Regression analysis with linked data

Rachel Anderson*

This Version: October 15, 2019

Abstract

This paper compares different methods for estimating parametric models with linked data, i.e. when 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 the matching step introduce measurement error that must be accounted for in subsequent inference; however, additional concerns about sample selection arise when these errors are correlated with unobservables that affect x or y.

1   Introduction

Ask any economist how to handle multiple matches when performing a one-to-one merge, and you will get a different answer. Some select among the possible matches those that they believe most likely to be correct. Some estimate the same model using multiple configurations of matched data, or using methods that allow for multiple matches without double counting observations. Others avoid the issue entirely, by ignoring all observations that may be associated with multiple matches. While there are myriad ways to handle multiple matches

in practice, there appears to be no one-size-fits-all solution, nor a formal theory about how subjective decisions about merging impact subsequent inference and estimation in economic analyses.

Although it is frequently encountered in economics, the "multiple matches" problem also appears in fields such as statistics, computer science, operations research, and epidemiology, under names such as record linkage, data linkage, entity resolution, instance identification, de-duplication, merge/purge processing, and entity reconciliation. Myriad books etc. ... theory is well developed (give credit where it's due!) But these issues are especially challenging for economists, who are often interested in causal estimates, and where errors in matching may be correlated with the parameters of interest.

In economics, significant interest in data combination techniques has emerged recently as the use and availability of large administrative datasets has risen across the profession. Indeed, a number of recent papers compare the performance of popular matching techniques used in economic history on the representativeness and accuracy of the datasets that they produce (Abramitzky et al., 2019, 2018; Bailey et al., 2017); however work on data combination dates back even further (**?**).

Despite the interest, and despite the work, there is no conclusive answer on what to do during estimation and inference! Nor is there consensus on whether to allow multiple matches or drop them or pick one randomly. The purpose of this paper is to show that inference using linked data requires making joint assumptions for the matching and estimation steps, such as whether multiple matches are allowed, and, if so, how these matches should be accounted for when using standard estimation techniques.

In the past, authors have handled multiple matches by generating a "composite match" equal to the average of the linked observations (Bleakley and Ferrie, 2016), constructing bounds on the parameter of interest using different configurations of matched data (Nix and Qian, 2015), or using methods that allow for multiple outcomes by Anderson et al. (2019)

using weighted least squares. Alternatively, if probabilistic record linkage methods are used, the bias introduced during the matching step may be removed using robust OLS estimators in Lahiri and Larsen (2005), or prior-informed imputation in Goldstein et al. (2012).

What is known is that:

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, generally, there is a tradeoff between reducing either one of the two (Abramitzky et al., 2019; **?**). Also representativeness is an issue. However, the impact of data pre-processing choices on estimation is still not well understood.

This paper adds to this literature by comparing methods that allow for multiple matches; to dtermine whether multiple matches are even optimal! It also makes practical suggestions for choosing which methods best suit a given setting. My hypothesis: if you use deterministic matching, you should allow for multiple matches. If you use probabilistic record linkage, you need to bias correct a la SW. Estimated probabilities can be noisy, so use agnostic weights if uncertain. Incorporate multiple matches to avoid problem that L_i is correlated with outcomes.

In order to illustrate the techniques studied in this paper, Section 2 introduces a numerical example that is used to demonstrate the matching and estimation techniques described in Sections 3 and 4. Section 5 provides details about the implementation of the methods and data generating processes. Section 6 contains the results, and Section 7 concludes.

2 Setup & Numerical Example

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 link observations in order to learn about the joint distribution of (x, y).

Perhaps the most straightforward way to estimate β in this setting involves first identifying which (x, y) pairs refer to the same underlying units, 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, we again consider the setup of Aizer et al. (2016), where 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 can be represented as 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 additionally contain a common set of variables w, including first and last name, and year of birth. Since no combination of these variables is necessarily unique, estimating φ in this setting would likely require distinguishing among multiple observations with identical w.

3 Numerical Example

The purpose of this section is to introduce a synthetic dataset that will be referenced throughout this paper. The benefits of using a synthetic dataset are threefold: first, I can control which variables in (1) are correlated with errors in the identifiers w; second, I can compare the performance of the matching algorithms and estimation techniques to a "ground truth" dataset containing all of the true matches; third, I can control the degree of overlap and similarity among observations, which is imporrtant for manipulating the number of possible matches if multiple matches are desired. To make the data as real as possible, I choose the identifiers and their corresponding transcription error rates to mimic those reported in the 1940 Census data, as reported by Abramitzky et al. (2019).

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, $x_{1i} \overset{i.i.d}{\sim} \text{Bernoulli}(0.5)$, 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 $(\beta_0, \beta_1, \beta_2, \sigma^2) = (2, 0.5, 1, 2)$, so that estimating the correctly specified linear regression model yields an R^2 value of approximately 0.50. Each observation is also assigned the identifiers, w_i, which include include a first and last name drawn at random from a list of

5

first and last names[1], and a random birthday between January 1, 1900, and December 31, 1925. The resulting dataset looks like the observations in the top panel of Figure 1.

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
			$\vdots$			
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
			$\vdots$			
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

Next, I split the ground truth dataset into an x- and y-datafile, which contain (x_1, x_2, w) and (y, w) values respectively. To construct the x-datafile, I select 400 observations at random from the ground truth dataset, and introduce random errors in their corresponding identifiers. These errors include deleting characters (e.g., "Anderson" becomes "Andersn"), exchanging vowels (e.g., "Rachel" becomes "Rachal"), and swapping English phonetic equivalents (e.g. "Ellie" becomes "Elie"). I also add normally distributed errors to the birth day, month, and year. The probabilities of introducing an error are set to match the transcription error rates reported in the 1940 Census by Abramitzky et al. (2019); for example, 7% of observations have misreported first names and 17% of observations have misreported last

---

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

names. The bottom panel of Figure 1 illustrates how the x- and y-datafiles are split visually.

The y-datafile includes all 1,000 values from the ground truth data, and does not contain any errors in the identifiers w. As a result, it will be likely that some x will be matched to multiple y. In later sections, I will consider versions of this synthetic dataset, where errors in w are correlated with x_1 or y.

4   Record Linkage Methods

Recall that the data consist of an x-datafile, denoted $X \equiv \{(x_i, w_i) : i = 1, \ldots, N_x\}$, and a y-datafile, denoted $Y \equiv \{(y_j, w_j) : j = 1, \ldots, N_y\}$, and the goal of record linkage is to use w_i and w_j to determine which $i \in \{1, \ldots, N_x\}$ and $j \in \{1, \ldots, N_y\}$ refer to the same individual.

