Regression analysis with linked data

Rachel Anderson*

This Version: October 1, 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,

$$\varphi : \{1, \ldots, N_x\} \to \{1, \ldots, N_y\} \cup \varnothing \tag{2}$$

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... 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 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 will now 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.

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]. The x- and y-datafiles are represented in the bottom panels of Figure 2.

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):

[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

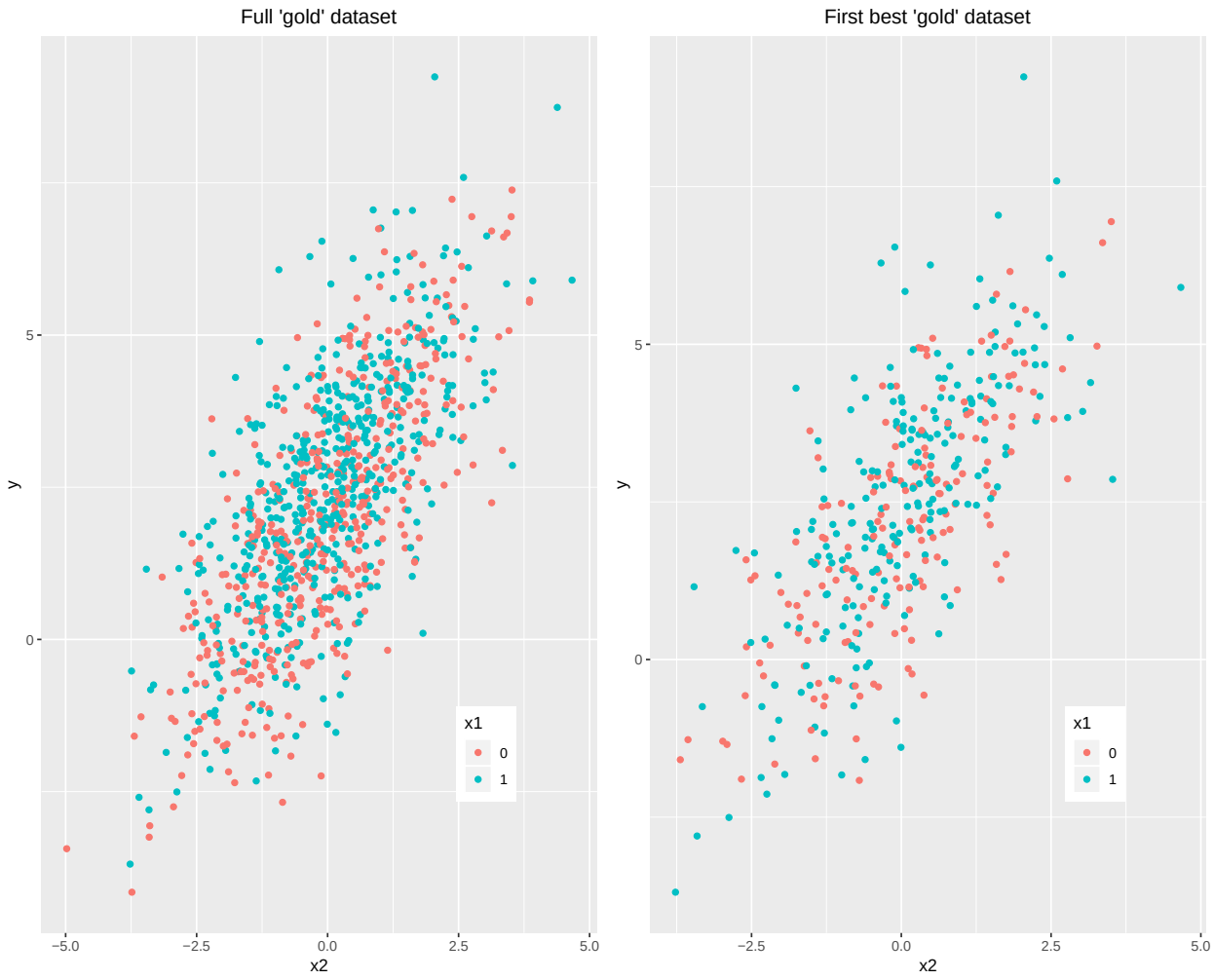

Figure 1: default

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

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

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.

Merging datasets with imperfect identifiers occurs frequently in 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 indi-

viduals' 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[4]. 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).

Example 1. Section 2 introduces the general problem that this paper seeks to address and outlines a common framework for comparing the techniques in Table **??**. 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.

3 Record Linkage Methods

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.

[4]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.

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).

2. Restrict the sample to people who are unique by first and last name, and year and place of birth

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.

I alter the algorithm slightly so that Step 2 becomes restrict the sample to people who are unique by first and last name, year, place of birth, and (x_1, x_2) values[5]. When allowing for

[5]This becomes y values when I repeat the algorithm linking y to the x-datafile

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.

A 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.

Example 1 (cont'd).

3.2 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 X and Y 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 X and Y 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 X and Y refer to the same entities.

