STATISTICA, anno LXXV, n. 1, 2015

REGRESSION ANALYSIS WITH LINKED DATA: PROBLEMS AND POSSIBLE SOLUTIONS.

Andrea Tancredi [1]
Dipartimento di metodi e Modelli per l'Economia, il territorio e la Finanza, Sapienza Università di Roma, Roma, Italia

Brunero Liseo
Dipartimento di metodi e Modelli per l'Economia, il territorio e la Finanza, Sapienza Università di Roma, Roma, Italia

1. Introduction

From a methodological statistical perspective, the operation of merging two (or more) data sets can be important for two different and complementary reasons:

(i) per sé, i.e. to obtain a larger reference data set or frame, suitable to perform more accurate statistical analyses;

(ii) to calibrate statistical models via the additional information which could not be extracted from either one of the two single data sets.

If the merging step can be accomplished without errors (maybe because a clear identification key is available and it can be used to match units in different datasets), there are no specific consequences on the statistical procedures undertaken in both the situations. In practice, however, identification keys are rarely available and linkage between statistical records is usually performed under uncertainty. This issue has caused a very active line of research among the statistical and the machine learning communities, named "record linkage", where the possibility to make wrong matching decisions must be accounted for, especially when the result of the linking operation, namely the merged data set, must be used for further statistical analyses.

To briefly explain what record linkage is, let us suppose we have two data sets, say F_1 and F_2, whose records respectively relate to statistical units (e.g. individuals, firms, etc.) of partially overlapping samples (or populations), say S_1 and S_2. Records in each data set consist of several fields, or variables, either quantitative or categorical, which may be observed together with a potential amount of noise. For example, in a file of individuals, fields could be *surname, age, sex*, and so on.

[1] Corresponding Author. E-mail: andrea.tancredi@uniroma1.it

The goal of a record linkage procedure is to detect all the pairs of units (j, j'), with $j \in S_1$ and $j' \in S_2$, such that j and j' actually refer to the same unit, and this is performed by the use of the information provided by the observed records in the two datasets. If the main goal of the record linkage process is the former outlined above (case (i)), a new data set is created by merging together three different subsets of units: those which are present in both data sets, those belonging to S_1 only and those belonging to S_2 only. Of course, information regarding the first group of individuals will be richer. Appropriate statistical data analyses may be then performed on the enlarged data set. Since the linkage step is done with uncertainty, the efficiency of the statistical analysis may be jeopardized by *i*) the presence of duplicate units and *ii*) a loss of power, mainly due to erroneous matching in the merging process.

On the other hand, the latter situation (case (ii)), which is more important for the scope of this paper, is even more challenging, both from a practical and from a methodological perspectives. Let us denote the observed variables in F_1 by $(Y, V_1, V_2, \ldots, V_h)$ whereas the observed variables in F_2 are $(X, V_1, V_2, \ldots, V_h)$. One might be interested in performing a linear regression analysis (or any other more complex association model) between Y and X, restricted to those pairs of records which are declared matches after a record linkage analysis based on variables $(V_1, \ldots, V_h)$. The intrinsic difficulties in such a simple problem are well documented in Neter *et al.* (1965) and deeply discussed in Scheuren and Winkler (1993), Scheuren and Winkler (1997) and Lahiri and Larsen (2005). In the regression example, it might be easily seen that the presence of false matches (that is, matching record pairs which do not actually refer to the same statistical unit) reduces the observed level of association between Y and X and, as a consequence, they introduces a bias effect towards zero when estimating the slope of the regression line. Similar biases may appear in every statistical procedure and, in most of the cases, the bias takes a specific direction. As another example, when linkage procedures are used for estimating the size N of a population through a capture-recapture approach, the presence of false matches may severely reduce the final estimate of N.

One should also note, at this point, that in the practical use of record linkage, it is quite usual that the linker (the researcher who matches the two files) and the analyst (the one which performs the statistical analysis) are two different persons, working separately. However, as Scheuren and Winkler (1993) states *"... it is important to conceptualize the linkage and analysis steps as part of a single statistical system and to devise appropriate strategies accordingly"*.

Following such a suggestion, and putting it into a broader perspective, let us assume we observe variables $(Y_1, Y_2, \ldots, Y_k, V_1, V_2, .., V_h)$ on n_1 units in file F_1 and variables $(X_1, X_2, \ldots, X_p, V_1, V_2, \ldots, V_h)$ on n_2 units in file F_2. In this set-up we consider the two-fold objective of *i*) using the key variables $V_1, V_2, \ldots, V_h$ to infer about the common units between sources F_1 and F_2 and, at the same time, of *ii*) adopting a model $\mathcal{M}$ to perform a statistical analysis based on the variables Ys and Xs (or even including the common variables V's), restricted to those records which have been recognized as matches. In order to pursue this double goal, we propose a fully Bayesian analysis which is able - in a very natural way - to

- improve the performance of the linkage step through the use of the extra information contained in the Y's and X's. This happens because pairs of records which do not adequately fit the model $\mathcal{M}$ will be automatically down-weighted in the matching process;
- allow to account for matching uncertainty in the estimation procedure related to model $\mathcal{M}$ involving Y's and X's.
- improve the accuracy of the estimators of the parameters of model $\mathcal{M}$ in terms of bias.

A first attempt to frame the statistical problem of record linkage from a Bayesian perspective can be found in Fortini *et al.* (2001). In that paper the likelihood function arising from the set of multiple comparisons among different records in the two datasets - comparisons which may involve several different variables - was used to estimate the matching configuration through the use of a specific Markov Chain Monte Carlo (MCMC) technique. That approach, together with the one outlined in Larsen (2005), can be interpreted as a Bayesian alternative to the classic record linkage approach, formalized by Jaro (1989), which followed the seminal paper by Fellegi and Sunter (1969). Recently, Tancredi and Liseo (2011) have proposed a different Bayesian matching procedure, particularly suited for categorical variables. They explicitly model the fully observed records through a particular measurement error models, inspired by the so called "hit-and-miss" strategy proposed by Copas and Hilton (1990). In the same paper, the problem of uncertainty in population size estimation based on capture-recapture models with linkage uncertainty was discussed in detail. In addition, Liseo and Tancredi (2011) have introduced a record linkage model for continuous data based on a multivariate normal model with measurement error.

