistent - the asymptotic bias is bounded by $1 - \lambda$
- They are able to remove almost all the bias of the face-value OLS provided the false linkage rate is low

Categorical data

Zhang, L.-C. and Tuoto, T. (2024). Categorical linkage-data analysis. *Submitted*

- Logistic regression: $p = \Pr(Y = 1|x) = \{1 + \exp(-x^\top \beta)\}^{-1}$
  - the Linkage Adjusted Estimating Equation

$$H(\beta; \lambda) = t_\lambda - \sum_{i \in D} p_i x_i + n \bar{p}_D \bar{x}_D = 0$$

    with $t_\lambda = \lambda^{-1} \sum_{i \in D} y_i^* x_i - \lambda^{-1} n \bar{y}_D^* \bar{x}_D$, is asymp. unbiased
  - Newton-Raphson method and sandwich-type variance
- Contingency table: categorical (x_i, y_i^*), linkage-data joint η_{gh}, true joint θ_{gh}, marginal (p_g, q_h)

$$\Pr(x_i = g, y_i^* = h \mid \ell_i = 1) = \lambda \theta_{gh} + (1 - \lambda) p_g q_h = \eta_{gh}$$

Estimation of joint distribution in contingency table

- Relative Bias:

$$\frac{\eta_{gh}}{\theta_{gh}} - 1 = (1-\lambda)\left(\frac{p_g q_h}{\theta_{gh}} - 1\right)$$

$$\text{RB}_{gh} = \text{FLR} \cdot \beta_{gh}$$

with β_{gh} cell-specific discrepancy from independence
RB is product of FLR and true dependence

- Odds Ratios: $\alpha^* = \frac{\eta_{gh}\eta_{ij}}{\eta_{ih}\eta_{gj}}$ and $\alpha = \frac{\theta_{gh}\theta_{ij}}{\theta_{ih}\theta_{gj}}$

$$\frac{\alpha^*}{\alpha} = \frac{1 + \text{RB}_{gh} + \text{RB}_{ij} + \text{RB}_{gh}\text{RB}_{ij}}{1 + \text{RB}_{ih} + \text{RB}_{gj} + \text{RB}_{ih}\text{RB}_{gj}} = \frac{1 + \text{FLR} \cdot (\beta_{gh} + \beta_{ij}) + \text{FLR}^2 \cdot \beta_{gh}\beta_{ij}}{1 + \text{FLR} \cdot (\beta_{ih} + \beta_{gj}) + \text{FLR}^2 \cdot \beta_{ih}\beta_{gj}}$$

Bias less obvious compared to linear association measure
$Cov(x_i, y_i^*) = \lambda_i Cov(x_i, y_i)$

Test for independence

- A test for $H_0 : \theta_{gh} = p_g q_h$ is given by

$$\mathcal{X}^* = \sum_{g,h} \frac{(n^*_{gh} - n_g n^*_h/n)^2}{n_g n^*_h/n} \sim \chi^2_{(G-1)(H-1)}$$

- Standard test for independence is invariant: given NILE, noise induced by the linkage errors has the same distribution as the matched (X, Y) under H_0
- Regardless H_0, we have

$$E(\mathcal{X}^*) \approx \lambda^2 E(\mathcal{X})$$

- $E(\mathcal{X}^*)$ less discriminant under departures from H_0 and $\mathcal{X}^*$ is less powerful than $\mathcal{X}$ due to linkage errors
- Helpful to estimate of $\mathcal{X}$ even when the distribution is unknown

Is NILE acceptable?

Diagnostic test for NILE in linear regression

- $\widehat{\beta}_G$ OLS estimate based on Gold linkage, not false links, only missing matches
- Both $\widehat{\beta}_G$ and $\widehat{\beta}_P$ are consistent, we have $\widehat{\beta}_G - \widehat{\beta}_P \xrightarrow{P} 0$, as $N_G^* \to \infty$

$$t = (\widehat{\beta}_G - \widehat{\beta}_P)^\top V (\widehat{\beta}_G - \widehat{\beta}_P)^{-1} (\widehat{\beta}_G - \widehat{\beta}_P) \sim \chi^2_p$$

for H_0 : NILE vs. H_1 : not NILE.

Concluding Remarks and Future research

- Data linkage is one step of the statistical analysis, and we have tools and methods to deal with it
- Linkage errors are only one piece of errors, and we have tools and methods to deal with them
- What next?
  - Informativeness of linkage errors, and the ways it can arise
  - Other useful models for linkage errors?
  - Extension to linkage of multiple files
  - More complex models: mixed models for small area estimation, survivor analysis, research is ongoing
  - Frequentist vs Bayesian approach to secondary analysis
  - ...

Thank you for your attention

Some references I


Chambers R. (2009). Regression analysis of probability-linked data. *Official Statistics Research Series, Vol. 4.*


Copas, J. B., and Hilton, F. J. (1990). Record linkage: statistical models for matching computer records. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 153(3), 287–312.


Enamorado T, Fifield B, Imai K. (2019) Using a Probabilistic Model to Assist Merging of Large-Scale Administrative Records. *American Political Science Review*. 113(2):353–371.


Fellegi, IP and Sunter, AB. (1969) A theory for record linkage. *Journal of the American Statistical Association*. 64,(328):1183–1210


Fortini, M., Liseo, B., Nuccitelli, A., Scanu, M. (2001). On Bayesian record linkage. *Research in official statistics*, 4(1), 185-198.


Han, Y. and Lahiri, P. (2018) Statistical analysis with linked data. *International Statistical Review*, 87, S139–S157.

Some references II


Hejblum, B., Weber, G., Liao, K. et al. (2019) Probabilistic record linkage of de-identified research datasets with discrepancies using diagnosis codes. *Nature Scientific Data* 6(1), 1–11.


Jaro, M.A. (1989) Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association*, 84, 414-420.


Lahiri, P. and Larsen, M.D. (2005). Regression analysis with linked data. *Journal of the American Statistical Association*, **100**, 222–230.


Larsen, M. D. (2005). Hierarchical Bayesian record linkage theory. *Iowa State University, Statistics*. 401.


Sadinle, M. (2017). Bayesian estimation of bipartite matchings for record linkage. *Journal of the American Statistical Association*, 112(518), 600-612.


Scheuren, F., and Winkler, W. E. (1993). Regression analysis of data files that are computer matched. *Survey Methodology*, 19(1), 39-58.

Some references III


Steorts, R. C., Hall, R., and Fienberg, S. E. (2016). A Bayesian approach to graphical record linkage and deduplication. *Journal of the American Statistical Association*, 111(516), 1660–1672.


Steorts, R., Tancredi, A., and Liseo, B. (2018). "Generalized Bayesian record Linkage and Regression with Exact Error Propagation." In *Proceedings of the International Conference on Privacy in Statistical Databases*, PSD2018 , 297-–313


Tancredi A. and Liseo B (2016) Regression Analysis with linked data: Problems and possible solutions *Statistica*,5(1), 19–35.


Tancredi, A., Steorts, R., Liseo, B. (2020). A unified framework for de-duplication and population size estimation (with discussion). *Bayesian Analysis*, 15(2), 633–682.


Zhang, L.-C. and Tuoto, T. (2021). Linkage-data linear regression. *Journal of the Royal Statistical Society, Series A*, **184**, 522–554.


Zhang, L.-C. and Tuoto, T. (2024). Categorical linkage-data analysis. *Submitted*