The effects of matching algorithms and estimation methods using linked data

Rachel Anderson*

This Version: August 26, 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 Abramitzky et al. (2018) and Abramitzky et al. (2019), who seek to understand how different matching algorithms impact the quality of inference. 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 as described in **?**, and Abramitzky et al. (2018), 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 real data consists of the unmerged files from Aizer et al. (2016), which I will preprocess using the practices developed by Abramitzky et al. (2018). 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

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

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.

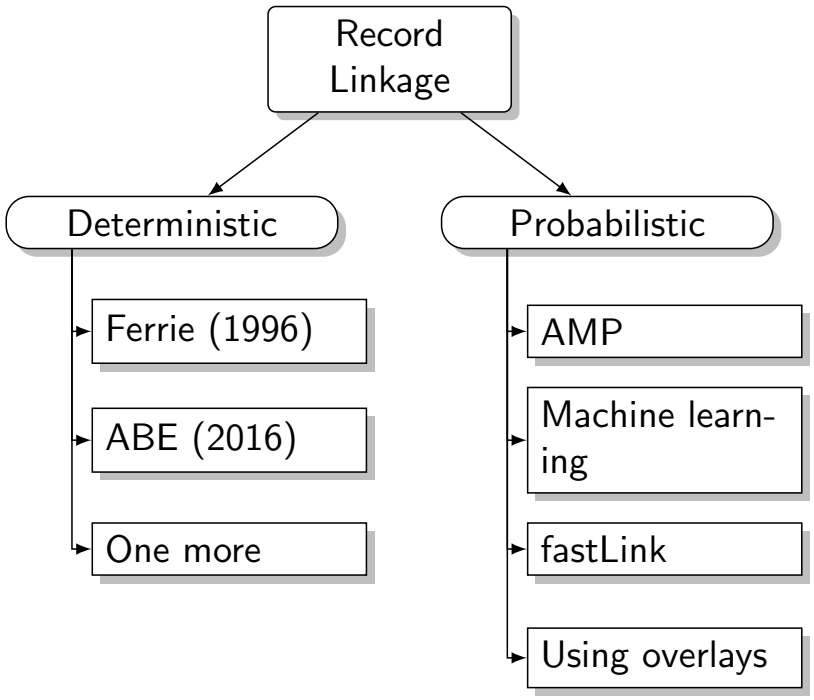
P. Lahiri and M. D. Larsen, "Regression analysis with linked data," *Journal of the American Statistical Association*, vol. 100, no. 469, pp. 222–230, 2005. [Online]. Available: http://www.jstor.org/stable/27590532

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

- Deterministic

- Probabilistic (see Winkler 2006 for survey)
  - E-M Algorithm
  - Training sample (Ruggles and Feigenbaum)
  - 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

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) say automated methods perform quite poorly
- Survey paper from handbook of econometrics

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