Methods for analyzing linked data

Rachel Anderson[*]

This Version: October 28, 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.

[*]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).

1 Introduction

When analyzing multiple data sources with overlapping units, automated record linkage procedures offer the least cost solution for merging data. These methods allow the researcher to specify a set of matching variables that are recorded across all of the datasets, and a decision rule for linking record pairs, in order to obtain a matched dataset in a matter of minutes. When the files to be linked are large, the time saved relative to manual linking or automated linking with clerical review is immense.

In the social sciences, interest in automated record linkage methods emerged in response to the increasing availability of administrative datasets, including recently digitized historical complete count population censuses. Although these techniques originated in other fields – primarily statistics, computer science, operations research, and epidemiology – social scientists have developed their own linking methods in response to concerns about the accuracy and representativeness of data that are matched using imperfect identifiers.

This new research agenda is driven by economic historians and historical demographers, who use data with identifiers that are prone to typographical, duplication, enumeration, and digitization error. Additionally, the identifiers may be misreported – for example, ages may be rounded to integers ending with a 0 or 5 – or repeated within a sample, as might happen if several people born in the same year have the same name. Historical record linkage procedures primarily differ according to how they address these issues of data quality.

Examples of this literature include recent papers by Abramitzky et al. (2019) and Bailey et al. (2017), who compare the performance of popular historical record linkage methods to datasets matched by hand-linking or to simulated "ground truth" datasets. Although they implement different methods, both papers document a tradeoff between the false positive rate and the (true) match rate across procedures, and seem to agree that using more conservative algorithms leads to more representative data.

Other contributions to this literature include papers by Abramitzky et al. (2018) and Enamorado et al. (2019), who demonstrate how to apply probabilistic record linkage methods from statistics to match historical and large-scale survey data. These methods offer an advantage over the deterministic methods studied by Abramitzky et al. (2019) and Bailey et al. (2017), in that they can quantify the uncertainty about the matched data. Furthermore, this extra information can be used to correct for bias introduced by false matches in the Ordinary Least Squares (OLS) estimator, using methods proposed by Scheuren and Winkler (1993) and Lahiri and Larsen (2005).

The ability to use probabilistic record linkage to correct for bias in the OLS estimator illustrates how the *outputs* of different matching procedures determine which estimation methods are available for subsequent analysis. Similarly, Anderson et al. (2019) develop methods for consistently estimating Generalized Method of Moments (GMM) models using linked data that include multiple matches per observation. Unfortunately, none of these estimation methods is acknowledged in the survey papers by Abramitzky et al. (2019) and Bailey et al. (2017); however, it seems natural that the choice of which matching procedure to use should be informed by which estimation methods are available.

The goal of this paper is to build a bridge between the matching and estimation steps in the analysis of linked data. I compare the performance of different methods that incorporate different types of information, as well as extend the methods from Anderson et al. (2019) to incorporate probabilities as may be outputted from a record linkage procedure.

By comparing the estimation methods, this gives suggestion to which record linkage procedure should be used (or, better yet, which outputs of the record linkage procedure are necessary for optimal estimation). The main result is that unless probabilities can be estimated accurately, knowledge about them should not be incorporated in the estimation step.

My analysis in the paper supports support the following suggestions for analyzing linked data: if you use deterministic matching, you should allow for multiple matches and use the estimator in Anderson et al. (2019). If you use probabilistic record linkage, you should choose the match with the highest probability of being correct if it exceeds a certain threshold; otherwise you should use multiple matches because the estimated probabilities can be noisy, and can result in large weights on observations with small $\pi_{i\ell}$. When it doubt, implement all methods and compare the results!

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

In this section, I describe a simplified version of the estimation problem described in Anderson et al. (2019). Whereas Anderson et al. (2019) study how to incorporate multiple matches in a GMM framework, this paper focuses on estimating β in the linear regression model,

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

where x_i and y_i are recorded in different datasets, and must be linked using auxiliary variables that are contained in both data sources.

Formally, the data consist of observations $\{x_i, w_i\}_{i=1}^{N_x}$ in the x-datafile, and observations $\{y_j, w_j\}_{j=1}^{N_y}$ in the y-datafile. I assume that $N_y \geq N_x$, and that every x_i has a unique match in the y-datafile that satisfies the relationship in (1), but the index j that corresponds with the match is unknown. Some y_j may not correspond to any observation in the x dataset, nor satisfy the relationship in (1) for some unobserved x_j. Hence, estimating the model in

(1) requires identifying which (x_i, y_j) pairs refer to the same individuals by comparing w_i and w_j and designating matches according to some matching procedure.

Example 1. To fix ideas, consider the work of Aizer et al. (2016), who seek to estimate the impact of providing cash transfers to single mothers on the life expectancy of their children. The x-datafile consists of mothers' welfare program applications, where x_i includes a binary variable equal to 1 if person i's mother received a cash transfer, and other demographic variables. The y-datafile is a universal database of death records, which includes y_j, person j's age at death for all deaths reported to the Social Security Administration after 1965. Both of the x- and y-datafiles also contain identifiers w_i and w_j, which include first name, middle initial, last name, day, month, and year of birth, so that individuals with common names are not identified uniquely.

For the purposes of this paper, a matching procedure is defined as a set of rules used to construct a (potentially multi-valued) linking function, $\varphi : \{1, \ldots, N_x\} \to \{1, \ldots, N_y\} \cup \varnothing$, where $\varphi(i) = j$ if the ith observation in the x-datafile is matched to the jth observation in the y-datafile, and $\varphi(i) = \varnothing$ if i is not assigned a match. If w_i and w_j identify individuals uniquely and without error, then setting $\varphi(i) = j$ if and only if $w_i = w_j$, and $\varphi(i) = \varnothing$ otherwise, will correctly identify all true matches contained in the data. In most applications, however, w_i and w_j cannot be used to unambiguously differentiate true matches, so that assigning φ may include linking false matches or excluding true matches. In this case, φ may be constructed using estimates of π_{ij}, which denotes the probability that an (i, j) pair refers to a match.

Example 1 (cont'd). Aizer et al. (2016) construct φ using a deterministic linking method that allows for errors in strings and dates of birth. To account for changes in spelling and typographical errors, they convert all names into sounds using a phonetic algorithm, and measure the similarity between two individual's phonetically-spelled names using a string distance metric called SPEDIS. They assign as matches all pairs of individuals whose birthdates fall within a predetermined range and whose SPEDIS score falls below a threshold. Notably, their procedure does not enforce unique matches, so that some individuals are matched to multiple death records, and φ is a correspondence.

Although I will discuss a few of the most popular methods used to construct φ in later sections, this is not the focus of this paper. Like Anderson et al. (2019), I take φ and the matched dataset it produces as given. The matched dataset can be written in the general form

$$\mathcal{D}_n \equiv \left(x_i, w_i, \{y_{i\ell}\}_{\ell=1}^{L_i}, \{\pi_{i\ell}\}_{\ell=1}^{L_i}\right)_{i=1}^{N} \tag{2}$$

where x_i is a vector of covariates for a single observation in the x-datafile, and $\{y_{i\ell}\}_{\ell=1}^{L_i}$ are the outcomes from the y-datafile to which it is linked based on the identifying variables w_i. The variables $\{\pi_{i\ell}\}_{\ell=1}^{L_i}$ correspond to the conditional probability that a particular $y_{i\ell}$ refers to the correct match, so that $\sum_{\ell=1}^{L_i} \pi_{i\ell} = 1$. If φ was constructed without estimating $\pi_{i\ell}$, as in a deterministic matching procedure, I assume that $\pi_{i\ell} = \frac{1}{L_i}$.

Additionally, I assume that (i) the true match y_i is included among the possible matches $\{y_{i\ell}\}_{\ell=1}^{L_i}$ for all i, and that (ii) the observed x_i and $\{y_{i\ell}\}_{\ell=1}^{L_i}$ are i.i.d. draws from their marginal distributions conditional on the identifying variables w_i. Assumption (i) is necessary for identification, and Assumption (ii) is a selection on observables assumption, which requires that all individuals with the same identifying information are equally likely to be included in $\mathcal{D}_n$.

Assumption (i) needs to be verified empirically; however the simulations in Section X suggest that it is reasonable when multiple matches are allowed. Assumption (ii) requires thats all individuals with the same identifying information (such as name, age, Census block) 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 individuals are matched by income); but the OLS estimator using perfectly linked data would also be biased because of unobserved sample selection.

3 Estimation methods for linked data

In this section, I review methods from Scheuren and Winkler (1993) and Anderson et al. (2019) (henceforth referred to as the SW and AHL estimators) for estimating regression models using matched data with the structure (2), and establish conditions under which they are equivalent. Although there are other methods for analyzing linked data, none is compatible with the setup described in Section 2. For example, Lahiri and Larsen (2005) develop a version of the SW estimator based on the assumption that each observation in the y-datafile is generated by model in (1), and that its corresponding value of x appears in the x-datafile. This differs from the problem considered in this paper, which allows for $N_y > N_x$, and assumes that the relationship in (1) holds only for observations in the x-datafile.

3.1 Scheuren and Winkler (1993)

Building upon the work Neter et al. (1965), Scheuren and Winkler (1993) study how to correct for bias introduced using incorrectly linked data in linear regression models. Their methods assume that the data consist of observations $(x_i, z_i)_{i=1}^{N}$, so that each x_i is linked with a single outcome z_i that may or may not correspond to the true y_i. Specifically,

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

and $\sum_{j=1}^{N_y} q_{ij} = 1,\ i = 1, \ldots, N$, where N_y is the size of the y-datafile and N is the size of the matched dataset. Estimating (1) using z_i as the dependent variable yields the naive least squares estimator,

$$\hat{\beta}_N = (X'X)^{-1}X'z \tag{3}$$

which is biased, because $E[z_i] = E\left[q_{ii}y_i + \sum_{j\neq i} q_{ij}y_j\right] \neq E[y_i]$ if $q_{ii} \neq 1$ for some i. Denoting $q_i = (q_{i1}, \ldots, q_{iN_y})'$, Scheuren and Winkler (1993) derive the bias of $\hat{\beta}_N$ conditional on the observed values of y,

$$\text{bias}(\hat{\beta}_N|y) = E[(\hat{\beta}_N - \beta)|y] = (X'X)^{-1}X'B \tag{4}$$

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 true response y_i.

Observing (4), Scheuren and Winkler (1993) propose estimating $\hat{B}$ using the first and second highest elements of q_i, and their corresponding values y_j to compute

$$\hat{B}_i^{TR} = (q_{ij_1} - 1)y_{ij_1} + q_{ij_2}y_{ij_2} \tag{5}$$

for each i, and then using it to correct for the bias in $\hat{\beta}_N$ as follows,

$$\hat{\beta}_{SW} = \hat{\beta}_N - (X'X)^{-1}X'\hat{B}^{TR} \tag{6}$$

In principle, $\hat{B}^{TR}$ can incorporate any number of elements of q_i, however Scheuren and Winkler (1993) show that if q_{ij1} is sufficiently high for all i, then the truncation with two links results in a very small bias.

