A unified approach to inference with linked data

Rachel Anderson*

This Version: September 26, 2019

Abstract

This paper studies the joint impact of matching algorithms and estimation methods for linked data on the quality of the analyses that they produce. In particular, I compare the bias and precision of the parameter estimates generated by combining different record linkage and estimation techniques in a variety of econometric models. I also compare how these methods perform when errors in the matching variables are correlated with the variables of interest, and the true match is not included among the potential matches.

1 Introduction

This paper concerns the case where the goal is to estimate a parametric model $y|x \sim f_\theta$ using observations (y_i, x_i), but y_i and x_i are contained in different datasets that lack unique identifiers. This scenario is interesting because, prior to estimating any model, the researcher must attempt to identify which observation pairs refer to the same individual. In the best case, this matching step introduces random measurement error (that is not usually reflected in standard errors); but, in the worst case, if auxiliary matching variables are correlated

*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).

with unobservables that affect the variables of interest, this process may introduce sample selection.

How to perform this match – representativeness, etc – has been a focus of recent work by X and Y, but usually the matching step is merely a means to an end.

Linking multiple data sources without unique IDs is common practice for projects that use historical U.S. data sources prior to the introduction of Social Security Numbers. For example, Aizer et al. (2016) link children listed on Mothers' Pension program welfare applications from 1911-1935 with Social Security Death Master File records from 1965-2012 using individuals' names and dates of birth. Although the authors are able to match 48 percent of children to a unique death record, 4 percent match to multiple possible records, and 48 percent remain unmatched.[1] To avoid dropping the 52 percent of observations with zero or multiple matches, Aizer et al. (2016) estimate hazard models using methods from Anderson et al. (2019) that allow observations to be associated with multiple, equally likely, outcomes.

The methods used by Aizer et al. (2016) illustrate how inference using linked data requires joint assumptions for the matching and estimation steps. Under different assumptions, the authors could have generated a "composite match" equal to the average of the linked observations (Bleakley and Ferrie, 2016), or constructed bounds on the parameter of interest using different configurations of matched data (Nix and Qian, 2015). This example also shows how the outputs of the matching process determine which estimation tools are available. Had the authors used probabilistic record linkage methods to link the data, they could have used the robust OLS estimators from Lahiri and Larsen (2005), or prior-informed imputation for missing records proposed by Goldstein et al. (2012).

Although there are a number of recent papers that separately compare the performance of different matching algorithms (Bailey et al., 2017; Abramitzky et al., 2018) and estimation

[1]The authors estimate that at least 32 percent of individuals in the Mothers' Pension program data died before 1965, and therefore should have no match in the 1965-2012 data.

methods for linked data (Harron et al., 2014), little is known about the *joint* impact of matching and estimation on the quality of inference with linked data. This paper fills this gap by comparing how different *combinations* of matching and estimation techniques affect parameter estimates and their confidence intervals in standard econometric models, with the hope of helping researchers choose which methods best suits their projects' needs.

In order to illustrate the techniques studied in this paper, I introduce a running example based on synthetic datasets that imitate historical U.S. Census data, yet offer the benefit that each observation's true match is known.

Example 1. The "ground truth" dataset consists of 1000 observations of $(x_{1i}, x_{2i}, y_i, w_i)$, where x_{1i} and x_{2i} are mutually independent, $x_{1i} \overset{i.i.d}{\sim}$ Bernoulli(0.5), and $x_{2i} \overset{i.i.d}{\sim} \mathcal{N}(0, 2)$. The value y_i generated by the relationship,

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i, \quad \varepsilon_i \mid x_{1i}, x_{2i} \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2) \tag{1}$$

with $(\beta_0, \beta_1, \beta_2, \sigma^2) = (2, 0.5, 1, 2)$. Given these parameter values, estimating a correctly specified linear regression model yields an R^2 value of approximately 0.50 (see Figure 1(a)). Each observation is associated with a vector of identifying variables, w_i, that consists of a first and last name drawn randomly from a list of first and last names[2], and a random birthday between January 1, 1900, and December 31, 1925, so that the full synthetic dataset resembles Table 1. The number of possible names is smaller than the number of observations to ensure that there are multiple observations with the same name.

