Regression analysis with linked data

Rachel Anderson[*]

This Version: October 2, 2019

Abstract

This paper studies what happens when the goal is to estimate a parametric model using observations (x, y), but x and y are observed in distinct datasets with imperfect identifiers. This setup requires that the researcher must attempt to identify which observations in the x- and y-datafiles refer to the same individual, prior to performing inference about the joint or conditional distributions of x and y. At a minimum, random errors in this matching step introduce measurement error that must be accounted for in subsequent analyses; however, concerns about sample selection arise when these errors are correlated with unobservables that affect x or y.

1 Introduction

Consider estimating β in a linear regression model,

$$y_i = x_i'\beta + \varepsilon_i,\ \ E[\varepsilon|x_i] = 0,\ \ E[\epsilon_i^2] = \sigma^2 \tag{1}$$

but, instead of observing (x, y) pairs directly, x and y are recorded in separate datasets. Additionally, both datasets contain a set of common variables w, that can be used to learn about the joint distribution of (x, y).

[*] 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).

Perhaps the most straightforward way to estimate β in this setting involves first identifying which (x, y) pairs refer to the same underlying units – the matching step – and then applying standard methods to estimate (1) using the matched pairs. Formally, for data $\{x_i, w_i\}_{i=1}^{N_x}$ and $\{y_j, w_j\}_{j=1}^{N_y}$, the matching step consists of estimating a function,

$$\varphi : \{1, \dots, N_x\} \rightarrow \{1, \dots, N_y\} \cup \varnothing \tag{2}$$

where $\varphi(i) = j$ if individual i in the x-datafile and individual j in the y-datafile refer to the same entity, and $\varphi(i) = \varnothing$ if i does not have a match in y-datafile. Note that if w identifies individuals uniquely and without error, then $\varphi(i) = j$ if and only if $w_i = w_j$, and $\varphi(i) = \varnothing$ otherwise. However, if w is not unique or recorded with error, then φ needs to be estimated, and inference about β may need to be adjusted accordingly.

To fix ideas, suppose that the goal is to estimate the effect of providing cash transfers to single mothers on the life expectancy of their children. Mathematically, the parameter of interest is β_1 in the regression model,

$$y_i = \beta_0 + x_{1i}\beta_1 + x_{2i}'\beta_2 + \varepsilon_i \tag{3}$$

where x_{1i} is a binary variable equal to 1 if person i's mother received a cash transfer, and x_{2i} includes all other demographic variables that are recorded on the welfare program applications (the x-datafile). The outcome y_i is person i's age at death, as reported in a universal database of death records (the y-datafile). The two data sources contain a common set of variables w, which include first and last name, and date of birth; however, the x- or y-datafile may contain additional variables such as place of death or ethnicity that are potentially correlated with elements in w, but only appear in one of the files. Since w contains only a few variables, individuals with common names are likely to be linked with multiple y; and so the estimated φ may need to allow for multiple possible matches.

In statistics, the task of recovering φ is called *record linkage.* A standard record linkage procedure consists of a set of decisions about (i) selecting and standardizing observations w_i and w_j, (ii) choosing which (x, y) pairs to consider as potential matches[1], (iii) defining which patterns of (w_i, w_j) constitute (partial) agreements, and (iv) designating (x, y) pairs as matches. For example, the following steps constitute a (deterministic) record linkage procedure for the setting above:

(i) Use a phonetic algorithm to standardize the first and last names in both datasets;

(ii) Consider as potential matches all (x, y) pairs whose phonetically standardized names begin with the same letter, and whose birth years are within ±2 years;

(iii) Measure the distance between any two names using Jaro-Winkler string distance, and the distance between any two birth dates as a difference in months;

(iv) Designate as matches all (x, y) pairs with Jaro-Winkler scores exceeding a pre-determined cut-off; and, if a record x has multiple possible matches that exceed the cut-off, then choose the corresponding y with the highest score (or pick one match at random if there is a tie).

Another record linkage procedure could be defined using the same rules for steps (i)-(iii), but replacing (iv) with a probabilistic matching rule that does not enforce one-to-one matching, such as

(iv*) Use the Expectation-Maximization algorithm to compute "match weights" for each (x, y) pair; then, designate as matches all pairs with match weights exceeding a threshold that is set to reflect specific tolerances for Type I and Type II error.

