The effects of matching algorithms and estimation methods using linked data

Rachel Anderson[*]

This Version: August 22, 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.

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 by matching individuals who have the same name and date of birth. Since name and date combinations are not necessarily unique and can be prone to typographical errors, the authors match some individuals to multiple death records, all of which seem equally likely to be the true match. Estimation proceeds using techniques from AHL (2019), which allow for observations to have multiple linked outcomes.

The methods in AHL (2019) consider the matched dataset as given, however the authors hypothesize that there could be efficiency gains if additional information about match quality is available. Specifically, if the probability that a record pair refers to the same individual

[*] 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).

is known, then this knowledge can be used to achieve a reduction in mean-squared error. Conveniently, these probabilities are estimable using probabilistic record linkage procedures developed by [**?**] and [**?**].

Thus, the goal of this paper is to study the effects of different matching algorithms and estimation procedures for linked data on the quality of the estimates that they produce. First, I will examine different matching algorithms to determine whether they produce different configurations of matched data. With multiple matched versions of the data in hand, I will then use various estimation methods to estimate the same parameter of interest – the average treatment effect of a conditional cash transfer program on recipients' children's longevity. I will compare these methods first theoretically, then with simulated data, and, finally, with the original, unmatched data from [].

Matching techniques include deterministic record linkage as described in [], and multiple implementations of probabilistic record linkage – the fastLink, machine learning, etc. Estimation techniques include AHL (2019), Lahiri and Larsen (2005), and a fully Bayesian approach, that is described in this paper.

Currently, little is known about how data pre-processing impacts subsequent inference in the economics literature, and especially those projects that rely on matching historical datasets with imperfect identifiers. This paper adds to a recent series of papers by Abramitzky, Boustan, etc. in its effort to understand how these decisions impact the quality of inference.

The general outline will be as follows:

1 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.

- 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.

2 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
- Semi-automated Feigenbaum, Ruggles et al