Methods for analyzing linked data

Rachel Anderson[*]

This Version: October 31, 2019

Abstract

This paper compares methods for estimating linear regression models when the variables of interest are recorded in separate datasets that must be merged prior to analysis. When it is not possible to identify the true matches with certainty, random errors in matching cause standard estimators to be biased (Neter, Maynes, and Ramanathan, 1965). In this context, Scheuren and Winkler (1993) and Anderson, Honoré, and Lleras-Muney (2019) propose methods to correct for the bias, however each procedure requires distinct assumptions on how the data are linked. This paper establishes conditions under which these methods are equivalent, and compares their performance using data that are matched by different record linkage procedures. The results suggest that the estimator from Anderson, Honoré, and Lleras-Muney (2019) is unbiased under weaker assumptions, and without sacrificing efficiency. The paper concludes with practical suggestions for analyzing linked data and discusses future areas of theoretical study.

[*] Mailing Address: Department of Economics, Julis Romo Rabinowitz Building, Princeton, NJ 08544. Email: rachelsa@princeton.edu. I am grateful to Bo Honoré for guidance on this project, and to Melissa Plotsky and Bryan Anderson for their expert copy editing. 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 a computationally efficient solution for merging data. These methods allow the researcher to specify a set of matching variables and a decision rule for linking observations to obtain a matched dataset in a matter of minutes. If 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 has emerged in response to the increasing availability of administrative datasets, including recently digitized historical complete count population censuses. These data often contain identifiers that are prone to typographical, duplication, enumeration, and mis-reporting errors. Although methods for record linkage in this context have been developed in statistics, computer science, operations research, and epidemiology, economic historians and historical demographers have developed their own linking algorithms out of concern about the accuracy and representativeness of data that are matched using imperfect identifiers (Ferrie, 1996; Abramitzky, Boustan, and Eriksson, 2012, 2014, 2019).

Recent papers by Abramitzky, Boustan, Eriksson, Feigenbaum, and Perez (2019) and Bailey, Cole, Henderson, and Massey (2017) 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, Mill, and Perez (2018) and Enamorado, Fifield, and Imai (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, Boustan, Eriksson, Feigenbaum, and Perez (2019) and Bailey, Cole, Henderson, and Massey (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, Honoré, and Lleras-Muney (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, Boustan, Eriksson, Feigenbaum, and Perez (2019) and Bailey, Cole, Henderson, and Massey (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 therefore to connect the existing literature on matching and estimating linear regression models with linked data. I compare the performance of different methods that incorporate different types of information, as well as extend the methods from Anderson, Honoré, and Lleras-Muney (2019) to incorporate probabilities as may be outputted from a record linkage procedure. Using these estimation methods for a variety of matched datasets, that are obtained by applying different matching procedures on the same raw data, allows me to conclude which *combinations* of matching and estimation procedures perform best in practice. The theoretical and empirical analyses in this paper suggest that deterministic matching methods which allow for multiple links, combined with the estimator from Anderson, Honoré, and Lleras-Muney (2019), may be a promising method for increasing the match rate, without introducing bias or sacrificing efficiency if the sample size is sufficiently large.

The rest of this paper is laid out as follows. Section 2 describes a simplified version of the setup from Anderson, Honoré, and Lleras-Muney (2019), which I call the linked data regression problem. Section 3 describes two existing methods for estimating linear regression models with matched data – specifically, the estimators proposed by Scheuren and Winkler (1993) and Anderson, Honoré, and Lleras-Muney (2019) – and describes under what conditions they are equivalent. Section 4 explores how knowledge of match probabilities may be used to improve the efficiency of the estimator from Anderson, Honoré, and Lleras-Muney (2019). Section 5 describes in detail how I test these methods in a Monte Carlo study that involves simulating and linking datasets with a variety of record linkage techniques, Section 6 reports the results, and Section 7 concludes.

2 Setup

In this section, I describe a simplified version of the estimation problem described in Anderson, Honoré, and Lleras-Muney (2019). Here, the goal is to estimate β in the linear regression model,

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

and I consider estimation techniques based on the following assumptions.

Assumption 1. The variables of interest are x_i and y_j, which are recorded in separate datafiles, along with the sets of identifiers w_i and w_j. Formally, the raw data consist of an x-datafile with observations $\{x_i, w_i\}_{i=1}^{N_x}$, and a y-datafile with observations $\{y_j, w_j\}_{j=1}^{N_y}$. I assume that $N_y \geq N_x$, and that every x_i has a corresponding value in the y-datafile that generates the relationship in (1), but the identity of this match is unknown. Furthermore, some y_j may not correspond to any value in the x-datafile, so that estimation requires identifying which (x_i, y_j) pairs refer to the same individuals.

Example 1. To fix ideas, consider the work of Aizer, Eli, Ferrie, and Lleras-Muney (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 may not be identified uniquely.

Assumption 2. The raw data are linked using w_i and w_j, resulting in a matched dataset of the form,

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

so that some values of x_i may be linked to multiple possible matches $\{y_{i\ell}\}_{\ell=1}^{L_i}$, and $\{\pi_{i\ell}\}_{\ell=1}^{L_i}$ is a vector of (estimated) probabilities that reflects how likely each of the $y_{i\ell}$ is to be the true match. Also, because $N_x \geq N$ and L_i is free to vary across i, I assume that $\{y_{i\ell}\}_{\ell=1}^{L_i}$ includes the unique true match for all of the observations in $\mathcal{D}_n$.

Example 1 (cont'd). Aizer, Eli, Ferrie, and Lleras-Muney (2016) link observations using a deterministic method that accounts for potential errors in w_i and w_j. 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 (SAS, 2019). They assign as matches all pairs of individuals whose SPEDIS scores and differences in birthdates fall within a pre-specified range. Their procedure does not enforce unique matches, so some individuals with common names are matched to multiple death records.