In the last years, several Authors have considered the problem of estimating the parameters of a regression model using linked data. Extending the pioneering works of Scheuren and Winkler (1993) and Scheuren and Winkler (1997) and under the somewhat restrictive assumption that the two data sets represent a permutation of the same list of units, Lahiri and Larsen (2005) have proposed an estimator (LL) of the regression coefficients which is unbiased, conditionally on the matching probabilities provided by the record linkage process. Their approach has been extended by Hof and Zwinderman (2012) to handle more complex and realistic linkage scenarios and logistic regression problems. Generalizations of the LL estimator have been also provided by Kim and Chambers (2012) which proposed a method based on estimating equations. A different approach is outlined in Goldstein *et al.* (2012); here the Authors consider the probabilities of being a match - provided by the record linkage algorithm - as an ingredient to be used within a multiple imputation scenario. Finally, a Bayesian procedure that jointly models the record linkage and the association between variables in two different data sets files has been proposed by Gutman *et al.* (2013). In that paper, the Authors consider the (at least computationally) simpler situation where the number of records to match in the two data sets is relatively small; this is obtained after a large and informative blocking step. They shows that their joint model both improves the matching procedure and the accuracy of the estimation of the

regression parameters in a real data example concerning the "end-of life costs". Another potential limitation of this paper is that the Authors assume a specific matching pattern; in fact, for each singe block of comparisons, all cases in the smaller list are present in the other list.

Practical applications of inference with linked data are very common in biostatistics and epidemiology. Recent examples include, for instance, Hof and Zwinderman (2015) who estimated the association between pregnancy duration of the first and second born children from the same mother from a register without mother identifier and Harron *et al.* (2013) where a data set comprising pediatric intensive care admission records has been linked with blood-stream infection surveillance data in order to evaluate the association between this kind of infection and specific risk factors due to pediatric intensive care.

In the next section we will briefly recall the standard approach to record linkage and then we will propose a simplified version of the Bayesian model described in Tancredi and Liseo (2011). We will also provide some details on a possible simulation strategy for the resulting posterior distribution. In Section 3 we will consider a generalization of the method in order to include the regression model. In Sections 4 and 5 we will illustrate our proposals with simulated and real data sets.

2. Record linkage models

In this section we sketch the probabilistic framework for setting up record linkage models. We first introduce the standard model for record linkage and then we discuss a different way of modelling the comparisons among units, which is more amenable to include the inference model $\mathcal{M}$.

2.1. A brief review of the standard record linkage approach

Suppose we have two matrices of record, say V_1 and V_2 of different sizes n_1 and n_2 respectively. Here

$$V_1 = (v_{11}, \dots v_{1n_1}) \qquad \text{and} \qquad V_2 = (v_{21}, \dots, v_{2n_2})$$

and each single v_{ij} can be represented as $v_{ij} = (v_{ij1}, \dots, v_{ijh})$, that is V_{ij} contains the observed values of a categorical random vector $v = (v_1, \dots, v_h)$ whose support is

$$\mathcal{V} = \{v_{s_1 s_2, \dots, s_h} = (s_1, \dots, s_h) \quad s_1 = 1 \dots, k_1; \dots; s_h = 1, \dots k_h\}.$$

Also, consider the sets M and U of "true matches" and "true non matches" respectively. More precisely,

$$M = \{(j, j') : \text{ record } j \in V_1 \text{ and } j' \in V_2 \text{ refer to the same unit}\},$$

and, of course, $U = M^c$, the complementary set. The main goal of any record linkage technique is to identify which pair of records should be assigned to M; notice that, in any application, no matter what is the overlapping of the two files of records, the cardinality of U is always much larger than the cardinality of

M. The statistical model for a record linkage analysis is built upon the so called comparison vectors $q_{jj'} = (q_{jj'1}, \cdots, q_{jj'h})$, where

$$q_{jj'l} = \begin{cases} 1 & v_{1jl} = v_{2j'l} \\ 0 & v_{1jl} \neq v_{2j'l} \end{cases}, \qquad l = 1, \ldots, h.$$

The comparison vectors $q_{jj'}$ are assumed to be independent and identically distributed random vectors with a distribution given by the following mixture

$$p(q_{jj'}|m, u, w) = w \prod_{l=1}^{h} m_l^{q_{jj'\,l}} (1 - m_l)^{1 - q_{jj'\,l}} + (1 - w) \prod_{l=1}^{h} u_l^{q_{jj'\,l}} (1 - u_l)^{1 - q_{jj'\,l}}. \quad (1)$$

In the previous formula, w represents the marginal probability that a random pair of records belong to the same unit. In other words, w may be interpreted as the percentage of overlapping of the two data sets. The quantities m_l and u_l, $l = 1, \ldots, h$, are the parameter of the two multinomial distributions associated with the two set of comparisons M and U, that is

$$m_l = \Pr(q_{jj'\,l} = 1 | j, j' \in M) \qquad u_l = \Pr(q_{jj'\,l} = 1 | j, j' \in U)$$

Notice that the independence assumption of the comparison vectors q_{jj}'s is, strictly speaking, untenable from a probabilistic perspective. Consider the following example; after comparing record A_1 with records B_1 and B_2, and then record A_2 with B_1 only, the result of the comparison between A_2 and B_2 is often already known. Also, in the standard model, the key variables are assumed independent of each other. Several extensions of this basic set-up have been proposed, mainly by introducing potential interactions among key variables, see for example Winkler (1995) and Larsen and Rubin (2001).

To test whether a given pair should be allocated to M or U, one may consider either the likelihood ratio

$$\lambda = \frac{P(q_{jj'}|(j, j') \in M)}{P(q_{jj'}|(j, j') \in U)} = \frac{\prod_{l=1}^{h} m_l^{q_{jj'l}} (1 - m_l)^{1 - q_{jj'l}}}{\prod_{l=1}^{h} u_l^{q_{jj'l}} (1 - u_l)^{1 - q_{jj'l}}}$$

or - in a Bayesian setting - the posterior probability that a single pair is a match $p((j, j') \in M|q_{jj'})$. In general, a pair of records with a likelihood ratio λ - or a posterior probability - above a fixed threshold, is declared a match. In practice, the choice of the threshold can be problematic, as illustrated, for example, in Belin and Rubin (1995). In this context, optimization techniques may be helpful to rule out the multiple matches issue, that is the possibility that a single unit in data set A is linked with more than one unit in data set B.

