Data Integration in Social Sciences: the earnings intergenerational mobility problem

Integrazione di dati nelle Scienze Sociali: il problema della mobilità intergenerazionale dei redditi

Veronica Ballerini, Francesco Bloise, Dario Briscolini and Michele Raitano

Abstract Merging operations on two or more datasets has become an usual need for Statistical Institutes in the last few decades. Nowadays social sciences offer the opportunity of a new application of data integration techniques. In this short paper we deal with the problem related to the estimation of the intergenerational earnings elasticity when proper datasets are not available. We compare the classical Two Samples Two Stages Least Squares with “Record Linkage” and “Matching” procedures.

Abstract *Abstract in Italian* La fusione di due o più dataset è ormai da qualche decennio una necessità per gli Istituti di Statistica. Oggi le scienze sociali offrono un’opportunità di nuova applicazione delle tecniche di integrazione di dati. In questo lavoro verrà trattato il problema legato alla stima dell’elasticità intergenerazionale del reddito quando datasets adeguati non sono disponibili. In particolare verrà confrontato l’approccio del Two Samples Two Stages Least Squares con le procedure di Record Linkage e Matching.

Key words: TSTSLS, Record Linkage, Matching

Veronica Ballerini
Sapienza Università di Roma, Via del Castro Laurenziano 9, Roma 00161,Italy
e-mail: veronica.ballerini@uniroma1.it

Francesco Bloise
Sapienza Università di Roma, Via del Castro Laurenziano 9, Roma 00161, Italy
e-mail: francesco.bloise@uniroma1.it

Dario Briscolini
Università di Padova, Via Giustiniani 2, Padova 35128, Italy
e-mail: dario.briscolini@unipd.it

Michele Raitano
Sapienza Università di Roma, Via del Castro Laurenziano 9, Roma 00161, Italy
e-mail: michele.raitano@uniroma1.it

1 Introduction

Empirical studies on economic mobility are usually intended to evaluate to what extent economic opportunities of children are associated with those of their parents, by estimating the so called intergenerational income elasticity (IGE) . This measure of mobility is the estimated coefficient of an OLS regression in which the dependent variable Y is the logarithm of permanent earnings (or incomes) of children and the regressor X is the logarithm of permanents earnings of parents. Unfortunately, data which allow to link earnings of sons and their fathers when the two generations were aged approximately the same are barely available in many countries. This the reason why scholars exploit another methodological approach which is called two-samples-two-stages least squares (TSTSLS) (see Björklund and Jäntti [1]). According to this methodology, it is possible to estimate IGEs by exploiting a sample of sons who report some socio-economic characteristics of their fathers and an independent sample of pseudo-fathers i.e. generic individuals taken in the pasts with approximately the same age of actual fathers. The problem of TSTSLS is that it may be biased. In this short paper we propose two different approaches to this issue, "Record Linkage" and Matching". The paper is organized as follows. The second section describes the formal statistical framework; the third reviews the TSTSLS methodology; the fourth and the fifth illustrate our alternatives to the classical approach. Finally we show an application on a real dataset.

2 The basic statistical framework

Let us suppose we have two data sets, say F_1 of size $N_1 \times (h+1)$ and F_2 of size $N_2 \times (h+p)$. Records in each data set consist of several variables which may be observed together with a potential amount of measurement error. Let us denote the observed variables in F_1 by $(Y, Z_{*1}, Z_{*2}, \ldots, Z_{*h})$, whereas the observed variables in F_2 are $(X_1, X_2, \ldots, X_p, Z_{*1}, Z_{*2}, \ldots, Z_{*h})$. Also suppose that one is interested in performing a linear regression analysis of Y on $X_1, X_2, \ldots, X_p$ restricted to those pairs of records which are declared matches after a record linkage or a matching analysis based on variables $(Z_{*1}, \ldots, Z_{*h})$. The goal of the "merging procedures" is to detect all the pairs of units (j, j'), with $j \in F_1$ and $j' \in F_2$, such that j and j' actually refer to the same unit or j' is the unit in F_2 mostly similar to j .

3 The classical approach

The TSTSLS consists of two sequential steps. In the first stage, earnings of pseudo fathers are regressed on socio-economic characteristics in the main sample according to the following regression:

$$X_{j'} = \alpha + \theta_1 Z^{pf}_{j'*} + v_{j'} \tag{1}$$

where $X_{j'}$ are earnings of pseudo-father j', $Z^{pf}_{j'*}$ is a vector of the socio-economic characteristics of j', α is the intercept and $v_{j'}$ is the usual disturbance. The estimated coefficient $\hat{\theta}_1$ is then used to predict missing fathers earnings by merging the two samples according to child-reported characteristics of actual fathers. The intergenerational earnings (or income) elasticity β is thus estimated in the second stage:

$$Y_j = \alpha + \beta \hat{X}_j + \varepsilon_j \tag{2}$$

where Y_j is the logarithm of son j earnings and $\hat{X}_j = \hat{\theta}_1 Z^{f}_{j*}$ is the prediction of the logarithm of his father earnings (see Jerrim et al. [6]). The more the socio-economic characteristics perform well at predicting fathers economic status, the less estimated elasticities will be biased. More specifically, when one tries to impute fathers economic status, he is likely to make some errors in measuring their income. This reduces estimated elasticities under the assumption of classical measurement error. Moreover, if the set of socio-economic characteristics is not able to capture other characteristics of individuals which are positively correlated across generations, then the elasticity will be again downward biased.

