# Deduplication, record linkage and inference with linked data


Brunero Liseo

MEMOTEF, Sapienza Università di Roma

*brunero.liseo@uniroma1.it*

joint work with with R.C. Steorts (Duke University) and A. Tancredi (Sapienza)


ITACOSM 2019,

Firenze, June 2019

# Summary

# Introduction

Linking two or more data sets can be important for different and complementary reasons:

(i) per sé, i.e. to obtain a larger reference data set or frame
  - Useful for administrative tasks
  - To overcame confidentiality constraints
  - More accurate summary statistics

(ii) to calibrate statistical models via the additional information which could not be extracted from either one of the two single data sets.
  - Linear and logistic regression
  - Survival analysis
  - Capture recapture
  - ...

# Introduction

Linking two or more data sets can be important for different and complementary reasons:

(i) per sé, i.e. to obtain a larger reference data set or frame
  - Useful for administrative tasks
  - To overcame confidentiality constraints
  - More accurate summary statistics

(ii) to calibrate statistical models via the additional information which could not be extracted from either one of the two single data sets.
  - Linear and logistic regression
  - Survival analysis
  - Capture recapture
  - ...

Here we focus on the methodological aspects of (ii) in the linear regression case and we will argue that the additional information may be helpful also for the record linkage (RL) process

# RL history:major steps

- Fellegi and Sunter (1969) A theory for record linkage. JASA, 64 11831210. (One to one comparison and testing strategy)

# RL history:major steps

- Fellegi and Sunter (1969) A theory for record linkage. JASA, 64 11831210. (One to one comparison and testing strategy)
- Jaro (1989) Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida, JASA, 84, 414–420. (formalization as a mixture model, with EM strategy)

# RL history:major steps

- Fellegi and Sunter (1969) A theory for record linkage. JASA, 64 11831210. (One to one comparison and testing strategy)
- Jaro (1989) Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida, JASA, 84, 414–420. (formalization as a mixture model, with EM strategy)
- Belin and Rubin (1995) A method for calibrating false - match rates in record linkage, JASA, 90, 694–707. (FDR influence)
- Larsen and Rubin (2001). Iterative automated record linkage using mixture models. JASA, 96, pag. 32–41 (Mixture models with interaction among ket variables through a log-linear model)

# Bayesian methods

Necessary to account for uncertainty in the matching step.

- Fortini et al. (2001) On Bayesian record linkage, *Research in Official Statistics*, 4, 185–198.
- Tancredi & Liseo (2011) A hierarchical Bayesian approach to record linkage and population size estimation. *Annals of Applied Statistics*, 5, 1553–1585.
- Steorts, Hall & Fienberg (2016), A Bayesian approach to graphical record linkage and de-duplication. (JASA),Volume 111, 2016 - Issue 516, 1660–1672

# Inference with linked data

- F. Scheuren, W. E. Winkler (1993). Regression analysis of data files that are computer matched. *Survey Methodology*, 19, pp. 39–58.
- P. Lahiri, M. D. Larsen (2005). Regression analysis with linked data. *JASA*, 100, pp. 222–230. 3
- G. Kim, R. Chambers (2012). Regression analysis under incomplete linkage. *CSDA*, 56, no. 9, pp. 2756–2770.
- Tancredi & Liseo (2016) Regression Analysis with linked data: Problems and possible solutions *Statistica*, 75,1, 19–35.
- . . . many more in the last years ..

# Linked data: the bias effect

- Assume we observe $Y, V_1, \ldots, V_h$ in a file and $X, V_1, \ldots, V_h$ in the other one. It is likely that many statistical uits are present in both files, maybe more than once . . .
- Consider a regression of $Y$ on $X$ based on pairs which we declare as matches after a RL analysis based on $(V_1, \ldots, V_h)$ (Scheuren & Winkler, *Srv. Mth*, '93 - Larsen & Lahiri, *JASA, '05*)
- The presence of false matches reduces the observed level of association between $Y$ and $X$.
  - bias effect towards zero when estimating the slope of the regression line.

# Linked data: the bias effect

- Assume we observe $Y, V_1, \ldots, V_h$ in a file and $X, V_1, \ldots, V_h$ in the other one. It is likely that many statistical uits are present in both files, maybe more than once ...
- Consider a regression of $Y$ on $X$ based on pairs which we declare as matches after a RL analysis based on $(V_1, \ldots, V_h)$ (Scheuren & Winkler, *Srv. Mth*, '93 - Larsen & Lahiri, *JASA, '05*)
- The presence of false matches reduces the observed level of association between $Y$ and $X$.
  - ◇ bias effect towards zero when estimating the slope of the regression line.
- Similar biases may appear in any statistical procedure: for example, false matches reduces the final estimate of $N$ when RL methods are used in capture-recapture models for estimating population size.

# Linked data

Consider the setting

**Data set A**

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $Y_1$ | $Y_2$ | $\dots$ | $Y_h$ | $X_1^{(A)}$ | $X_2^{(A)}$ | $\dots$ | $X_k^{(A)}$ | | | | |
| $y_{11}$ | $y_{12}$ | $\dots$ | $y_{1h}$ | $X_{11}^{(A)}$ | $X_{12}^{(A)}$ | $\dots$ | $X_{1k}^{(A)}$ | | | | |
| $\dots$ | $\dots$ | $\dots$ | $\dots$ | $\dots$ | $\dots$ | $\dots$ | $\dots$ | $\dots$ | $\dots$ | $\dots$ | $\dots$ |
| $y_{n1}$ | $y_{n2}$ | $\dots$ | $y_{nh}$ | $X_{n1}^{(A)}$ | $X_{n2}^{(A)}$ | $\dots$ | $X_{nk}^{(A)}$ | | | | |