For the purposes of this paper, I define a record linkage procedure as a set of decisions about (i) selecting and standardizing the identifying variables in w_i and w_j, (ii) choosing which (i, j) pairs to consider as potential matches, (iii) defining which patterns of (w_i, w_j) constitute (partial) agreements, and (iv) designating (i, j) pairs as matches.[2]

Step (i) addresses the fact that differences may arise in w_i and w_j because of transcription error or misreporting, even when observations i and j refer to the same individual. In practice, this step consists of removing spaces and non-alphabetic characters from string variables and processing names with phonetic algorithms to account for potential misspellings; common nicknames may also be replaced with full names.

Step (ii) reduces the computational burden of a matching procedure when $N_x \times N_y$ is large by partitioning $X \times Y$ into "blocks." Only records within the same block are attempted to be matched, while records in different blocks are assumed to be non-matches. Blocking

---

[2]By contrast, Bailey et al. (2017) categorize record linkage procedures according to the set of assumptions that motivate their use.

variables should be recorded with minimal error, otherwise blocking may adversely affect the Type II error rate.

Step (iii) defines a metric for quantifying the similarity between non-numeric variables, such as Jaro-Winkler distances for strings. For more details, see Abramitzky et al. (2018).

Finally, Step (iv) is where record linkage procedures differ in the most meaningful ways; hence, this step will be the focus of my analysis. Consider the following (deterministic) record linkage procedure as an example:

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

(ii) Consider as potential matches all (i, j) 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 (i, j) pairs with Jaro-Winkler scores exceeding a pre-determined cut-off; and, if a record i has multiple possible matches that exceed the cut-off, then choose the corresponding j 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 steps (i)-(iii), but replacing (iv) with a probabilistic matching rule that does not enforce one-to-one matching:

(iv*) Use the Expectation-Maximization algorithm to compute "match weights" for each (i, j) 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 switching (iv) and (iv*) will differ, if only because the former method matches each x with at most one y, the latter potentially matches the same x with multiple y. This example also illustrates the

difference between deterministic and probabilistic record linkage methods: while (iv) uses pre-determined rules to designate pairs as matches, (iv*) uses statistical theory to inform the selection of the decision rule. Probabilistic record linkage also involves the estimation of match weights, which can be incorporated in subsequent estimation steps.

Below I will discuss two record linkage methods – one deterministic and one probabilistic – that I will use in my analysis. Each method will be implemented twice: first, requiring unique matches, and then allowing for multiple matches. While these methods are by no means exhaustive, they are intended to be representative of the most commonly used methods in economics. For a detailed survey of record linkage techniques, please refer to books by Harron et al. (2015); Christen (2012) or Herzog et al. (2007), or any of the references in this paper.

4.1 Deterministic

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

1. Clean names in the 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 in the x-datafile with unique first name, last name, and birth year 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.

9

(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. Repeat Step 3 for each record in the y-datafile, searching for matches in the x-datafile; then designate as matches all record pairs in the intersection of the two matched samples.

An interesting quirk of this algorithm is that an individual with multiple matches is dropped from the sample only if those matches occur before a unique match is found in Step 3. That is, a person with a unique, same-year match, and multiple matches with birth years within one year, will not be dropped from the sample. If the same-year match were not included in the dataset, then that same individual would be dropped. This has significant implications for bootstrapping standard errors; notably, the nonparametric bootstrap will fail.

Note that this quirk only occurs when the algorithm enforces unique matches. When allowing for multiple matches, I designate as a match any pair that satisfies any of the categories in Step 3.

4.2   Probabilistic Record Linkage

The probabilistic record linkage technique implemented in this paper is based on the canonical model by Fellegi and Sunter (1969), which views record linkage as a classification problem, where every record pair belongs either to the set of *matches* (M) or *non-matches*

10

(U):

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

To determine whether a record pair (i, j) belongs to M or U, the pair is evaluated according to K different comparison criteria. These comparisons are represented in a *comparison vector*,

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

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 $\boldsymbol{\gamma_{ij}}$ can be modeled as arising from the mixture distribution:

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

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

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

Thus, if we can estimate p_M, $P(\boldsymbol{\gamma_{ij}}|M)$ and $P(\boldsymbol{\gamma_{ij}}|U)$, 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).

One difficulty arises from the fact that there are at least $2^K - 1$ possible configurations of $\boldsymbol{\gamma_{ij}}$[3]. While in principle we could model $P(\boldsymbol{\gamma_{ij}}|M)$ and $P(\boldsymbol{\gamma_{ij}}|U)$ as

$$(\gamma_{ij}^1, \ldots, \gamma_{ij}^K) \mid M \sim \text{Dirichlet}(\boldsymbol{\delta_M})$$

$$(\gamma_{ij}^1, \ldots, \gamma_{ij}^K) \mid U \sim \text{Dirichlet}(\boldsymbol{\delta_U})$$

but the parameters $\boldsymbol{\delta_M}$ and $\boldsymbol{\delta_U}$ 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 will assume conditional independence to ease computation, and because the matching variables in the synthetic dataset are generated independently of each other.

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 $P(\boldsymbol{\gamma_{ij}}|M)$ and $P(\boldsymbol{\gamma_{ij}}|U)$. The parameters can be estimated using the expectation-maximization (EM), first applied to record linakge by Larsen and Rubin (2001). For this paper, I use the `fastLink` algorithm developed by Enamorado et al. (2019).

---

[3]There are more, if any of the comparison criteria are non-binary

5 Estimation with linked data

When performing estimation with linked data, a few obvious comparison metrics arise: data can be partitioned into three parts - identfied links, nonlinks and potential links. Could repeat the analysis for each group or for subsets of these groups. They use nonlinks to adjust the potential links, and thereby, gain an additional perspective that could lead to reductions in MSE over statistics calculated only from the linked data. Benchmark is OLS with one-to-one-matching, or using observations assigned $L_i = 1$ matches.

Using only data from pairs of records that are highly likely to be links might mean throwing away additional information from potentiallyl inked pairs, which could contain true links. Additionally, we could bias results because confidently linked pairs may differ from potentially linked piars. For example, considering affirmative actiona nd income questison, certain records may be harder to match. For deterministic methods, people reweight on observables..