4 Record Linkage

The issues raised in the previous section suggest the necessity of alternative strategies and perspectives. One of this possibilities is represented by direct modeling of the linkage uncertainty. In contrast to heuristic methods of data integration the introduction of a probabilistic model on the linkage step allows to evaluate the quality of the results. Moreover it may represent a more natural and realistic way to deal with the absence of joint information on the variables of interest.

Regression with linked data is well documented in Lahiri and Larsen [7]. Tancredi and Liseo [9] have proposed a Bayesian approach for Record Linkage (see also Briscolini et al. [2]). Let $\tilde{z}_{ijl}$ be the true latent value for field l of record j in data set Z_i and let $\tilde{Z}_i$ $(i = 1,2)$ be the corresponding unobserved data matrix. Assume the “hit and miss’ model by Copas and Hilton [3]:

$$p(Z_1, Z_2 | \tilde{Z}_1, \tilde{Z}_2, \nu) = \prod_{ijl} p(z_{ijl} | \tilde{z}_{ijl}, \nu_l) = \prod_{ijl} \left[\nu_l I(z_{ijl} = \tilde{z}_{ijl}) + (1 - \nu_l) \xi(z_{ijl}) \right] \tag{3}$$

The above expression is a mixture of two components: the former is degenerate at the true value while the latter can be any distribution whose support is the set of all possible values of the variable Z_{*l}, with $l = 1,2,...,h$.

As in Tancredi and Liseo [9], 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.

Assume that each data set does not contain replications. Also, assume that the joint distribution of $\tilde{Z}_1$ and $\tilde{Z}_2$ depends on C and on a probability vector θ which describes the distribution of the true values of Z_1 and Z_2. In detail, assume that

$$p(\tilde{Z}_1,\tilde{Z}_2|C,\theta) = \prod_{j:C_{jj'}=0,\forall j'} p(\tilde{z}_{1j}|\theta) \prod_{j':C_{jj'}=0,\forall j} p(\tilde{z}_{2j'}|\theta) \prod_{jj':C_{jj'}=1} p(\tilde{z}_{1j},\tilde{z}_{2j'}|\theta) \quad (4)$$

where $p(\tilde{z}_{ij}|\theta)$ and $p(\tilde{z}_{1j},\tilde{z}_{2j'}|\theta)$ are specific probability distributions depending on θ. The record linkage model has to be combined with the regression model in order to estimate the IGE. To complete the model, a prior distribution must be elicited for the matrix C, ν, θ and the regression parameters. The posterior distribution is not available in closed form: MCMC samples are required.

5 Statistical Matching

A valid alternative to the record linkage procedures described above is statistical matching. Although this integration technique may seem similar to record linkage, they address two different problems. On the one hand, contrarily to record linkage, matching procedures deal with observed units that are not overlapping. On the other hand, statistical matching does not take into account the possibility of measurement errors. To the aim of this work, we only focus on the so called micro approach of statistical matching (see D'Orazio [4]), i.e. the approach whose purpose is the generation of a synthetic dataset with complete information on both the variables observed only in one of the files and those observed in common. In practice, through the micro approach we are able to match records according to the related key variables. We assume that the samples belonging to the different files are generated from the same unknown distribution $f(Y,X,Z)$. In this framework, an example of statistical matching method belonging to the micro approach is the Distance Hot Deck (among others, see Rodgers [8]). In the two samples case, the first step consists in assigning to one of the files the role of recipient, i.e. the file that receives information from the so called donor. Secondly, for each record in the recipient file we measure the distance d among the h matching variables of each record in the donor; the pairs of records with the minimum distance become matches. In other words, (j,j'), $j \in F_1$,$j' \in F_2$, is a match if:

$$d_{(jj')} = \min_{j'} |z_{1jl} - z_{2j'l}|, \; j = 1,...,N_1, j' = 1,...,N_2, l = 1,...,h \quad (5)$$

This is valid in the case of continuous matching variables. Ordinal categorical might be associated to continuous variables, yet considering a proper weighting system, accordingly to the meaning and the role of the variable. Instead, distances between non ordinal may be computed assigning 1 if the value of the key variable of the recipient is equal to that of the donor, and 0 otherwise. At the end of the procedure, we obtain a complete dataset. According to the times each record in the donor file can be used as donor, we distinguish the methods of unconstrained (more than once) and constrained (only once) distance hot deck, whose main advantage is to maintain the marginal distribution of the imputed variable.