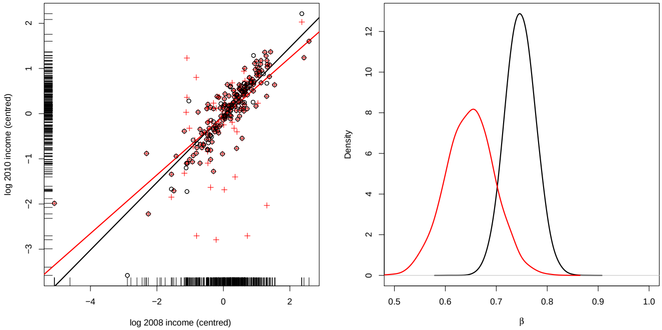
**Data set B**

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | $X_1^{(B)}$ | $X_2^{(B)}$ | $\dots$ | $X_k^{(B)}$ | $Z_1$ | $Z_2$ | $\dots$ | $Z_p$ |
| | | | | $X_{11}^{(B)}$ | $X_{12}^{(B)}$ | $\dots$ | $X_{1k}^{(B)}$ | $z_{11}$ | $z_{12}$ | $\dots$ | $Z_{1p}$ |
| $\dots$ | $\dots$ | $\dots$ | $\dots$ | $\dots$ | $\dots$ | $\dots$ | $\dots$ | $\dots$ | $\dots$ | $\dots$ | $\dots$ |
| | | | | $X_{m1}^{(B)}$ | $X_{m2}^{(B)}$ | $\dots$ | $X_{mk}^{(B)}$ | $z_{m1}$ | $z_{m2}$ | $\dots$ | $Z_{mp}$ |

## Example: Italian survey of household and income wealth (SHIW)

- Data set $A$: 2008 income for a single block (434 units)
- Data set $B$: 2010 income for the same block (355 units)
- 203 panel individuals

A slight modification of the matching configuration (deleting 10% of true matches and adding 5% of false matches) may produce strongly different regression analyses




Posterior distribution of the slope (black=true, red=noised)

# Standard RL methods

- Records (or transformations thereof) are compared among each other

# Standard RL methods

- Records (or transformations thereof) are compared among each other
- Some metric is used to measure “distance” between pairs

# Standard RL methods

- Records (or transformations thereof) are compared among each other
- Some metric is used to measure "distance" between pairs
- A decision (either based on a test or a posterior probability) is taken.

# Standard RL methods

- Records (or transformations thereof) are compared among each other
- Some metric is used to measure “distance” between pairs
- A decision (either based on a test or a posterior probability) is taken.
- Output: few matches and a huge number of non matches.

# Standard RL methods

- Records (or transformations thereof) are compared among each other
- Some metric is used to measure “distance” between pairs
- A decision (either based on a test or a posterior probability) is taken.
- Output: few matches and a huge number of non matches.
- Curse of dimensionality; difficult to generalize to $k$ files

# RL, duplications, $k$ lists and the hit-and-miss model

New approach for record linkage based on Steorts et al. (2016): $k$ lists and $N$ latent individuals

- $k$ files sharing a set $V_1, \ldots, V_p$ of categorical key variables
- $V_l \sim \{v_{l1} \ldots, v_{l M_l}; \theta_{l1} \ldots, \theta_{l M_l}\}$ $l = 1, \ldots p$
- $v_{ij} = (v_{ij1}, \ldots, v_{ijp})$ denotes the record $j$ in file $i$ $(j = 1, \ldots, r_i)$
- $\tilde{v}_{j'} = (\tilde{v}_{j'1} \ldots \tilde{v}_{j'p})$ is the *true record* for the latent individual $j'$ , $j' = 1, \ldots N$
- $\lambda_{ij} \in \{1 \ldots, N\}$ denotes the latent individual generating $v_{ij}$

# RL, duplications, $k$ lists and the hit-and-miss model

New approach for record linkage based on Steorts et al. (2016): $k$ lists and $N$ latent individuals

- $k$ files sharing a set $V_1, \dots, V_p$ of categorical key variables
- $V_l \sim \{v_{l1} \dots, v_{l M_l}; \theta_{l1} \dots, \theta_{l M_l}\}$ $l = 1, \dots p$
- $v_{ij} = (v_{ij1}, \dots, v_{ijp})$ denotes the record $j$ in file $i$ ($j = 1, \dots, r_i$)
- $\tilde{v}_{j'} = (\tilde{v}_{j'1} \dots \tilde{v}_{j'p})$ is the *true record* for the latent individual $j'$ , $j' = 1, \dots N$
- $\lambda_{ij} \in \{1 \dots, N\}$ denotes the latent individual generating $v_{ij}$

$$\lambda_{ij_1} = \lambda_{ij_2} \quad \Rightarrow \quad \text{a duplication in the same list}$$
$$\lambda_{i_1 j_1} = \lambda_{i_2 j_2} \quad \Rightarrow \quad \text{a match between two lists}$$

- The hit-and-miss model (Copas and Hilton (1990) JRSSA)

$$p(V_{ijl} = v_{ijl} | \lambda_{ij}, \tilde{v}, \alpha_l) = (1 - \alpha_l)\delta_{\tilde{v}_{\lambda_{ij}l}, v_{ljl}} + \alpha_l \theta_{l\, v_{ijl}}$$

  is the *conditional* generating processes of the key variables:
  - the true value is correctly generated with probability $1 - \alpha_l$
  - a value is generated from $V_l$ with probability $\alpha_l$.

- The hit-and-miss model (Copas and Hilton (1990) JRSSA)

$$p(V_{ijl} = v_{ijl} | \lambda_{ij}, \tilde{v}, \alpha_l) = (1 - \alpha_l)\delta_{\tilde{v}_{\lambda_{ij} l}, v_{ljl}} + \alpha_l \theta_{l\, v_{ijl}}$$

  is the *conditional* generating processes of the key variables:
  - the true value is correctly generated with probability $1 - \alpha_l$
  - a value is generated from $V_l$ with probability $\alpha_l$.
- Conditional independence among all the observed records given their respective unobserved true records

$$p(v|\lambda, \tilde{v}, \alpha) = \prod_{ijl} p(v_{ijl}|\tilde{v}, \lambda, \alpha) = \prod_{ijl} [(1 - \alpha_l)\delta_{\tilde{v}_{\lambda_{ij} l}, v_{ijl}} + \alpha_l \theta_{l\, v_{ijl}}]$$

- $\tilde{V}_{j'l} \sim V_l$ independently for $j' = 1, \ldots N$ and $l = 1 \ldots p$

# Prior distributions and other assumptions

- Steorts et al. (2016) assume a uniform prior on the set $\Lambda$, $\pi(\Lambda) = \prod_{ij} \pi(\lambda_{ij}) = \prod_{ij} \frac{1}{N}$
  - in the $k$ -lists framework: $k$ independent simple random samples with replacement from a population of $N$ labels
- $\alpha_l \overset{i.i.d}{\sim} Beta(p, q)$ or exchangeable
- Probabilities $\theta_{l1} \ldots \theta_{l M_l}$ are hard to be estimated. Simplifying assumption: they are equal to the corresponding population *or sample* frequencies (Empirical Bayes step).