Some methods specifically attempt to correct for the bias introduced by the matching step. Seminal work by Neter, Maynes and Ramanathan (1965) shows that if matching errors are moderate then regression coefficeints can be severely biased. This work is formalized by Scheuren and Winkler (1993).

The primary examples are Lahiri and Larsen (2005) and Scheuren and Winkler (1993). Scheuren and Winkler (1993) presuppose that the linker has provided a combined data file consisting of pairs of records (one from each input file) along with the match probability and the link status – either link, nonlink, or potential link – of each pair. They assume that the file of linked cases has been augmented so that every record on the smaller of the two files has been paired with two records of the larger file having the highest matching weights. Some cases will consist of (link, nonlink) combinations or (nonlink, nonlink) combinations, but they rule out settings where more than one true link could occur, so that (link link)

combinations are ruled out.

Formally Scheuren and Winkler (1993) 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 potential matches. Hence,

$$
z_i = \begin{cases} y_i & \text{with probability } q_{ii} \\[2ex] y_j & \text{with probability } q_{ij} \text{ for } j \neq i, \ j = 1, \ldots, n \end{cases}
$$

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

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

14

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 $\psi = \{p_M, P(\boldsymbol{\gamma_{ij}}|M), \text{ and } P(\boldsymbol{\gamma_{ij}}|U)\}$. 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.

The work of Anderson et al. (2019) is like a generalized version of Scheuren and Winkler (1993). They 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..]

6 Estimation with linked data and probabilities

H

7 Optimality of the AHL Estimator

Consider the problem of estimating the mean of a random variable $X \sim F_X(\mu; \sigma^2)$ using two observations X_1 and X_2. With probability π, X_1 is drawn from the true distribution F_X and X_2 is noise drawn from the distribution $F_Y(\kappa, \omega^2)$. With probability $1 - \pi$, X_2 is drawn from the correct distribution and X_1 is noise. Under this specification, exactly one of X_1 or X_2 is drawn from the distribution of interest at all times.

Observe that if π is known, we can construct an unbiased estimator using only X_1,

$$\hat{\mu}_1 = \frac{X_1}{\pi} - \frac{1 - \pi}{\pi}\kappa \tag{13}$$

Similarly, we can construct an unbiased estimator using only X_2,

$$\hat{\mu}_2 = \frac{X_2}{1 - \pi} - \frac{\pi}{1 - \pi}\kappa \tag{14}$$

Compare these to an estimator that uses both X_1 and X_2,

$$\hat{\mu} = a_1 X_1 + a_2 X_2 - a_3 \kappa \tag{15}$$

which has the following expectation,

$$E[\hat{\mu}] = (a_1\pi + a_2(1 - \pi))\mu + (a_1(1 - \pi) + a_2\pi - a_3)\kappa$$

so that unbiased, requires

$$a_1\pi + a_2(1 - \pi) = 1 \implies a_2(a_1) = \frac{1}{1 - \pi} - \frac{a_1\pi}{1 - \pi} \tag{16}$$

$$a_1(1 - \pi) + a_2\pi = a_3 \implies a_3(a_1) = \frac{\pi}{1 - \pi} + \frac{a_1 - 2a_1\pi}{1 - \pi} \tag{17}$$

Hence we can write $\hat{\mu}$ as a function of a_1,

$$\hat{\mu}(a_1) = a_1 X_1 + \left(\frac{1}{1-\pi} - \frac{a_1 \pi}{1-\pi} \right) X_2 - \left(\frac{\pi}{1-\pi} + \frac{a_1 - 2a_1 \pi}{1-\pi} \right) \kappa$$

When $a_1 = \frac{1}{\pi}$, then $\hat{\mu} = \hat{\mu}_1$; and if $a_1 = 0$ then $\hat{\mu} = \hat{\mu}_2$.

We can write:

$$\hat{\mu} = (a_1 \pi)\hat{\mu}_1 + (1 - a_1 \pi)\hat{\mu}_2 + a_1(1-\pi)\kappa + \frac{\pi}{1-\pi}(1 - a_1)\kappa - \left(\frac{\pi}{1-\pi} + \frac{a_1 - 2a_1 \pi}{1-\pi} \right)\kappa$$

$$= (a_1 \pi)\hat{\mu}_1 + (1 - a_1 \pi)\hat{\mu}_2 - (a_1 \pi)\kappa$$

Hence any unbiased estimator $\hat{\mu}$ that uses X_1 and X_2 can be written as a linear combination of estimators using only X_1 or X_2. The problem of finding the minimum variance, unbiased estimator $\hat{\mu}$ reduces to finding d^* that solves

$$\min_d \; \mathrm{Var}\left(d\hat{\mu}_1 + (1-d)\hat{\mu}_2 \right)$$

which is solved by $d^* = 0$ or $d^* = 1$ depending on whether $\mathrm{Var}(\hat{\mu}_1)$ or $\mathrm{Var}(\hat{\mu}_2)$ is smaller.

I now show that $\mathrm{Var}(\hat{\mu}_1) > \mathrm{Var}(\hat{\mu}_2)$, without loss of generality, except when $\pi = 0.5$, or $\sigma^2 = \omega^2 = (\mu - \kappa)^2$. Observe that,

$$\mathrm{Var}(\hat{\mu}_1) = \frac{\mathrm{Var}(X_1)}{\pi^2} = \frac{1}{\pi^2}\left(\pi\sigma^2 + (1-\pi)\omega^2 + \pi(1-\pi)(\mu - \kappa)^2 \right)$$

$$\mathrm{Var}(\hat{\mu}_2) = \frac{\mathrm{Var}(X_2)}{(1-\pi)^2} = \frac{1}{(1-\pi)^2}\left((1-\pi)\sigma^2 + \pi\omega^2 + \pi(1-\pi)(\mu - \kappa)^2 \right)$$

This follows from the law of total variance, with the random variable $D = 1$ if X_1 is drawn from the correct distribution (and X_2 is drawn from the incorrect distribution), and

18

$D = 0$ otherwise.

$$\text{Var}\,(X_1) = E[\text{Var}\,(X_1|D)] + \text{Var}\,(E[X_1|D])$$

$$= P(D = 1)\sigma^2 + P(D = 0)\omega^2 + \text{Var}\,(\mu D + \kappa(1 - D))$$

$$= \pi\sigma^2 + (1 - \pi)\omega^2 + \pi(1 - \pi)(\mu - \kappa)^2$$

Similarly,

$$\text{Var}\,(X_2) = (1 - \pi)\sigma^2 + \pi\omega^2 + \pi(1 - \pi)(\mu - \kappa)^2$$