Next, I split the "ground truth" dataset into an x- and y-datafile, which contain (x_{1i}, x_{2i}, w_i) and (y_j, w_j) values respectively. The x-datafile contains values for 400 observations, selected at random from the ground truth dataset; the y-datafile includes all 1,000 observations, with w_j identical to the ground truth data. For each w_i in the x-datafile, I

[2]The first and last name lists contain 41 and 24 names, respectively, and can be found in the replication files.

Figure 1: Creation of Synthetic Datasets

ID	y	x_1	x_2	First Name	Last Name	Birthday
1	y_1	$x_{1,1}$	$x_{2,1}$	Tyler	Ashenfelter	1915-05-13
2	y_2	$x_{1,2}$	$x_{2,2}$	Brandon	Christensen	1904-06-27
				⋮		
195	y_{195}	$x_{1,195}$	$x_{2,195}$	Samantha	Andersen	1914-08-18
196	y_{196}	$x_{1,196}$	$x_{2,196}$	Victoria	Andersen	1918-11-25
				⋮		
1000	y_{500}	$x_{1,500}$	$x_{2,500}$	Vicky	Anderson	1915-04-14

x-Datafile

ID	x	Name	Birthday
2	$(x_{1,2}, x_{2,2})$	Branden Christenson	1905-06-27
		...	
195	$(x_{1,195}, x_{2,195})$	Samantha Anderson	1914-08-21
198	$(x_{1,198}, x_{2,198})$	Jon Smyth	1918-12-20
		...	
1000	$(x_{1,1000}, x_{2,1000})$	Vic Andersn	1915-04-14

y-Datafile

ID	y	Name	Birthday
1	y_1	Tyler Ashenfelter	1915-05-13
2	y_2	Brandon Christensen	1904-06-27
		...	
195	$y_{1,195}$	Samantha Anderson	1914-08-18
		...	
1000	y_{1000}	Vicky Anderson	1915-04-14

randomly add noise to the original names, by deleting characters (e.g., "Anderson" becomes "Andersn"), exchanging vowels (e.g., "Rachel" becomes "Rachal"), or swapping English phonetic equivalents (e.g. "Ellie' becomes "Elie") according to pre-specified probabilities. I also add normally distributed errors to the birth day, month, and year, according to pre-specified probabilities. These files are represented in Figure 2.

To be clear, the unit of analysis in this study is a matching/estimation/econometric model combination. First, I implement several of the most commonly used deterministic and probabilistic record linkage procedures to generate multiple matched versions of the data. Then, for each matched version of the data, I estimate different types of econometric models, using different estimation methods, and compare the parameter estimates and confidence intervals that these combinations produce. One iteration of this process is represented in Figure 2, and Table 1 contains a comprehensive list of the methods that are tested.