# Prior on the partition space

- A uniform prior on $\Lambda$ space can also be interpreted in terms of partitions. Let $k$ be the number of blocks in a given partition.
- Then, for fixed population size $N$, $\pi(\Lambda) \propto 1$ gives the same prior to all partitions with the same $k$, namely (Pitman, 2006)

$$\pi(k|N) = \frac{N!S(n,k)}{(N-k)!N^n}$$

with $S(n,k)$ the 2nd type Stirling numbers[1].

- Easy to see that

$$\mathbb{E}(k|N) = N(1-(1-1/N)^n)$$

and

$$\lim_{n\to\infty} \mathbb{E}(k|N) = N; \quad \lim_{n\to\infty} \mathbb{V}(k|N) = 0$$

Also

$$\lim_{N\to\infty} \mathbb{E}(k|N) = n; \quad \lim_{N\to\infty} \mathbb{V}(k|N) = 0$$

[1] $S(n,k)$ is the number of ways to partition a set of $n$ objects into $k$ non-empty subsets

# An alternative Bayesian nonparametric prior

- However, the latent model of Steorts et al. (2016) suggests a clustering process of the records around $N$ latent units ...
- In particular, record linkage models typically create a large number of small clusters ( the micro-clustering issue) (Miller et al. 2015, Johndrow et al. 2018)

# An alternative Bayesian nonparametric prior

- However, the latent model of Steorts et al. (2016) suggests a clustering process of the records around $N$ latent units ...
- In particular, record linkage models typically create a large number of small clusters ( the micro-clustering issue) (Miller et al. 2015, Johndrow et al. 2018)
- Bayesian analysis for these problems is generally based on the use of a prior process on the random partitions.

Then, a more flexible process is deemed necessary in order to induce a micro-clustering effect ...

# Pitman-Yor Process

Assume the first $j$ records of the $i$-th file and all the records of the first $i-1$ lists are classified into $k_{i,j}$ clusters, identified by labels $j'_1, \ldots, j'_{k_{i,j}}$ with sizes $n_1, n_2, \ldots, n_{k_{i,j}}$ respectively.

Let $N_{i,j} = \sum_{l=1}^{i-1} N_l + j$.

Suppose the next record label $\lambda_{i,j+1}$ identifies a new cluster with probability

$$P\left(\lambda_{i,j+1} = \text{“new”} \,|\lambda_{1,1}, \ldots, \lambda_{i,j}\right) = \frac{k_{i,j}\sigma + \vartheta}{N_{i,j} + \vartheta}, \left[= \frac{\vartheta}{N_{i,j} + \vartheta}\right]$$

with $\sigma \in [0,1)$ with $\vartheta > -\sigma$ or $\sigma < 0$ with $\theta = m|\sigma|$ for some integer $m$.

Also $\lambda_{i,j+1}$ takes an already existing label $j'_g$ with a cluster of size $n_g$ with probability

$$P\left(\lambda_{i,j+1} = j'_g | \lambda_{1,1}, \ldots, \lambda_{i_1,j_1} =\right) \frac{n_g - \sigma}{N_{i,j} + \vartheta} \quad g = 1, \ldots, k_{i,j}.$$

# Prior Modelling

It can be proved that the mean number of occupied clusters after $n$ arrivals is

$$E(K_n) = \sum_{i=1}^{n} \frac{(\theta+\sigma)_{(i-1)\uparrow}}{(\theta+1)_{(i-1)\uparrow}} = \begin{cases} \sum_{i=1}^{n} \frac{\theta}{\theta+i-1} & \sigma = 0 \\ \frac{(\theta+\sigma)_{n\uparrow}}{\sigma(\theta+1)_{(n-1)\uparrow}} - \frac{\theta}{\sigma} & \sigma \neq 0 \end{cases}$$

with $(x)_{n\uparrow} = \frac{\Gamma(x+n)}{\Gamma(x)} = x(x+1)\cdots(x+n-1)$.

- This might help in the elicitation of the hyper-parameters.
- The value of $\sigma$ characterizes the asymptotic behavior of $K_n$. Positive values of $\sigma$ induces an infinite number of clusters. If $-1 < \sigma < 0$, the number of clusters remains bounded.

# Hit-and-miss model and clustering

- For a given $\lambda$ observed records clusterize:

$$C_{j'} = \{(i,j); \lambda_{ij} = j'\} \quad v_{C_{j'}} = (v_{ij} : \lambda_{ij} = j') \quad v_{C_{j'}l} = (v_{ijl} : \lambda_{ij} = j')$$

# Hit-and-miss model and clustering

- For a given $\lambda$ observed records clusterize:

$$C_{j'} = \{(i,j); \lambda_{ij} = j'\} \quad v_{C_{j'}} = (v_{ij} : \lambda_{ij} = j') \quad v_{C_{j'}l} = (v_{ijl} : \lambda_{ij} = j')$$

The distribution of the data $\mathbf{v}$ is the product of the record cluster distributions

$$p(v|\tilde{v}, \lambda, \alpha) = \prod_{j'=1}^{N} p(v_{C_{j'}}|\alpha, \tilde{v}_{j'}) = \prod_{j'=1}^{N} \prod_{l=1}^{p} p(v_{C_{j'}l}|\alpha_l, \tilde{v}_{j'l})$$

# Hit-and-miss model and clustering

- For a given $\lambda$ observed records clusterize:

$$C_{j'} = \{(i,j); \lambda_{ij} = j'\} \quad v_{C_{j'}} = (v_{ij} : \lambda_{ij} = j') \quad v_{C_{j'}l} = (v_{ijl} : \lambda_{ij} = j')$$

The distribution of the data $\mathbf{v}$ is the product of the record cluster distributions

$$p(v|\tilde{v}, \lambda, \alpha) = \prod_{j'=1}^{N} p(v_{C_{j'}}|\alpha, \tilde{v}_{j'}) = \prod_{j'=1}^{N} \prod_{l=1}^{p} p(v_{C_{j'}l}|\alpha_l, \tilde{v}_{j'l})$$