2.2. An alternative Bayesian record linkage model

A different approach can be obtained by directly modelling the observed data matrices V_1 and V_2 of the key variables, rather than the mutual comparisons. In this way one can take into account both the potential measurement error and the matching constraints. Let $\tilde{v}_{ijl}$ be true unobserved value for the field l of the

record j on data set V_i and let $\tilde{V}_i$ be the corresponding unobserved data matrix. We assume that

$$p(V_1, V_2|\tilde{V}_1, \tilde{V}_2, \gamma) = \prod_{ijl} p(v_{ijl}|\tilde{v}_{ijl}, \gamma_l) = \prod_{ijl} [\gamma_l I(v_{ijl} = \tilde{v}_{ijl}) + (1 - \gamma_l) q(v_{ijl})] .$$

Notice that v_{ijl} is a mixture of two components, the former is degenerate on the true value while the latter can be any distribution whose support is the set of all possible values of the variable V_l; we believe that - in absence of specific information - taking a uniform distribution for the second component of the mixture is a reasonable assumption. This way, $q(v_{ijl}) = 1/k_l$. Also notice that, in this new model, γ_l is the probability that the variable V_l is observed without noise. This model, known as "hit and miss", was introduced in the record linkage literature by Copas and Hilton (1990) and recently adapted in the Bayesian framework by Tancredi and Liseo (2011) and Hall *et al.* (2013); Other examples of finite mixture models with uniform background has been discussed in Banfield and Raftery (1993) for clustering continuous data in presence of noise.

To build a model for true values $\tilde{v}_{ijl}s$ we need to introduce a matching matrix C. In particular, let C be a $n_1 \times n_2$ matrix whose unknown entries are either 0 or 1, where $C_{jj'} = 1$ represents a match, $C_{jj'} = 0$ denotes a non-match. We assume that each data set does not contain replication of the same unit so that $\sum_{j'} C_{jj'} \leq 1$, and $\sum_j C_{jj'} \leq 1$. Green and Mardia (2006) have used a similar matching matrix in slight different context, i.e. in the problem of alignment of unlabelled points for reconstructing molecular shapes. We assume that the joint distribution for $\tilde{V}_1$ and $\tilde{V}_2$ depends either on the entries of the matching matrix C and on the probability vector $\theta = (\theta_{s_1 \ldots s_h}, s_1 = 1 \ldots, k_1; \ldots; s_h = 1 \ldots, k_h)$ which describes the distribution of the true values one can observe on each sample. More precisely, we assume that

$$p(\tilde{V}_1, \tilde{V}_2|C, \theta) = \prod_{j: C_{jj'}=0\,, \forall j'} p(\tilde{v}_{1j}|\theta) \prod_{j': C_{jj'}=0\,, \forall j} p(\tilde{v}_{2j'}|\theta) \prod_{jj': C_{jj'}=1} p(\tilde{v}_{1j}, \tilde{v}_{2j'}|\theta), \tag{2}$$

where

$$p(\tilde{v}_{ij}|\theta) = \prod_{s_1 \ldots s_h} \theta_{s_1,\ldots,s_h}^{I(\tilde{v}_{ij}=(s_1,\ldots,s_h))},$$

and

$$p(\tilde{v}_{1j}, \tilde{v}_{2j'}|\theta) = \begin{cases} 0 & \text{if } \tilde{v}_{1j} \neq \tilde{v}_{2j'} \\ \prod_{s_1 \ldots s_h} \theta_{s_1,\ldots,s_h}^{I(\tilde{v}_{ij}=(s_1,\ldots,s_h))} & \text{if } \tilde{v}_{1j} = \tilde{v}_{2j'} \end{cases}$$

It should be noticed that the above model can be considered a simplified version of the one proposed in Tancredi and Liseo (2011), where an additional layer - introducing a super-population model - was added at the top of the hierarchy. This simplest version, already used in Hall *et al.* (2013), can be easily obtained by integrating out the additional layer of hierarchy, under specific prior assumptions. Following Hall *et al.* (2013), we also assume that the key variables are independent. In symbols, setting $\theta_{l,s_l} = p(\tilde{v}_{ijl} = s_l|\theta_l)$, with $\theta_l = (\theta_{l1}, \ldots, \theta_{l,k_l})$, we assume that

$$\theta_{s_1,\ldots,s_h} = \prod_{l=1}^{k} \theta_{l,s_l}.$$

To complete the model we need to specify a distribution for the matching matrix C and the prior distributions for the parameters γ_l and θ_l, $l = 1, \ldots, h$. For these latter quantities the standard assumptions of independent Beta distributions for the probabilities γ_l and independent Dirichlet distributions for the vectors θ_l can be adopted. With respect to C, the prior can be elicited in two stages. The first stage consists of a prior distribution $p(t)$, $t = 0, 1, 2, \ldots n_1 \wedge n_2$ on the random variable T: "number of matched pairs in the two data sets". At this stage, the researcher can easily collect information, looking at previous experiences or at the statistical characteristics of the data sets (e.g. if the two data sets refer respectively to a census and a sample, we can expect a large number of matched pairs). At the second stage we define a conditional prior distribution for the configuration matrix C *given* the number of matches. We take the natural noninformative choice of a uniform conditional prior on the set $C^{(t)} = \{C : \sum_{jj'} C_{j,j'} = t\}$ Note also that the cardinality of $C^{(t)}$ is $|C^{(t)}| = \binom{n_1}{t}\binom{n_2}{t}t!$ and that a uniform unconditional prior for C, that will be our choice throughout this paper, can be obtained by assuming $p(t) \propto |C^{(t)}|$ and the aforementioned uniform conditional prior for $p(C|T)$.

2.3. MCMC estimation

The model just outlined cannot be analyzed in a closed form and some form of simulation from the posterior distribution is necessary. In particular, we have implemented a Metropolis within Gibbs algorithm where the updating of parameters γ_l and θ_l can be easily performed by simulating from their respective full conditional distributions, for $l = 1, \ldots, h$. On the other hand, the updating of the matching matrix C and the true values $\tilde{V}_1$ and $\tilde{V}_2$ is jointly obtained In particular, we propose - via a Metropolis-Hastings step - a new matching matrix C, by adding or deleting one matches or switching two matches. Conditionally on the acceptance of the proposed value for C, a Gibbs step is used for the updating of the elements of $\tilde{V}_1$ and $\tilde{V}_2$.

