The effects of matching algorithms and estimation methods using linked data

Rachel Anderson*

This Version: August 30, 2019

Abstract

This paper studies the effect of different matching algorithms and estimation procedures for linked data on the quality of the estimates that they produce.

1 Introduction

In applied microeconomics, identifying a common set of individuals appearing in two or more datasets is often complicated by the absence of unique identifying variables. For example, Aizer et al. (2016) link children listed on their mother's welfare program applications with their death records using individuals' names and dates of birth. However, since name and date combinations are not necessarily unique (and may be prone to typographical error), the authors identify cases where multiple death records seem to refer to the same individual. Instead of dropping these observations from their analysis, they use estimation techniques from Anderson et al. (2019) that allow for observations to have multiple linked outcomes.

*Mailing Address: Department of Economics, Julis Romo Rabinowitz Building, Princeton, NJ 08544. Email: rachelsa@Princeton.edu. This project received funding from the NIH (Grant 1 R01 HD077227-01).

The methods in Anderson et al. (2019) assume that each of the linked outcomes is equally likely to be the true match; however, the authors describe how to construct more efficient estimators if additional information about match quality is available. Specifically, if the researcher can estimate the probability that each individual-outcome pair is a true match, then this knowledge can be used to achieve a reduction in mean-squared error. Such probabilities are outputted by probabilistic record linkage procedures, first developed by Fellegi and Sunter (1969) in the statistics literature, but only recently applied to economics Abramitzky et al. (2018). Hence, any discussion of best practices for using linked data should address how to choose matching algorithms and estimation procedures jointly.

The goal of this paper is therefore to study the effects of different combinations of matching algorithms and estimation procedures for linked data on the quality of the estimates that they produce. First, I will compare how different matching algorithms perform in terms of the representativeness of the matched data they produce and their tolerance for type I and type II errors. Next, with multiple matched versions of the data in hand, I will compute point estimates and confidence intervals for the same parameter of interest using methods that vary by whether they allow for multiple matches, incorporate the matching probabilities, and are likelihood-based in their approach. In total, I will perform the above analysis twice – with simulated data and with real data that the simulated data are generated to imitate.

To the best of my knowledge, how data pre-processing impacts subsequent inference in economics research is not well understood. This paper adds to a recent series of papers by Bailey et al. (2017) and Abramitzky et al. (2018, 2019a), which evaluate the performance of common matching algorithms for historical data in a variety of real and simulated data settings, and offer informal discussions about the impact of matching on subsequent inference. This paper pushes these ideas further, by measuring the *joint* effects of matching and estimation; with a focus on developing estimation techniques that incorporate information from the matching process to improve efficiency and accurately reflect uncertainty.

The matching techniques that will be studied in this paper include deterministic record linkage procedures developed by Ferrie (1996) and Abramitzky et al. (2012); abe; Abramitzky et al. (2019b), and Aizer et al. (2016), and multiple implementations of probabilistic record linkage, specifically the fastLink Enamorado et al. (2019), and machine learning approaches Feigenbaum (2016). Estimation techniques will include Anderson et al. (2019), Lahiri and Larsen (2005), and a fully Bayesian approach that I will develop in this paper.

The data used in this paper consist of the unmerged files from Aizer et al. (2016), which I will pre-process using the practices developed by Abramitzky et al. (2018). The parameter of interest is the average treatment effect of receiving a cash transfer on the long-term outcomes of children in poor families.

Section 2 describes the general problem that this paper seeks to address. Section 3 provides an overview of the matching techniques that I will study. Section 4 describes the estimation techniques. Section 5 will have simulations and results. Section 6 will have real data and results.

2 General problem

Suppose that the researcher would like to estimate θ_0 in a parametric model of the form

$$E[m(\mathbf{z_i}; \theta_0)] = 0 \tag{1}$$

where $m(\cdot)$ is a moment function and $\mathbf{z_i} = (x_i, y_i)$ is the data associated with an individual i sampled at random from the population of interest; however, instead of observing (x_i, y_i) pairs directly, the researcher observes two datasets. The first dataset contains variables x_i and identifiers w_i for individuals $i = 1, \ldots, N_0$ recorded at time 0. The second dataset contains outcomes y_j and identifiers w_j for individuals $j = 1, \ldots, N_T$, observed at time T far

in the future.

To estimate (1) with standard econometric methods, the researcher must identify which of the x_i and y_j refer to the same individuals. That is, she needs to recover the matching function $\varphi : \{1, \ldots, N_0\} \to \{1, \ldots, N_T\}$, where $\varphi(i) = j$ if individual i observed at time 0 and individual j observed at time T refer to the same entity. If w_i and w_j identify individuals uniquely and do not change over time, then $\varphi(i) = j$ if and only if $w_i = w_j$. However, if the identifiers are non-unique or prone to errors introduced by the record-generating process, then φ needs to be estimated, and inference about θ needs to be adjusted accordingly.

In statistics, the task of recovering φ is called *record linkage*. A record linkage procedure consists of a set of decisions about (i) selecting and standardizing the identifiers w_i and w_j, (ii) choosing which records to consider as potential matches (especially when $N_0 \times N_T \times \dim(w_i)$ is large), (iii) defining what patterns of (w_i, w_j) constitute (partial) agreements, and (iv) designating (i, j) pairs as matches. Each step of the record linkage process introduces the possibility that a true match is overlooked (Type II error), or that a false match is designated as correct (Type I error). [Abramitzky et al. (2019a) illustrate a tradeoff between the two...].