Thus, $\text{Var}\,(\hat{\mu}_1)$ and $\text{Var}\,(\hat{\mu}_2)$ can be written as functions of σ^2, ω^2, and $(\mu - \kappa)^2$,

$$g(\sigma^2, \omega^2, (\mu - \kappa)^2, x) \equiv \frac{1}{x^2}\left(x\sigma^2 + (1 - x)\omega^2 + x(1 - x)(\mu - \kappa)^2\right)$$

$$\text{Var}\,(\hat{\mu}_1) = g(\sigma^2, \omega^2, (\mu - \kappa)^2, \pi)$$

$$\text{Var}\,(\hat{\mu}_2) = g(\sigma^2, \omega^2, (\mu - \kappa)^2, 1 - \pi)$$

Importantly,

$$\frac{\partial g(\sigma^2, \omega^2, (\mu - \kappa)^2, x)}{\partial x} = \frac{\omega^2(x - 2) - x(\sigma^2 + (\mu - \kappa)^2)}{x^3} < 0, \ x \in (0, 1)$$

and so $\text{Var}\,(\hat{\mu}_\ell)$ is strictly decreasing in the probability that the observation ℓ is drawn from the correct distribution, and $\text{Var}\,(\hat{\mu}_1) \neq \text{Var}\,(\hat{\mu}_2)$ unless $\pi = 0.5$. Thus, the minimum variance unbiased estimator is equal to $\hat{\mu}_\ell$ for the observation ℓ that has the highest probability of being correct.

The above result holds also for L observations $X_1, \ldots, X_L$ with corresponding probabilities $\pi_1, \ldots, \pi_L$; that is, the minimum variance unbiased estimator $\hat{\mu}$ will use only X_ℓ with the highest π_ℓ and apply inverse probability weighting.

19

Now consider a sample of N sets of i.i.d. observations, $\left\{\{X_{i\ell}\}_{\ell=1}^{L_i}\right\}_{i=1}^{N}$. If the values

$$\pi_{i\ell} = \Pr(X_{i\ell} \text{ is drawn from the correct distribution}), \sum_{\ell=1}^{L_i} \pi_{i\ell} = 1$$

are known for all i, then the optimal estimator is

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} \hat{\mu}_i = \frac{1}{N} \sum_{i=1}^{N} \frac{X_{i\ell_i}}{\pi_{i\ell_i}} - \frac{1 - \pi_{i\ell_i}}{\pi_{i\ell_i}} \kappa$$

where $\hat{\mu}_i$ is the minimum variance estimator constructed for observations $\{X_{i\ell}\}_{\ell=1}^{L_i}$.

Since $\hat{\mu}$ is an inverse probability weighting estimator, small values of $\pi_{i\ell}$ may be detrimental for its finite sample performance. Rather than dropping observations whose maximal $\pi_{i\ell}$ is small, however, it is possible to give these observations equal weights for all $\{X_{i\ell}\}$, as in the AHL estimator. I hypothesize that if $\pi_{i\ell} < 0.5$ (or another threshold) for all ℓ, then it is better to give equal weights to all $X_{i\ell}$ associated with i, even when the $\pi_{i\ell}$ are known.

In practice, $\pi_{i\ell}$ and κ will most likely need to be estimated. Imprecise estimates of $\hat{\pi}_{i\ell}$ may introduce large finite-sample bias, as can seen via simulation. Similarly the variance of $\hat{\kappa}$ may not be insignificant. The advantage of the AHL estimator is that it does not require knowledge about $\pi_{i\ell}$, and may even have best worst case performance in terms of asymptotic MSE; however, the bias correction term $\hat{\kappa}$ still needs to be estimated.

8   Monte Carlo Study

Following the same procedure for simulating the empirical example described in Section 2, I generate 1,000 random x- and y- dataset pairs. I implement four types of matching procedures using each dataset pair: (i) deterministic matching with unique matches (ABE Single), (ii) deterministic matching with multiple matches (ABE Multi), (iii) probabilistic

matching with unique matches (PRL Single), and (iv) probabilistic matching with multiple matches (PRL Multi). Allowing for multiple matches means that a single observation in the x- datafile may be matched to multiple observations in the y-datafile.

Each matching method produces a distinct matched dataset, so that the matching step produces a total of 4,000 linked datasets. For each of these datasets, I then compute a variety of estimates of β, primarily the Scheuren-Winkler bias-correction estimator, and the AHL estimator that allows for multiple matches. I additionally compute three standard OLS estimators; the first using only observations that are assigned a single match by the record linking procedure, and the second using only the correctly matched pairs. The third OLS estimator is the benchmark estimator is $\hat{\beta}^{opt}$, which is calculated using the correctly linked version of the x- and y-dataset pair. Details on the implementation of these algorithms and estimation procedures can be found in the appendix.

8.1   Matching results

To evaluate the matching procedures, I compute the following statistics for each linked dataset, reported in Table 1:

- the proportion of observations in the x-datafile that are linked to at least one observation in the y-datafile (match rate),

- the total number of links made by the matching algorithm,

- the proportion of links that are incorrect (Type I error rate),

- the proportion of correct (x, y) links that are not found by the matching algorithm (Type II error rate),

- the proportion of observations whose links include the true match

For the linked datasets that contain multiple matches per observation, I report also the

21

average number of links per observation, and how often those links include the true match (Table 2).

Table 1: Average performance and SD for matching algorithms

Method	Match Rate	# Matches	Type I	Type II	P(Contains True)
ABE (Single)	0.71 (0.02)	356.50 (10.60)	0.03 (0.01)	0.26 (0.02)	0.97 (0.01)
ABE (Multi)	0.79 (0.02)	505.08 (17.30)	0.23 (0.02)	0.20 (0.02)	0.99 (0.01)
PRL (Single)	0.74 (0.02)	369.15 (9.65)	0.11 (0.02)	0.15 (0.03)	0.89 (0.02)
PRL (Multi)	0.74 (0.02)	435.65 (14.94)	0.18 (0.02)	0.23 (0.02)	0.97 (0.01)

Note: Based on 1,000 simulations. Standard deviations are reported in parentheses.