As an example, we illustrate the acceptance probabilities for the specific move in which we "add" a match: when proposing a move from $C_{jj'} = 0$ to $C_{jj'} = 1$, we accept it with probability

$$1 \quad \wedge \quad \frac{q(C|C')}{q(C'|C)} \frac{p(V_1, V_2|C', \theta, \gamma)}{p(V_1, V_2|C.\theta, \gamma)} \frac{p(C')}{p(C)},$$

where $q(C|C')$ is the probability of proposing the reversible "deleting match" move, $q(C'|C)$ is the probability of proposing the "adding match" move. Finally,

$$\frac{p(V_1, V_2|C', \theta, \gamma)}{p(V_1, V_2|C, \theta, \gamma)} = \frac{p(v_{1j}, v_{2j'}|\theta, \gamma)}{p(v_{1j}|\theta, \gamma)p(v_{2j'}|\theta, \gamma)} = \tag{3}$$

$$= \frac{\prod_{l=1}^{h} \left(\gamma_l^2 \theta_{l v_{1jl}} I(v_{1jl} = v_{2j'l}) + \gamma_l(1-\gamma_l)(\theta_{l\, v_{1jl}} + \theta_{l\, v_{2j'l}})/k_l + (1-\gamma_l)^2/k_l^2\right)}{\prod_{l=1}^{h} \left([\gamma_l \theta_{l\, v_{1jl}} + (1-\gamma_l)/k_l][\gamma_l \theta_{l\, v_{2jl}} + (1-\gamma_l)/k_l]\right)}.$$

When the move is accepted, we then propose new values $(\tilde{v}_{1j}, \tilde{v}_{2j'})$ by sampling

from their full conditional distributions given $C_{jj'} = 1$, that is

$$\begin{aligned} p(\tilde{v}_{1jl}, \tilde{v}_{2j'l} | \theta, \gamma, v_{1jl}, v_{2j'l}, C_{jj'} = 1) &\propto \theta_l{}^{\tilde{v}_{1jl}} [\gamma_l I(v_{1jl} = \tilde{v}_{1jl}) + (1 - \gamma_l)/k_l] \\ &\times [\gamma_l I(v_{2jl} = \tilde{v}_{2jl}) + (1 - \gamma_l)/k_l] \end{aligned}$$

if $\tilde{v}_{1jl} = \tilde{v}_{2jl'}$ and 0 otherwise. Similar expressions can be easily obtained for the other possible moves, that is deleting a match or switching matches. Notice that the ratio (3), which appears in the above acceptance probability, is the Bayes factor for comparing the hypothesis that the pair (j, j') is a match versus the alternative hypothesis that it is not a match: see for example Lindley (1977) and Liseo and Tancredi (2011) for similar expressions involving Gaussian distributions.

After that a reasonably large sample has been drawn from the posterior distribution, we estimate the matching configuration via the following - rather natural - point estimate of the matrix C, namely

$$\widehat{C}_{ij} = \begin{cases} 1 & \text{if } p(C_{ij} = 1 | V_1, V_2) \geq \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$$

Some remarks are necessary here. First, the estimate $\widehat{C}$ is in some sense suggested by simple decision theoretic considerations (see Tancredi and Liseo (2011)). Notice that, in this situation, the posterior mean cannot be used since it would provide useless real numbers between 0 and 1. Second, the estimated matrix $\widehat{C}$ should only be used when the linkage procedure is the final goal of the statistical analysis and a set of potential matches must be declared. If the merged data set is the starting point of a new statistical analysis, then one should try to account for the uncertainty on C provided by the posterior distribution of the matrix itself. This is what we describe next in the particular, although very common, case of linear multiple regression model.

3. Bayesian Regression with linked data

Consider the situation where the first data set is a $n_1 \times (h + 1)$ matrix consisting of the variables (y, V_1), and the other data set is a $n_2 \times (h + p)$ matrix, including variables $(V_2, \mathbf{X})$ where $\mathbf{X} = (X_1, \ldots, X_p)$. Also, let $\tilde{\mathbf{X}}$ be the matrix containing the true (unobserved) covariate values for Y. $\tilde{\mathbf{X}}$ has dimension $n_1 \times p$. Conditionally on $\tilde{\mathbf{X}}$ and on the true matching variables $\tilde{V}_1$ and $\tilde{V}_2$, we assume a Gaussian linear regression model for y, that is,

$$y | \tilde{\mathbf{X}}, \boldsymbol{\beta}, \sigma^2 \sim N(\tilde{\mathbf{X}}\boldsymbol{\beta}, \sigma^2 I). \tag{4}$$

In addition, given the matrices of true values $\tilde{V}_1$ and $\tilde{V}_2$, we assume, for V_1 and V_2, the "hit and miss" model as illustrated in §2.2.

Conditionally on the matching matrix C, we also assume that the actual covariate values for y_j are given by the vector $x_{j'}$ (the x-part of the j'-th row of data set B) only if $C_{jj'} = 1$; otherwise, when the j-th row of C is a string of 0's, we assume that the true covariate values for y_j are unknown with a specific distribution $p(\tilde{x})$. This way the covariates for the non-matches pairs are treated as missing

variables. The choice of $p(\cdot)$ is often not crucial and, in general, a multivariate Gaussian distribution will be used. More precisely, we have

$$p(\tilde{\mathbf{X}}|C) = \prod_{j,j':C_{jj'}=1} \delta_{x_{j'}}(\tilde{x}_j) \prod_{j:\sum_{j'} C_{jj'}=0} p(\tilde{x}_j).$$