Except in rare cases, the estimated matching functions obtained by using (iv) and (iv*) will differ, if only because the former matches each x with at most one y, while the latter potentially matches the same x with multiple y. The second method also produces estimates

[1] This is primarily to reduce computation when $N_x \times N_y$ is large

of the probability that each of the associated y values refers to the true match, which can be combined with estimation techniques such as those in Lahiri and Larsen (2005).

Each step of the record linkage process introduces the 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). However, the above example shows that not only do the estimates of β likely depend on the estimates of φ, but also the *methods* for estimating β may also differ. It is henceforth the goal of this paper to study the *joint* impact of matching and estimation on the quality of inference with linked data.

A number of recent papers compare the performance of different matching algorithms (Bailey et al., 2017; Abramitzky et al., 2018) and estimation methods for linked data (Harron et al., 2014). This paper adds to this literature by comparing how different *combinations* of matching and estimation techniques affect parameter estimates and their confidence intervals in standard econometric models. It also makes practical suggestions for choosing which methods best suit a given setting.

In order to illustrate the techniques studied in this paper, the next section introduces 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. I will then use this synthetic data set to illustrate the matching and estimation techniques described in Sections 3 and 4. 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 Empirical Example

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 y_i values are generated according to the linear relationship,

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i, \quad \varepsilon_i \mid x_{1i}, x_{2i} \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2) \tag{4}$$

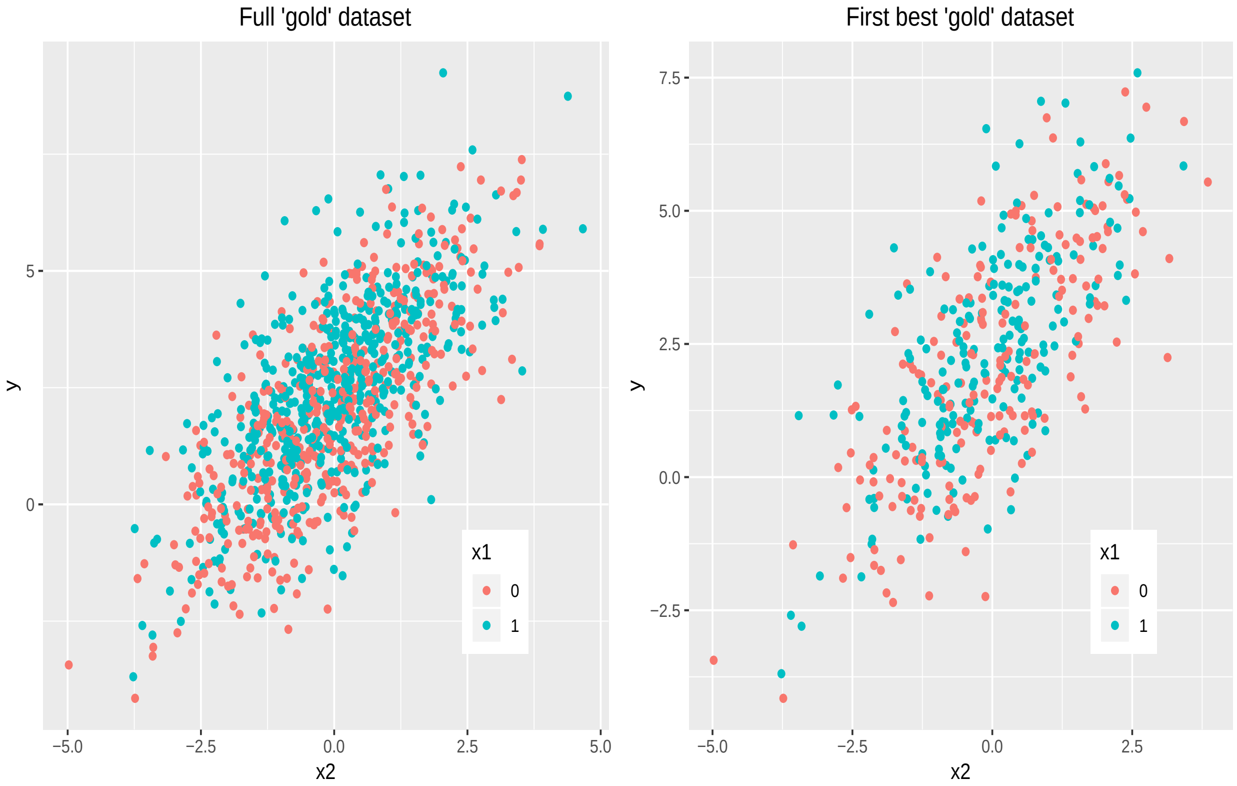
with $(\beta_0, \beta_1, \beta_2, \sigma^2) = (2, 0.5, 1, 2)$. Given these parameter values, estimating a correctly specified linear regression model yields an R^2 value of approximately 0.50 (see Figure 1(a)). Each observation is associated with a vector of identifying variables, w_i, that consists 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 the top panel in Figure 2. Note that 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 the x- and y-datafiles, which contain (x_1, x_2, w_x) and (y, w_y) values respectively. The y-datafile is identical to the ground truth data, except that it excludes the variables x_1 and x_2. The x-datafile contains values for 400 observations, selected at random from the full dataset. To construct w_x, I modify the corresponding w_y 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"). I also add normally distributed errors to the birth day, month, and year. The probability of introducing an error to any one element of w_x is set to mimic real-world data [3]. Figure 2 illustrates how the x- and y-datafiles are split visually.