One can also integrate out the $\tilde{v}_{j'}$'s within each cluster. The new sampling model now only depends on $\lambda$ and $\alpha$,

$$p(v|\lambda, \alpha) = \prod_{j'=1}^{N} p(v_{C_{j'}}|\alpha) = \prod_{j'=1}^{N} \prod_{l=1}^{p} p(v_{C_{j'}l}|\alpha_l)$$

Some expressions:

- Cluster with a single record $C_{j'} = \{(i\,j)\}$

$$P(v_{C_{j'}}|\alpha) = \prod_{l=1}^{p} p(v_{C_{j'}l} = v_{ijl}|\alpha) = \prod_{l=1}^{p} \theta_{l\,v_{ijl}}$$

- Cluster with two records $C_{j'} = \{(i_1 j_1), (i_2 j_2)\}$

$$P(v_{C_{j'}}|\alpha) = \prod_{l=1}^{p} \left[\delta_{v_{i_1 j_1 l}, v_{i_2 j_2 l}} \theta_{l\,v_{i_1 j_1 l}} (1-\alpha_l)^2 + (2\alpha_l - \alpha_l^2)\theta_{l\,v_{i_1 j_1 l}} \theta_{l\,v_{i_2 j_2 l}}\right]$$

- A recursive formula for a cluster $C_{j'} = \{(i_1 j_1), \ldots, (i_n j_n)\}$

$$p(v_{C_{j'}l}|\alpha_l) = p(v_{C_{j'}\setminus(i_n j_n) l})\alpha_l \theta_{v_{i_n j_n l}} + (1-\alpha_l)\theta_{v_{i_n j_n l}} \prod_{h=1}^{n-1} \left[(1-\alpha_l)\delta_{v_{i_h j_h l}, v_{i_n j_n l}} + \alpha_l \theta_{l v_{i_h j_h l}}\right]$$

# Computation

- Steorts (2015) proposes a Gibbs sampler driven by an additional set of binary latent variables $z_{ijl}$'s: a latent variable is added for each component of the vector of observations in each record of each files.
  $z_{ijl}$ indicates whether the $l$-th variable, on the $j$-th record of $i$-th file, is distorted.
  $\rightarrow$ Gibbs sampling very straightforward to implement;
  $\rightarrow$ The huge number of correlated latent variables jeopardizes the mixing of the resulting Markov chain.
- A Gibbs sampler can also be easily obtained for simulating $p(\lambda, \tilde{v}, \alpha | v)$ when the true values are not integrated out
- We propose to simulate $p(\lambda, \alpha | v)$ via a Metropolis within Gibbs algorithms with an exact step for $\lambda$ and a Metropolis step for $\alpha$

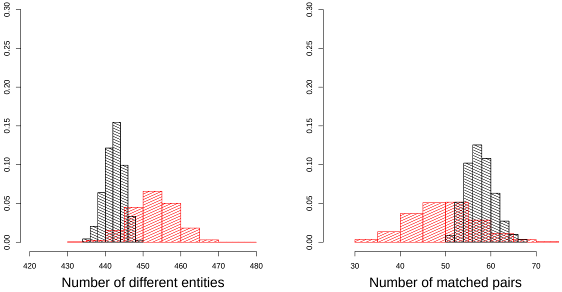
# A break: `RLdata500`

It contains artificial personal data for the evaluation of RL procedures.

- Synthetic data set with 500 records: first name, family name and date of birth
- 50 records have been duplicated and distorted
- Single list with $n = 450$ different entities.

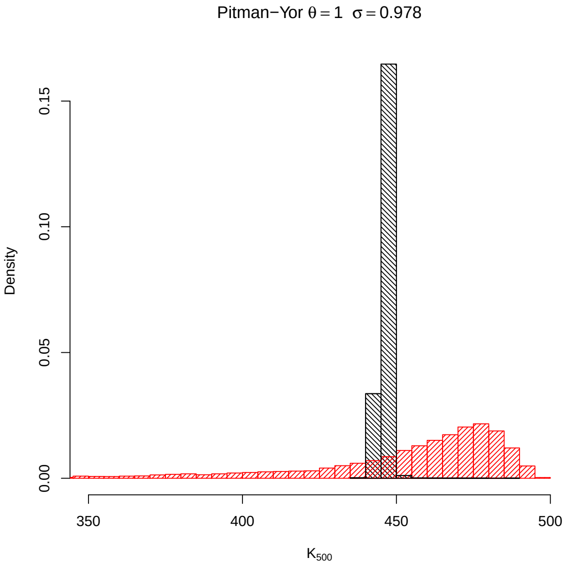
| | fname_c1 | fname_c2 | lname_c1 | lname_c2 | by | bm | bd |
|---|---|---|---|---|---|---|---|
| 1 | CARSTEN | | MEIER | | 1949 | 7 | 22 |
| 2 | GERD | | BAUER | | 1968 | 7 | 27 |
| 3 | ROBERT | | HARTMANN | | 1930 | 4 | 30 |
| 4 | STEFAN | | WOLFF | | 1957 | 9 | 2 |
| 5 | RALF | | KRUEGER | | 1966 | 1 | 13 |
| ⋮ | | | | | | | |
| 43 | GERD | | BAUERH | | 1968 | 7 | 27 |
| ⋮ | | | | | | | |
| 58 | FRANK | | MUELLDR | | 1978 | 5 | 20 |
| ⋮ | | | | | | | |
| 148 | FRANK | | MUELLER | | 1978 | 5 | 20 |
| ⋮ | | | | | | | |

- In order to apply the model we transform name and surname via the SOUNDEX algorithm. Year of birth has been split into 4 fields.
- We set $N = 2500$ so that the prior mean of the number of pairs in a file with 500 records is $(1/N)\binom{500}{2} = 49.9$
- Independent beta priors for $\alpha$ with mean 0.01 (we expect that 1% of the fields have been distorted)




Prior (red) and posterior (black) distribution for the number of matches and the number of different elements (hit-and-miss model).