For the matrices of true values $\tilde{V}_1$ and $\tilde{V}_2$ we will adopt the same model described in (2). The posterior simulation can be easily conducted via a Metropolis-Hastings within Gibbs algorithm where the matching matrix C, the true values $\tilde{V}_1$ and $\tilde{V}_2$ and the true covariates $\tilde{\mathbf{X}}$ are jointly updated, in a way very similar to what we have described in the previous section. In particular, in this case, the "add-one-match" move is accepted with probability

$$1 \quad \wedge \quad \frac{q(C|C')}{q(C'|C)} \frac{p(y, V_1, V_2|C', \theta, \gamma, \boldsymbol{\beta}, \sigma^2)}{p(y, V_1, V_2|C, \theta, \gamma, \boldsymbol{\beta}, \sigma^2)} \frac{p(C')}{p(C)},$$

where

$$\frac{p(y, V_1, V_2|C', \theta, \gamma, \boldsymbol{\beta}, \sigma^2)}{p(y, V_1, V_2|C, \theta, \gamma, \boldsymbol{\beta}, \sigma^2)} = \frac{\phi(y_j; x_{j'}^T \beta, \sigma^2)}{\int \phi(y_j; \tilde{x}^T \beta, \sigma^2) p(\tilde{x}) d\tilde{x}} \frac{p(v_{1j}, v_{2j'}|\theta, \gamma)}{p(v_{1j}|\theta, \gamma) p(v_{2j'}|\theta, \gamma)}, \quad (5)$$

with $\phi(\cdot; \mu, \sigma^2)$ representing the density of a Normal variable with mean μ and variance σ^2. There are several important remarks and comments concerning the acceptance probability (5).

i) Formula 5 points out that distribution of the matching matrix C is dependent on the values of the $\boldsymbol{\beta}$ parameters and makes explicit the feed-back effect between the parameters of the regression model and the matching process. This effect has the role to modify the matching estimation leading both to a possible improvement for the record linkage process and to a "bias-correction" effect for the regression estimates.

ii) A closed-form expression for

$$p(y_j; \boldsymbol{\beta}, \sigma^2) = \int \phi(y_j; \tilde{x}^T \boldsymbol{\beta}, \sigma^2) p(\tilde{x}) d\tilde{x}$$

can be obtained, for example, by assuming a multivariate normal for $p(\tilde{x})$. In fact, assuming that $\tilde{x} \sim N(\mu_0, \Sigma_0)$, a simple calculation shows that

$$y_j|\boldsymbol{\beta}, \sigma^2 \sim N(\mu_0^T \boldsymbol{\beta}, \sigma^2 + \boldsymbol{\beta}^t \Sigma_0 \boldsymbol{\beta}).$$

iii) When the "add-one-match" move is accepted, one updates the true values $(\tilde{x}_j, \tilde{v}_{1j}, \tilde{v}_{2j'})$ by drawing a value from their full conditional distributions conditionally on the new status $C_{jj'} = 1$

We also observe that, in order to update the regression parameters $\boldsymbol{\beta}$ and σ^2, we only need to consider the likelihood provided by the regression model (4); hence a standard Gibbs move can be adopted.

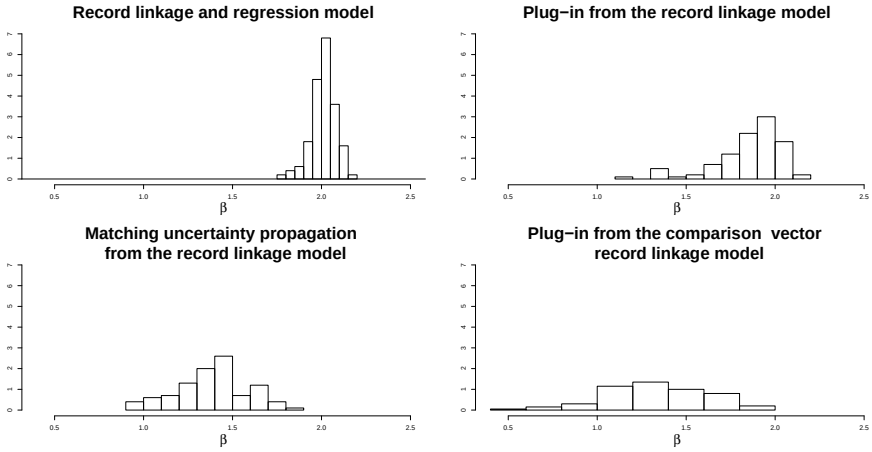

Figure 1 – Simulation study. The posterior distributions of β's obtained under the "record linkage and regression" model, the "plug-in" approach from the record linkage model, the "matching uncertainty propagation" approach from the record linkage model and the plug-in approach from the "comparison vector record linkage" model (Fortini *et al.*, 2001). The true value of β is 2.

4. Simulation study

We now evaluate our hierarchical model for regression analysis with linked data via a simulation study. In particular we have generated 100 pairs of data sets with sizes $n_1 = 100$ and $n_2 = 80$. The number of true matches T is Binomial with size 80 and probability 0.75 and the distribution of the matching matrix C given T is uniform. Each pair of data sets shares 3 independent key variables $\tilde{V}_j$, $j = 1, \ldots, 3$ with 5, 10 and 50 categories, respectively, and different probability distributions. The probability of correctly observing the true values is $\gamma_l = 0.95 \;\; l = 1, \ldots, 3$. For the regression model, we assume a single covariate X whose values are generated from a Normal distribution with mean $\mu_x = 100$ and variance $\sigma_x^2 = 20^2$ and for the response variable we assume that $y|x$ is Normal with mean $\alpha + \beta x$ with $\alpha = 3$ and $\beta = 2$, and variance $\sigma^2_{y|x} = 10^2$.

For each simulated pair of data sets we run the MCMC sampler both for the joint "record linkage and regression" model outlined in Section 3 and for the "record linkage only" discussed in Section 2.2. For both models we have initialized the algorithms via a matching matrix without matches and we have drawn 55000 iterations of the algorithm with a burn-in of 5000. With the "record linkage and regression" model the natural estimate of β is the posterior mean $E(\beta|V_1, V_2, y, X)$. The upper left panel of Figure 1 reports the histogram of the 100 posterior means. In order to estimate β with the "record linkage only" model. one can consider a plug-in approach by using the posterior mean of β, conditionally on the matching configuration provided by the point estimate $\widehat{C}$. Alternatively, in order to better propagate the matching uncertainty in the regression estimates, one can adopt a hybrid approach where the MCMC estimates $\hat{\beta}$ provided by the

simulated matching matrix at each MCMC iteration of the record linkage model, are averaged. The resulting estimates of β are reported respectively in the upper right and lower left panels of Figure 1. For each simulated pair of data sets we also provide an estimate of the matching matrix via the standard approach outlined in Section 2.1. The lower right panel of Figure 1 shows the corresponding estimates of β obtained conditionally on the estimated matching matrix.