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

[3] add a footnote here

Figure 1: From full data set to x dataset



Full 'gold' dataset

First best 'gold' dataset

Figure 2: 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

$\downarrow$

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

3 Record Linkage Methods

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 β_0 that satisfies the relationship in (1), that can be consistently estimated using standard econometric techniques if φ_0 is known.

In total, I implement two record linkage techniques, each of which I implement while allowing multiple or enforcing single matches. Here I provide an overview of those techniques.

3.1 Deterministic

The deterministic matching algorithm described herein is based upon Abramitzky et al. (2012). It consists of the following steps

1. Clean names in x and y datafiles to remove any non-alphabetic characters and account for common mis-spellings and nicknames (e.g., so that Ben and Benjamin would be considered the same name). This step usually involves the use of phonetic algorithms, such as NYSIIS or SOUNDEX.

2. Restrict the sample to people in the x datafile with unique first name, last name, and birth date combinations

3. For each record in the x-datafile, look for records in the y-datafile that match on first name, last name, place of birth, and exact birth year. At this point there are three possibilities

   (a) If there is a *unique* match, this pair of observations is considered a match.

   (b) If there are multiple potential matches in the y-datafile with the same year of birth, the observation is discarded.

   (c) If there are no matches by exact year of birth, the algorithm searches for matches within $\pm$ 1 year of reported birth year, and if this is unsuccessful, it looks for matches within $\pm$ 2 years. In each of these steps, only unique matches are accepted. If none of these attempts produces a unique match, the observation is discarded.

4. Step 3 is repeated for each record in the y-datafile, after which the intersection of the two matched samples is taken.

An interesting quirk of this algorithm is that one person could have a unique exact year match, but then multiple matches with birth years off by 1; this person is included when a unique match is desired. But if the unique match with zero year difference were not present, then the observation would be dropped.

3.2 Probabilistic Record Linkage

The probabilistic record linkage technique implemented in this paper are based on seminal work by Fellegi and Sunter (1969), which views record linkage as a classification problem. As before, let $\{x_i, w_i\}$, $i = 1, \ldots, N_x$ denote the observations in a dataset X; and $\{y_j, w_j\}$, $j = 1, \ldots N_y$, denote the observations in a dataset Y. The space of record pairs $X \times Y$ can be divided into two disjoint sets, *matches* (M) and *non-matches* (U), defined as

$$M = \{(i, j) \in X \times Y : j \in \varphi(i)\}$$
$$U = \{(i, j) \in X \times Y : j \notin \varphi(i)\}$$

To determine whether a record pair belongs to M or U, the pair is evaluated according to K different comparison criteria, which result from comparing w_i and w_j. These comparisons are represented in a *comparison vector*,

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

where each comparison field γ_{ij}^k may be binary-valued, as in "i and j have the same birthday" and "i and j have the same last name," or use ordinal values to indicate partial agreement between strings.