Figure 2: Research design

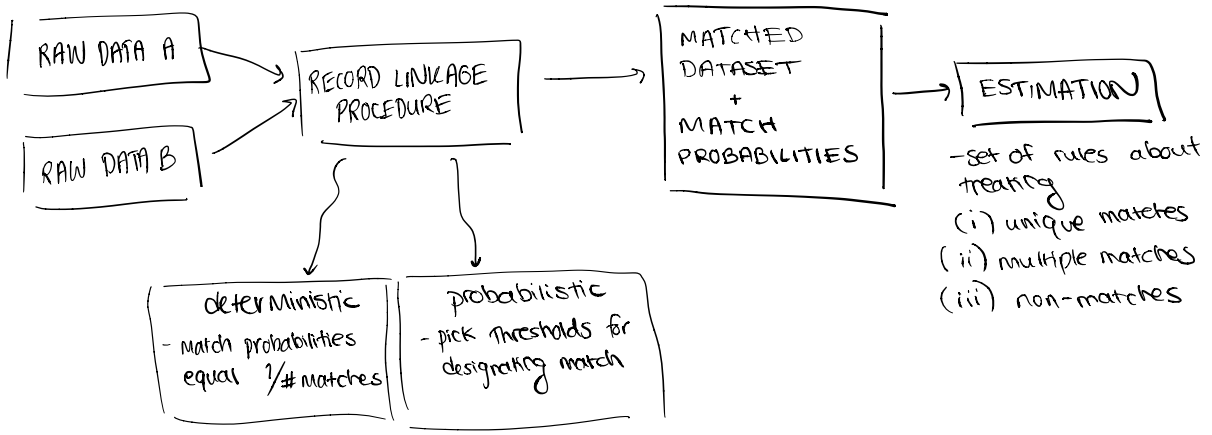

Table 1: List of methods to be implemented

Matching Methods		Estimation Methods	
Deterministic	Probabilistic	Uses match weights	No match weights
Abramitzky et al. (2012)	**?**	Lahiri and Larsen (2005)	Anderson et al. (2019)
		Goldstein et al. (2012)	Nix and Qian (2015)
		Anderson (2019)	Bleakley and Ferrie (2016)

At first glance, it's not obvious that the methods studied in this paper should be directly comparable, or whether certain combinations of matching and estimation techniques are possible to implement at the same time; however, I will show that, with a few adjustments, both of these tasks are possible.[3] Specifically, I will show that it is possible to compare the performance of deterministic and probabilistic matching methods by setting thresholds for each method that make the Type I and Type II error rates equal.

Importantly, the estimation techniques studied in this paper differ according to whether they require estimating the probability that a record pair refers to a match. When implementing methods that require this extra information in combination with deterministic matching techniques that *do not* output match probabilities, I will treat the matches as

[3]maybe even a theoretical result showing equivalence across classes of matching methods and estimators?

being equally likely. Similarly, if an estimation procedure does not use match probabilities, but a probabilistic matching algorithm outputs them, then the probabilities will be ignored.