Table 2: Average performance and SD for multiple match procedures

L	1	2	3	4	5	6+
ABE Multi						
Pr(Contains True)	0.99 (0.01)	0.99 (0.01)	0.99 (0.02)	0.99 (0.07)	0.99 (0.10)	1.00 (0.00)
Pr(L=ℓ)	0.52 (0.15)	0.35 (0.16)	0.11 (0.11)	0.03 (0.03)	0.02 (0.03)	0.02 (0.01)
PRL Multi						
Pr(Contains True)	0.97 (0.01)	0.98 (0.02)	0.98 (0.06)	0.98 (0.12)	0.99 (0.05)	1.00 (0.00)
Pr(L=ℓ)	0.59 (0.21)	0.33 (0.22)	0.07 (0.11)	0.02 (0.05)	0.03 (0.06)	0.01 (0.01)

Note: Based on 1,000 simulations. Standard deviations are reported in parentheses.

Here are some observations that need to be turned into paragraphs:

- As seen in Table 1, the average match rates range between 71 and 79 percent across the various matching procedures. I also plot the distribution of match rates across replications in Figure 2. The distribution for ABE Multi lies almost entirely to the right of the mean match rate for the other procedures. The PRL Single and PRL Multi have about the same match rate, which suggests that allowing for multiple matches does not match new individuals, but instead adds additional noise. ABE Multi, on the other hand, increases match rates by matching new observations.

- ABE Multi appears to be the least discriminating of the matching procedures, with higher match rates across all replications, and with the highest number of matches.

Figure 2: Match Rates by Linking Procedure
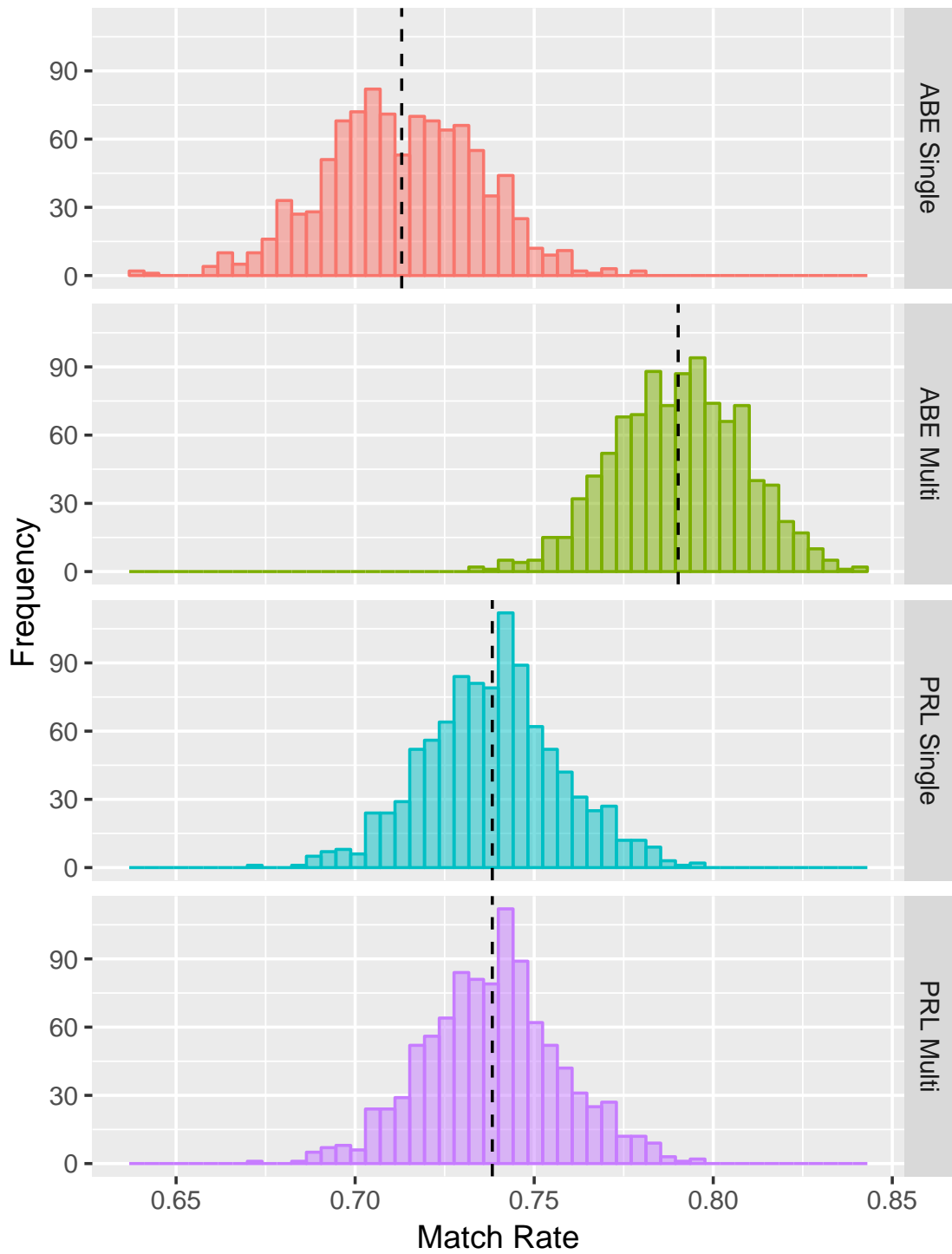
*Based on 1,000 simulations. Vertical line indicates the sample mean.

- PRL Single matches more observations than ABE Single, but those extra matches seem more often to be wrong (check with a regression?) because the Type I error rate is higher.

- The multi-match methods have higher Type I error rates by construction; despite adding false matches, they seem to improve upon the unique-match methods because they are more likely to include the true match among the possible matches.

- ABE Single is the most conservative of all the matching algorithms, with the lowest match rate and lowest type I error. Of course, the tradeoff is that it fails to identify a quarter of all possible matches.

- It seems like PRL Single is the worst performing algorithm – despite higher match rates than ABE single it seems that those matches are not likely to be correct. This may be an artifact of the threshold used to assign the matches. A higher threshold may result in matches more similar to ABE Single. It may be interesting to calculate whether these two algorithms are systematically matching different types of record pairs.

- In Table 2, we see that PRL Multi is more likely to match observations to a unique outcome; at the price of sometimes excluding the true match. ABE Multi is more likely to assign more matches, however the true match is contained with high probability. Few observations are matched with more than three different outcomes. [4]

- Based on these results, I hypothesize that ABE Multi seems to be the best performing match procedure, if multiple matches are desired. This is because the methods described in this paper typically require that the true match be included among the matches. But another question is how much noise do extra matches introduce.