Note that in terms of the inference for β, the "record linkage and regression" model outperforms all other three estimation strategies. In particular the sampling distribution of the estimator $E(\beta|V_1, V_2, y, X)$ is centred around the true value $\beta = 2$ (the observed average is 2.01) while all other estimators are strongly biased towards 0. The bias elimination effect provided by the "record linkage and regression" model is mainly due to the low value of likelihood that the false matches, leading to independent y and x pairs, receive from the regression part of the model. We also notice that the sampling variability of $E(\beta|V_1, V_2, y, X)$ is much smaller when compared with the other approaches. In fact, the introduction of the information provided by the linear relationship between y and x in the "record linkage model" has also had the effect of improving the record linkage quality via a reduction of the matching uncertainty. This feedback effect is confirmed by the true positive matches rate

$$TPR = \frac{\sum_{jj'} C_{jj'} \hat{C}_{jj'}}{\sum_{jj'} \hat{C}_{jj'}}$$

distribution reported in Figure 2 together with the distribution of the declared matches $\hat{T} = \sum_{jj'} \hat{C}_{jj'}$. With respect to the other approaches, the "record linkage and regression" model has, on average, the higher and less variable true positive matches rate and a distribution of $\hat{T}$ more concentrated on the correct number of matches.

5. An Illustration: Italian Survey of Household Income and Wealth

In this section we illustrate an application of the proposed methods using data from the Italian Survey on Household Income and Wealth. The Italian Survey on Household Income and Wealth (SHIW) is a sample survey conducted by the Bank of Italy every 2 years. The 2010 survey covers 7,951 households composed of 19,836 individuals. Panel households and individuals represent 58% of the data. From the 2010 survey we consider the individual net disposable income as the response variable Y of our regression model and the following matching variables: sex, age, marital status, employment status, working sector. From the 2008 survey we consider, in addition to the matching variables, the 2008 net disposable income which is assumed as the covariate X of the regression model. The aim of the application is to calibrate a regression model of Y on X, which is based on those pairs of records which are declared matches by the record linkage procedure.

Before illustrating the results of the matching model in this particular example, where the linkage structure is known for each pair of records, we first note how a slight modification of the matching configuration, for example by deleting 10%

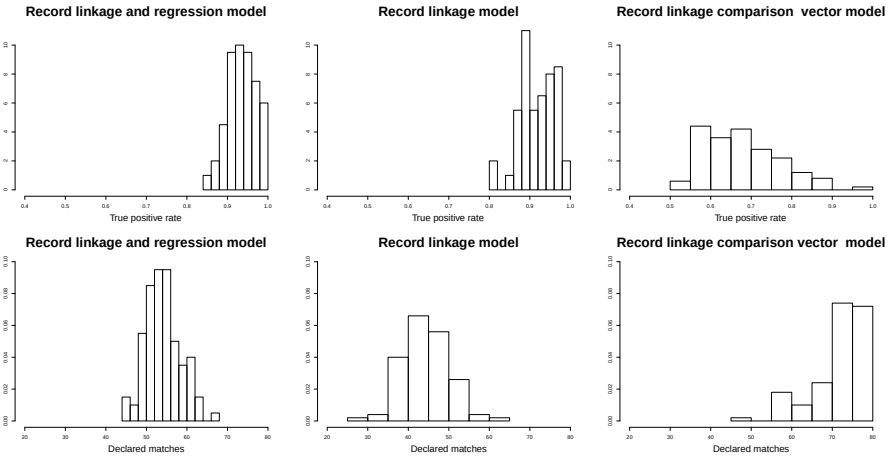

Figure 2 – Simulation study. True positive rate and declared matches distributions obtained with the "record linkage and regression" model, the "record linkage only" model and the "comparison vector record linkage" model.

of the true matches and adding 10% of false matches, may lead to dramatically different regression analyses. This is illustrated in Figure 3 for the single block provided by the Friuli region. In the left panel the circle dots represent the true matching configuration while the cross dots represent the perturbed one. In the left panel, the two ordinary least square regression lines, obtained with the two different data sets, are also reported; as expected, the line with the smaller slope refers to the perturbed data set. In the right panel we show the posterior distributions of the slope coefficient with the two data sets and the usual noninformative prior for (α, β, σ). namely $\pi(\alpha, \beta, \sigma) \propto \sigma^{-1}$. Note that the the two distributions provide quite different credible intervals.

To illustrate the results of our methods we focus on the single Friuli block by comparing different regression analyses. The first one is based on a subset of six key variables for the matching part of the model and the raw income data for the regression fitting. In the upper panels of Figure 4 we show the regression estimates obtained (i) by

i. fitting a Bayesian regression model directly on the true matching configurations (203 matches);

ii. applying our regression and matching model;

iii. fitting a Bayesian linear regression via the matching matrix estimated by the record linkage model, i.e. the plug-in approach;

iv. repeating the analysis on the true matching configurations without the two very influential observations with 2008 income level larger than 150000 Euros.

It is interesting to note that the integrated model produces inferences which are very similar to those obtained by using the true matches, but *without* the two

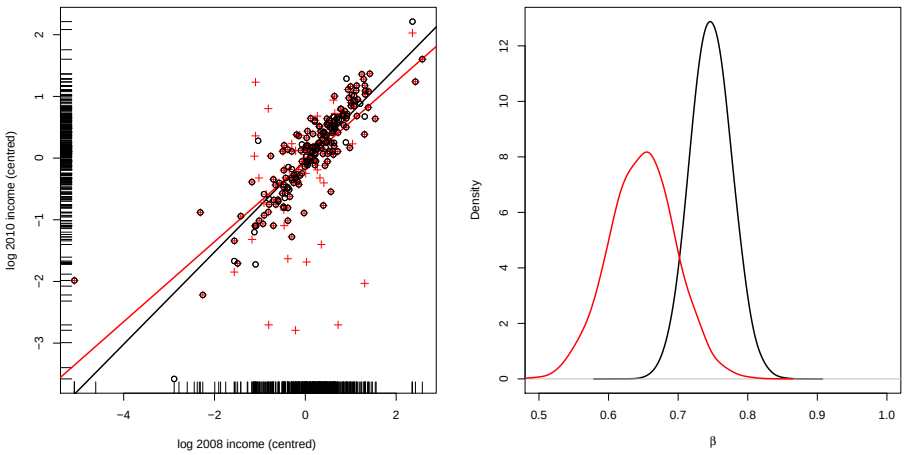

Figure 3 – SHIW data (Friuli block, $n_1 = 434, n_2 = 355$). Regression analysis with the 2010 individual income as the response variable and the 2008 individual income as a covariate. *Left panel*: $\circ$ =true matches, $+$ =declared matches after a perturbation procedure. *Right panel*: posterior distributions for the regression coefficientes with the true matches (black line) and the declared matches.