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

$$P(\boldsymbol{\gamma}_{\mathbf{ij}}) = P(\boldsymbol{\gamma}_{\mathbf{ij}}|M)p_M + P(\boldsymbol{\gamma}_{\mathbf{ij}}|U)p_U \tag{5}$$

where $P(\boldsymbol{\gamma}_{\mathbf{ij}}|M)$ and $P(\boldsymbol{\gamma}_{\mathbf{ij}}|U)$ are the probabilities of observing the pattern $\boldsymbol{\gamma}_{\mathbf{ij}}$ 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 $\boldsymbol{\gamma}_{\mathbf{ij}}$,

$$P(M|\boldsymbol{\gamma}_{\mathbf{ij}}) = \frac{p_M P(\boldsymbol{\gamma}_{\mathbf{ij}}|M)}{P(\boldsymbol{\gamma}_{\mathbf{ij}})} \tag{6}$$

Thus, if we can estimate p_M, $P(\boldsymbol{\gamma}_{\mathbf{ij}}|M)$ and $P(\boldsymbol{\gamma}_{\mathbf{ij}}|U)$, then we can estimate the probability that any two records refer to the same entity using (6). These probabilities can then be used to designate pairs as matches, or to estimate the false positive rate associated with a particular match configuration using the formulas in Fellegi and Sunter (1969).

Note that when X or Y is large, constructing comparison vectors for all $N_x \times N_y$ possible pairs can be computationally intensive. In practice, researchers solve this problem by partitioning $X \times Y$ into "blocks," such that comparison vectors are only constructed for records within the same block, and records in different blocks are assumed to be non-matches. Importantly, the blocking variables should be recorded with minimal error, otherwise blocking may adversely affect the Type II error rate.

Example 1 (cont'd). 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.

Another difficulty arises from the fact that there are at least $2^K - 1$ possible configurations of $\boldsymbol{\gamma}_{\mathbf{ij}}$. While in principle we could model $P(\boldsymbol{\gamma}_{\mathbf{ij}}|M)$ and $P(\boldsymbol{\gamma}_{\mathbf{ij}}|U)$ as

$$(\gamma_{ij}^1, \ldots, \gamma_{ij}^K) \mid M \sim \text{Dirichlet}(\boldsymbol{\delta}_{\mathbf{M}})$$
$$(\gamma_{ij}^1, \ldots, \gamma_{ij}^K) \mid U \sim \text{Dirichlet}(\boldsymbol{\delta}_{\mathbf{U}})$$

but the parameters $\boldsymbol{\delta}_{\mathbf{M}}$ and $\boldsymbol{\delta}_{\mathbf{U}}$ may be high-dimensional. However, if the comparison fields γ_{ij}^k are independent across k conditional on match status, then the number of parameters used to describe each mixture class can be reduced to K by factoring:

$$P(\gamma_{ij}|C) = \prod_{k=1}^{K} P(\gamma_{ij}^k|C)^{\gamma_{ij}^k}(1 - Pr(\gamma_{ij}^k|C))^{1-\gamma_{ij}^k} \qquad C \in \{M, U\} \tag{7}$$

Alternatively, dependence between fields can be modeled using log-linear models; however, I assume conditional independence to ease computation.

Example 1 (cont'd). Errors are constructed to satisfy this assumption

4 Estimation Methods

This section provides an overview of the estimation methods I compare for analyzing the matched datasets.

4.1 OLS Bias Correction

Scheuren and Winkler (1993) and Lahiri and Larsen (2005) propose techniques for correcting the bias from mismatched pairs in linear regression. They assume that the matching procedure produces n pairs (x_i, z_i), where z_i may or may not correspond to y_i, yet the true

y_i is included among the matches. Hence,

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

and $\sum_{j=1}^{n} q_{ij} = 1,\ i = 1, \ldots, n$. Estimating (1) using z_i as the dependent variable yields the naive least squares estimator,

$$\hat{\beta}_N = (X'X)^{-1}X'z \tag{8}$$

which is biased. Denoting $q_i = (q_{i1}, \ldots, q_{in})'$ and $Q = (q_1, \ldots, q_n)'$, we can write the bias of $\hat{\beta}_N$ as

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

since $E[z_i] = E[q_i'y] = q_i'X\beta = \sum_{j=1}^{n} q_{ij}x_j'\beta$.