The SW estimator has two important caveats. The first is that it requires that the y value associated with the largest element in q_i corresponds with the true outcome y_i, so that errors in z_i result from random assignment rules or requiring that no two values x_i and x_j are linked to the same value of y. The second is that constructing $\hat{\beta}_{SW}$ requires knowledge of q_i and the corresponding elements of y, and so it cannot be applied to data linked with deterministic methods. Even if estimates of q_{ij} are available, $\hat{\beta}_{SW}$ will be biased if the estimates $\widehat{q_{ij}}$ are correlated with x or y, which occurs if x or y is correlated with errors in the matching variables. This assumption is different from Assumption (ii) in Section 2, and often fails in economic application, such as in Nix and Qian (2015), where y measures whether a person's recorded ethnicity changes between Census years, and changes in first and last name (the matching variables) are strongly correlated with y.

3.2 Anderson et al. (2019)

Unlike the SW estimator, the estimator proposed in Anderson et al. (2019) is unbiased under Assumptions (i) and (ii) from Section 2, and does not require estimates of q_{ij}. Recall that the assumptions are that (i) the true match of x_i is included among the matches $\{y_{i\ell}\}_{\ell=1}^{L_i}$ for all i, and (ii) x_i and y_i are random samples conditional on the matching variables. Assumption (ii) allows x and y to be correlated with errors in matching, so long as that dependence is fully captured by w; hence the AHL estimator would be valid in the Nix and Qian (2015) example if ethnicity is included among the matching variables.

The goal of Anderson et al. (2019) is to estimate θ_0 that satisfies the model

$$E_0\left[m\left(y_i, x_i; \theta_0\right)\right] = 0 \tag{7}$$

where y_i and x_i are vectors or scalars of data for an individual i, the function $m(\cdot)$ is known, and the expectation is taken with respect to the joint distribution $f_0(y, x)$. The data consist of observations $\left(x_i, w_i, \{y_{i\ell}\}_{\ell=1}^{L_i}\right)_{i=1}^{N}$, where the x_i and $y_{i\ell}$ are recorded in distinct datasets and matched according to the identifier w_i. They assume $L_i > 1$ for some i, so that the identity of the outcome that generates the relationship in (7) is unknown.

Under assumptions (i) and (ii) in Section 2, (7) can be rewritten,

$$E_0[m(y_i, x_i; \theta)] = E\left[\sum_{\ell=1}^{L_i} m(y_{i\ell}, x_i; \theta)\right] - E\left[(L_i - 1)g(w_i, L_i; x_i; \theta)\right]$$

$$g(w_i, L_i, x_i; \theta) = E\left[m(y_i, x_i; \theta)|\ w_i, L_i\right]$$

so if g is known or can be estimated consistently, a sample version of (7) can be constructed as follows,

$$\overline{m}_n(\theta, \hat{g}) = \frac{1}{N}\sum_{i=1}^{N}\sum_{\ell=1}^{L_i} m(y_{i\ell}, x_i; \theta) - \frac{1}{N}\sum_{i=1}^{N}(L_i - 1)\hat{g}(w_i, L_i, x_i; \theta) \tag{8}$$

which is in general a two-step procedure, where $\hat{g}$ is estimated using nonparametric methods such as k-Nearest Neighbors or local polynomial regression in the first step. The GMM estimator applied to (8) is consistent and asymptotically normal under the regularity conditions described in Anderson et al. (2019).

For the linear regression model in (1), the AHL estimator can be computed by applying OLS to the transformed regression model,

$$\sum_{\ell=1}^{L_i} y_{i\ell} - (L_i - 1)\hat{g}(w_i, L_i) = x_i'\beta + u_i \tag{9}$$

where $\hat{g}(w_i, L_i)$ is a (possibly nonparametric) estimator of $E[y_{i\ell}|w_i, L_i]$, $u_i = \varepsilon_i + \sum_{\ell=1}^{L_i} \nu_{i\ell}$, and $\nu_{i\ell} = y_{i\ell} - \hat{g}(w_i, L_i)$. If, additionally, $E[\varepsilon_i^2|x_i, w_i, L_i] = \sigma_\varepsilon^2$ and $E[\nu_{i\ell}^2|x_i, w_i, L_i] = \sigma_\nu^2$ then the efficient estimator is weighted least squares,

$$\hat{\beta}^{WLS} = \left(\sum_{i=1}^{N}\frac{x_i x_i'}{\sigma(X_i)}\right)^{-1}\left(\sum_{i=1}^{N}\frac{x_i}{\sigma(X_i)}\left(\sum_{\ell=1}^{L_i} y_{i\ell} - (L_i - 1)g(w_i, L_i)\right)\right) \tag{10}$$

where $\sigma(X_i) = \sigma_\varepsilon^2 + (L_i - 1)\sigma_\nu^2$.

Unfortunately, the performance of the AHL estimator depends on the accuracy of

$\hat{g}(w_i, L_i)$, which is a function of a potentially high-dimensional vector w_i that may contain string or categorical variables. However, if consider adding an assumption (iii) that $E[m(y_i, x_i; \theta)|L_i = \ell] = E_0[m(y_i, x_i; \theta)]$, then

$$E_0[m(y_i, x_i; \theta)] = E\left[\sum_{\ell=1}^{L_i} m(y_{i\ell}, x_i; \theta)\middle| L_i = 2\right] - E\left[\sum_{\ell=1}^{L_i} m(y_{i\ell}, x_i; \theta)\middle| L_i = 3\right] \tag{11}$$

and

$$E_0[m(y_i, x_i; \theta)] = E[m(y_{i\ell}, x_i; \theta)|L_i = 1] \tag{12}$$

So that we apply GMM to (11) and (12) as the new moment conditions. Applying this to the linear regression model (1), yields the following moment conditions,

$$\begin{bmatrix} \frac{2}{N_{L_2}} \sum_{i=1}^{N_{L_2}} (y_{i1} + y_{i2} - x_i'\beta)x_i - \frac{1}{N_{L_3}} \sum_{i=1}^{N_{L_3}} (y_{i1} + y_{i2} + y_{i3} - x_i'\beta)x_i \\ \frac{1}{N_{L_1}} \sum_{i=1}^{N_{L_1}} (y_i - x_i'\beta)x_i \end{bmatrix} = 0 \tag{13}$$

where N_{L_ℓ} is the number of observations matched to ℓ observations. In practice, the precision of the estimator depends on the number of observations that are linked to two and three outcomes, however the estimator based on these moments should be asymptotically consistent.

3.3 Comparing $\hat{\beta}_{SW}$ and $\hat{\beta}_{AHL}$

There is a natural parallel between $\hat{\beta}_{SW}$ and $\hat{\beta}_{AHL}$ when we consider data of the form in (2), and we assume that L_i is the number of links whose q_{ij} exceed a threshold $\bar{q}$, and that the conditional probabilities $\pi_{i\ell} = \frac{q_{i\ell}}{\sum_{\ell=1}^{L_i} q_{i\ell}}$. Let also $y_{i\ell^*}$ refer to the element of $\{y_{i\ell}\}_{\ell=1}^{L_i}$ that is associated with the highest value of $\{\pi_{i\ell}\}_{\ell=1}^{L_i}$. Then, we can write $\hat{\beta}_{SW}$ as the OLS estimator for the model

$$z_i - \hat{B}_i = x_i'\beta + \varepsilon_i \tag{14}$$

with $\hat{B}_i = \sum_{\ell=1}^{L_i} \pi_{i\ell} y_{i\ell} - y_{i\ell^*}$. Note that the Scheuren and Winkler (1993) method assumes that the $y_{i\ell}$ with the highest value of $\pi_{i\ell}$ is the true match.[1]

The AHL estimator can be written in the form (14) with

$$\hat{B}_i = z_i - \sum_{\ell=1}^{L_i} y_{i\ell} + (L_i - 1)\hat{g}(w_i, L_i) \tag{15}$$

so that $\hat{\beta}^{AHL}$ and $\hat{\beta}^{SW}$ differ only in their choice of B_i. Alternatively, $\hat{\beta}_{SW}$ can be written in the form of $\hat{\beta}_{AHL}$, as in equation (7), by setting

$$\hat{g}(w_i, L_i) = \frac{1}{L_i - 1}\left(\sum_{\ell \neq \ell^*} y_{i\ell} + y_{i\ell^*}\right) \tag{16}$$

Since $\hat{g}(\cdot)$ as written in (16) ignores information about w_i, and assumes that $y_{i\ell^*}$ is the correct match, the AHL estimator may perform better if $y_{i\ell^*}$ is not the true match, informative $\pi_{i\ell}$ are not available, or w_i contains information about the conditional mean of the $y_{i\ell}$ drawn from the incorrect distribution. However, if reliable estimates of $\pi_{i\ell}$ are available, it may be possible to improve the AHL estimator by incorporating this information. I explore this possibility in the next section.

4 Incorporating probabilities in the AHL estimator

I begin by considering a simplified version of the AHL (2019) problem, based on the observation that

$$E[m(y_{i\ell}, x_i; \theta)] = \begin{cases} 0 & \text{if } y_{i\ell} = y_i \\ g(w_i, L_i, x_i; \theta) & \text{if } y_{i\ell} \neq y_i \end{cases}$$

[1] There are reasons that $z_i \neq y_{i\ell^*}$. This happens, for example, if z_i is assigned at random according to the probabilities $\pi_{i\ell}$ or if a one-to-one matching is enforced such that no single y is assigned to two distinct values of x

If $\hat{g}$ can be estimated consistently, or $g(\cdot)$ is a constant function, this problem can be reduced to estimating the mean using observations $\{X_{i\ell}\}_{\ell=1}^{L_i}$, where each $X_{i\ell}$ is drawn from the correct distribution with probability $\pi_{i\ell}$ and drawn from the incorrect distribution with probability $(1-\pi_{i\ell})$. Under Assumption (i), exactly one of the $X_{i\ell}$ is drawn from the correct distribution, so that $\sum_{\ell=1}^{L_i} X_{i\ell} = \mu + (L_i - 1)\kappa$, where $\mu = 0$ in the above example, and $\kappa = g(\cdot)$.