To fix ideas, suppose that θ_0 is the long-run impact of cash transfers on the longevity of children raised in poor families. The vector x_i includes family and child characteristics as observed in welfare program applications; and the outcomes y_j are constructed by calculating (day of death − day of birth) for all of the observations in a set of death records. For all i and j, the identifiers w_i and w_j include the individual's first and last name, and date of birth. Additionally, w_i includes i's place of birth; w_j includes j's place of death; and some w_i and w_j contain the individual's middle name or middle initial.

In this setting, an example of a (deterministic) record linkage procedure consists of:

(i) using a phonetic algorithm to standardize all string variables;

(ii) considering as potential matches only (i, j) pairs whose phonetically standardized

names begin with the same letter, and whose birth years are within ± 2 years;

(iii) measuring agreements among names using Jaro-Winkler string distances, and weighing disagreements in birth year more than differences in birth month (and more than differences in birth day),

(iv) designating as matches all (i, j) pairs with scores calculated using the metrics in (iii) exceeding a pre-specified cutoff; and, if a record i has multiple possible matches j that exceed the cut-off, then choosing the match with the highest score (or picking a random match if there is a tie).

Another record linkage procedure could be defined using the same rules for steps (i)-(iii), but replace (iv) with a probabilistic matching rule that does not enforce one-to-one matching:

(iv*) use the Expectation Maximization algorithm to estimate the probability that each record pair is a match, and then designate matches using the Fellegi and Sunter (1969) rule for pre-specified tolerances for Type I and Type II errors

Except in rare cases, the matchings produced by the preceding record linkage procedures will be distinct. Whereas the first procedure associates each x_i with at most one matched y_j, the second procedure may associate the same x_i with multiple y_j (or zero!). In the former case, estimation proceeds as a standard GMM problem; in the latter, estimating θ requires methods that allow φ to take on multiple values (Anderson et al., 2019). This illustration shows that not only are the point estimates of θ likely to depend on the estimates of φ, but also the methods for estimating θ are likely to be different.

As observed in Bailey et al. (2017), record linkage procedures differ by the set of assumptions that motivate their use. However, all of the procedures discussed in this paper will be studied under the following, common set of assumptions:

1. (De-duplication) Within a given dataset, each observation refers to a distinct entity.

That is, if two observations share the same identifier, they represent two different individuals.

2. (No unobserved sample selection) The observed x_i and y_j are random samples conditional on w_i and w_j, respectively. This means that all individuals with the same identifying information have equal probability of appearing in the sample.

3. There exists a unique $\theta_0 \in \Theta$ that satisfies the relationship in (1), that can be consistently estimated using standard econometric techniques if φ_0 is known.

The next section discusses in detail the exact record linkage techniques that will be studied, and their motivating assumptions.

References

A. Aizer, S. Eli, J. Ferrie, and A. Lleras-Muney, "The long-run impact of cash transfers to poor families," *American Economic Review*, vol. 106, no. 4, pp. 935–71, April 2016. [Online]. Available: http://www.aeaweb.org/articles?id=10.1257/aer.20140529

R. Anderson, B. Honore, and A. Lleras-Muney, "Estimation and inference using imperfectly matched data," *Working paper*, August 2019. [Online]. Available: http://www.github.com/rachelsanderson/ImperfectMatching

I. P. Fellegi and A. B. Sunter, "A theory for record linkage," *Journal of the American Statistical Association*, vol. 64, pp. 1183–1210, 1969.

R. Abramitzky, R. Mill, and S. Perez, "Linking individuals across historical sources: a fully automated approach," National Bureau of Economic Research, Working Paper 24324, February 2018. [Online]. Available: http://www.nber.org/papers/w24324

M. Bailey, C. Cole, M. Henderson, and C. Massey, "How well do automated linking methods perform? lessons from u.s. historical data," National Bureau

of Economic Research, Working Paper 24019, November 2017. [Online]. Available: http://www.nber.org/papers/w24019

R. Abramitzky, L. Boustan, K. Eriksson, J. Feigenbaum, and S. Perez, "Automated linking of historical data," *NBER Working Paper*, 2019.

J. P. Ferrie, "A new sample of males linked from the public use microdata sample of the 1850 u.s. federal census of population to the 1860 u.s. federal census manuscript schedules," *Historical Methods*, vol. 29, no. 4, pp. 141–156, 1 1996.

R. Abramitzky, L. P. Boustan, and K. Eriksson, "Europe's tired, poor, huddled masses: Self-selection and economic outcomes in the age of mass migration," *American Economic Review*, vol. 102, no. 5, pp. 1832–56, May 2012. [Online]. Available: http://www.aeaweb.org/articles?id=10.1257/aer.102.5.1832

R. Abramitzky, L. P. Boustan, and K. Eriksson, "To the new world and back again: Return migrants in the age of mass migration," *ILR Review*, vol. 72, no. 2, pp. 300–322, 2019. [Online]. Available: https://doi.org/10.1177/0019793917726981

T. Enamorado, B. Fifield, and K. Imai, "Using a probabilistic model to assist merging of large-scale administrative records," *American Political Science Review*, vol. 113, no. 2, p. 353?371, 2019.

J. J. Feigenbaum, "A machine learning approach to census record linking ?" 2016.

P. Lahiri and M. D. Larsen, "Regression analysis with linked data," *Journal of the American Statistical Association*, vol. 100, no. 469, pp. 222–230, 2005. [Online]. Available: http://www.jstor.org/stable/27590532