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

$$\text{bias}(\hat{\beta}_N|y) = E[(\hat{\beta}_N - \beta)|y] = (X'X)^{-1}X'B \tag{9}$$

where $B = (B_1, \ldots, B_n)'$ and $B_i = (q_{ii} - 1)y_i + \sum_{j \neq i} q_{ij}y_j = q_i'y - y_i$, which is the difference between a weighted average of responses from all observations and the actual response y_i. The authors suggest estimating (9) using the first and second highest elements of the vector q_i, so that $\hat{B}_i^{TR} = (q_{ij_1} - 1)z_{j_1} + q_{ij_2}z_{j_2}$, and

$$\hat{\beta}_{SW} = \hat{\beta}_N - (X'X)^{-1}X'\hat{B}^{TR} \tag{10}$$

The estimator can incorporate any number of elements of q_i, but, if the probability is high that the best candidate link is the true link, then the truncation results in a very small bias.

Alternatively, Lahiri and Larsen (2005) use the fact that $E(z_i) = w_i'\beta$, where $w_i = q_i'X_i\beta$, to construct the unbiased estimator:

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

where $W = (w_1, \ldots, w_N)'$. To construct $\hat{\beta}_U$ in practice, they also recommend using a truncated version of W, with $w_i^{TR} = q_{ij_1}x_{j_1} + q_{ij_2}x_{j_2}$.

For both methods, the values q_{ij} are typically calculated using (6) and parameter values $\psi = \{p_M, P(\boldsymbol{\gamma}_{\mathbf{ij}}|M)$, and $P(\boldsymbol{\gamma}_{\mathbf{ij}}|U)\}$. Thus, we can write the estimators $\hat{\beta}_{SW} = \hat{\beta}_{SW}(\psi)$ and $\hat{\beta}_U = \hat{\beta_U}(\psi)$. In practice, ψ is unknown, and a reasonable estimator $\hat{\psi}$ must be used. Details on how to estimate $\hat{\psi}$ are provided in the next section. Importantly, $\hat{\beta}_U(\hat{\psi})$ is unbiased whenever $\hat{\psi}$ is independent of z, which occurs if errors in the matching variables (which determine the distribution of $\hat{\psi}$ are independent of the response variable y. Unfortunately, this assumption is unlikely to hold in many economic applications, such as Nix and Qian (2015), where y indicates whether a person's recorded ethnicity changes between survey years, but data quality significantly differs for individuals with different values of y.

4.2 Multiple match WLS

Anderson et al. (2019) propose a GMM estimator that uses data where each observation x_i is linked to L_i equally likely, potential outcomes, denoted $\{y_{i\ell}\}_{\ell=1}^{L_i}$. Importantly, their methods require that (i) the true outcome is included among the set of possible matches, (ii) each of the possible matches is equally likely to be the true match, and (iii) that the observations x_i and $\{y_{i\ell}\}_{\ell=1}^{L_i}$ are random samples from their marginal distributions conditional on (w_i, L_i).

Under these assumptions, the authors show how to construct an unbiased and consistent estimator $\hat{\beta}$ by considering the smoothed regression:

$$\sum_{\ell=1}^{L_i} y_{i\ell} - (L_i - 1)g(w_i, L_i) = x_i'\beta + u_i \tag{11}$$

where $g(w_i, L_i) = E[y_{i\ell}|w_i, L_i]$, $u_i = \varepsilon_i + \sum_{\ell=1}^{L_i} \nu_{i\ell}$, and $\nu_{i\ell} = y_{i\ell} - E[y_{i\ell}|w_i, L_i]$.