# A (more diffuse) Pitman & Yor prior and posterior

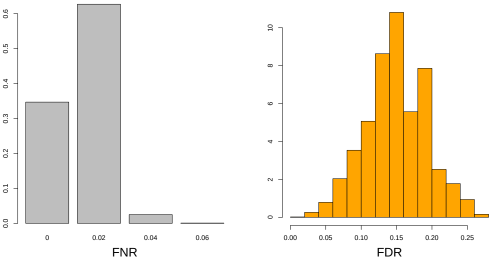


Set

$$\Delta_{j_1,j_2} = \begin{cases} 1 & \lambda_{j_1} = \lambda_{j_2} \\ 0 & \lambda_{j_1} \neq \lambda_{j_2} \end{cases}$$

Linkage performance can be evaluated through

$$FNR = \frac{\sum_{j_1 j_2} (1 - \hat{\Delta}_{j_1 j_2}) \Delta_{j_1 j_2}}{\sum_{j_1 j_2} \Delta_{j_1 j_2}} \qquad FDR = \frac{\sum_{j_1 j_2} \hat{\Delta}_{j_1 j_2} (1 - \Delta_{j_1 j_2})}{\sum_{j_1 j_2} \hat{\Delta} j_1 j_2}$$




Hit-and-miss model: FNR and FDR posterior distribution.
The model introduces some false matches, $E(FDR|v) \approx 0.148$, but almost all the true matches are spotted, $E(FNR|v) \approx 0.014$.

# RL, duplications, $k$ lists and regression

Consider a linear regression model $Y = \tilde{X}\beta + \varepsilon$. Assume $Y$ and $X$ are observed across the lists: two different scenarios

**Partial regression**

$$
\begin{array}{ccccc}
y_{11} & v_{111} & \cdots & v_{11p} & \\
 & & \vdots & & \\
y_{1r_1} & v_{1r_1 1} & \cdots & v_{1r_1 p} & \\
 & & & & \\
\text{---} & & & & \\
 & v_{211} & \cdots & v_{21p} & x_{21} \\
 & & \vdots & & \\
 & v_{2r_2 1} & \cdots & v_{2r_2 p} & x_{2r_2} \\
 & & \vdots & & \\
\text{---} & & & & \\
 & v_{k11} & \cdots & v_{k1p} & x_{21} \\
 & & \vdots & & \\
 & v_{kr_k 1} & \cdots & v_{kr_k p} & x_{kr_k}
\end{array}
$$

**Complete scenario**

$$
\begin{array}{ccccc}
y_{11} & v_{111} & \cdots & v_{11p} & x_{11} \\
 & & \vdots & & \\
y_{1r_1} & v_{1r_1 1} & \cdots & v_{1r_1 p} & x_{1r_1} \\
 & & & & \\
\text{---} & & & & \\
y_{21} & v_{211} & \cdots & v_{21p} & x_{21} \\
 & & \vdots & & \\
y_{2r_2} & v_{2r_2 1} & \cdots & v_{2r_2 p} & x_{2r_2} \\
 & & \vdots & & \\
\text{---} & & & & \\
y_{k1} & v_{k11} & \cdots & v_{k1p} & x_{21} \\
 & & \vdots & & \\
y_{kr_k} & v_{kr_k 1} & \cdots & v_{kr_k p} & x_{kr_k}
\end{array}
$$

- Assume the $X$ variables are noisy measurements of the true covariates $\tilde{X}$. Let $\tilde{X}_{j'}$ the true value of $X$ for the cluster $C_j'$.
- Consider the complete scenario, a cluster $C_{j'} = \{(i,j))\}$ and, *to simplify*, a single covariate $X$ and a model without intercept. Assume that

$$\left[\begin{array}{c} Y_{ij} \\ X_{ij} \end{array}\right] \mid \tilde{X}_{j'} = \tilde{x}_{j'} \sim N_2 \left[ \left(\begin{array}{cc} \beta & 0 \\ 0 & 1 \end{array}\right) \left[\begin{array}{c} \tilde{x}_{j'} \\ \tilde{x}_{j'} \end{array}\right], \left(\begin{array}{cc} \sigma^2_{y|\tilde{x}} & 0 \\ 0 & \sigma^2_{x|\tilde{x}} \end{array}\right) \right]$$

- Assume the $X$ variables are noisy measurements of the true covariates $\tilde{X}$. Let $\tilde{X}_{j'}$ the true value of $X$ for the cluster $C'_j$.
- Consider the complete scenario, a cluster $C_{j'} = \{(i,j))\}$ and, *to simplify*, a single covariate $X$ and a model without intercept. Assume that

$$\left[\begin{array}{c} Y_{ij} \\ X_{ij} \end{array}\right] | \tilde{X}_{j'} = \tilde{x}_{j'} \sim N_2 \left[ \left( \begin{array}{cc} \beta & 0 \\ 0 & 1 \end{array} \right) \left[ \begin{array}{c} \tilde{x}_{j'} \\ \tilde{x}_{j'} \end{array} \right], \left( \begin{array}{cc} \sigma^2_{y|\tilde{x}} & 0 \\ 0 & \sigma^2_{x|\tilde{x}} \end{array} \right) \right]$$

- Also, center the $x$'s and assume $\tilde{X}_{j'} \sim N(0, \sigma^2_{\tilde{x}})$, then

$$\left[\begin{array}{c} Y_{ij} \\ X_{ij} \end{array}\right] \sim N_2 \left[ \left( \begin{array}{c} 0 \\ 0 \end{array} \right), \sigma^2_{\tilde{x}} \left( \begin{array}{cc} \beta^2 & \beta \\ \beta & 1 \end{array} \right) + \left( \begin{array}{cc} \sigma^2_{y|\tilde{x}} & 0 \\ 0 & \sigma^2_{x|\tilde{x}} \end{array} \right) \right]$$

conditionally on $(i,j) \in C_{j'}$. [ $X_{j'}$ is integrated out of the model]

- Now take a cluster $C_{j'} = \{(i_1, j_1), (i_2, j_2)\}$. Set $Z_{i_h j_h} = (Y_{i_h j_h}, X_{i_h j_h})'$ $h = 1, 2$.
- Conditionally on $\tilde{X}_{j'} = x_{j'}$, $Z_{i_1 j_1}$ and $Z_{i_2 j_2}$ are i.i.d.

$$N_2 \left[ \begin{pmatrix} \beta & 0 \\ 0 & 1 \end{pmatrix} \mathbf{1}_2 \tilde{x}_{j'}, \mathbf{\Sigma} \right]$$

with

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma^2_{y|\tilde{x}} & 0 \\ 0 & \sigma^2_{x|\tilde{x}} \end{pmatrix}$$