- I compare also ABE Multi and PRL Multi to determine whether the probability of

---

[4]Note PR(L=l) does not add up to one because these are average of the proportions, not the average proportions – i.e. I calculate what proportion $L = l$ for each monte carlo rep, and then take the average of the proportions.

containing a true match increases by allowing for multiple matches (Table 3). In both procedures, more than half of observations are matched to a unique outcome, and those outcomes are correct 99 and 97 percent of the time for ABE Multi and PRL Multi, respectively. Essentially, allowing for multiple matches in the deterministic procedure increases the probability that the true match is contained among the possible matches.

8.2 Estimation Results

Benchmark is ...

AHL Estimator performs well....

OLS with (L=1)

When only a single matching is allowed, it is not possible to use SW method because it requires that we carry over multiple matches for each observation.

Table 3: Median Absolute Deviations for Estimators

Parameter	AHL	SW	NaiveOLS	OLSTrue	OLS(L=1)
ABE Single					
β_0	0.113	0.113	0.113	0.109	0.113
β_1	0.150	0.150	0.150	0.152	0.150
β_2	0.057	0.057	0.057	0.054	0.057
ABE Multi					
β_0	0.115	0.155	0.103	0.100	0.120
β_1	0.155	0.201	0.149	0.148	0.159
β_2	0.055	0.077	0.064	0.050	0.061
PRL Single					
β_0	0.115	0.115	0.115	0.112	0.115
β_1	0.163	0.163	0.163	0.162	0.163
β_2	0.063	0.063	0.063	0.056	0.063
PRL Multi					
β_0	0.116	0.150	0.112	0.106	0.123
β_1	0.168	0.204	0.165	0.158	0.175
β_2	0.058	0.074	0.062	0.055	0.064

Figure 3: Comparing OLS with true matches produced by matching algorithm vs. matches with L=1
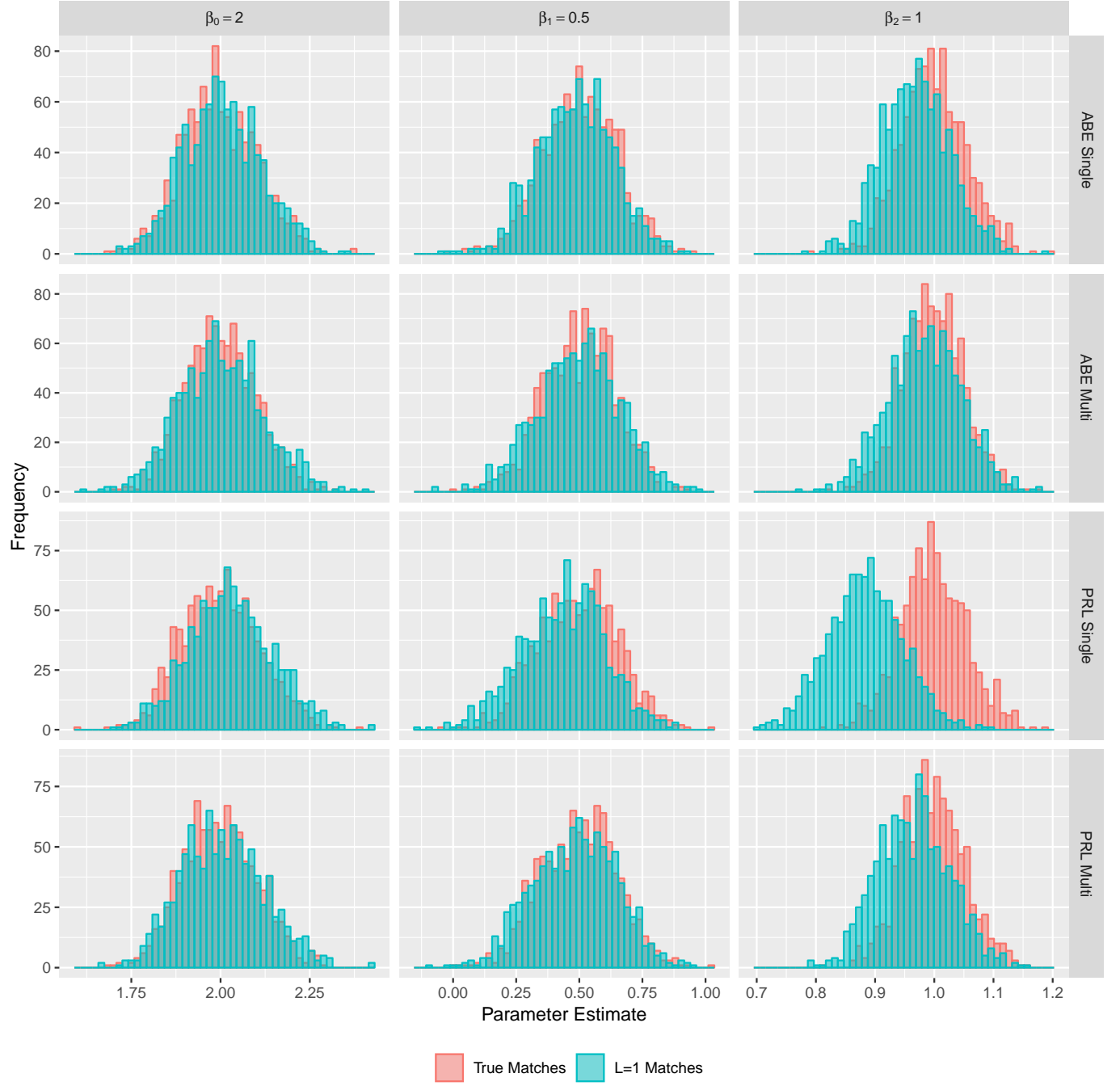
9 Discussion

To what extent my results generalize beyond the simulated data is unclear. Many arbitrary choices – name dictionary choice, uniform probability over those names – and how I introduced error in the identifying variables may impact my results in important ways. This needs to be studied in more detail, ideally with real data where a ground truth is also available.