If, additionally, $E[\varepsilon_i^2|w_i, L_i] = \sigma_\varepsilon^2$ and $E[\nu_{i\ell}|z_i, L_i] = \sigma_\nu^2$ then $\hat{\beta}$ can be estimated efficiently using weighted least squares, with $\sigma(X_i) = \sigma_\varepsilon^2 + (L_i - 1)\sigma_\nu^2$, and

$$\hat{\beta}^{WLS} = \left(\sum_{i=1}^{N} \frac{x_i x_i'}{\sigma(X_i)}\right)^{-1} \left(\sum_{i=1}^{N} \frac{x_i}{\sigma(X_i)} \left(\sum_{\ell=1}^{L_i} y_{i\ell} - (L_i - 1)g(w_i, L_i)\right)\right) \tag{12}$$

which can be estimated in two-steps, where the first step involves estimating $\hat{g}(\cdot)$ and $\hat{\sigma}(X_i)$. The resulting estimator is consistent and asymptotically normal under the regularity conditions described in Anderson et al. (2019).

Assumption (iii) rules out the possibility of unobserved sample selection, in the sense that all individuals with the same identifying information have equal probability of appearing in the sample. This assumption would be violated if, for example, higher income individuals have a greater probability of appearing in the sample (unless w_i includes income). However, unlike the OLS bias correction estimators, the methods here explicitly correct for any dependence between the outcome variable and the matching variables w_i and the parameters of the matching procedure, insofar as they are captured by L_i.

[This suggests that the AHL (2019) estimator may be more robust when L_i is correlated with x_i..]

5 Implementation

5.1 Matching Algorithms

Implementation Notes

When applying the algorithm with the synthetic data, I make the following alterations:

1. I standardize the names by using the nysiis function in R. I do not need to correct for nicknames, because of how I have generated the names.
2. I restrict the all observations with unique first name, last name, date of birth, and (x_1, x_2) combinations.
3. I only perform two-way matches – that is I repeat the process for all observations in the y datafile and take the intersection of the matches). Does this change my results (check)?
4. When allowing for multiple matches, I count as matches all record pairs with the same name, and the difference in recorded birth years is within two (or five) years. That is, I designate all potential matches that arise in Step 3 as matches.

Example 1 (cont'd). Here are the differences that arise with two-way vs. not, nysiis vs. not, etc.

PRL

Since membership to M or U is not actually observed, a convenient way of simultaneously estimating p_M, p_U and classifying record pairs as matches or non-matches is via mixture modeling, with mixture distributions $P(\boldsymbol{\gamma}_{\mathbf{ij}}|M)$ and $P(\boldsymbol{\gamma}_{\mathbf{ij}}|U)$.

For convenience, denote $p_{M\ell} = P(\gamma_{ij}^{\ell}|M)$ and $p_{U\ell} = P(\gamma_{ij}^{\ell}|U)$. Assuming conditional independence across ℓ (and global parameters that do not vary by block, if using blocked records), a convenient prior distribution is the product of independent Beta distributions,

$$\begin{gathered}
p_M \sim \text{Beta}(\alpha_M, \beta_M) \\
p_{M\ell} \sim \text{Beta}(\alpha_{M\ell}, \beta_{M\ell}),\ \ell = 1, \ldots, L \\
p_{U\ell} \sim \text{Beta}(\alpha_{U\ell}, \beta_{U\ell}),\ \ell = 1, \ldots, L
\end{gathered}$$

For $i = 1, \ldots, n_1 \in X_1$, $j = 1, \ldots, n_2 \in X_2$ define the parameter of interest as,

$$I_{ij} = \begin{cases} 1, & \text{if records } i \in X_1 \text{ and } j \in X_2 \text{ refer to the same entity;} \\ 0, & \text{otherwise} \end{cases}$$

If $(p_M, p_{M\ell}, p_{U\ell})$ are known, then $P(I_{ij} = 1|\boldsymbol{\gamma}_{\mathbf{ij}}(p_M, p_{M\ell}, p_{U\ell}))$ is distributed as in (6). Alternately, if the match indicators $\mathbf{I}$ were known, the posterior distributions of $(p_M, \mathbf{p_{M\ell}}, \mathbf{p_{U\ell}})$ would be:

$$p_M|I \sim \text{Beta}\left(\alpha_M + \sum_{(i,j)} I_{ij},\ \beta_M + \sum_{(i,j)} (1 - I_{ij})\right) \tag{13}$$

$$p_{M\ell}|I \sim \text{Beta}\left(\alpha_{M\ell} + \sum_{(i,j)} I_{ij}\gamma_{ij}^{\ell},\ \beta_{M\ell} + \sum_{(i,j)} I_{ij}(1 - \gamma_{ij}^{\ell})\right), \quad \ell = 1, \ldots, L \tag{14}$$

$$p_{U\ell}|I \sim \text{Beta}\left(\alpha_{U\ell} + \sum_{(i,j)} (1 - I_{ij})\gamma_{ij}^{\ell},\ \beta_{U\ell} + \sum_{(i,j)} (1 - I_{ij})(1 - \gamma_{ij}^{\ell})\right), \quad \ell = 1, \ldots, L \tag{15}$$