- Now take a cluster $C_{j'} = \{(i_1, j_1), (i_2, j_2)\}$. Set $Z_{i_h j_h} = (Y_{i_h j_h}, X_{i_h j_h})'$ $h = 1, 2$.
- Conditionally on $\tilde{X}_{j'} = x_{j'}$, $Z_{i_1 j_1}$ and $Z_{i_2 j_2}$ are i.i.d.

$$N_2 \left[ \begin{pmatrix} \beta & 0 \\ 0 & 1 \end{pmatrix} \mathbf{1}_2 \tilde{x}_{j'}, \mathbf{\Sigma} \right]$$

with

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma^2_{y|\tilde{x}} & 0 \\ 0 & \sigma^2_{x|\tilde{x}} \end{pmatrix}$$

- Standard calculations lead to

$$\begin{pmatrix} Z_{i_1 j_1} \\ Z_{i_2 j_2} \end{pmatrix} \sim N_4 \left( 0_4, \boldsymbol{I}_2 \otimes \mathbf{\Sigma} + \sigma^2_{\tilde{x}} \boldsymbol{J}_2 \otimes \boldsymbol{B} \right).$$

with

$$\boldsymbol{B} = \begin{pmatrix} \beta^2 & \beta \\ \beta & 1 \end{pmatrix}$$

- Now take a cluster $C_{j'} = \{(i_1, j_1), (i_2, j_2)\}$. Set $Z_{i_h j_h} = (Y_{i_h j_h}, X_{i_h j_h})' \ h = 1, 2$.
- Conditionally on $\tilde{X}_{j'} = x_{j'}$, $Z_{i_1 j_1}$ and $Z_{i_2 j_2}$ are i.i.d.

$$N_2 \left[ \begin{pmatrix} \beta & 0 \\ 0 & 1 \end{pmatrix} \mathbf{1}_2 \tilde{x}_{j'}, \mathbf{\Sigma} \right]$$

with

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma^2_{y|\tilde{x}} & 0 \\ 0 & \sigma^2_{x|\tilde{x}} \end{pmatrix}$$

- Standard calculations lead to

$$\begin{pmatrix} Z_{i_1 j_1} \\ Z_{i_2 j_2} \end{pmatrix} \sim N_4 \left( 0_4, \boldsymbol{I}_2 \otimes \mathbf{\Sigma} + \sigma^2_{\tilde{x}} \boldsymbol{J}_2 \otimes \boldsymbol{B} \right).$$

with

$$\boldsymbol{B} = \begin{pmatrix} \beta^2 & \beta \\ \beta & 1 \end{pmatrix}$$

This argument can be extended to any cluster size. When $|C_{j'}| = n$, the marginal distribution of $\boldsymbol{Z} = (Z_{i_1 j_1}, \ldots, Z_{i_n j_n}$ is again multivariate normal

$$\boldsymbol{Z} \sim N_{2n}\left(0_{2n}, \boldsymbol{I}_n \otimes \boldsymbol{\Sigma} + \sigma^2_{\tilde{x}} \boldsymbol{J}_n \otimes \boldsymbol{B}\right).$$

- The likelihood function for the partially observed scenario can be obtained by integrating out $X_{ij}$ (if $i = 1$) and/or $Y_{ij}$ (if $i > 1$)
- Set $(y, x)_{C_j'} = ((y_{ij}, x_{ij}) : \lambda_{ij} = j')$ the likelihood for $\lambda, \alpha, \beta, \sigma^2_{y|\tilde{x}}, \sigma^2_{x|\tilde{x}}$ is - in both cases -

$$p(y, x|\lambda, \beta, \alpha, \sigma^2_{x|\tilde{x}}, \sigma^2_{y|\tilde{x}}) = \prod_{j'=1}^{N} p((y, x)_{C_j'}|\beta, \sigma^2_{x|\tilde{x}}, \sigma^2_{y|\tilde{x}})$$

- Assumption: conditional independence between regression covariates and key variables. [not crucial. . . ]
Given $\lambda$, we can merge the regression and the hit-and-miss models into a broader model and then simulate from the joint posterior distribution

$$
\begin{aligned}
p(\lambda,\beta,\alpha,\sigma^2_{y|\tilde{x}},\sigma^2_{x|\tilde{x}}|v,x,y) &\propto p(v|\lambda,\alpha)p(y,x|\lambda,\beta,\sigma^2_{y|\tilde{x}},\sigma^2_{x|\tilde{x}}) \\
&\times p(\lambda,\alpha,\beta,\sigma^2_{y|\tilde{x}},\sigma^2_{x|\tilde{x}})
\end{aligned}
$$

- Computation via a Metropolis within Gibbs algorithms with exact step for the $\lambda$ updating.

# The general case

Set $|C_{j'}| = n$, $\boldsymbol{Y}_{C_{j'}}$ the response $n$-vector, $\boldsymbol{X}_{C_{j'}}$ the $n \times p$ design matrix, and

$$\boldsymbol{Z}_{C_{j'}} = \left( \boldsymbol{Y}_{C_{j'}}, \text{vec}(\boldsymbol{X}_{C_{j'}})' \right)'$$

# The general case

Set $|C_{j'}| = n$, $\boldsymbol{Y}_{C_{j'}}$ the response $n$-vector, $\boldsymbol{X}_{C_{j'}}$ the $n \times p$ design matrix, and

$$\boldsymbol{Z}_{C_{j'}} = \left(\boldsymbol{Y}_{C_{j'}}, \text{vec}(\boldsymbol{X}_{C_{j'}})'\right)'$$

Assume $\tilde{\boldsymbol{X}}_{j'} \sim N_p(0_p, \boldsymbol{\Sigma}_{\tilde{x}})$.
One has

$$\boldsymbol{Z}_{C_{j'}} | \tilde{\boldsymbol{X}}_{j'} \sim N_{n(p+1)}\left(\boldsymbol{\mu}, \boldsymbol{\Psi}\right),$$

with

$$\boldsymbol{\mu} = \left(\boldsymbol{I}_n \otimes \begin{pmatrix} \boldsymbol{\beta}' \\ \boldsymbol{I}_p \end{pmatrix}\right)\left(1_n \otimes \tilde{X}_{j'}\right)$$

and

$$\boldsymbol{\Psi} = \left(\boldsymbol{I}_n \otimes \begin{pmatrix} \sigma^2_{y|\tilde{X}} & 0_p \\ 0 & \boldsymbol{\Sigma}_{X|\tilde{X}} \end{pmatrix}\right)$$