outliers: since these two observations do not fit the regression model calibrated on the bulk of the matched pairs, they receive a low likelihood of being a match from the regression part of the model. As a consequence they are erroneously considered as non matches, thus removing, or at least mitigating, their effect on the regression estimates. Also, note that the plug-in approach with the record linkage model does not have this protection mechanism against outliers.

In the central panels of Figure 4 we show the results obtained by taking the logarithm of the income variables and repeating the four regression analyses listed above. After the log transformation, the two extreme observations do not produce any effect on the regression fitting: in this case the plug-in approach produces very similar estimates when compared to the true regression line, while the integrated model provides a slightly larger slope coefficient. Finally, in the bottom panels we show all the nine variables and the log-transformed regression variables. With the additional information provided by the three additional key variables the results of the plug-in approach are even more closer to the true fitting.

We conclude this Section by analysing the record linkage performance. In Table 1 we report the false negative rate, FNR, and the false positive rate FPR

$$FNR = \frac{\sum_{jj'}(1-\hat{C}_{jj'})C_{jj'}}{\sum_{jj'}\hat{C}jj'} \quad FPR = \frac{\sum_{jj'}\hat{C}_{jj'}(1-C_{jj'})}{\sum_{jj'}\hat{C}jj'}$$

of the several approach proposed, obtained by changing the number of key variables or by using the log trasformation for the regression variables. Note that the integrated "record linkage and regression" model, because of the feedback effect on the matching estimation, produces better record linkage performance, by lowering both the false negative rate and the false positive rate with respect to the

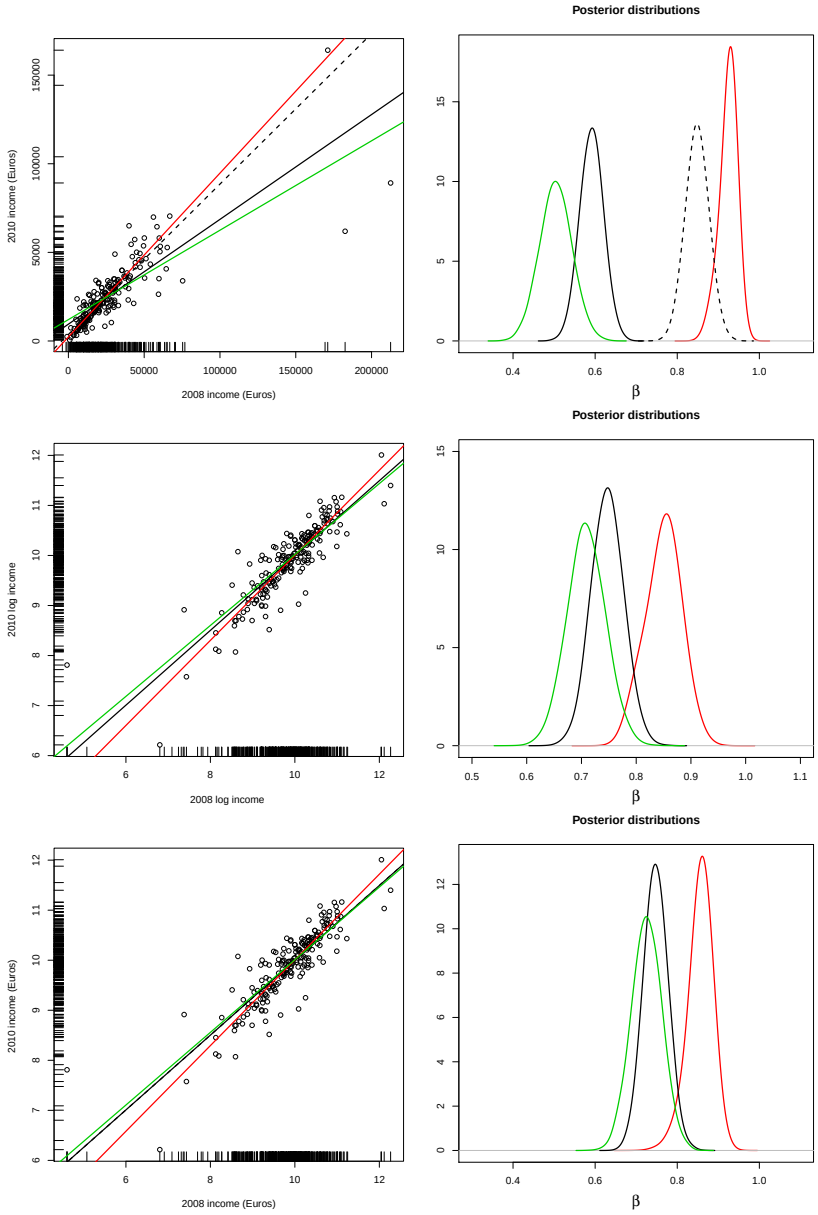

Figure 4 – Results for the SHIW data (Friuli block). Black line: true regression line using the 203 true matches. Black dashed line: true regression line without 2 very influential observations. Red line: Bayesian estimate with the "regression and record linkage". Green line: Bayesian estimate with the "record linkage only" model and posterior regression on the matched pairs. First row: six key variables, non transformed data. Second row: six key variables, log transformed response and covariate. Third row: nine key variables, log transformed response and covariate

TABLE 1
Record linkage performance for the SHIW data with the Friuli block.

model	RL	RL+REG	RL+REG (log)	RL	RL+REG (log)
key variables	6	6	6	9	9
FNR	0.43	0.33	0.33	0.27	0.22
FPR	0.33	0.30	0.28	0.29	0.30

simple Bayesian matching model. We also note that the record linkage results are affected by the use of the logarithm trasformation for the regression variables which, in this particular example, provide better performance. Such a behaviour is connected with the improved goodness of fit of the linear model on the log variables. As a general comment, one might argue that a perfect linear relationship between the response and the covariates would be practically equivalent to having a common identification key between the two files. This suggests that strong linear relationships are generally more informative for the matching process than weak relationships. For example, with the above data, the value of the R^2 index calculated with the true matches is equal to 0.68 on the original scale and 0.77 after the log trasformations.