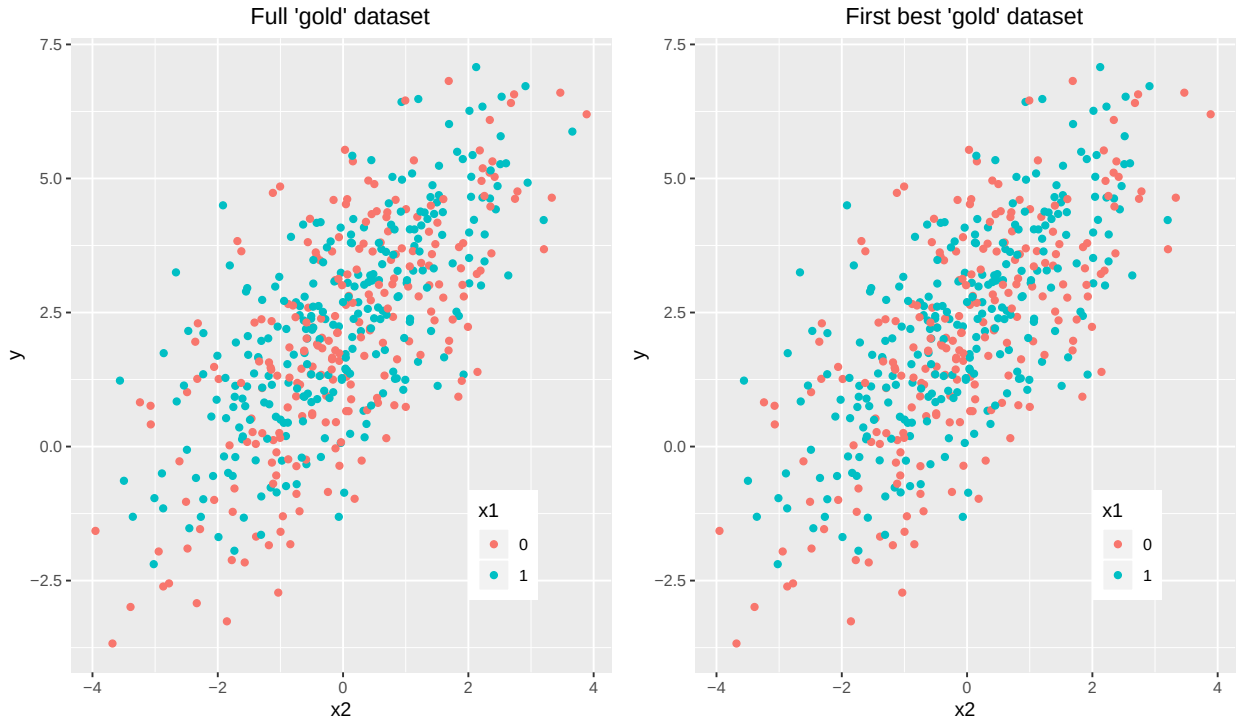

Figure 3: default

Section 2 introduces the general problem that this paper seeks to address and outlines a common framework for comparing the techniques in Table 1. Sections 3 and 4 describe each of the methods in detail. Section 5 describes the implementation of the methods and the data. Section 6 contains the results. Section 7 will be a real empirical application, and Section 8 will conclude.

2 General problem

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

$$E[m(\mathbf{z_i}; \theta_0)] = 0 \tag{2}$$

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 D_1 contains variables x_i and identifiers w_i for individuals $i = 1, \ldots, N_1$. The second dataset D_2 contains outcomes y_j and identifiers w_j for individuals $j = 1, \ldots, N_2$.

Example 1 (continued). The w_i includes first name, last name, and birth date (year, month, day). From the gold dataset, I generate the x datafile by selecting 350 observations at random and dropping their y values. Then I add noise to the names using pre-specified probabilities for certain types of transcription errors, and I add randomly distributed errors to the birthdates. The y dataset contains the original names and birth days, recorded without error, for all 500 initial observations.

To estimate (2) 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 $\varphi : \{1, \ldots, N_1\} \to \{1, \ldots, N_2\} \cup \varnothing$, where $\varphi(i) = j$ if individual i in dataset D_1 and individual j in dataset D_2 refer to the same entity; and $\varphi(i) = \varnothing$ if i does not have a match in D_2. If w_i and w_j identify individuals uniquely and do not change over time, then $\varphi(i) = j$ if and only if $w_i = w_j$; otherwise, $\varphi(i) = \varnothing$. 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 $N_1 \times N_2 \times \dim(w_i)$ is large), (iii) defining what patterns of (w_i, w_j) constitute (partial) agreements, and (iv) designating (i, j) pairs as matches.[4] Each step of the record linkage process introduces the

[4]Note that this is the author's own definition. By contrast, 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

possibility that a true match is overlooked (Type II error), or that a false match is designated as correct (Type I error), and there is generally a tradeoff between reducing either one of the two (Abramitzky et al., 2019; Doidge and Harron, 2018).

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

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.

matching:

(iv*) use the Expectation-Maximization algorithm to compute "match weights" for each record pair; then designate as matches all pairs that exceed thresholds that are set to reflect specific tolerances for Type I and Type II error.

Except in rare cases, the matching function outputted by replacing (iv) with (iv*) will be different. 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 (in technical terms, this implies φ is a correspondence). The former case might use a standard GMM model to estimate θ; while the latter requires methods that associates multiple values of y_j with each x_i (Anderson et al., 2019). This example shows that not only do the estimates of θ likely depend on the estimates of φ, but also the *methods* for estimating θ may be also differ.

As observed by 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 (with some departures later on):

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 (2), 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.

3 Record Linkage Methods

Research on record linkage appears in many fields, such as statistics, computer science, operations research, and epidemiology, under many names, such as data linkage, entity resolution, instance identication, de-duplication, merge/purge processing, and reconciliation. As such, there are several books devoted to its study (Harron et al., 2015; Christen, 2012; Herzog et al., 2007), and dozens of commercial and open source systems software developed for its implementation (Köpcke and Rahm, 2010). Although insights from other fields have been slow to reach economics, recent working papers examine the impact of different record linkage techniques on inference using historic U.S. census data (Abramitzky et al., 2019; Bailey et al., 2017); and the most recent version of the Handbook of Econometrics includes a chapter on the "Econometrics of Data Combination" (Ridder and Moffitt, 2007).[5]

Record linkage techniques are broadly categorized as deterministic or probabilistic[6]; however, every deterministic linkage method has an equivalent probabilistic version (Doidge and Harron, 2018). Furthermore, the first three steps of the record linkage task described above are identical for all procedures. I begin by discussing these steps, then introduce the PRL framework, and show how it can be used to express deterministic linking rules.