# Finally

The marginal distribution within the cluster is then

$$\boldsymbol{Z}_{C_{j'}} \sim N_{n(p+1)}\left(0_{n(p+1)}, \boldsymbol{\Omega}\right),$$

with

$$\boldsymbol{\Omega} = \left(1_n \otimes \begin{pmatrix} \boldsymbol{\beta}' \\ \boldsymbol{I}_p \end{pmatrix}\right) \boldsymbol{\Sigma}_{\tilde{X}} \left(1_n' \otimes \begin{pmatrix} \boldsymbol{\beta}' \\ \boldsymbol{I}_p \end{pmatrix}'\right).$$

When $\boldsymbol{\Sigma}_{\tilde{X}} = \boldsymbol{I}_p$

$$\boldsymbol{\Omega} = \boldsymbol{J}_n \otimes \begin{pmatrix} \boldsymbol{\beta}'\boldsymbol{\beta} & \boldsymbol{\beta}' \\ \boldsymbol{\beta} & \boldsymbol{I}_p \end{pmatrix}.$$

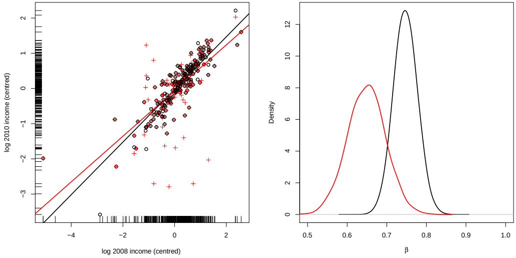
# An example: Italian survey of household and income wealth

- The Italian Survey on Household Income and Wealth (SHIW): sample survey made by Bank of Italy every 2 years
- 2010 survey covers 7,951 households (19,836 individuals).
- We consider the 2010 individual net disposable income ($Y$) and the matching variables: `sex`, `age`, `marital status`, `employment status`, `working sector`, `education`
- We consider the 2008 net disposable income as a covariate ($X$)

## Example: Italian survey of household and income wealth (SHIW)

- Data set $A$: 2008 income for a single block (434 units)
- Data set $B$: 2010 income for the same block (355 units)
- 203 panel individuals

A slight modification of the matching configuration (deleting 10% of true matches and adding 5% of false matches) may produce strongly different regression analyses




Posterior distribution of the slope (black=true, red=noised)

# Feed-back or not ? . . .

Question: Should we use the information in $Y$ and $X$ in the linkage step?
We certainly use the information in the key variables to improve the calibration of the regression model, BUT . . .

Is the reverse always convenient?

# Feed-back or not ? . . .

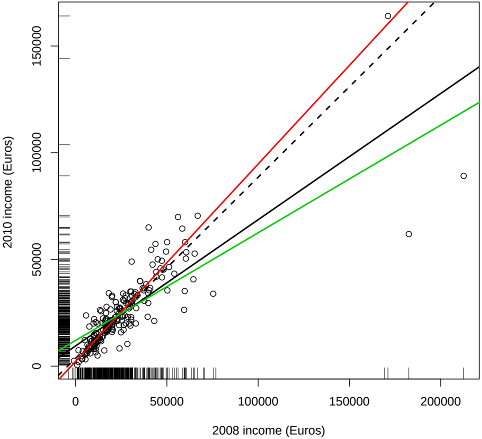
Question: Should we use the information in $Y$ and $X$ in the linkage step?
We certainly use the information in the key variables to improve the calibration of the regression model, BUT . . .

Is the reverse always convenient?
There is no a clear-cut answer to this question . . .
It depends on

- the reason why we link data sets
- data quality of $(Y, X)$
- . . .

# SHIV data: Friuli

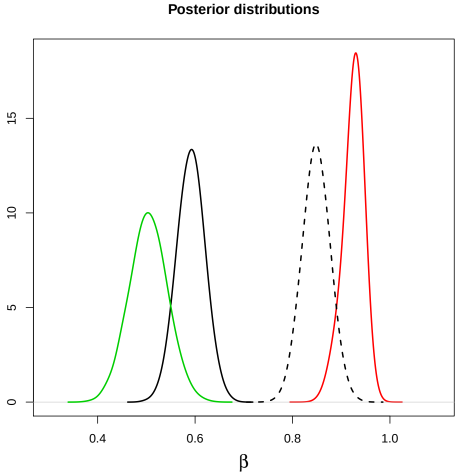
$$n_1 = 434, n_2 = 355$$




- Black line: true regression line given by the 203 true matches
- Black dashed line: true regression line without 2 very influential obs.
- Red line: Bayesian estimate via the regression AND linking model
- Green line: Bayesian estimate via the linking model and regression with a plug-in estimate of matched records.

# SHIV data: Friuli

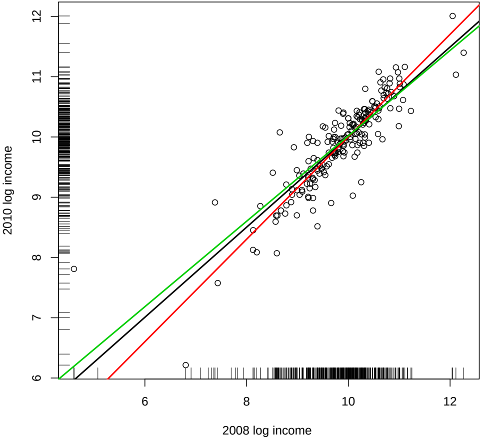
$n_1 = 434, n_2 = 355$




- Black line: posterior with the 203 true matches.
- Black dashed line: posterior without 2 very influential obs.
- Red line: Posterior density of $\beta$ via regression AND linking model
- Green line: Posterior density of $\beta$ via the linking model and regression with a plug-in estimate of matched records.

# SHIV data: Friuli

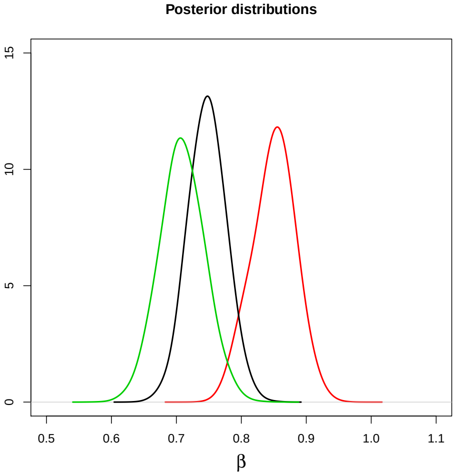
$n_1 = 434, n_2 = 355$




- Log transformation of the data
- Black line: true regression line (203 true matches)
- Red line: Bayesian estimate via regression AND linking model.
- Green line: Bayesian estimate via the linking model and regression with a plug-in estimate of matched records.