4.1 Estimating the mean

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,

$$\hat{\mu}_1 = \frac{X_1}{\pi} - \frac{1-\pi}{\pi}\kappa \tag{17}$$

and, similarly, we can construct an unbiased estimator using only X_2,

$$\hat{\mu_2} = \frac{X_2}{1-\pi} - \frac{\pi}{1-\pi}\kappa \tag{18}$$

Any unbiased linear estimator $\hat{\mu}$ that uses both X_1 and X_2 can be written as a combination of $\hat{\mu}_1$ and $\hat{\mu}_2$ (see Lemma 1 in the Appendix for a proof), so finding the minimum variance, unbiased linear estimator $\hat{\mu}$ requires minimizing

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

which is solved by

$$d^* = \frac{\text{Var}\,(\hat{\mu}_2) - \text{Cov}(\hat{\mu}_1, \hat{\mu}_2)}{\text{Var}\,(\hat{\mu}_1) + \text{Var}\,(\hat{\mu}_2) - 2\text{Cov}(\hat{\mu}_1, \hat{\mu}_2)} \tag{19}$$

where

$$\text{Var}\,(\hat{\mu}_1) = \frac{\text{Var}(X_1)}{\pi^2} = \frac{1}{\pi^2}\left(\pi\sigma^2 + (1-\pi)\omega^2 + \pi(1-\pi)(\mu-\kappa)^2\right) \tag{20}$$

$$\text{Var}\,(\hat{\mu}_2) = \frac{\text{Var}\,(X_2)}{(1-\pi)^2} = \frac{1}{(1-\pi)^2}\left((1-\pi)\sigma^2 + \pi\omega^2 + \pi(1-\pi)(\mu-\kappa)^2\right) \tag{21}$$

$$\text{Cov}(\hat{\mu}_1, \hat{\mu}_2) = \frac{\text{Cov}(X_1, X_2)}{\pi(1-\pi)} = \frac{1}{\pi(1-\pi)}\left((1-\pi^2-(1-\pi)^2)\mu\kappa - \pi(1-\pi)(\mu^2+\kappa^2)\right) \tag{22}$$

Derivations of these formulas are in the appendix.

Thus, the minimum variance unbiased estimator is

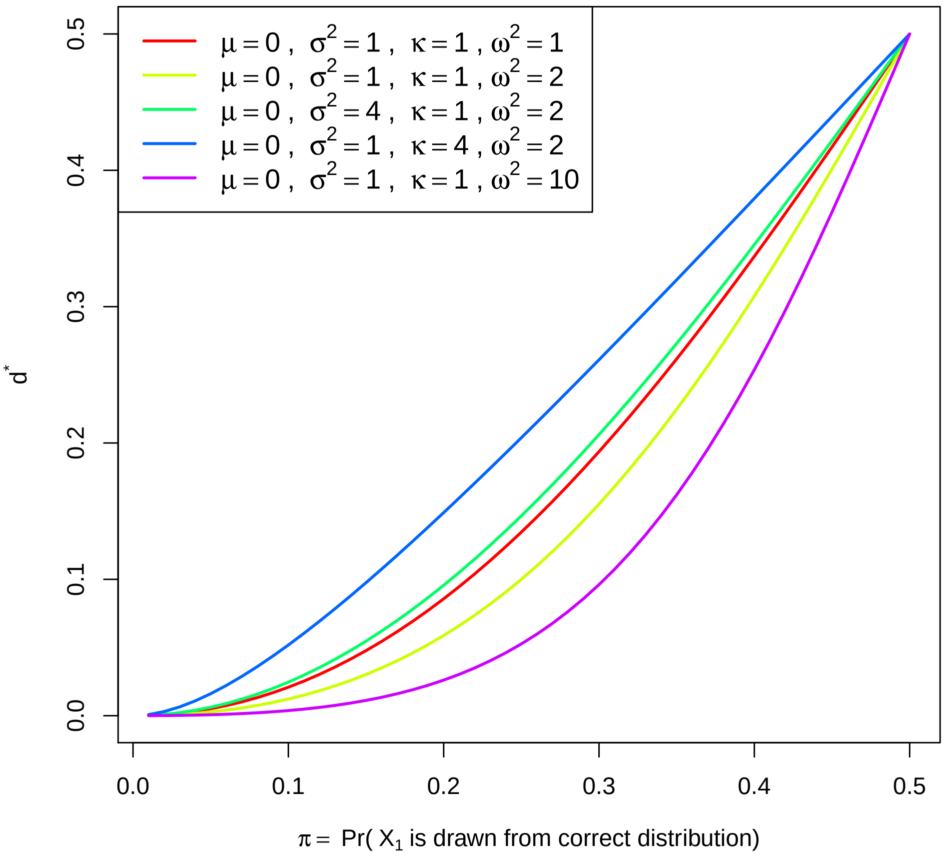
$$\hat{\mu}^* \equiv \hat{\mu}(d^*) = d^*\hat{\mu}_1 + (1-d^*)\hat{\mu}_2 \tag{23}$$

where d^* is defined as in (19). Note that d^* is strictly increasing in π, but at a rate that depends on σ^2, ω^2, μ, and κ. Intuitively, this means that the optimal estimator $\hat{\mu}^*$ puts more weight on the observation that is most likely to be correct.

Figure 1 plots the optimal d^* for $\pi \in [0, 0.5]$, since the solution is symmetric in π when $L = 2$. When $\pi = 0.5$, $\text{Var}\,(\hat{\mu}_1) = \text{Var}\,(\hat{\mu}_2)$ so that $d^* = 0.5$, regardless of the other parameter values. When the variance of both the correct and incorrect distributions are the same (i.e., $\sigma^2 = \omega^2$), then $\text{Var}\,(X_1) = \text{Var}\,(X_2)$, and differences in d^* reflect only changes in π. When $\sigma^2 \neq \omega^2$, the optimal d^* puts additional weight (relative to the equal variance case) on the estimator based on the X_i that is more likely to come from the lower variance distribution. The curve with $\sigma^2 = 1$ and $\omega^2 = 10$ is the extreme version of this scenario, and represents what may happen if κ is estimated imprecisely. The resulting d^* assigns very low weight to the observation that is more likely drawn from the incorrect distribution.

More generally, the estimator $\hat{\mu}^*$ can be computed for a sample of observations $(X_{i1}, X_{i2})_{i=1}^N$,

Figure 1: Optimal d^* as a function of π and $\sigma^2, \omega^2, \mu, \kappa$



where X_{i1} is drawn from F_X with probability π_i, and X_{i2} is drawn from F_X with probability $1-\pi_i$. In this case, d^* is calculated according to (19) using,

$$\hat{\mu}_1 = \frac{1}{N}\sum_{i=1}^{N}\frac{X_{i1}}{\pi_i} - \frac{1-\pi_i}{\pi_i}\kappa \qquad \text{Var}\left(\hat{\mu}_1\right) = \frac{1}{N^2}\sum_{i=1}^{N}\frac{g(\pi_i;\theta)}{\pi_i^2}$$

$$\hat{\mu}_2 = \frac{1}{N}\sum_{i=1}^{N}\frac{X_{i2}}{1-\pi_i} - \frac{\pi_i}{1-\pi_i}\kappa \qquad \text{Var}\left(\hat{\mu}_2\right) = \frac{1}{N^2}\sum_{i=1}^{N}\frac{g(1-\pi_i;\theta)}{(1-\pi_i)^2}$$

$$\text{Cov}(\hat{\mu}_1,\hat{\mu}_2) = \frac{1}{N^2}\sum_{i=1}^{N}\frac{\text{Cov}(X_{1i},X_{2i})}{\pi_i(1-\pi_i)}$$

where $g(p;\theta) = p\sigma^2 + (1-p)\omega^2 + p(1-p)(\mu-\kappa)^2$ is the variance of $X_{i\ell}$, for $\ell \in \{1,2\}$, that has probability p of being drawn from the correct distribution.

4.2 Errors in $\hat{\pi}$

The construction of $\hat{\mu}^*$ was based on the assumption that π was known; this section studies the performance of $\hat{\mu}^*$ when only an estimate $\hat{\pi}$ is available.

Suppose that we have an i.i.d. sample of observations $(X_{i1}, X_{i2})_{i=1}^{N}$, where X_{i1} is drawn from F_X with probability π, and X_{i2} is drawn from F_X with probability $1-\pi$. In the context of record linkage, X_{i1} and X_{i2} may refer to two possible matches for an observation, and π is the probability that X_{i1} is the true match. The estimated probabilities $\hat{\pi}$ may be obtained from a probabilistic record linkage procedure or reflect prior knowledge about the matching application[2].

As observed in Anderson et al. (2019), when π is unknown, it is possible to construct

[2] For example, $\hat{\pi}$ may reflect the econometrician's belief that "Alicia" is more likely than "Alex" to refer to the true match of an individual named "Ali".

an unbiased linear estimator of $\hat{\mu}$ by weighting all observations equally,

$$\hat{\mu}^{AHL} = \frac{1}{N}\sum_{i=1}^{N} X_{i1} + \frac{1}{N}\sum_{i=1}^{N} X_{i2} - \kappa \tag{24}$$

The variance of this estimator is

$$\text{Var}\left(\hat{\mu}^{AHL}\right) = \frac{\text{Var}\left(X_{1i} + X_{2i}\right)}{N} \tag{25}$$

Note that $\text{Var}\left(\hat{\mu}^{AHL}\right) = \text{Var}\left(\hat{\mu}(\pi)\right) = \text{Var}\left(\pi\hat{\mu}_1 + (1-\pi)\hat{\mu}_2\right)$, so that $\text{Var}\left(\hat{\mu}^{AHL}\right) \geq \text{Var}\left(\hat{\mu}^*\right)$ if π is known, with equality holding if and only if $\pi = 0.5$.

Since $\hat{\mu}^{AHL}$ is unbiased regardless of the beliefs $\hat{\pi}$, it is interesting to study whether $\hat{\mu}^*$ continues to minimize the mean squared error when beliefs about π are misspecified, i.e. $\hat{\pi} \neq \pi$. Unless $\hat{\pi} = 0.5$, the estimator $\hat{\mu}^*$ that uses $\hat{\mu}_1, \hat{\mu}_2$, and d^* based on incorrect beliefs $\hat{\pi}$ will be biased. For example, if $(\mu, \sigma^2, \kappa, \omega^2) = (0, 1, 1, 2)$ and $\pi = 0.6$, but the econometrician believes $\hat{\pi} = 0.9$, then

$$\hat{\mu}_1 = \frac{1}{N}\sum_{i=1}^{N} \frac{X_{i1}}{\hat{\pi}} - \frac{1-\hat{\pi}}{\hat{\pi}} = \frac{1}{N}\sum_{i=1}^{N} \frac{10}{9}X_1 - \frac{1}{9}$$
$$\hat{\mu}_2 = \frac{1}{N}\sum_{i=1}^{N} \frac{X_2}{1-\hat{\pi}} - \frac{\hat{\pi}}{1-\hat{\pi}} = \frac{1}{N}\sum_{i=1}^{N} 10X_2 - 9$$

both of which are biased, because $E[\hat{\mu}_1] = \frac{1}{3}$ and $E[\hat{\mu}_2] = -3$. Similarly, using $\hat{\pi}$ instead of π in (20)-(22) to calculate $\text{Var}\left(\hat{\mu}_1\right)$, $\text{Var}\left(\hat{\mu}_2\right)$, and $\text{Cov}(\hat{\mu}_1, \hat{\mu}_2)$ results in choosing $d^* = 0.987$, and $\text{Bias}(\hat{\mu}^*) = 0.292$ and $\text{Var}\left(\hat{\mu}^*\right) = \frac{1.94}{N}$.

