The effects of matching algorithms and estimation methods using linked data

Rachel Anderson*

This Version: August 27, 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, **?** 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 **?** 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 ? 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 ? in the statistics literature, but only recently applied to economics ?. 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 ? and ??, which compare how common matching algorithms for historical data produce different datasets, and offer informal discussion of the impact on inferences. This paper goes a step further, by testing also the effect of different estimation techniques that incorporate information from the matching process, and uses simulations to make more generalizable conclusions.

Matching techniques will include deterministic record linkage procedures developed by ? and ???, and ?, and multiple implementations of probabilistic record linkage, specifically

the fastLink **?**, and machine learning approaches **?**. Estimation techniques will include **?**, **?**, and a fully Bayesian approach that I will develop in this paper.

The real data consists of the unmerged files from **?**, which I will pre-process using the practices developed by **?**. The parameter of interest is the average treatment effect of a conditional cash transfer program on recipients' children's longevity.

By Friday, I will write a description of the model, the parameter of interest, and the set of assumptions that I will use. I will also provide a list of the matching and estimation techniques (with descriptions) that I will study, as well as a timeline for implementing each of them.

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.

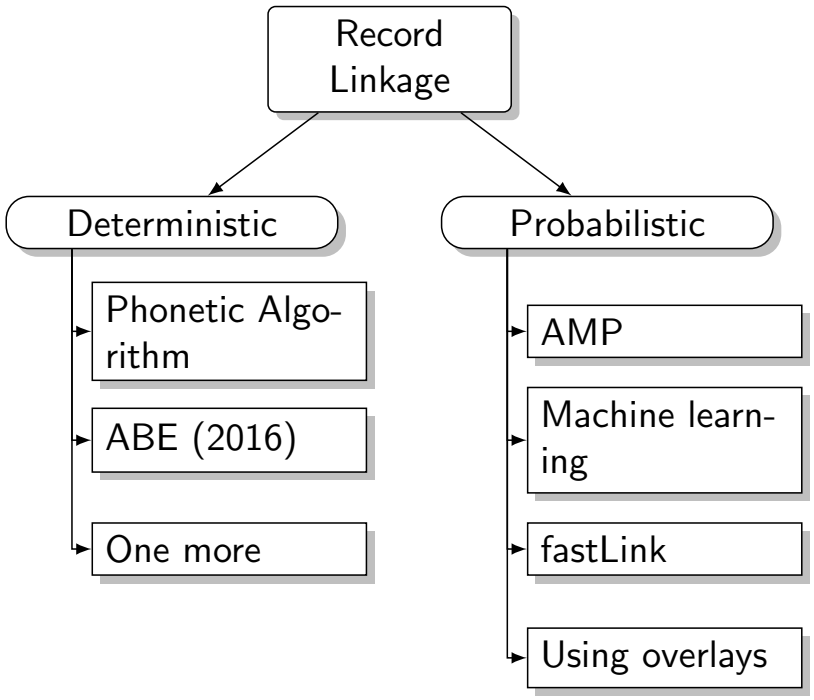
P. Lahiri and M. D. Larsen, "Regression analysis with linked data," *Journal of the American Statistical Association*, vol. 100, no. 469, pp. 222–230, 2005. [Online]. Available: http://www.jstor.org/stable/27590532

2 Matching Methods

A matching procedure is a set of choices about (i) selecting which variables to use when matching, (ii) defining a "distance" metric between said variables, (iii) blocking observations into non-overlapping groups for computational feasibility, and (iv) designating record pairs as matches if a one-to-one matching is desired.

Matching procedures can be divided into two categories, (i) deterministic approaches, where a fixed set of rules determine which records are matching and which are not; and (ii) probabilistic methods, which involve estimating the probability that each record pair refers to a match. COMPARE AND CONTRAST WEAKNESSES. Deterministic approaches are susceptible to X YZ see Enamorado, etc.