4 Estimation Methods

See the table in the appendix for overview of methods and their assumptions.

[5]Similar survey papers also exist in fields outside of economics, such as epidemiology and computer science (Doidge and Harron, 2018; Winkler, 1999). In fact, that record linkage is studied by many fields makes writing (and reading!) such surveys difficult, because authors are constantly writing the same things For example, Goldstein et al. (2012) prove similar results to those published in Hirukawa and Prokhorov (2018).

[6]define them here

5 Timeline for completion

- set meetings for September
- Oct 1 – first draft done
- Early Oct (discuss this) – present in student seminar
- Oct 28: final draft done
- Nov 4: Submit paper!

Please stop reading here!

5.1 Steps of the record linkage task

Here I will write a little bit more about each of the steps introduced in Section 2.

(i) selecting and standardizing the identifiers w_i and w_j

phonetic algorithms

(ii) choosing which records to consider as potential matches

blocking

(iii) defining what patterns of (w_i, w_j) constitute (partial) agreements

Jaro-winkler distances, for example

5.1.1 (iv) designating (i, j) pairs as matches

this is where the differences between deterministic/probabilistic really arise!

5.2 Probabilistic Record Linkage

In describing the record linkage techniques implemented in this paper, I use notation from Fellegi and Sunter (1969). As before, consider two datafiles D_1 and D_2 that record information from two overlapping sets of individuals or entities. The files originate from two record-generating processes that may induce errors and missing values, so that identifying which individuals are represented in both D_1 and D_2 is nontrivial. I assume that individuals

appear at most once in each datafile, so that the goal of record linkage is to identify which records in files D_1 and D_2 refer to the same entities.

Suppose that files D_1 and D_2 contain N_1 and N_2 records, respectively, and without loss of generality that $N_2 \geq N_1$. Denote also the number of entities represented in both files as N_M, so that $N_1 \geq N_M \geq 0$.

We say that the set of ordered record pairs $D_1 \times D_2$ is the union of two disjoint sets, *matches* (M) and *non-matches* (U):

$$M = \{(i, j) : i \in D_1, j \in D_2, i = j\}$$
$$U = \{(i, j) : i \in D_1, j \in D_2, i \neq j\}$$

Hence, the formal goal of record linkage is to identify whether an arbitrary record pair $(i, j) \in D_1 \times D_2$ belongs to M or U. Note that this task is identical to

To perform this task, each record pair is evaluated according to L different comparison criteria, which are the result of comparing data fields for records i and j. For example, if a record pair (i, j) represents two individuals, the pair may be evaluated according to whether they share a first name or have the same birthday. These comparisons are represented by a *comparison vector*,

$$\boldsymbol{\gamma}_{\mathbf{ij}} = (\gamma_{ij}^1, \ldots, \gamma_{ij}^\ell, \ldots, \gamma_{ij}^L)$$

where each comparison field γ_{ij}^ℓ may be binary-valued, as in "i and j have the same birthday," or use levels to account for partial agreement between strings (see **?**, for details). The models presented herein use only binary comparison vectors, however they may be extended to allow for partial agreement using the methods from **?**.

The probability of observing a particular configuration of $\boldsymbol{\gamma}_{\mathbf{ij}}$ can be modeled as arising

from the mixture distribution:

$$P(\boldsymbol{\gamma}_{\mathbf{ij}}) = P(\boldsymbol{\gamma}_{\mathbf{ij}}|M)p_M + P(\boldsymbol{\gamma}_{\mathbf{ij}}|U)p_U \tag{3}$$

where $P(\gamma_{ij}|M)$ and $P(\gamma_{ij}|U)$ are the probabilities of observing the pattern γ_{ij} conditional on the record pair (i, j) belonging to M or U, respectively. The proportions p_M and $p_U = 1-p_M$ are the marginal probabilities of observing a matched or unmatched pair. Applying Bayes' Rule, we obtain the probability of $(i, j) \in M$ conditional on observing $\boldsymbol{\gamma}_{\mathbf{ij}}$,

$$P(M|\boldsymbol{\gamma}_{\mathbf{ij}}) = \frac{p_M P(\boldsymbol{\gamma}_{\mathbf{ij}}|M)}{P(\boldsymbol{\gamma}_{\mathbf{ij}})} \tag{4}$$

so that if we can estimate the variables in (3), we can estimate the probability that any two records refer to the same entity in (4).