I evaluate the performance of the matching procedures and estimation procedures using the following criteria. For the matching procedures, I calculate
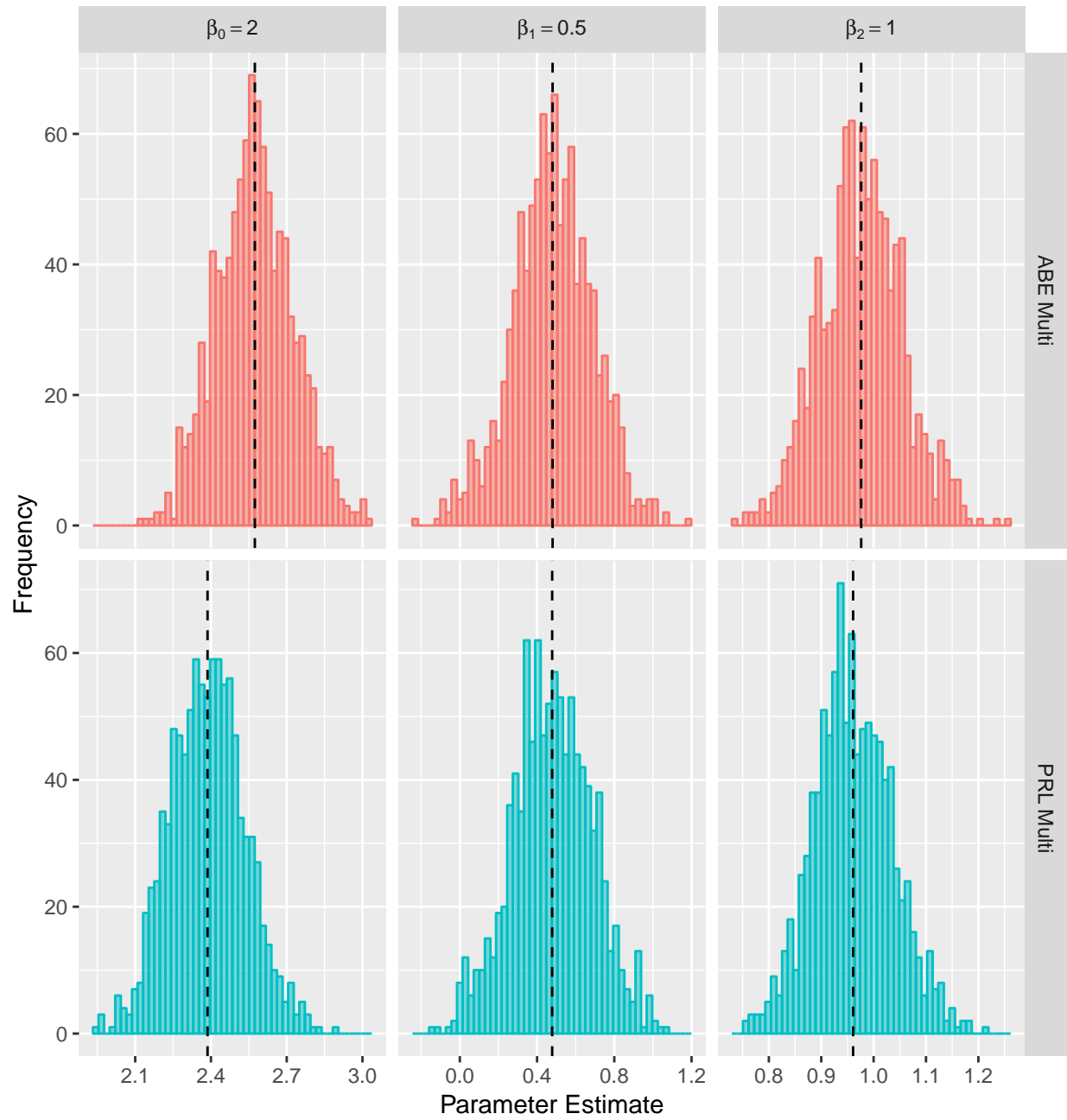
I compare estimators according to median absolute deviation. I plot a histogram of the estimators.
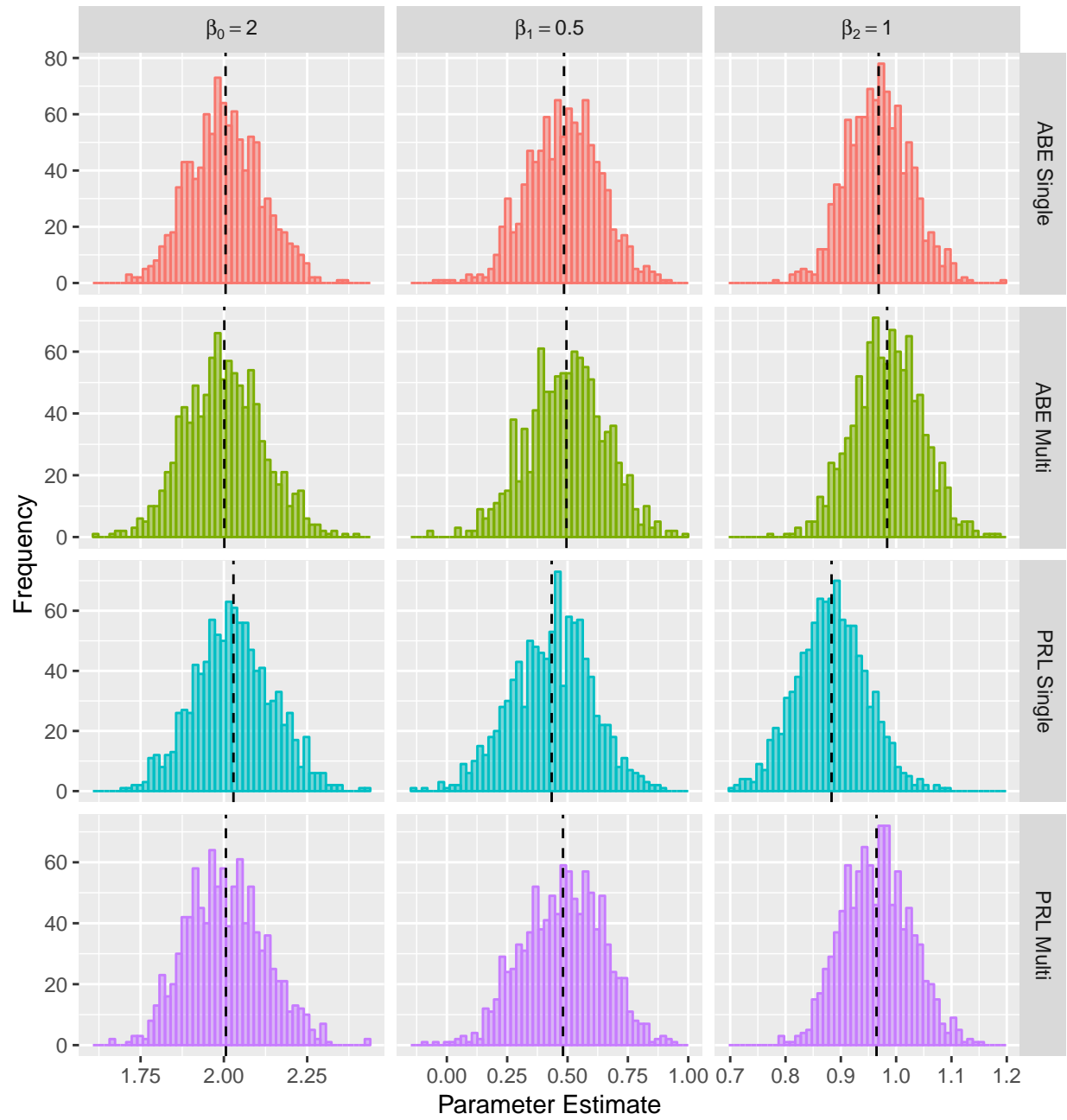
AHL Estimator
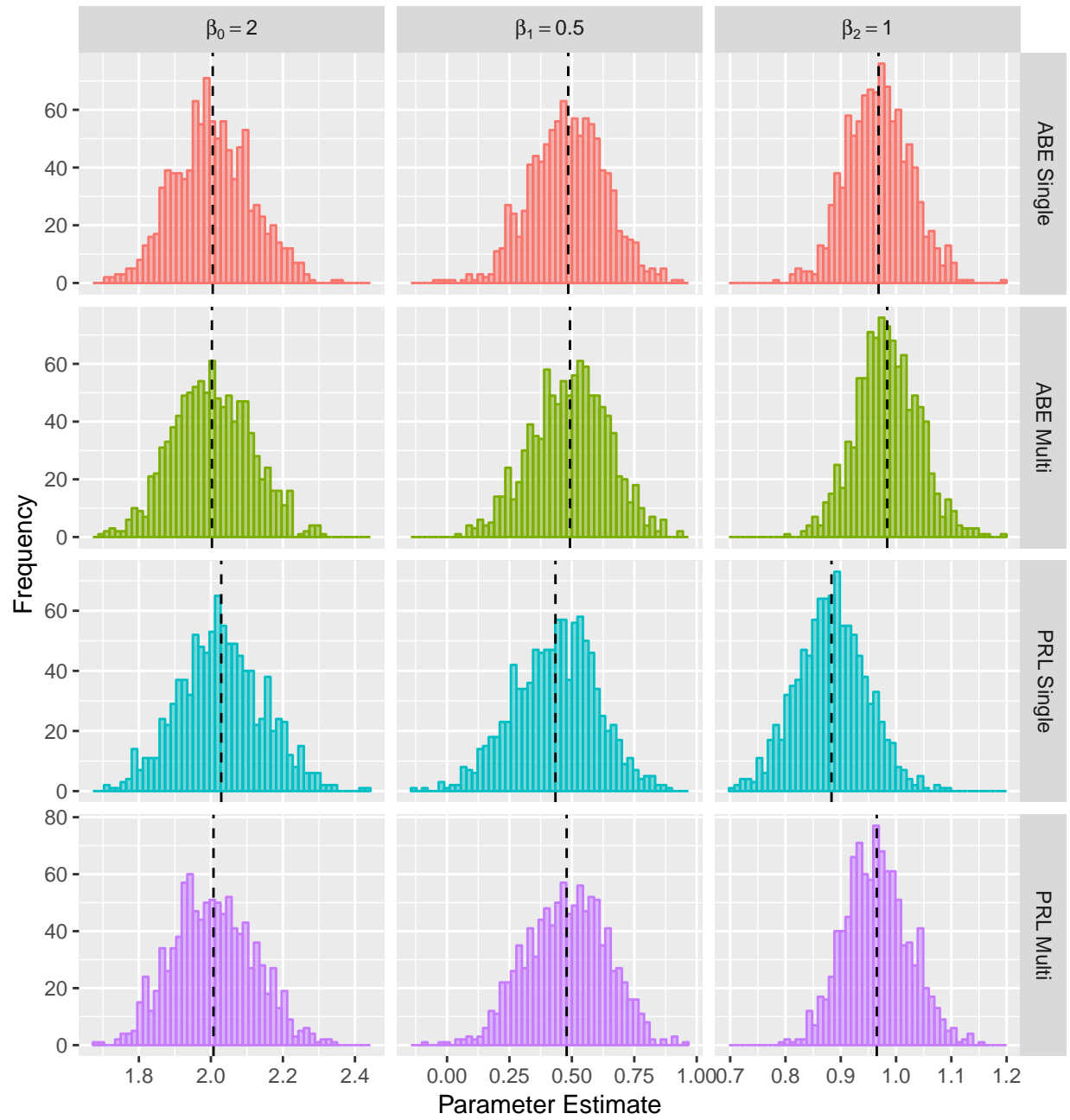
*Based on 1,000 simulations. Vertical line indicates the sample mean.