Figure 1: Overview of matching methods



INSERT A GRAPHIC WITH THE METHODS I WILL TEST

TABLE WITH SURVEY OF MATCHING METHODS used in literature. categoreis

are like enforces one to one matching, blocking, string comparator, phonetic algorithm (this is lit review essentially/annotated bibliography)

- Deterministic: Rely heavily on phonetic algorithms (i.e. SOUNDEX, NYSIIS, Metaphone, Spanish Metaphone) and string distance measures (SPEDIS, Jaro-Winkler). **?** discuss differences in these choices and call SOUNDEX outdated.

    1. **?**: Described in appendix, match uses first name, middle initial, last name, day, month and years of birth. Match allows for errors in strings (using SOUNDEX and SPEDIS) and in single digits for DOB.

    2. Abramtizky, Boustan and Eriksson (ABE) algorithms – discard observations that do not have unique matches; they typically perform multiple variations of the same algorithm to check robustness. Perform matching in both directions, and then take the intersection of the two matched samples. Implemented with abematch stata command.

    3. ABE-JW adjustment algorithm – block by place of birth, first letter of first and last names match. Jaro-winkler can be implemented with stata jarowinkler command; R package stirngdist. They also only accept one-to-one matching.

       Limitation of ABE algorithms is that "it is not clear how to appropriately weight differences in name spelling versus differences in age when comparing two records".

- Probabilistic (see Winkler 2006 for survey)

    – E-M Algorithm **?** still blocking by same place of birth, predicted year of birth, and first letter of first and last name. Still use decision rule to enforce one-to-one matching.

    – Training sample (Ruggles and Feigenbaum) – trained to NOT allow mutliple matches; generates a predicted probability of being a match for each pair of

records in A and B

– IPUMS linking method: trains support vector machine on training sample of manually classified records (like Feigenbaum 2016) In historical applications this is problematic due to sample attrition. The DGP changes, so a full likelihood is a good idea.

- Overview of matching methods

Important measurements: estimated type 1, type 2 errors; representativeness of sample, sample size, overlapping of samples - Comparison of matching methods from (a) theoretical perspective, (b) with simulated data, (c) with actual data

1. Estimation Methods

   - Anderson, Honore, Lleras-Muney (2019)
   - Lahiri Larsen
   - Scheuren Winkler

   - Overview of estimation methods

   - Comparison of estimation methods from (a) theoretical perspective, (b) with simulated data, (c) with actual data

   (3) Further investigation/follow-up simulations inspired by steps 1 and 2

Data problems in practice **?**

- "time-invariant" features not time-invariant – names misspelled, individual reported incorrectly, or individual changed name
- Age heaping – rounding ages to the nearest multiple of five
- digitization of handwritten manuscripts

String comparators: NYSIIS, Soundex, Jaro-Winkler, Levenstein, etc.

I will also allow for missing data.

3 Annotated bibliography

- Neter, Maynes, and Ramanathan (1965): small mismatch errors in finite population sampling can lead to a substantial bias in estimating the relationship between response errors and true values
- Scheuren and Winkler (1993): propose method for adjusting for bias of mismatch error in OLS
- SW (1997, 1991): iterative procedure that modifies regression and matching results for apparent outliers
- Lahiri and Larsen (2005): provides unbiased estimator directly instead of bias correction for OLS, by applying regression to transformed model
- Abramitzky, Mill, Pérez (2019): guide for researchers in the choice of which variables to use for linking, how to estimate probabilities, and then choose which records to use in the analysis. Created R code and stata command to implement the method
- Ferrie 1996, Abramitzky, BOustan and Eriksson (2012 2014 2017) are deterministic. Conservative methods require no other potential match with same name within a 5-year band
- Semi-automated Feigenbaum, Ruggles et al
- Abramitzky, Boustan, Eriksson, Feigenbaum, Pèrez (2019): evaluate different automated methods for record linkage, specifically deterministic (like Ferrie and ABE papers), machine learning Feigenbaum approach, and the AMP approach with the EM

algorithm. Document a frontier between type I and type II errors; cost of low false positive rates comes at cost of designating relatively fewer (true) matches. Humans typically match more at a cost of more false positives. They study how different linking methods affect inference – sensitivity of regression estimates to the choice of linking algorithm. They find that the parameter estimates are stable across linking methods. Find effect of matching algorithm on inference is small.