By comparison, $\text{Bias}(\hat{\mu}^{AHL}) = 0$ and $\text{Var}\left(\hat{\mu}^{AHL}\right) = \frac{3}{N}$, so we can solve for N such that $MSE_n(\hat{\mu}^{AHL}) < MSE_n(\hat{\mu}^*)$ for all $n \geq N$:

$$0.292^2 + \frac{1.94}{N} = \frac{3}{N} \implies N = 12.43$$

This example suggests that for fixed $\theta = (\mu, \sigma^2, \kappa, \omega^2)$ and N, we can compare the ratio of $MSE_n(\hat{\mu}^*; \theta)/MSE_n(\hat{\mu}^{AHL}; \theta)$ for different values of Bias$(\hat{\pi})$. Alternatively, for a fixed value of Bias$(\hat{\pi})$, we can calculate the minimum sample size N such that it is more efficient to use $\hat{\mu}^{AHL}$.

Figures 2 and 3 plot the bias and variance of $\hat{\mu}^*$ as a function of the mis-specified beliefs $\hat{\pi}$ for different values of θ. The bias is quadratic in $|\hat{\pi} - \pi|$, with zero bias at $\hat{\pi} = \pi$ and $\hat{\pi} = 0.5$. The variance of $\hat{\mu}^*$ is not minimized at $\hat{\pi} = \pi$, but at some value determined by σ^2, ω^2, $(\mu - \kappa)^2$, and Bias$(\hat{\pi})$. The variance term is less interesting than the bias, because $\text{Var}\,(\hat{\mu}^*) \to 0$ as $N \to \infty$, whereas the bias does not disappear.

This issue is reflected in Tables 1-3, which display the ratio of the $MSE_N(\hat{\mu}^{AHL}; \theta)/MSE_N(\hat{\mu}^*; \theta)$ for $N = 10, 100$, and 1000, when $\hat{\mu^*}$ is calculated for different values of $\hat{\pi}$. Although the values in these tables are calculated for $\theta = (\mu, \sigma^2, \kappa, \omega^2) = (0, 1, 1, 2)$ the same pattern of results appears for other parameter combinations included in the Appendix.

Although no economist would use $N = 10$ in practice, Table 1 illustrates how, even in small samples, the AHL estimator can be more efficient than $\hat{\mu}^*$ for incorrect beliefs such that $|\hat{\pi} - \pi| > 0.35$. For $N = 100$, this tolerance for error in $\hat{\pi}$ decreases to $|\hat{\pi} - \pi| > 0.15$; and, for $N = 1,000$, $\hat{\mu}^*$ outperforms $\hat{\mu}^{AHL}$ only if $\hat{\pi} = \pi$. This pattern may suggests that incorporating knowledge about π offers potential efficiency gains for estimators applied to small samples, but that the potential gains, as well as the tolerance for errors in $\hat{\pi}$, decrease with sample size. Before this conclusion can be generalized, however, the MSE ratio needs to be studied for heterogenous π_i and $\hat{\pi}_i$, and $L > 2$ observations.

4.3 Incorporating π in linear regression

Suppose we have two matches $\{y_{i1}, y_{i2}\}_{i=1}^N$ for each observation. We get the same conditions for unbiasedness of the OLS estimator if we consider using a linear combination of

Figure 2: Bias of $\hat{\mu}^*$ as a function of $\hat{\pi}$

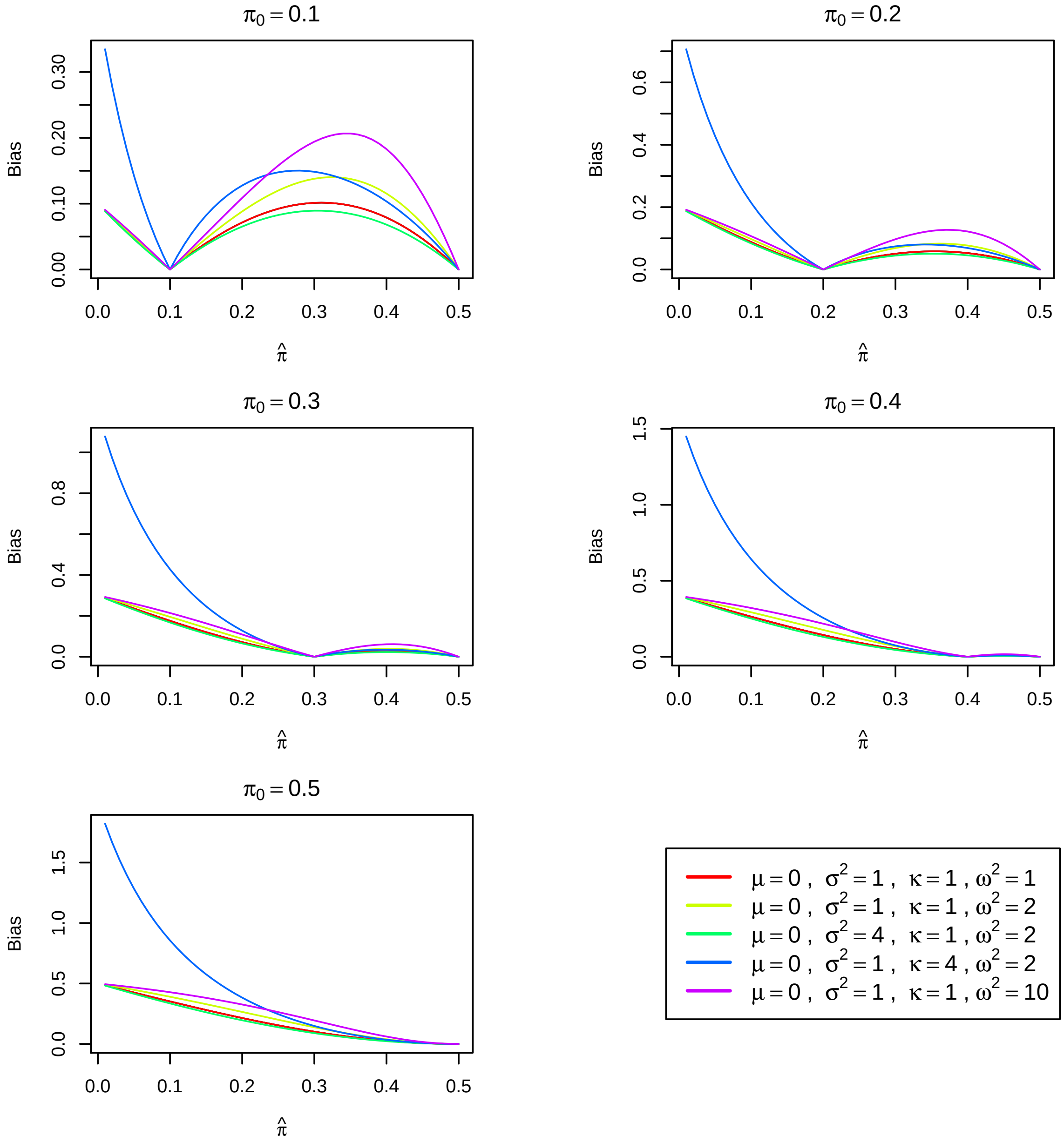


Figure 3: Variance of $\hat{\mu}^*$ as a function of $\hat{\pi}$ with $N = 1$

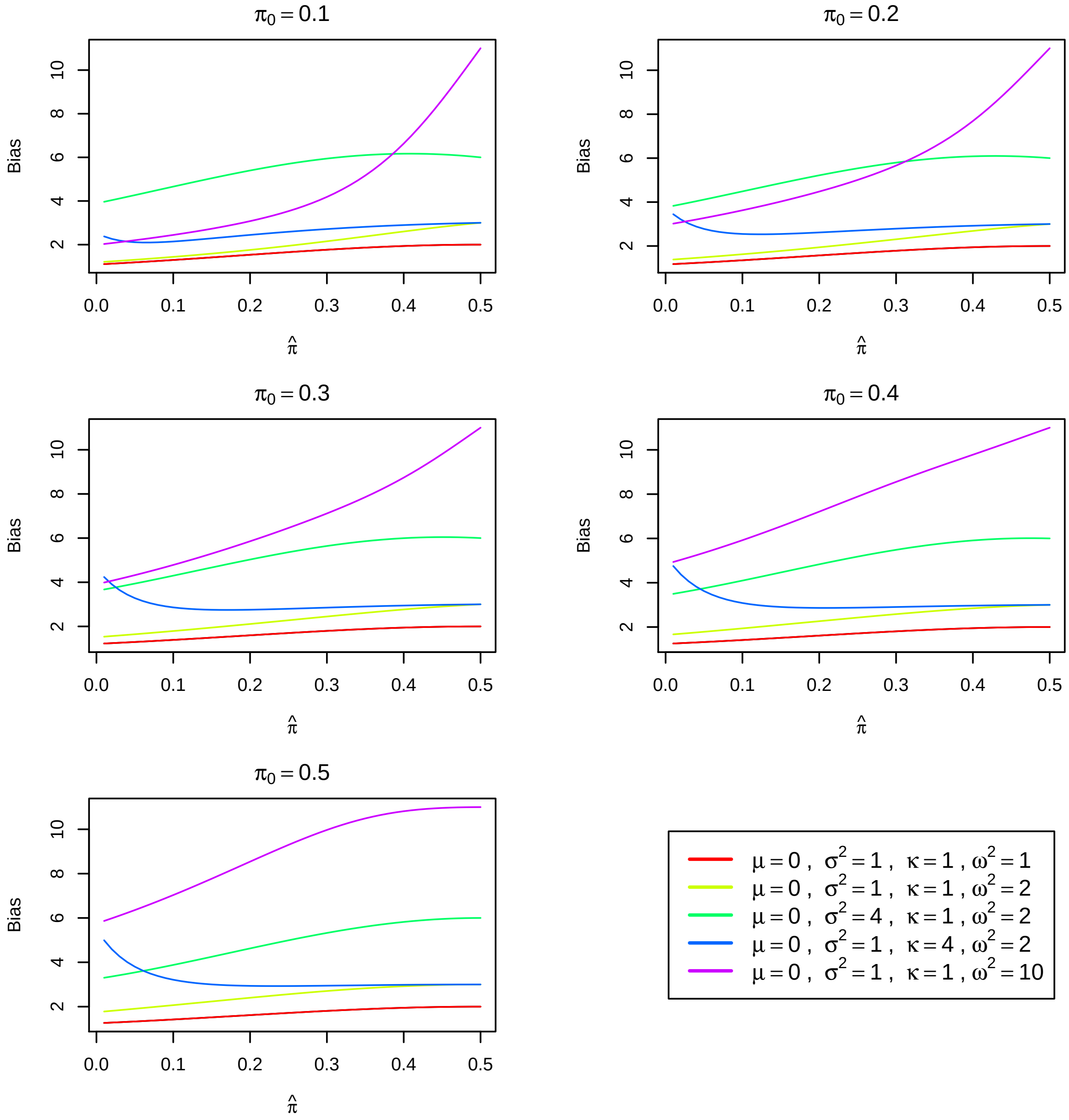


Table 1: MSE ratio for $\hat{\mu}^*$ and $\hat{\mu}^{AHL}$ for $N = 10$ and $(\mu, \sigma^2, \kappa, \omega^2) = (0,1,1,2)$