6. Discussion

We have described the possibility to deal with record linkage and a regression model for linked data within a common Bayesian framework. The resulting model has the twofold effect of propagating the matching uncertainty into the regression analysis and to account for the information provided by the linear relationships between the response and the covariates into the matching estimation. We have shown via simulated data that the latter effect may significantly improve the estimation process.

Anyway in real applications, where the linear dependence between a response variable and a set of covariates is no more than a model assumption, the matching results should be presented together with an additional sensitivity analysis with respect to the possible modeling alternatives and/or with an evaluation of the uncertainty associated with the selected model. In fact different set of covariates or transformations involving both the response and the covariates may affect the record linkage process. On the other hand, we also notice that classical record linkage techniques usually face similar problems regarding the transformation of the key variables and the modeling assumptions of the comparison vector. All in all, the additional sensitivity induced by the regression variables should be interpreted as the fair price to pay for the "creation" of an extra key variable comprised by the response variable on one dataset and the regression covariates on the other.

References

J. D. Banfield, A. E. Raftery (1993). *Model-based gaussian and non-gaussian clustering*. Biometrics, pp. 803–821.

T. Belin, D. Rubin (1995). *A method for calibrating false - match rates in record linkage*. Journal of the American Statistical Association, 90, pp. 694–707.

J. Copas, F. Hilton (1990). *Record linkage: statistical models for matching computer records*. Journal of the Royal Statistical Society, A, 153, pp. 287–320.

I. Fellegi, A. Sunter (1969). *A theory of record linkage*. Journal of the American Statistical Association, 64, pp. 1183–1210.

M. Fortini, B. Liseo, A. Nuccitelli, M. Scanu (2001). *On Bayesian record linkage*. Research in Official Statistics, 4, pp. 185–198.

H. Goldstein, K. Harron, A. Wade (2012). *The analysis of record-linked data using multiple imputation with data value priors*. Statistics in Medicine, 31, no. 28, pp. 3481–3493.

P. J. Green, K. V. Mardia (2006). *Bayesian alignment using hierarchical models, with application in protein bioinformatics*. Biometrika, 93, pp. 235–254.

R. Gutman, C. C. Afendulis, A. M. Zaslavsky (2013). *A Bayesian procedure for file linking to analyze end-of-life medical costs*. Journal of the American Statistical Association, 108, no. 501, pp. 34–47.

R. Hall, R. C. Steorts, S. E. Fienberg (2013). *Bayesian parametric and nonparametric inference for multiple record linkage*. Working paper, Carnagie Mellon University.

K. Harron, H. Goldstein, A. Wade, B. Muller-Pebody, R. Parslow, R. Gilbert (2013). *Linkage, evaluation and analysis of national electronic healthcare data: application to providing enhanced blood-stream infection surveillance in paediatric intensive care*. PloS one, 8, no. 12, p. e85278.

M. Hof, A. Zwinderman (2012). *Methods for analyzing data from probabilistic linkage strategies based on partially identifying variables*. Statistics in Medicine, 31, no. 30, pp. 4231–4242.

M. Hof, A. Zwinderman (2015). *A mixture model for the analysis of data derived from record linkage*. Statistics in medicine, 34, no. 1, pp. 74–92.

M. Jaro (1989). *Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida*. Journal of the American Statistical Association, 84, pp. 414–420.

G. Kim, R. Chambers (2012). *Regression analysis under incomplete linkage*. Computational Statistics & Data Analysis, 56, no. 9, pp. 2756–2770.

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

M. Larsen (2005). *Advances in record linkage theory: Hierarchical Bayesian record linkage theory.* Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 3277–3283.

M. D. Larsen, D. Rubin (2001). *Iterative automated record linkage using mixture models.* Journal of the American Statistical Association, 96, pp. 32–41.

D. Lindley (1977). *A problem in forensic science.* Biometrika, 64, pp. 207–213.

B. Liseo, A. Tancredi (2011). *Bayesian estimation of population size via linkage of multivariate normal data sets.* Journal of Official Statistics, 27, pp. 491–505.

J. Neter, E. S. Maynes, R. Ramanathan (1965). *The effect of mismatching on the measurement of response errors.* Journal of the American Statistical Association, 60, no. 312, pp. 1005–1027.

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

F. Scheuren, W. E. Winkler (1997). *Regression analysis of data files that are computer matched, Part II.* Survey Methodology, 23, pp. 157–165.

A. Tancredi, B. Liseo (2011). *A hierarchical Bayesian approach to record linkage and population size problems.* Annals of Applied Statistics, 5, pp. 1553–1585.

W. Winkler (1995). *Matching and record linkage.* In *Buisness Survey Methods*, Wiley, New York, pp. 355–384. B. G. Cox, D. A. Binder, B. N. Chinnappa, A. Christianson, M.J. Colledge and P.S. Kott Editors.

Summary

In this paper we have described and extended some recent proposals on a general Bayesian methodology for performing record linkage and making inference using the resulting matched units. In particular, we have framed the record linkage process into a formal statistical model which comprises both the matching variables and the other variables included at the inferential stage. This way, the researcher is able to account for the matching process uncertainty in inferential procedures based on probabilistically linked data, and at the same time, he/she is also able to generate a feedback propagation of the information between the working statistical model and the record linkage stage.
We have argued that this feedback effect is both essential to eliminate potential biases that otherwise would characterize the resulting linked data inference, and able to improve record linkage performances. The practical implementation of the procedure is based on the use of standard Bayesian computational techniques, such as Markov Chain Monte Carlo algorithms. Although the methodology is quite general, we have restricted our analysis to the popular and important case of multiple linear regression set-up for expository convenience.

Keywords: Bayesian regression; Hit-miss model; Metropolis-Hastings algorithm; Record linkage