6 The application

We use the dataset AD-SILC that has been built merging the cross sectional 2005 wave of IT-SILC (the Italian component of the Eu-SILC) with the administrative records - collected by the Italian National Social Security Institute (INPS) - about working episodes and earnings of all individuals interviewed in IT-SILC 2005 since the beginning of their career. To our aims, IT-SILC 2005 includes a specific section about intergenerational mobility where many aspects of family background of the respondents are recorded. Sons, selected in the period 1970-1974, are followed since age 35 up to 39 in the period 2005-2013. Pseudo-fathers are selected in the period 1980-1988 and followed since age 40 up to 44: they were born in the period 1940-1944. Earnings of both generations are averaged. According to the notation introduced in section 2, let F_1 and F_2 be the matrices of records of the younger and the older generation respectively. The sizes of the files are $1509 \times (1+5)$ and $17245 \times (1+5)$ respectively. For each individual belonging to F_1, his full wage (Y), and a set $Z = (Z_{*1}, Z_{*2}, ..., Z_{*5})$[1] of characteristics about his father are recorded. Yet for each individual of F_2, his full wage (X) is recorded along with 5 personal characteristics - the same of Z.

$$F_1 = \begin{pmatrix} Y_1 & Z_{*1,1} & \cdots & Z_{*1,5} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{1509} & Z_{*1509,1} & \cdots & Z_{*1509,5} \end{pmatrix}, F_2 = \begin{pmatrix} X_1 & Z_{*1,1} & \cdots & Z_{*1,5} \\ \vdots & \vdots & \ddots & \vdots \\ X_{17245} & Z_{*17245,1} & \cdots & Z_{*17245,5} \end{pmatrix}.$$

Z_{*1} is used as blocking variable in record linkage steps and, similarly, as donor class in matching procedure (as suggested by D'Orazio [5]). In the latter we assign to Z_{*4} a double role, given its structure: it is first a stratification variable and then it has been used to compute distances. Firstly, the ISCO-08 macro classification of occupation categories is used to create a blocking variable that assumes 9 values. Secondly, through Z_{*4} we compute distances within the macro categories. Z_{*2}, Z_{*3} and Z_{*5} are matching variables.

7 Discussion

As we expected, the estimates obtained through the standard approach are not the same as those provided by the data integration techniques (table 1). Differently from

[1] Z_{*1}: region of residence (ISTAT codification), Z_{*2}: year of birth, Z_{*3}: level of education (ISCED classification), Z_{*4}: occupation category (according to ISCO-08 classification), Z_{*5}: dummy assuming value 1 if the individual is self employed and 0 otherwise.

Table 1 Comparison of different estimators of the regression coefficient IGE

Method	Point Estimate	Standard deviation
TSTSLS	0.499	0.055
Record Linkage	0.272	0.040
Matching	0.256	0.023

the proposed methods, TSTSLS does not preserve the whole variability of X; the regression coefficient may be overestimated. Since X is only a proxy of the true fathers' income, further analyses including the measurement error seem opportune.

Certainly, record linkage allows to better account for linkage uncertainty with respect to the TSTSLS. Furthermore, it deals with measurement error in the matching variables; this is relevant since sons could have misreported some key information about their fathers. From this point of view, record linkage should be preferred also to statistical matching. On the other hand, as already said in section 5, matching procedures apply to non-overlapping samples. With this respect, statistical matching seems to be more appropriate than linkage procedures. Nevertheless, improving the information provided by the matching variables increases the efficiency of data integration techniques.

References

1. Björklund A., Jäntti M. (2011). *Intergenerational income mobility and the role of family background*. The Oxford Handbook of Economic Inequality.
2. Briscolini D., Di Consiglio L., Liseo B., Tancredi A., Tuoto T. (2017). *New methods for small area estimation with linkage uncertainty*. To appear on International Journal of Approximate Reasoning (IJAR).
3. Copas, J., Hilton, F. (1990). *Record linkage: statistical models for matching computer records*. Journal of the Royal Statistical Society, A, 153, pp. 287–320.
4. D'Orazio M., Di Zio M., Scanu M. (2004) *Statistical matching: theory and practice*, Wiley Series in Survey Methodology, John Wiley & Sons, Ltd.
5. D'Orazio M. (2017) *Statistical matching and imputation of survey data with StatMatch for the R environment*, R package vignette https:cran.r-project.orgwebpackages StatMatchvignettesStatistical_Matching_with_StatMatch.pdf
6. Jerrim, J., Choi, A. and Rodriguez Simancas, R. (2016) *Two-Sample Two-Stage Least Squares (TSTSLS) estimates of earnings mobility: how consistent are they?* Survey Research Methods, Vol. 10, No. 2, pp. 85-102
7. Lahiri P., Larsen, M.D. (2005). *Regression Analysis With Linked Data.* Journal of the American Statistical Association, 100, 222–230.
8. Rodgers W.L. (1984) *An evaluation of statistical matching*, Journal of Business and Economic Statistics 2, pp. 91102
9. Tancredi, A. and Liseo, B. (2015) *Regression Analysis with linked data: Problems and possible solutions*. Statistica, 75,1:19–35.