	$\hat{\pi}$									
π	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
0.1	2.253	2.085	1.863	1.635	1.436	1.280	1.170	1.097	1.047	1
0.2	1.756	1.744	1.668	1.545	1.406	1.275	1.169	1.093	1.040	1
0.3	1.325	1.381	1.398	1.371	1.308	1.225	1.144	1.078	1.031	1
0.4	1.001	1.073	1.132	1.165	1.167	1.141	1.098	1.054	1.020	1
0.5	0.768	0.836	0.905	0.967	1.014	1.036	1.036	1.022	1.006	1
0.6	0.600	0.658	0.725	0.796	0.866	0.924	0.963	0.983	0.991	1
0.7	0.479	0.527	0.586	0.656	0.734	0.814	0.885	0.938	0.974	1
0.8	0.389	0.428	0.479	0.543	0.622	0.713	0.807	0.891	0.955	1
0.9	0.321	0.354	0.397	0.454	0.529	0.623	0.732	0.842	0.935	1

Table 2: MSE ratio for $\hat{\mu}^*$ and $\hat{\mu}^{AHL}$ for $N = 100$ and $(\mu, \sigma^2, \kappa, \omega^2) = (0,1,1,2)$

	$\hat{\pi}$									
π	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
0.1	1.930	2.085	1.658	1.183	0.886	0.740	0.703	0.763	0.910	1
0.2	0.808	1.164	1.502	1.545	1.317	1.079	0.944	0.915	0.966	1
0.3	0.383	0.536	0.763	1.039	1.230	1.225	1.115	1.029	1.004	1
0.4	0.216	0.286	0.394	0.558	0.776	0.981	1.072	1.054	1.017	1
0.5	0.137	0.173	0.230	0.319	0.457	0.651	0.855	0.977	1.003	1
0.6	0.094	0.116	0.148	0.200	0.285	0.423	0.622	0.837	0.966	1
0.7	0.068	0.082	0.103	0.136	0.190	0.285	0.446	0.683	0.909	1
0.8	0.052	0.061	0.075	0.098	0.135	0.201	0.325	0.546	0.839	1
0.9	0.041	0.047	0.057	0.073	0.100	0.148	0.243	0.435	0.764	1

Table 3: MSE ratio for $\hat{\mu}^*$ and $\hat{\mu}^{AHL}$ for $N = 1000$ and $(\mu, \sigma^2, \kappa, \omega^2) = (0,1,1,2)$

	$\hat{\pi}$									
π	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
0.1	0.794	2.085	0.790	0.314	0.184	0.142	0.141	0.188	0.394	1
0.2	0.126	0.269	0.753	1.545	0.807	0.425	0.322	0.349	0.565	1
0.3	0.047	0.075	0.138	0.303	0.774	1.225	0.890	0.706	0.793	1
0.4	0.024	0.034	0.052	0.090	0.178	0.409	0.862	1.054	0.987	1
0.5	0.015	0.019	0.027	0.041	0.070	0.138	0.311	0.682	0.975	1
0.6	0.010	0.012	0.017	0.024	0.037	0.066	0.137	0.337	0.769	1
0.7	0.007	0.009	0.011	0.015	0.023	0.038	0.075	0.183	0.545	1
0.8	0.005	0.006	0.008	0.011	0.015	0.025	0.047	0.112	0.380	1
0.9	0.004	0.005	0.006	0.008	0.011	0.017	0.032	0.075	0.271	1

the y's, as in the model:

$$a_1 y_{i1} + a_2 y_{i2} - \kappa = x_i'\beta + \varepsilon_i, \qquad \mathrm{Var}\left(\varepsilon | x_i\right) = \sigma^2 \tag{26}$$

Then the OLS estimator is

$$\hat{\beta} = \left(\frac{1}{N}\sum_{i=1}^{N} x_i x_i'\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N} x_i'(a_1 y_{i1} + a_2 y_{i2} - \kappa)\right)$$

and

$$\begin{aligned} E\left[\hat{\beta}\right] &= E[x_i x_i']^{-1} E[x_i'(a_1 y_{i1} + a_2 y_{i2} - \kappa] \\ &= \beta(a_1\pi + a_2(1-\pi)) + E[x_i x_i']^{-1} E[x_i](a_2\pi + (1-\pi)a_1 - a_3)\kappa \end{aligned}$$

Unbiasedness requires the same conditions on a_1, a_2, and a_3 as derived in Lemma 1, i.e.

$$\begin{aligned} a_2(a_1) &= \frac{1 - \pi a_1}{1 - \pi} \\ a_3(a_1) &= \frac{\pi}{1 - \pi} + \frac{a_1 - 2\pi a_1}{1 - \pi} \end{aligned}$$

which means that any unbiased linear estimator $\hat{\beta}$ can be written as a linear combination of unbiased estimators that use only y_{i1} or y_{i2},

$$\hat{\beta}_1 = \left(\frac{1}{N}\sum_{i=1}^{N} x_i x_i'\right)^{-1} \frac{1}{N}\sum_{i=1}^{N} \frac{x_i y_{i1}}{\pi} - \frac{1-\pi}{\pi}\kappa$$
$$\hat{\beta}_2 = \left(\frac{1}{N}\sum_{i=1}^{N} x_i x_i'\right)^{-1} \frac{1}{N}\sum_{i=1}^{N} \frac{x_i y_{i2}}{1-\pi} - \frac{\pi}{1-\pi}\kappa$$

and so the minimum variance estimator is $\hat{\beta}^* = d^*\hat{\beta}_1 + (1-d^*)\hat{\beta}_2$, where

$$d^* = \frac{\text{Var}\left(\hat{\beta}_2\right) - \text{Cov}\left(\hat{\beta}_1, \hat{\beta}_2\right)}{\text{Var}\left(\hat{\beta}_1\right) + \text{Var}\left(\hat{\beta}_2\right) - 2\text{Cov}\left(\hat{\beta}_1, \hat{\beta}_2\right)} \tag{27}$$

and we can repeat the exercise in the previous sections, comparing variance and bias for misspecified beliefs $\hat{\pi}$ and different parameter combinations. The choice of the weights d^* that give the optimal $\hat{\beta}^*$ is now complicated by the fact that it depends on the second moments of X_i, however the formulas for Var $\left(\hat{\beta}_i\right)$ are the same as in (20)-(22), but replacing μ, σ^2, κ, and ω^2 with:

4.4 Generalizing results to $L_i > 2$

5 Monte Carlo Study

In order to compare how these methods perform in practice, I conduct a Monte Carlo study where each replication consists of (i) generating an x- and y-datafile, (ii) linking the x- and y-datafiles to obtain matched data of the form (2), and (iii) estimating (1) using the matched datasets and the techniques described in Sections 3 and 4. Since the performance of the estimators depends on whether multiple matches or estimated probabilities are available, Step (ii) involves applying four record linkage procedures, each of which outputs a distinct dataset. The remainder of this section describes in detail how I generate data for a single replication of the Monte Carlo study, with a special focus on the record linkage procedures implemented in Step (ii).

5.1 Generating the x- and y-datafiles

I begin by constructing a "ground truth" dataset with 1000 observations of $(x_{1i}, x_{2i}, y_i, w_i)$, where x_{1i} and x_{2i} are mutually independent, i.i.d draws from Bernoulli(0.5) and Normal(0,2) distributions, respectively.s The y_i values are generated according to

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i, \quad \varepsilon_i \mid x_{1i}, x_{2i} \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2) \tag{28}$$

where ε_i are independent draws from a Normal(0,2) distribution. I chooses $(\beta_0, \beta_1, \beta_2, \sigma^2) = (2, 0.5, 1, 2)$, so that estimating the correctly specified linear regression model yields an R^2 value of approximately 0.50.

The vector of identifying variables, w_i, includes a first name, last name, and birth year. In total, there are 960 unique first and last name combinations, so multiple observations will be assigned the same first and last name. The birth years are drawn at random from a uniform distribution over the set of integers between 1900 and 1925. The resulting dataset resembles the top panel of Figure 4.

Next, I split the ground truth dataset into the x- and y-datafiles, as in the bottom panel of Figure 4. The x-datafile contains values of (x_{i1}, x_{i2}) for 500 observations selected at random from the ground truth dataset. The identifiers in the x-datafile are equal to the original w_i plus some random transcription error. The probability of introducing a certain type ofs typographical error is equal to that reported for the 1940 Census data in Abramitzky et al. (2019)[3] These errors include deleting characters (e.g., "Anderson" becomes

[3] For example, 7% of observations have misreported first names and 17% of observations have misreported

Figure 4: 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

"Andersn"), exchanging vowels (e.g., "Rachel" becomes "Rachal"), and swapping English phonetic equivalents (e.g. "Ellie" becomes "Elie"). For half of the observations, I introduce random errors in the birth year drawn from a Normal(0, 2.5) distribution and rounded to the nearest integer.

The y-datafile includes all 1,000 values of y_i from the ground truth data, along with the original identifiers w_i. The aim of this construction is to make it likely that some observations in the x-datafile will be linked to multiple values of y. The next section describes the record linkage methods used to link the x- and y-datafiles.

last names.

5.2 Linking the x- and y-datafiles

Taking the x- and y-datafiles as given, I implement four record linkage procedures to obtain four configurations of matched data with the form (2). 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.

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.

The identifiers in the simulated data do not include non-alphabetic characters, but do include misspelled names, so I standardize all of the first and last names using the New York State Identification and Intelligence (NYSIIS) phonetic algorithm. Other popular phonetic algorithms include Soundex[4] and Metafone; however, I assume that the NYSIIS algorithm performs sufficiently well for the purposes of my analysis, given that the names I use are selected from among the most common names in the United States.