Suppose that files X and Y contain N_x and N_y records, respectively, and without loss of generality that $N_y \geq N_x$. Denote also the number of entities represented in both files as N_{xy}, so that $N_x \geq N_{xy} \geq 0$.

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

$$M = \{(i, j) : i \in X, j \in Y, i = j\}$$

$$U = \{(i, j) : i \in X, j \in Y, i \neq j\}$$

Hence, the formal goal of record linkage is to identify whether an arbitrary record pair $(i, j) \in X \times Y$ belongs to M or U.

To perform this task, each record pair is evaluated according to K 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*,

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

where each comparison field γ_{ij}^k may be binary-valued, as in "i and j have the same birthday," or use levels to account for partial agreement between strings (see **?**, for details).

Example 1 (cont'd).

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 $P(\gamma_{ij}|M)$ 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 $\boldsymbol{\gamma}_{\mathbf{ij}}$,

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

so that if we can estimate the variables in (5), we can estimate the probability that any two records refer to the same entity in (6). These probabilities can then be used to designate pairs as matches, and to estimate the false positive rates associated with any potential match

configuration.

Let $\mathbf{\Gamma} \equiv \{\boldsymbol{\gamma}_{\mathbf{ij}} : (i,j) \in X \times Y\}$ denote the set of comparison vectors for all records pairs $(i,j) \in X \times Y$. Note that $\mathbf{\Gamma}$ contains potentially $N_x \times N_y$ elements, so that calculating $\mathbf{\Gamma}$ may be computationally expensive when X or Y is large. In practice, researchers partition $X \times Y$ 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 non-matches. Importantly, the blocking variables should be recorded without error, and sometimes there are none available.

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 important simplifying assumption is that of conditional independence. While in principle we can model,

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

there are $2^K - 1$ possible configurations of each $\boldsymbol{\gamma}_{\mathbf{ij}}$, so that $\boldsymbol{\delta}_{\mathbf{M}}$ and $\boldsymbol{\delta}_{\mathbf{U}}$ may be high-dimensional. Instead, if the comparison fields are structured so that the γ_{ij}^k are independent across k conditional on match status, then,

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

and the number of parameters used to describe each mixture class is reduced to K. This assumption can be relaxed using log-linear models, but for now 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 might produce a very small bias.

Using $E(z_i) = w_i'\beta$, where $w_i = q_i'X_i\beta$, Lahiri and Larsen (2005) propose the unbiased estimator:

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

where $W = (w_1, \ldots, w_N)'$. They also suggest using a truncated version of W, with $w_i^{TR} = q_{ij_1}x_{j_1} + q_{ij_2}x_{j_2}$, where values of q_{ij} are estimated by using a probabilistic record linkage procedure.

For both methods, standard errors are calculated according to a parametric bootstrap procedure/formulas in the appendix.

These procedures are valid if the estimated probabilities in $\hat{Q}$ are 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. However, this assumption is unlikely to hold in important economics applications, such as the racial "passing" example from Nix and Qian (2015).

4.2 Multiple match WLS

Anderson et al. (2019) consider estimating a GMM model for data $(x_i, \{y_{i\ell}\}_{\ell=1}^{L_i}, w_i)_{i=1}^{N}$, so that each observation x_i is linked to L_i equally likely, potential outcomes. Importantly, they assume that the true outcome y_i is included among the possible matches, and that the observations x_i and $\{y_{i\ell}\}$ are random samples 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:

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

where $g(w_i, L_i) = E[y_{i\ell}|w_i, L_i]$, which can be estimated nonparametrically. This results in a weighted least squares setup.

Let $M_i = \{y_{i\ell}\}_{\ell=1}^{L_i}$, the set of outcomes linked to observation x_i. If M_i contains the true match y_i, then the estimator is unbiased:

$$\begin{aligned}
E[\hat{\beta}] &= E[x_i x_i']^{-1} E\left[x_i \left(\sum_{\ell=1}^{L_i} y_{i\ell} - (L_i - 1)g(w_i, L_i)\right)\right] \\
&= \beta + E[x_i x_i']^{-1} E\left[x_i \left(\sum_{M_i/y_i} y_{i\ell} - (L_i - 1)g(w_i, L_i)\right)\right] \\
&= \beta + E[x_i x_i']^{-1} E\left[E[x_i|w_i, L_i] E\left[\sum_{M_i/y_i} y_{i\ell} - (L_i - 1)g(w_i, L_i)|w_i, L_i\right]\right] \\
&= \beta
\end{aligned}$$

The second to last line follows from the assumption that x_i and $y_{i\ell}$ are random samples conditional on w_i, L_i.

5 Simulation Results

To refresh, there are two DGPs.

The initial (no correlation) and then one with correlation between x_1 and the probability of an error.

5.1 Matching step

Figure 3: Matches

Table 1: Match rate for matching algorithms

	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

5.2 Estimation Results

Table 2: Parameter estimates for different matched datasets and estimation procedures

6 Conclusion

Borrow from the Bayesian Record Linkage.

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.*