As shown by Fellegi and Sunter (1969), it is possible to use the estimated probabilities to construct an "optimal" matching, given any threshold for false positive and false negative match rates. Conversely, the probabilities also allow us to estimate the false positive rate for any configuration of matches (**?**). Given the usefulness of the quantities in (4), the next sections will introduce two methods for estimating them.

5.3 Simplifying assumptions

Let $\boldsymbol{\Gamma} \equiv \{\boldsymbol{\gamma}_{\mathbf{ij}} : (i, j) \in X_1 \times X_2\}$ denote the set of comparison vectors for all records pairs $(i, j) \in X_1 \times X_2$. Note that $\boldsymbol{\Gamma}$ contains potentially $n_1 \times n_2$ elements, so that calculating $\boldsymbol{\Gamma}$ may be computationally expensive when X_1 or X_2 is large. In practice, researchers partition $X_1 \times X_2$ into "blocks," such that only records belonging to the same block are attempted to be linked, and records belonging to different blocks are assumed to be nonmatches. For example, postal codes and household membership are often used to define blocks when

linking census files (**?**). Importantly, the blocking variables should be recorded without error, and sometimes there are none available. This paper assumes that no blocking is used; or, alternatively, that records are already divided into blocks that can be analyzed independently using the methods outlined below.

Conditional independence

In principle, we can model,

$$\boldsymbol{\gamma_{ij}} \mid M \sim \text{Dirichlet}(\boldsymbol{\delta_M})$$
$$\boldsymbol{\gamma_{ij}} \mid U \sim \text{Dirichlet}(\boldsymbol{\delta_U})$$

However, there are $2^L - 1$ possible configurations of each $\boldsymbol{\gamma_{ij}}$, so that $\boldsymbol{\delta_M}$ and $\boldsymbol{\delta_U}$ may be very high-dimensional if we want to allow weights to vary across different comparison criteria.

A common assumption in the literature is that the comparison fields ℓ are defined so that γ_{ij}^{ℓ} are independent across ℓ conditional on match status. This implies:

$$P(\gamma_{ij}|C) = \prod_{\ell=1}^{L} P(\gamma_{ij}^{\ell}|C)^{\gamma_{ij}^{\ell}}(1 - Pr(\gamma_{ij}^{\ell}|C))^{1-\gamma_{ij}^{\ell}} \qquad C \in \{M, U\} \tag{5}$$

Hence the number of parameters used to describe each mixture class is reduced to L.

? have shown how to relax this assumption using log-linear models, but for now I assume conditional independence to ease computation.

5.4 Measuring record linkage performance

(could also summarize) They are evaluated according to matching rates, type I and type II error rates; robustness to selection/attrition?

List of record linkage methods Link of estimation methods

6 Estimation Methods

See the attached table for an overview (please disregard the text below, it is lifted from my own notes).

6.1 Anderson et al. (2019)

6.2 Nix and Qian (2015)

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[7], 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.

[7]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

6.3 Bleakley and Ferrie (2016)

Bleakley and Ferrie (2016) assign equal probability of winning (matched variable equal to $1/n$) to all n individuals matched to the same winner. The goal is to estimate the treatment effect of winning a parcel in the lottery by comparing mean outcomes for winners and losers in a simple bivariate regression with a dummy varialbe for winning a parcel on the right-hand side. Here, winning the lottery is coded as 0 or $1/n$, where n is the number of matches for person i. [Think about how this compares to ahl method]

6.4 Lahiri and Larsen (2005)

Lahiri and Larsen (2005) take as input two files are linked by a computerized record linkage technique (CRL). The true data pairs (x_i, y_i) are not observable; instead, the CRL produces pairs (x_i, z_i) in which z_i may or may not correspond to y_i. The (true) regression model is:

$$y_i = x_i'\beta + \epsilon_i,\ E[\epsilon_i] = 0,$$
$$\text{var}(\epsilon_i) = \sigma^2,\ \text{cov}(\epsilon_i\epsilon_j) = 0$$

but the researcher estimates this model with z_i as the dependent variable, where

$$z_i = \begin{cases} y_i & \text{with probability } q_{ii} \\ y_j & \text{with probability } q_{ij} \text{ for } j \neq i,\ j = 1, \ldots, n \end{cases}$$

and $\sum_{j=1}^{n} q_{ij} = 1,\ i = 1, \ldots, n$. Define $\mathbf{q_i} = (q_{i1}, \ldots, q_{in})'$. The naive least squares estimator of β, which ignores mismatch errors, is given by:

$$\hat{\boldsymbol{\beta}}_N = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{z}$$