Step (ii) reduces the computational burden of a matching procedure when $N_x \times N_y$ is large, by dividing observations into non-overlapping "blocks" based on their values of w_i or w_j. For example, if observations are divided into blocks based on their place of birth, only individuals born in the same state will be considered as potential matches, and all pairs of individuals who are born in different states are assumed to be non-matches. Given the size of my simulated datafiles, I do not impose any blocking rule. If the datasets were larger, I

[4] This is the phonetic algorithm used in Aizer et al. (2016) from Example 1

could block observations according to the first letter of an observation's standardized first or last name, at the cost of possibly increasing the Type II error rate if these variables are recorded with error.

Step (iii) defines a metric for quantifying the similarity between non-numeric variables, such as Jaro-Winkler distances for strings. Other metrics include the Levenshtein "edit" distance, however I use the Jaro-Winkler method because it was developed specifically for record linkage by Jaro and Winkler. The metric measures the number of matching characters and required transpositions between two names, and gives a higher weight to discrepancies in the first part of the string. Importantly, having identical NYSIIS codes does not imply that the Jaro-Winkler distance between two strings will be 1, so the methods I use apply the Jaro-Winkler distance to the phonetically-spelled names.

Finally, Step (iv) is where record linkage procedures differ in the most meaningful ways; hence, this step will be the focus of my analysis. In fact I hold Step (i)-(iii) as described above, and define differences in the record linkage procedures according to differences in Step (iv).

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.

5.3 Deterministic Method

The deterministic matching algorithm I implement is based upon the matching procedure used in Abramitzky et al. (2012). My version is as follows,

1. Use the NYSIIS phonetic algorithm to obtain phonetically-spelled versions of the names in the x- and y- datafiles.
2. Restrict the sample to people in the x-datafile with unique first name, last name, birth year, and (x_{i1}, x_{i2}) combinations.
3. For each record i in the x-datafile, search for a record j in the y-datafile whose phonetically spelled first and last names and birth year match exactly.
   (a) If there is a *unique* match, designate the records (i, j) as a match.
   (b) If there are multiple possible matches in the y-datafile, drop the observation i from the sample.
   (c) If there are no observations in the y-datafile that match i's exact year of birth, search for a match within ±1 year of i's reported birth year; and, if this is unsuccessful, search for a match within ±2 years. If i matches to multiple observations at any point, or if none of these attempts produces a match, then the observation is discarded.
4. Repeat Steps 2 and 3 for each record in the y-datafile, searching for matches in the x-datafile. Set the matched sample equal to the intersection of the two sets of matches.

To allow for multiple matches, I replace Step 3 with,

3.* Designate as a match any observation in the y-datafile that matches i's phonetically spelled first and last name exactly, and whose birth year falls within ±2 years of i's birth year.

5.4 Probabilistic Method

Fellegi and Sunter (1969) developed the canonical framework for probabilistic record linkage by posing record linkage as a classification problem in which each (i, j) record pair belongs either to the set of matches (M), or non-matches (U).

To determine whether (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*,

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

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:

$$P(\boldsymbol{\gamma}_{\mathbf{ij}}) = P(\boldsymbol{\gamma}_{\mathbf{ij}}|M)p_M + P(\boldsymbol{\gamma}_{\mathbf{ij}}|U)p_U \tag{29}$$

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 $\boldsymbol{\gamma}_{\mathbf{ij}}$,

$$P(M|\boldsymbol{\gamma}_{\mathbf{ij}}) = \frac{p_M P(\boldsymbol{\gamma}_{\mathbf{ij}}|M)}{P(\boldsymbol{\gamma}_{\mathbf{ij}})} \tag{30}$$

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 (30). 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 $\boldsymbol{\gamma}_{\mathbf{ij}}$[5]. 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:

$$P(\gamma_{ij}|C) = \prod_{k=1}^{K} P(\gamma_{ij}^k|C)^{\gamma_{ij}^k}(1 - Pr(\gamma_{ij}^k|C))^{1-\gamma_{ij}^k} \qquad C \in \{M, U\} \tag{31}$$

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 $P(\boldsymbol{\gamma}_{\mathbf{ij}}|M)$ and $P(\boldsymbol{\gamma}_{\mathbf{ij}}|U)$. 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).

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

6 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. Using each of the linked datasets, I then compute (i) naive OLS estimator (using all observations and also with observations assigned ($L_i = 1$), (ii) the Scheuren and Winkler (1993) bias-corrected estimator, and (iii) the AHL estimator that assigns equal weights to multiple matches. As a benchmark, I also compute the OLS estimator that uses only the correctly matched pairs produced by the matching algorithm, and the OLS estimator applied to all 500 correctly linked record pairs. Details on the implementation of these algorithms and estimation procedures can be found in the appendix.

Table 4: Summary of matching algorithm performance

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 5: Performance of multiple match methods by value of L_i

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.

6.1 Matching results

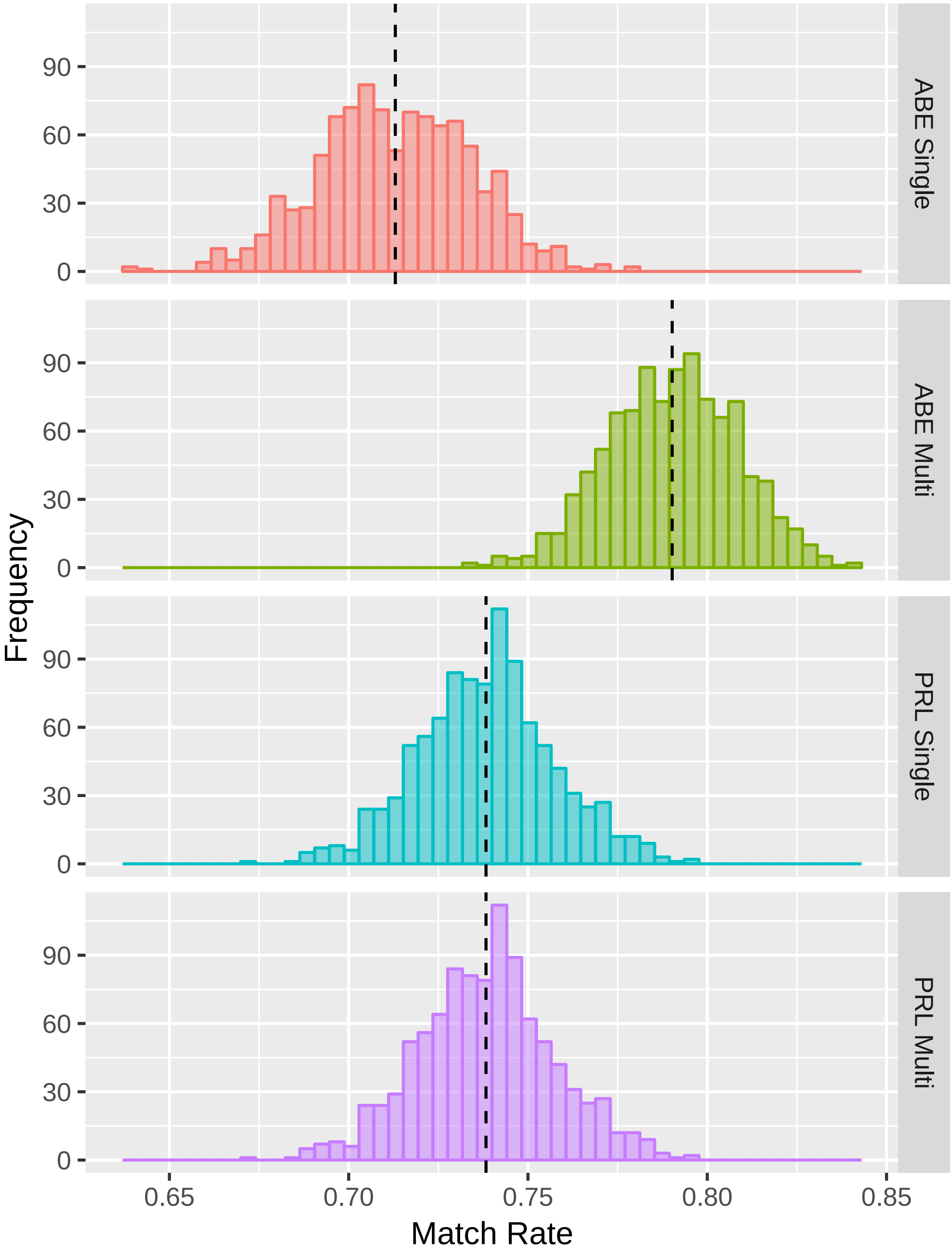
To evaluate the matching procedures, I compute the following statistics for each linked dataset, reported in Table 4:

- 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 average number of links per observation, and how often those links include the true match (Table 5).

As seen in Table 4, the average match rates range between 71 and 79 percent across the various matching procedures, but plotting the distribution of match rates in Figure 5shows that ABE Multi consistently matches more observations than any other procedure. PRL Single and PRL Multi have about the same match rate, which suggests that allowing

Figure 5: Match Rates by Linking Procedure



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

for multiple matches adds additional matches per observations rather than matching new individuals (however, this may be an artifact of how my PRL implementation). ABE Multi, on the other hand, seems to increase match rates by matching new observations relative to ABE Single.

When comparing Type I error rates, it is important to note that multiple-match methods will produce more false links by construction. Therefore, it is best to compare multi-match methods by measuring the proportion of observations whose matches contain the true link. In this metric, both ABE Multi and PRL Multi perform very well. Furthermore, we can compute these values for each value of L_i, as in Table 5, which shows that allowing multiple matches improves the accuracy of the ABE algorithm. Table 5 also shows that ABE Multi and PRL Multi rarely assign more than three matches to any given observation.

Comparing ABE Single and PRL Single in Table 4, demonstrates the usual tradeoff between Type I and Type II errors. ABE Single is more conservative, produces incorrect matches only 3 percent of the time, but failing to identify 26 percent of all matches. PRL Single is less conservative, missing only 15 percent of matches, but at the cost of matching false links 11 percent of the time.

Based on these results, ABE Multi seems to perform well if multiple matches are desired. The linked datasets produced by ABE Multi are very likely to include the true match, which is required for all of the estimation methods described in this paper. It is also easier in terms of computation, because it does not require linear sum assignment programs or thresholds to determine which record pairs should be designated as matches.

6.2 Estimation Results

I compare the estimators according to median absolute deviation, and plot histograms of the estimated values in Figures 6-7. In implementing the AHL (2019) estimator, I set

$\hat{g}(w_i, L_i) = \sum_{j=1}^{N_y} y_j$, the mean of all y observations, to reduce the computational burden and because I have generated y such that it is independent of the identifiers w_i. The AHL estimator will probably perform better in scenarios where w_i has predictive power in estimating the conditional mean of y_i.

