Annotated bibliography

Rachel Anderson[*]

This Version: September 3, 2019

Historical Matching Overview

Primary variable for matching is an individual's name, which are rarely unique within a given population. Low literacy rates mean names are often misspelled. Misspelling also arises due to geographical relocation and regional variation in names. "For example, an illiterate individual from Louisiana with the surname of Thibideaux, who chooses to move to another state, would likely have his name spelled phonetically as Tibido." (Nix and Qian 2015). Researchers use phonetic algorithms to account for these differences

(Meta) MatchingPapers

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 rep-

[*] 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).

resentative samples; machine linking has high number of false links and may introduce bias into analyses.

Matching Methods

Bailey et al. (2017) categorize historical linking algorithms (that match observations using name and age only) according to how they treat candidate pairs in the following four categories:

- M1: A perfect, unique match in terms of name and age similarity
- M2: A single, similar match that is slightly different in terms of age, name, or both
- M3: Many perfect matches, leading to problems with match disambiguation
- M4: Multiple similar matches that are slightly different in terms of age, name or both

Historical linking algorithms generally treat M1 cases as matches, but differ in how they treat M2, M3, and M4 candidate pairs. To account for differences in M2, **??** and **?** search within fixed-year bands. Multiple matches are either (i) ignored, (ii) picked at random, or (iii) equally weighted; or used a weighted combination.

Bleakley and Ferrie (2016) Nix and Qian (2015) AHL AMP (2018)

- 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.

- Treatment of equally likely – equal probability weighting of tied candidates (Bleakley and Ferrie 2016); weighted combo of linking features to ehlp disambiguate potential matches. Ferrie 96; Old ABE, new ABE, and Feigenbaum.

- 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).

Estimation Papers

Important Applications

Nix and Qian (2015) study racial passing by linking individual U.S. census records to determine whether an individual's recorded race changed from one census to the next. To achieve higher match rates than those of previous studies[1], the authors develop methods for including individuals with multiple potential matches. These methods include selecting one match at random, and selecting the match that produces an upper/lower bound for estimating the "passing" rate.

Nix and Qian (2015) also use the unmatched individuals from their data to calculate absolute bounds for the population passing rates. For a given algorithm, the absolute upper bound is obtained by using the "upper bound" configuration of data, combined with assuming that all unmatched individuals passed. The lower bound is obtained in the same way, assuming that none of the excluded individuals passed.

Nix and Qian (2015) argue that increasing the match rate improves the bounds around any true population statistic, even though their methods introduce random measurement error in the estimand.

[1] The authors match 61-67 percent of individuals. ABE (2012), Hornbeck and Naidu (2014), Long and Ferrie (2013), Mill and Stein (2012) have match rates around 30, 24, 22, 11-34 percent respectively

Simulation Idea

Could use only simulation data, with variety of possible biases, motivated by the applications above. For example, motivated by N-Q, introduce error with geographical relocation. Then test which techniques are robust to these types of sample selection/error.

References

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

E. Nix and N. Qian, "The fluidity of race: ?passing? in the united states, 1880-1940," National Bureau of Economic Research, Working Paper 20828, January 2015. [Online]. Available: http://www.nber.org/papers/w20828