An alternative to this naive estimator is one that minimizes the sum of absolute deviations, which decrease the influence of outliers and hence should decrease the impact of erroneously paired predictor and response values.

Note that $E(z_i) = \mathbf{w_i'}\boldsymbol{\beta}$, with $\mathbf{w_i} = \mathbf{q_i'X_i} = \sum_{j=1}^{n} q_{ij}x_j'$, and so the bias of $\hat{\boldsymbol{\beta}}_N$ is given by

$$\text{bias}(\hat{\beta}_N) = [(X'X)^{-1}X'QX - I]\beta$$

Hence, if $Q = I$, then $\hat{\beta}_N$ is unbiased. This is equivalent to giving all potential matches the same weight (i.e. treating all matches as equally likely to be correct), as discussed in Anderson et al. (2019).

In order to reduce the bias of $\hat{\beta}_N$, Scheuren and Winkler (1993) observed that

$$\text{bias}(\hat{\beta}_N|y) = E[(\hat{\beta}_N - \beta)|y] = (X'X)^{-1}X'B,$$

where $B = (B_1, \ldots, B_n)'$ and $B_i = (q_{ii} - 1)y_i + \sum_{j \neq i} q_{ij}y_j = \mathbf{q_i'y} - y_i$, which is the difference between a weighted average of responses from all observations and the actual response y_i. Thus, if an estimator $\hat{B}$ is available, the SW estimator is given by:

$$\hat{\beta}_{SW} = \hat{\beta}_N - (X'X)^{-1}X'\hat{B}$$

SW give a truncated estimator of B_i using the first and second highest elements of the vector q_i, $\hat{B}_i^{TR} = (q_{ij_1} - 1)z_{j_1} + q_{ij_2}z_{j_2}$, that can also be written more generally for an arbitrary number of elements of q_i. This means that $\hat{\beta}_{SW}$ is not unbiased, but, if the probability is high that the best candidate link is the true link, then the truncation might produce a very small bias.

Using $E(z) = W\beta$, Lahiri and Larsen (2005) propose an exactly unbiased estimator of

β:

$$\hat{\beta}_U = (W'W)^{-1}W'z$$

where $W = (w_1, \ldots, w_N)'$, $w_i = q_i'X_i$ as above. They suggest using a truncated version of w_i, $w_i^{TR} = q_{ij_1}x_{j_1} + q_{ij_2}x_{j_2}$. Lahiri and Larsen (2005) use estimates of Q obtained from applying the Fellegi-Sunter/EM procedure, and observe that replacing Q with $\hat{Q}$ yields unbiased estimates of β whenever $\hat{Q}$ can be assumed to be independent of z. They argue that this is expected to be true in most applications, because the distribution of matching variables (e.g. first and last name, age), which determines the distribution of $\hat{Q}$, is usually independent of the response variable y (e.g. income), and hence of z.

Importantly, this assumption does not hold in some economics applications, such as the racial "passing" example from Nix and Qian (2015).

Lahiri and Larsen (2005) conclude that in simulations, least median regression is not sufficient to guard against matching errors, whereas the method of SW (1003) made a useful adjustment. Their method performed well across a range of situations,a nd the bootstrap procedure is useful for reflecting uncertainty due to matching.

6.5 Goldstein et al. (2012)

Prior-informed imputation (PII), proposed in Goldstein et al. (2012), aims to select the correct value for variables of interest, rather than accepting a single complete record as a link. Information from match probabilities in candidate linking records (the prior) is combined with information from unequivocally linked records.