# SHIV data: Friuli

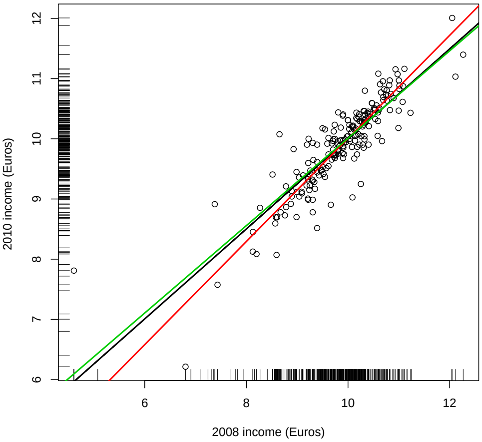
$n_1 = 434, n_2 = 355$



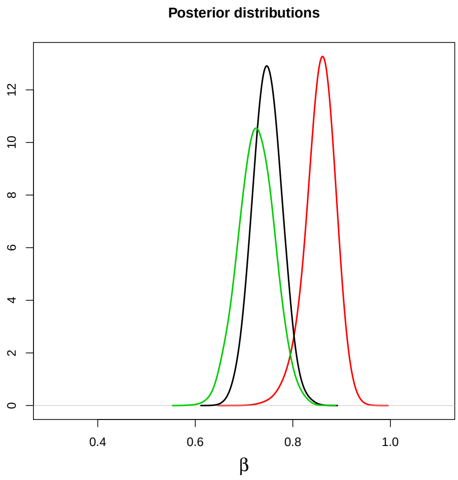

- Log transformation of the data
- Black line: “true” posterior density of $\beta$ (203 true matches)
- Red line: Posterior density of $\beta$ via regression AND linking model
- Green line: Posterior density of $\beta$ via the linking model and regression with a plug-in estimate of matched records.

# Same as before, 9 key variables

$n_1 = 434, n_2 = 355$




- Log transformation of the data
- Black line: true regression line (203 true matches)
- Red line: Bayesian estimate via regression AND linking model.
- Green line: Bayesian estimate via the linking model and regression with a plug-in estimate of matched records.




- Log transformation of the data
- Black line: "true" posterior density of $\beta$ (203 true matches)
- Red line: Posterior density of $\beta$ via regression AND linking model
- Green line: Posterior density of $\beta$ via the linking model and regression with a plug-in estimate of matched records.

# Discussion

- We obtained improvements both for the $\beta$ estimation and for the matching process in a single *partially* simulated data set...,
- Similar results can also be obtained in large scale simulations and real data sets
- Current research: prior calibration
- Problems may arise when the regression model does not hold
- More robust estimates assuming heavy tails for the regression error

# Discussion

- We obtained improvements both for the $\beta$ estimation and for the matching process in a single *partially* simulated data set...,
- Similar results can also be obtained in large scale simulations and real data sets
- Current research: prior calibration
- Problems may arise when the regression model does not hold
- More robust estimates assuming heavy tails for the regression error
- **The joint hit-and-miss and regression model can also be seen as a "new" Record Linkage model which is able to handle both categorical and continuous key variables.**

# Some references


- B. Liseo & A. Tancredi (2011). Bayesian estimation of population size via linkage of multivariate normal data sets. *Journal of Official Statistics,* 27(3), pp. 491–505.
- J. Pitman (2006). *Combinatorial Stochastic Processes*. Ecole d'Eté de Probabilités de Saint-Flour XXXII, Lecture Notes in Mathematics, vol. 1875, Berlin, Springer.
- R.C. Steorts, R. Hall & S. Fienberg (2016). A Bayesian approach to graphical record linkage and de-duplication. Journal of the American Statistical Association, Volume 111, 2016 - Issue 516, 1660–1672.
- R. C. Steorts. (2015) Entity resolution with empirically motivated priors. *Bayesian Analysis,* 10(4), pp. 879–875.
- A. Tancredi & B. Liseo (2011). A hierarchical Bayesian approach to record linkage and population size estimation. *Annals of Applied Statistics,*5, pp. 1553–1585.
- A. Tancredi, A., Liseo, B. (2015). Regression Analysis with linked data: Problems and possible solutions. Statistica, 75,1, pp. 1935.
- A. Tancredi, Steorts, R.C., Liseo, B. (2018). Generalized Bayesian Record Linkage and Regression with Exact Error Propagation International Conference on Privacy in Statistical Databases PSD 2018: Privacy in Statistical Databases, pp. 297–313.

# Some references


- B. Liseo & A. Tancredi (2011). Bayesian estimation of population size via linkage of multivariate normal data sets. *Journal of Official Statistics*, 27(3), pp. 491–505.
- J. Pitman (2006). *Combinatorial Stochastic Processes*. Ecole d'Eté de Probabilités de Saint-Flour XXXII, Lecture Notes in Mathematics, vol. 1875, Berlin, Springer.
- R.C. Steorts, R. Hall & S. Fienberg (2016). A Bayesian approach to graphical record linkage and de-duplication. Journal of the American Statistical Association, Volume 111, 2016 - Issue 516, 1660–1672.
- R. C. Steorts. (2015) Entity resolution with empirically motivated priors. *Bayesian Analysis*, 10(4), pp. 879–875.
- A. Tancredi & B. Liseo (2011). A hierarchical Bayesian approach to record linkage and population size estimation. *Annals of Applied Statistics*,5, pp. 1553–1585.
- A. Tancredi, A., Liseo, B. (2015). Regression Analysis with linked data: Problems and possible solutions. Statistica, 75,1, pp. 1935.
- A. Tancredi, Steorts, R.C., Liseo, B. (2018). Generalized Bayesian Record Linkage and Regression with Exact Error Propagation International Conference on Privacy in Statistical Databases PSD 2018: Privacy in Statistical Databases, pp. 297–313.


**THANK YOU!!!**