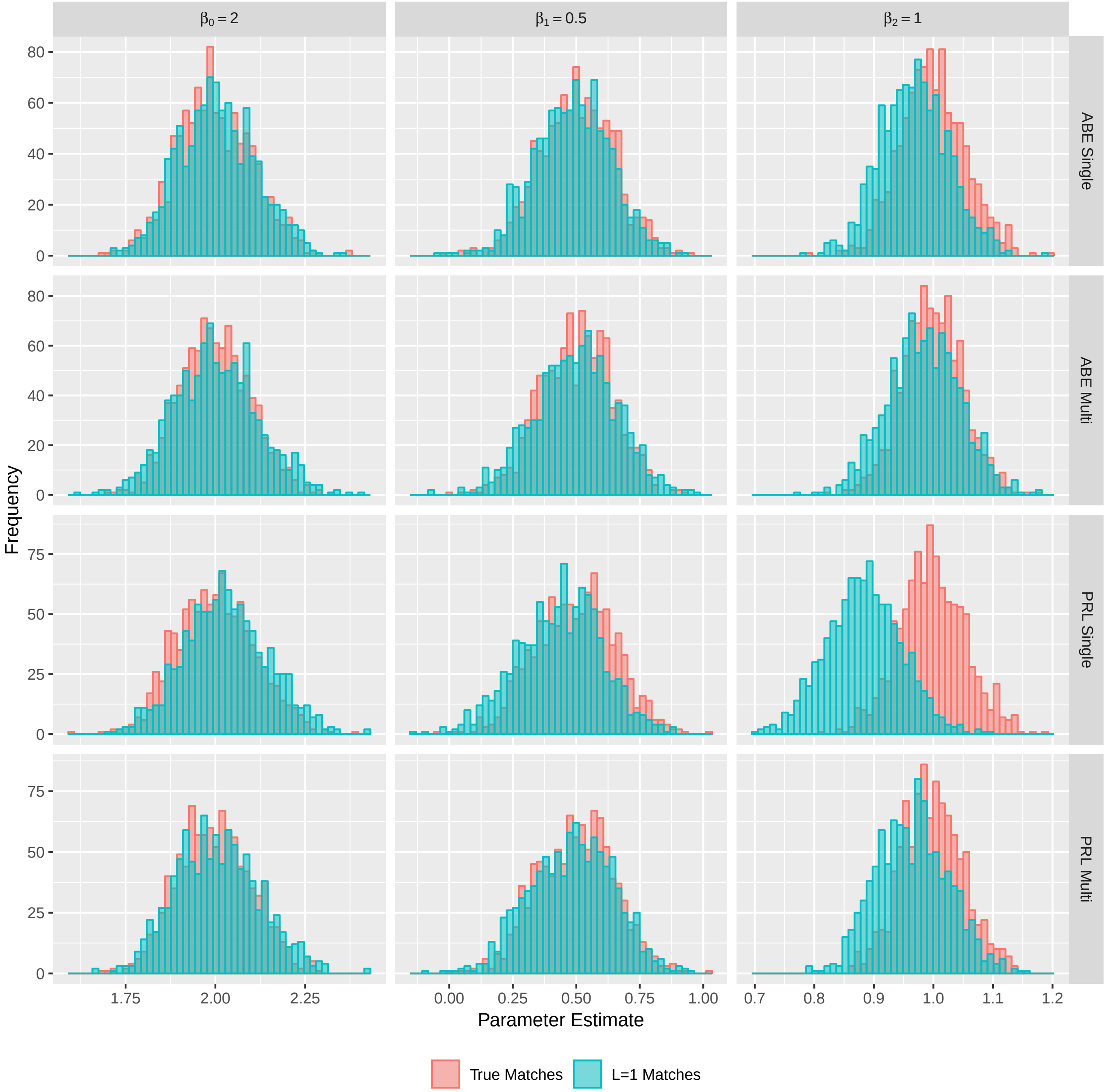
Table 6: Median Absolute Deviations for Estimators

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

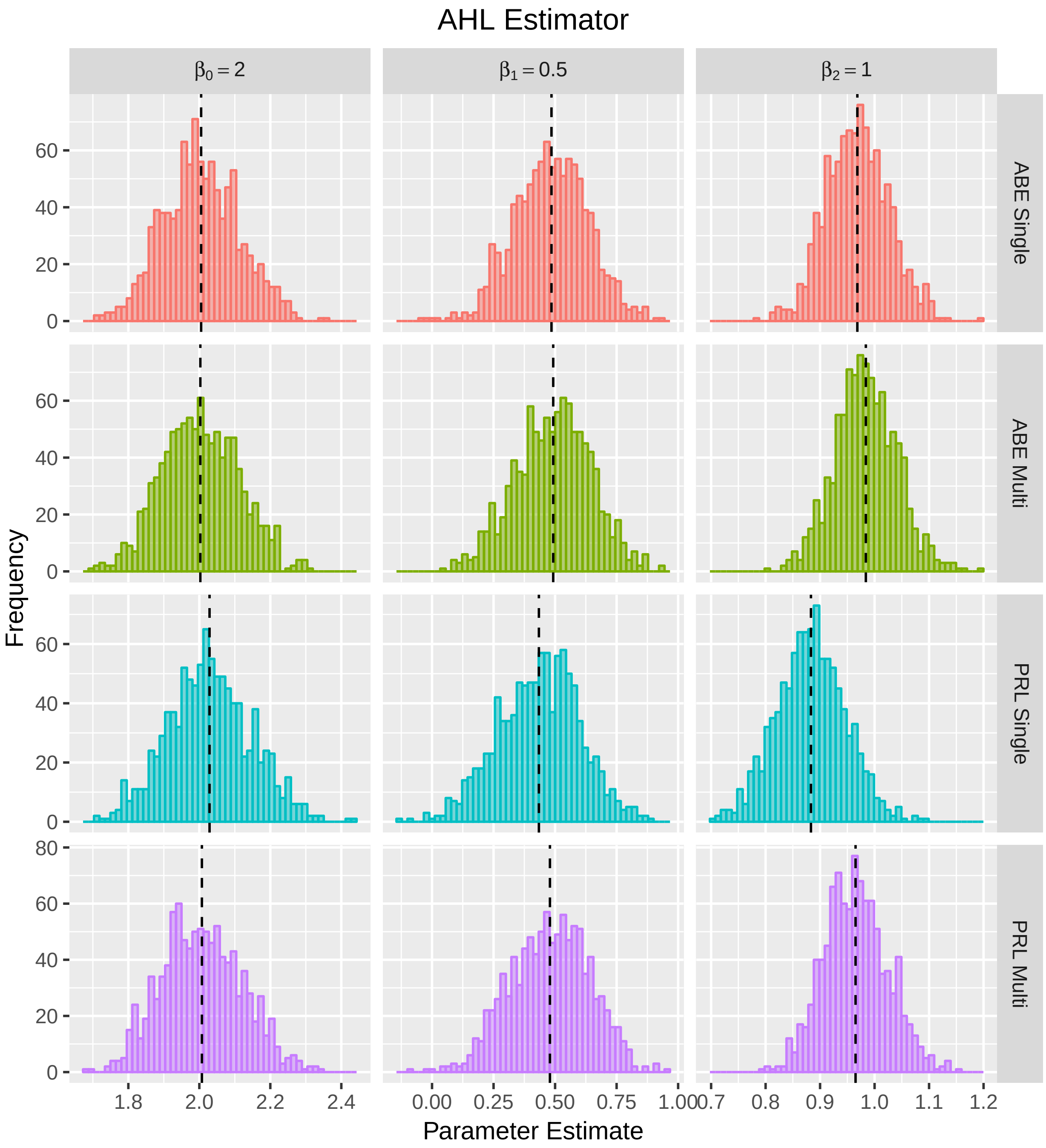
7 Discussion/Conclusion

To what extent my results generalize beyond the simulated data is unclear, as I made many arbitrary choices while generating the synthetic data – such as the dictionary of names and the structure of the typographical errors that I introduce in the x-datafile – that may impact my results in important ways. However, my theoretical results suggest that (i) using the match that is most likely to be correct, and bias correcting based on the probability that it is correct is optimal, (ii) if weights are estimated imprecisely, or if no match has a high

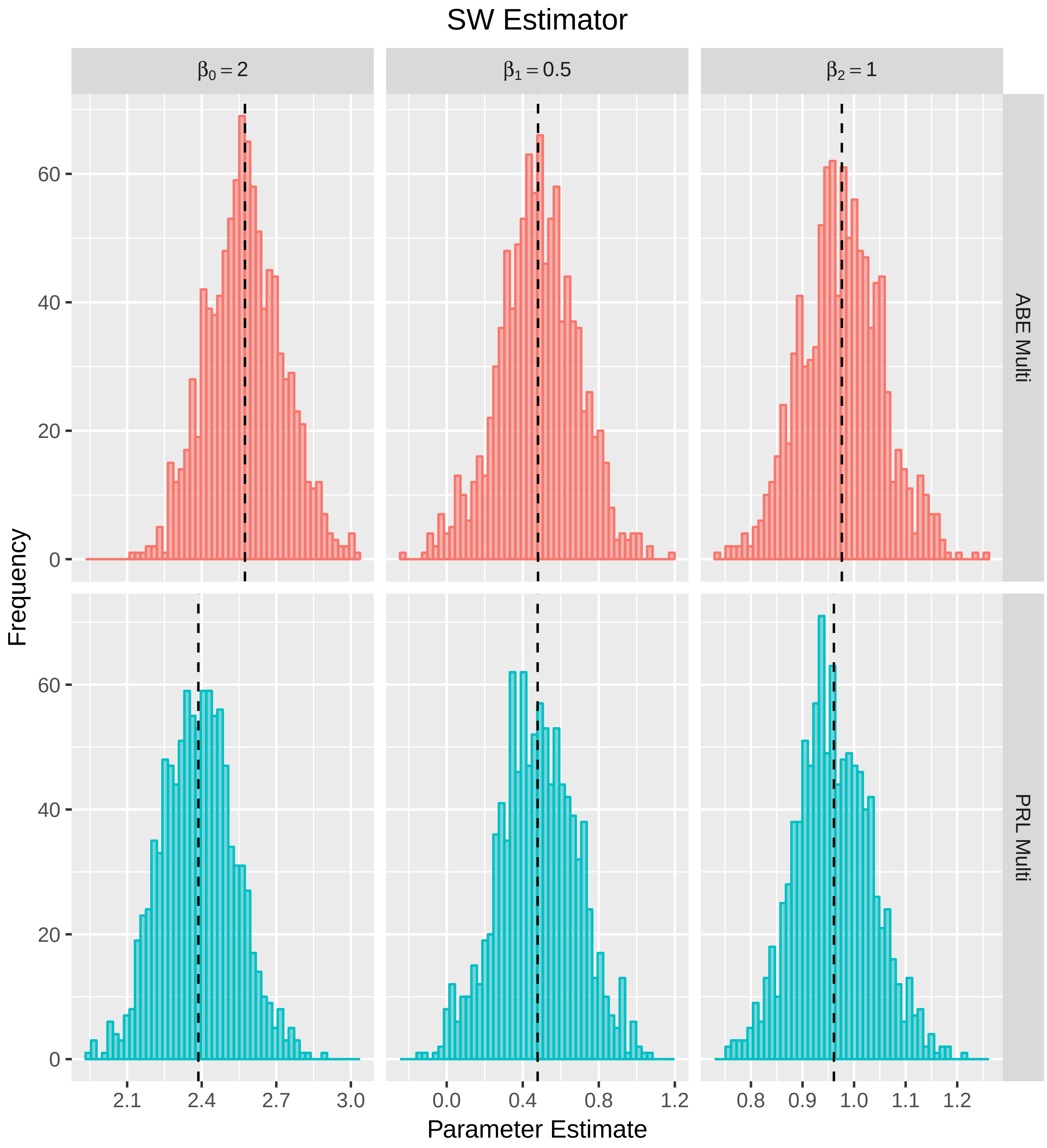
Figure 6: Comparing OLS with true matches produced by matching algorithm vs. matches with L=1