Based on these ideas, **?** proposed a Bayesian version of record linkage for the mixture model approach, that uses a Gibbs Sampling scheme[4] for simulating the posterior distribution of $I, (p_M, p_{M\ell}, p_{U\ell})$.

5.2 Estimation Algorithms

For LL, z requires single matches. So I pick at random one of the highest posterior matches and call that z.

Standard errors for $\hat{\beta}_{SW}$ are calculated using the formula. Standard errors for $\hat{\beta}_{LL}$ are calculated via the parametric bootstrap described in Lahiri and Larsen (2005).

For AHL (2019), I use nearest neighbors applied to the datasets $\{x_i, \{y_{i\ell}\}_\ell, w_i\}$ that are outputted by the matching algorithms above. That means, I implement the method as if the researcher only observes the matched dataset.

6 Results

I test three DGPs. The first was described above. The second allows for correlation between x_1 and the probability of an error. The third will allow for correlation between y and the probability of an error.

The y are generated according to the same DGP described above; that is the true $\beta_0 = (2, 0.5, 1)$. I compare my results to an oracle linkage method ("first best"), which would successfully link all 400 x observations to their correct y. This produces $\hat{\beta}_{FB}$, which is shown in the table below:

Table 1: Naive OLS for all of the matchings

	Dependent variable:				
			y		
	First Best	ABE (Single)	ABE (Multi)	PRL (Single)	PRL (Multi)
x1	0.408***	0.413**	0.369**	0.587***	0.357**
	(0.143)	(0.172)	(0.153)	(0.180)	(0.168)
x2	1.096***	1.010***	0.905***	0.975***	0.930***
	(0.052)	(0.062)	(0.056)	(0.065)	(0.062)
Constant	2.079***	2.106***	2.100***	1.932***	2.091***
	(0.102)	(0.125)	(0.109)	(0.131)	(0.122)
Observations	400	327	454	311	373
R^2	0.534	0.455	0.367	0.429	0.384
Adjusted R^2	0.532	0.452	0.364	0.425	0.381
Note:					*p<0.1; **p<0.05; ***p<0.01

Table 2: Match rate for matching algorithms

method	nMatches	pCorrect	nUniqueX
ABE (Single)	327	0.95	327
ABE (Multi)	454	0.78	360
PRL (Single)	311	0.85	311
PRL (Multi)	373	0.79	311

6.1 Matching step

6.2 Estimation Results

7 Conclusion

Borrow from the Bayesian Record Linkage.

Merging datasets with imperfect identifiers occurs frequently in projects that use his-

[4] See Appendix **??**

Figure 3: Distribution of number of matches produced by multi-matching algorithms