- Bailey et al. (2017) review literature on historical record linkage in US and examines performance of automated record linkage algorithms with two high-quality historical datasets and one synthetic ground truth. They conclude that no method consistently produces representative samples; machine linking has high number of false links and may introduce bias into analyses
- Survey paper from handbook of econometrics
- For example, Goeken et al. (2017) document that in two enumerations of St. Louis in the 1880 Census, nearly 46 percent of first names are not exact matches, and the Early Indicators project notes that 11.5 percent of individuals in the Oldest Old sample have a shorter first name in pension records than in the original Civil War enlistment records (Costa et al. 2017).

Overall, high variability in performance of matching methods depending on choice of variables, string comparators used.

4 Empirical Application

4.1 Data

What are the long-run effects of cash transfers to poor families? Specifically, are there lifelong benefits for children raised in poor families? Transfers may not help poor children if the amounts are insufficent, parents mis-allocate funds, or the transfers induce behavioral changes that are detrimental to the child.

Evaluating whether cash transfers improve outcomes necessitates identifying a plausible counterfactual. **?** collect administrative records from the MOthers' Pension program (1911-1935), which was the first US government-sponsored welfare program for poor mothers with dependent children. The intent of the MP program was to imrpove the conditions of "young children that have become dependent through the loss or disability of the breadwinner" (Abbott 1993, p 1). The transfers generally represented 12-25 percent of family income, and typically lasted for three years.

The authors measure the impact in terms of longetivity and other outcomes of the children whose mothers applied to the program. These data include information on thousands of accepted and rejected applciants born between 1900 and 1925, most of whom had died by 2012. The idetnifying information in the application records allows them to link children with other datasets to trace their lifetime outcomes.

Identification comes from comparing children of mothers who applied for transfers and who were initially deemed eligible, but were denied upon further investigation. This is a standard strategy in program evaluation that has been used successfully in studies of disability insurance. The validity of this strategy hinges on the assumpotion that accepted and rejected mothers and their children do not differ on unobservable characteristics. Rejected mothers were on average slightly better-off, based on observable characteristics at the time

of the applciation. Authors say that the outcomes for boys of rjeected mothers provide a best-case scenario (upper bound) for what could be expected of beneficiaries in the absence of transfers.

Data collected on over 16,000 boys from 11 states who were born between 1900 and 1925, and whose mothers applied to the MP program; find that transfers increased longevity by about 1 year. Poorest families in the sample had longevity increased by 1.5 years of life. Results are robust to alternative function form specifications, counterfactual comparisons, and treatment of attrition. They interpret their results as the effect of cash transfers alone.

The authors match also a subset of the records to WWII enlistment and 1940 census records; cash transfers reduced the probability fo being underweight by half, increased educational attainment by 0.34 years, and increased income in early adulthood by 14 percent. They say that these mechanisms are responsible for 75 percent of the increase in longevitiy

4.2 Empirical Model

Basic empirical model

$$\log(\text{age at death})_{ifts} = \theta_0 + \theta_1 MP_f + \theta_2 X_{if} + \theta_3 Z_{st} + \theta_c + \theta_t + \epsilon_{if} \quad (1)$$

Model to address attrition and mutliple matches

$$P(\text{survived to age } a = 1)_{iftcs} = f(\theta_0 + \theta_1 MP_f + \theta_2 X_{if} + \theta_3 Z_{st} + \theta_c + \theta_t + \epsilon_{if}) \quad (2)$$

(i) above using unique matches – baseline , (ii) above using multiple matches – ahl

baseline, (iii) above in bayesian framework – bayesian baseline, (iv) above using probabilities of matches (freq + bayesian versions)

Replicate Table 4 with different matching; different estimation techniques

Do a basic Bayesian log normal model for different states – pooled v. separate multiple matching