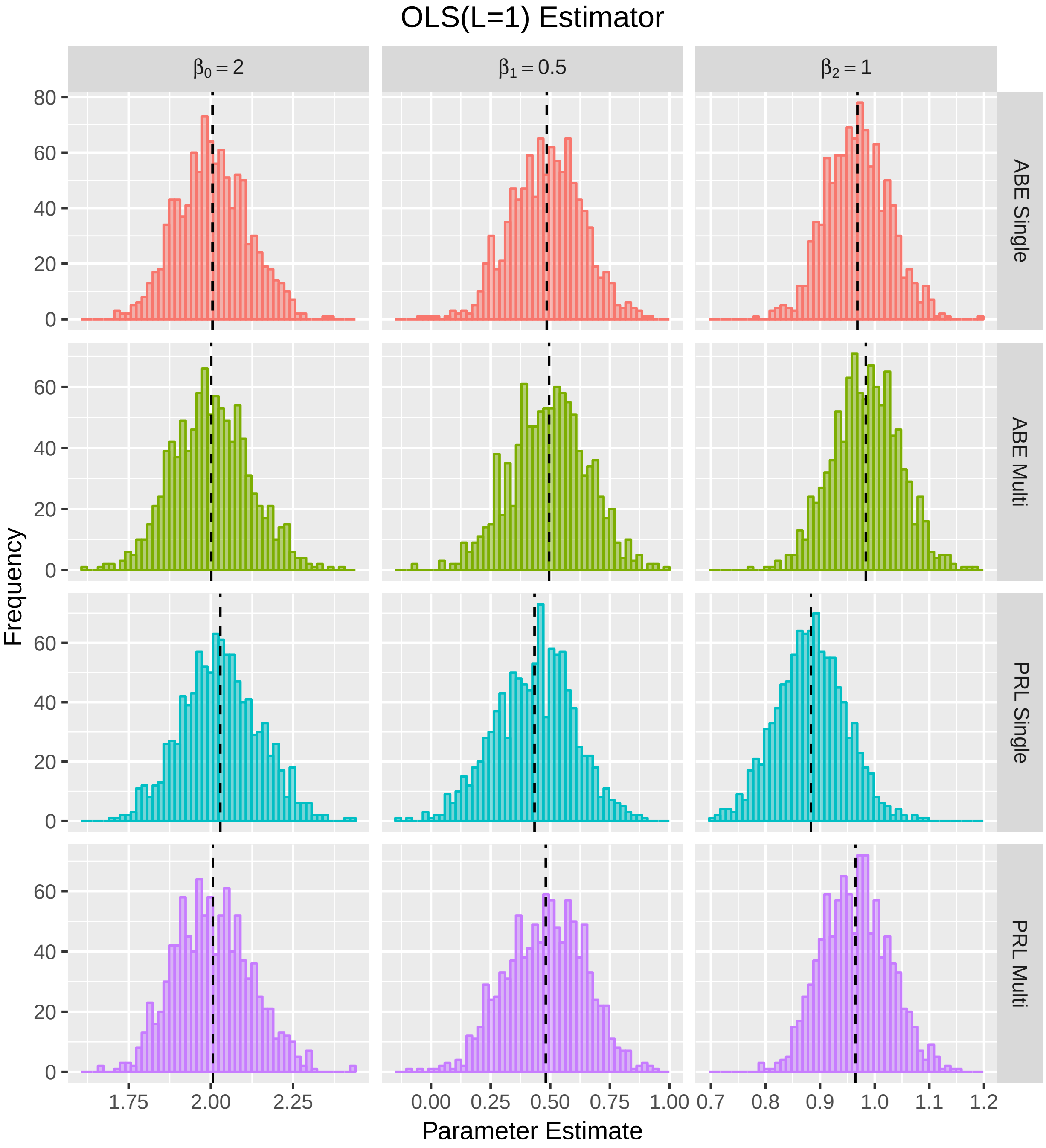
probability of being correct, then the it is better to assign equal weights to multiple matches. This result needs to be studied using more general models, and ideally applied to real data.



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



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



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

8 Appendix

8.1 Proofs

Lemma 1. Any linear unbiased estimator of μ of the form

$$\hat{\mu} = a_1 X_1 + a_2 X_2 - a_3 \kappa \tag{32}$$

with a_1 and $a_2 > 0$, can be written as a linear combination of

$$\begin{aligned} \hat{\mu}_1 &= \frac{X_1}{\pi} - \frac{1-\pi}{\pi}\kappa \\ \hat{\mu}_2 &= \frac{X_2}{1-\pi} - \frac{\pi}{1-\pi}\kappa \end{aligned}$$

Proof. The expectation of $\hat{\mu}$ is

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

so that unbiasedness requires choosing a_1, a_2 and a_3 that satsify,

$$a_1\pi + a_2(1-\pi) = 1 \implies a_2(a_1) = \frac{1}{1-\pi} - \frac{a_1\pi}{1-\pi} \tag{33}$$

$$a_1(1-\pi) + a_2\pi = a_3 \implies a_3(a_1) = \frac{\pi}{1-\pi} + \frac{a_1 - 2a_1\pi}{1-\pi} \tag{34}$$

Rewriting $\hat{\mu}$ as a function of a_1,

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

which, after some rearranging, can be written as

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

which completes the proof.

Lemma 2. Suppose the data consist of $\{X_\ell\}_{\ell=1}^L$, where each X_ℓ is drawn from the correct distribution with probability π_ℓ and from the incorrect distribution with probability $1 - \pi_\ell$, and exactly one X_ℓ is drawn from the correct distribution. Then, any unbiased estimator of μ that places positive weight on all of the $\{X_\ell\}$ can be written as a linear combination of $\hat{\mu}_\ell$, the unbiased estimator of μ that only uses X_ℓ,

$$\hat{\mu}_\ell = \frac{X_\ell}{\pi_\ell} - \frac{1 - \pi_\ell}{\pi_\ell}\kappa, \quad \ell = 1, \ldots, L$$

Proof. The proof follows by induction. Lemma 1 proves the base case for $L = 2$. Assume that $\hat{\mu}^{(L-1)} = \sum_{\ell=1}^{L-1} b_\ell \hat{\mu}_\ell$. Construct $\hat{\mu}^{(L)} = a_1 \hat{\mu}^{(L-1)} + a_2 X_L - a_0 \kappa$, so that

$$E[\hat{\mu}^{(L)}] = (a_1 + a_2 \pi_L)\mu + (a_2(1 - \pi_L) - a_0)\kappa$$

Unbiasedness requires that

$$\begin{aligned} a_2(a_1) &= \frac{1 - a_1}{\pi_L} \\ a_3(a_1) &= \left(\frac{1 - \pi_L}{\pi_L}\right)\left(\frac{1 - a_1}{\pi_L}\right) \end{aligned}$$

Plugging this into $\hat{\mu}^{(L)}$ yields,

$$\begin{aligned} \hat{\mu}^{(L)} &= a_1 \hat{\mu}^{(L-1)} + (1 - a_1)\underbrace{\left(\frac{X_L}{\pi_L} - \frac{1 - \pi_L}{\pi_L}\right)}_{\hat{\mu}_L} \\ \implies \hat{\mu}^{(L)} &= \sum_{\ell=1}^{L-1} a_1 b_\ell \hat{\mu}_\ell + (1 - a_1)\hat{\mu}_\ell \end{aligned}$$

which completes the proof.

8.2 Variance formulas

$\text{Var}\,(X_1)$ and $\text{Var}\,(X_2)$ are calculated using the law of total variance, using the random variable $D = 1$ if X_1 is drawn from the correct distribution (and X_2 is drawn from the incorrect distribution), and $D = 0$ otherwise:

$$\begin{aligned}
\text{Var}\,(X_1) &= E[\text{Var}\,(X_1|D)] + \text{Var}\,(E[X_1|D]) \\
&= P(D=1)\sigma^2 + P(D=0)\omega^2 + \text{Var}\,(\mu D + \kappa(1-D)) \\
&= \pi\sigma^2 + (1-\pi)\omega^2 + \pi(1-\pi)(\mu-\kappa)^2
\end{aligned}$$

The same trick can be applied to calculate $\text{Var}\,(X_2)$

8.3 Variance of alternative AHL estimator

$$\text{Var}\left(\frac{2}{N_{L_2}}\sum_{i=1}^{N_{L_2}}(y_{i1}+y_{i2}-x_i'\beta)x_i\middle| L_i = 2\right) = \frac{4}{N_{L_2}}\big(\text{Var}\,(x_i\varepsilon_i|L_i=2) + E[x_i^2|L_i=2](\omega^2+\kappa^2) - E[x_i|L$$

$$\text{Var}\left(\frac{1}{N_{L_3}}\sum_{i=1}^{N_{L_3}}(y_{i1}+y_{i2}+y_{i3}-x_i'\beta)x_i\middle| L_i = 3\right) = \frac{1}{N_{L_3}}\big(\text{Var}\,(x_i\varepsilon_i|L_i=3) + 2\,\big(E[x_i^2|L_i=3](\omega^2+\kappa^2) - E[x$$

8.4 implementation notes

Let's talk about what I need to include here. Here are some ideas:

Formulas for SE of SW estimator.

Details about implementation of fastLink algorithm. Also

- 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.

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.

Neter, John, E. Scott Maynes, and R. Ramanathan, "The Effect of Mismatching on the Measurement of Response Errors," *Journal of the American Statistical Association*, 1965, *60* (312), 1005–1027.

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.*