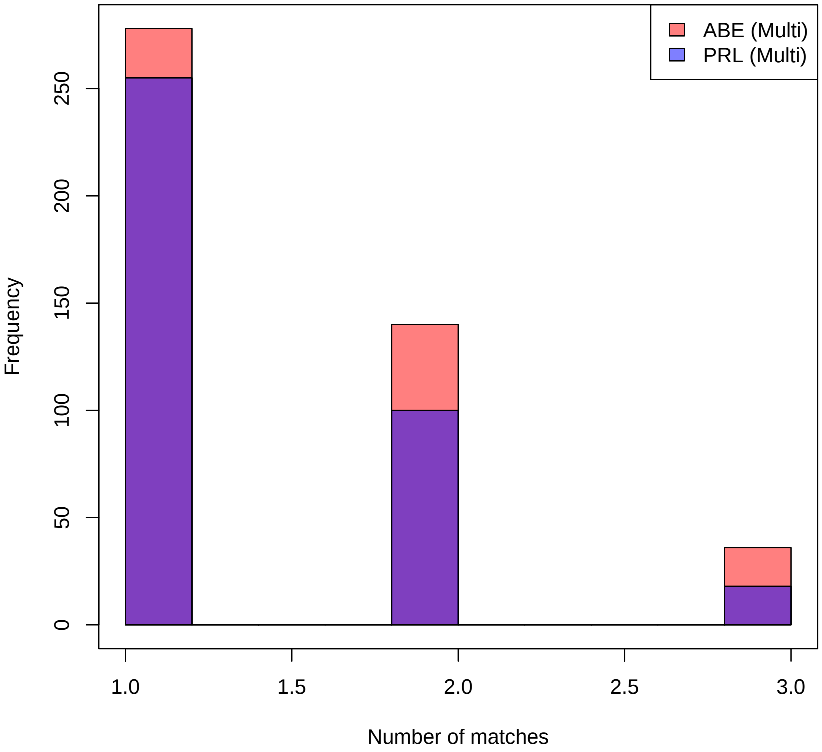


Figure 4: default

Figure 5: Matches

torical 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 match 48 percent of children to a unique death record, and 4 percent to multiple possible records, 48 percent of observations remain unmatched[5]. 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.

[5] 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.

Table 3:

	method	nMatches	pCorrect	nUniqueX
1	abe_single	338	0.96	338
2	abe_multi	479	0.77	375
3	prl_single	335	0.87	335
4	prl_multi	397	0.79	335

Table 4: Parameter estimates for different matched datasets and estimation procedures

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).

References

Abramitzky, Ran, Leah Boustan, Katherine Eriksson, James Feigenbaum, and Santiago Perez, "Automated Linking of Historical Data," *NBER Working Paper*, 2019.

_ , Leah Platt Boustan, and Katherine Eriksson, "Europe's Tired, Poor, Huddled Masses: Self-Selection and Economic Outcomes in the Age of Mass Migration," *American Economic Review*, May 2012, *102* (5), 1832–56.

_ , Roy Mill, and Santiago Perez, "Linking Individuals Across Historical Sources: a Fully Automated Approach," Working Paper 24324, National Bureau of Economic Research February 2018.

Aizer, Anna, Shari Eli, Joseph Ferrie, and Adriana Lleras-Muney, "The Long-Run Impact of Cash Transfers to Poor Families," *American Economic Review*, April 2016, *106* (4), 935–71.

Anderson, Rachel, Bo Honore, and Adriana Lleras-Muney, "Estimation and inference using imperfectly matched data," *Working paper*, August 2019.

Bailey, Martha, Connor Cole, Morgan Henderson, and Catherine Massey, "How Well Do Automated Linking Methods Perform? Lessons from U.S. Historical Data," Working Paper 24019, National Bureau of Economic Research November 2017.

Bleakley, Hoyt and Joseph Ferrie, "Shocking Behavior: Random Wealth in Antebellum Georgia and Human Capital Across Generations," *The Quarterly Journal of Economics*, 2016, *131* (3), 1455–1495.

Doidge, James and Katie Harron, "Demystifying probabilistic linkage," *International Journal for Population Data Science*, 01 2018, *3.*

Fellegi, I. P. and A. B. Sunter, "A Theory for Record Linkage," *Journal of the American Statistical Association*, 1969, *64*, 1183–1210.

Goldstein, Harvey, Katie L Harron, and Angela Mills Wade, "The analysis of record-linked data using multiple imputation with data value priors.," *Statistics in medicine*, 2012, *31 28*, 3481–93.

Harron, Katie, Angie Wade, Ruth Gilbert, Berit Muller-Pebody, and Harvey Goldstein, "Evaluating bias due to data linkage error in electronic healthcare records," *BMC medical research methodology*, 03 2014, *14*, 36.

Lahiri, P. and Michael D. Larsen, "Regression Analysis with Linked Data," *Journal of the American Statistical Association*, 2005, *100* (469), 222–230.

Nix, Emily and Nancy Qian, "The Fluidity of Race: Passing in the United States, 1880-1940," Working Paper 20828, National Bureau of Economic Research January 2015.

Scheuren, Fritz and William Winkler, "Regression analysis of data files that are computer matched," *Survey Methodology*, 01 1993, *19.*