The method is more or less as follows: use the usual FS/EM procedure to estimate probabilities of candidate pairs referring to a match. The result is each record i is associated with $\{y_{ij}\}$ and probabilities p_{ij}. Assume, wlog, that all variables follow a joint multivariate

normal distribution[8] A lower threshold can be chosen so that records with probabilities less than a threshold are ignored. In practice, they recommend ignoring records that have no match on any matching variable; regard these records as having missing variables and use standard MI.

Denote distribution of variables in linking datafile A (y in ahl framework), conditional on variables in the file of interest B that they are linking to, by $f(Y^{A|B})$. Conditioning is on responses and any covariates in the imputation model, includes variables from A that are treated as auxiliary predictor variables in the imputation model. This conditional distribution is also multivariate normal.

For each record i, we compute a modified prior probability π_{ij} which is the likelihood component $f(Y^{A|B})$ multiplied by the prior p_{ij}, so that $\pi_{ij} \propto f(Y^{A|B})p_{ij}$. The normalized set π_i comprises the modified probability distribution for each i record in A.

Set a lower threshold for accepting a record as a true link, and if any records exceeds this, we choose that with the largest probability. If no record exceeds the threshold, then we regard the data values as missing and use standard multiple imputation. "The largest gain can be expected to arise when the probability of a link is associated with the values of the variables to be transferred. When the MAR assumption discussed earlier holds, then given a high enough threshold, the proposed procedure will produce unbiased estimates" Incorporating the likelihood component can be expected to lead more often to the correct record exceeding a given threshold.

For missing observation: Assume we have an estimate of mortality rate of individuals in A, π_d. If a proportion of the A file $\pi_m < \pi_d$ are unequivocally matched, then the probability that a randomly chosen remaining record in B is not a death from the A file is $\pi_r = 1 - (\pi_d - \pi_m)$. Therefore multiply p_i by $1 - \pi_r$ and add an extra pseudo-record with

[8]If this is not the case for some of the observed variables, then a joint MVN distribution can be obtained using the latent MVN trick (for categorical variables, imputed values are back-transformed to their original scales.

probability π_r with an associated code for a surviving patient.

6.6 This paper

7 Data and Implementation

8 Results

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

H. Bleakley and J. Ferrie, "Shocking behavior: Random wealth in antebellum georgia and human capital across generations," *The Quarterly Journal of Economics*, vol. 131, no. 3, pp. 1455–1495, 2016. [Online]. Available: https://doi.org/10.1093/qje/qjw014

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

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

H. Goldstein, K. L. Harron, and A. M. Wade, "The analysis of record-linked data using multiple imputation with data value priors." *Statistics in medicine*, vol. 31 28, pp. 3481–93, 2012.

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, 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

K. Harron, A. Wade, R. Gilbert, B. Muller-Pebody, and H. Goldstein, "Evaluating bias due to data linkage error in electronic healthcare records," *BMC medical research methodology*, vol. 14, p. 36, 03 2014.

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

J. Doidge and K. Harron, "Demystifying probabilistic linkage," *International Journal for Population Data Science*, vol. 3, 01 2018.

K. Harron, H. Goldstein, and C. Dibben, *Methodological Developments in Data Linkage*. United States: John Wiley Sons Inc., 2015.

P. Christen, *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Springer Publishing Company, Incorporated, 2012.

T. N. Herzog, F. J. Scheuren, and W. E. Winkler, *Data Quality and Record Linkage Techniques*, 1st ed. Springer Publishing Company, Incorporated, 2007.

H. Köpcke and E. Rahm, "Frameworks for entity matching: A comparison," *Data Knowledge Engineering*, vol. 69, pp. 197–210, 02 2010.

G. Ridder and R. Moffitt, "The Econometrics of Data Combination," in *Handbook of Econometrics*, ser. Handbook of Econometrics, J. Heckman and E. Leamer, Eds. Elsevier, January 2007, vol. 6, ch. 75. [Online]. Available: https://ideas.repec.org/h/eee/ecochp/6b-75.html

W. Winkler, "The state of record linkage and current research problems," *Statist. Med.*, vol. 14, 10 1999.

M. Hirukawa and A. Prokhorov, "Consistent estimation of linear regression models using matched data," *Journal of Econometrics*, vol. 203, no. 2, pp. 344 – 358, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0304407617302464

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