SW Estimator

*Based on 1,000 simulations. Vertical line indicates the sample mean.

OLS(L=1) Estimator

*Based on 1,000 simulations. Vertical line indicates the sample mean.

10   Appendix: Implementation Notes

Exact matching procedure.

Formulas for estimators.

Implementation of fastLink algorithm.

Here I will talk about all the nit-picky stuff, like

- what threshold level I use for the fastLink algorithm (0.6)

- what nonparametric technique I use for AHL (nearest neighbor)

- how I choose z in LL when there are multiple matches (randomly)

- how I calculate standard errors for all of the estimators (using formulas for now)

- how I standardize the variables for matching (nysiis function in R)

- I change Step 2 in the ABE algorithm to restrict the all observations with unique first name, last name, date of birth, and (x_1, x_2) combinations.

- 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.

11   Conclusion

Will write when I have results.

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.

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

Enamorado, Ted, Benjamin Fifield, and Kosuke Imai, "Using a Probabilistic Model to Assist Merging of Large-Scale Administrative Records," *American Political Science Review*, 2019, *113* (2), 353?371.

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, Harvey Goldstein, and Chris Dibben, *Methodological Developments in Data Linkage*, United States: John Wiley Sons Inc., 2015.

Herzog, Thomas N., Fritz J. Scheuren, and William E. Winkler, *Data Quality and Record Linkage Techniques*, 1st ed., Springer Publishing Company, Incorporated, 2007.

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

Larsen, Michael D and Donald B Rubin, "Iterative Automated Record Linkage Using Mixture Models," *Journal of the American Statistical Association*, 2001, *96* (453), 32?41.

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.*