Assumption 3. The observations of x_i and $\{y_{i\ell}\}_{\ell=1}^{L_i}$ included in $\mathcal{D}_n$ comprise random samples conditional on the identifying variables w_i. More specifically, I assume that each x_i is drawn independently from $f_x(x|w_i)$, and its true match y_i is drawn from $f_y(y|x_i, w_i)$; and, the false matches y_j are drawn independently from $f_y(y|w_j)$.

Example 1 (cont'd). Assumption 3 requires that all individuals with the same identifying information are equally likely to be included in $\mathcal{D}_n$. This would be violated if, for example, higher income individuals had a greater probability of appearing in the sample, unless w_i and w_j were to include or proxy for income. Importantly, Assumption 3 does not rule out all forms of selection, as w_i could be correlated with selection into $\mathcal{D}_n$. This could occur if individuals with some names were easier (or harder) to match, and may pose additional challenges for analysis that are beyond the scope of this paper.

3 Estimation methods for linked data

In this section, I review methods from Scheuren and Winkler (1993) and Anderson, Honoré, and Lleras-Muney (2019) for estimating regression models using matched data, and establish conditions under which they are equivalent. Although there are other methods for analyzing linked data, such as those proposed by Lahiri and Larsen (2005) and Goldstein, Harron, and Wade (2012), they assume that each observation in the y-datafile has a unique match in the x-datafile, which is violated if $N_y > N_x$. Furthermore, the setup in Section 2 is more general than those considered previously, because it does not assume that all of the observations in the y-datafile are generated according to (1).

3.1 Scheuren and Winkler (1993)

Building upon the work Neter, Maynes, and Ramanathan (1965), Scheuren and Winkler (1993) demonstrate 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 inconsistent, 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,

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

which I henceforth refer to as the SW estimator. 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.

In addition to Assumptions 1-3, the SW estimator imposes two additional assumptions that complicate its implementation. The first is 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, which may not be available for 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 may fail in settings such as 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, Honoré, and Lleras-Muney (2019)

Anderson, Honoré, and Lleras-Muney (2019) consider estimating 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 1-3, (7) can be rewritten,

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

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, Honoré, and Lleras-Muney (2019).

Applied to the linear regression model in (1), the procedure above is equivalent to 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,

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

where $\sigma(X_i) = \sigma_\varepsilon^2 + (L_i - 1)\sigma_\nu^2$. I henceforth refer to (10) as the AHL estimator.

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 we consider adding another assumption that $E_0[m(y_i, x_i; \theta)|L_i = \ell] = E_0[m(y_i, x_i; \theta)]$, then we could construct the moments,

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

and

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

and apply GMM to (11) and (12) as the new moment conditions. In practice, the precision of the estimator depends on the number of observations that are linked to two and three outcomes. Additional work is necessary to evaluate the theoretical performance and feasibility of this estimator.

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

$$z_i - \hat{B}_i = x_i'\beta + \varepsilon_i \tag{13}$$

with $\hat{B}_i = \sum_{\ell=1}^{L_i} \pi_{i\ell} y_{i\ell} - y_{i\ell^*}$.

The AHL estimator can be written in the form (13) with

$$\hat{B}_i = z_i - \sum_{\ell=1}^{L_i} y_{i\ell} + (L_i - 1)\hat{g}(w_i, L_i) \tag{14}$$

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 $\hat{\beta}_{AHL}$, by setting

$$\hat{g}(w_i, L_i) = \frac{1}{L_i - 1}\left(\sum_{\ell \neq \ell^*} y_{i\ell} + y_{i\ell^*}\right) \tag{15}$$

Since $\hat{g}(\cdot)$ as written in (15) 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 problem in Anderson, Honoré, and Lleras-Muney (2019), based on the observation that

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

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 2, 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,

$$\hat{\mu}_1 = \frac{X_1}{\pi} - \frac{1-\pi}{\pi}\kappa \tag{16}$$

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

$$\hat{\mu_2} = \frac{X_2}{1-\pi} - \frac{\pi}{1-\pi}\kappa \tag{17}$$

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

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

which is solved by

$$d^* = \frac{\mathrm{Var}\left(\hat{\mu}_2\right) - \mathrm{Cov}(\hat{\mu}_1,\hat{\mu}_2)}{\mathrm{Var}\left(\hat{\mu}_1\right) + \mathrm{Var}\left(\hat{\mu}_2\right) - 2\mathrm{Cov}(\hat{\mu}_1,\hat{\mu}_2)} \tag{18}$$

where

$$\operatorname{Var}(\hat{\mu}_1) = \frac{\operatorname{Var}(X_1)}{\pi^2} = \frac{1}{\pi^2}\left(\pi\sigma^2 + (1-\pi)\omega^2 + \pi(1-\pi)(\mu-\kappa)^2\right) \tag{19}$$

$$\operatorname{Var}(\hat{\mu}_2) = \frac{\operatorname{Var}(X_2)}{(1-\pi)^2} = \frac{1}{(1-\pi)^2}\left((1-\pi)\sigma^2 + \pi\omega^2 + \pi(1-\pi)(\mu-\kappa)^2\right) \tag{20}$$

$$\operatorname{Cov}(\hat{\mu}_1, \hat{\mu}_2) = \frac{\operatorname{Cov}(X_1, X_2)}{\pi(1-\pi)} = \frac{1}{\pi(1-\pi)}\left((1-\pi^2-(1-\pi)^2)\mu\kappa - \pi(1-\pi)(\mu^2+\kappa^2)\right) \tag{21}$$

Derivations of these formulas are in the appendix.

Thus, the minimum variance unbiased estimator is

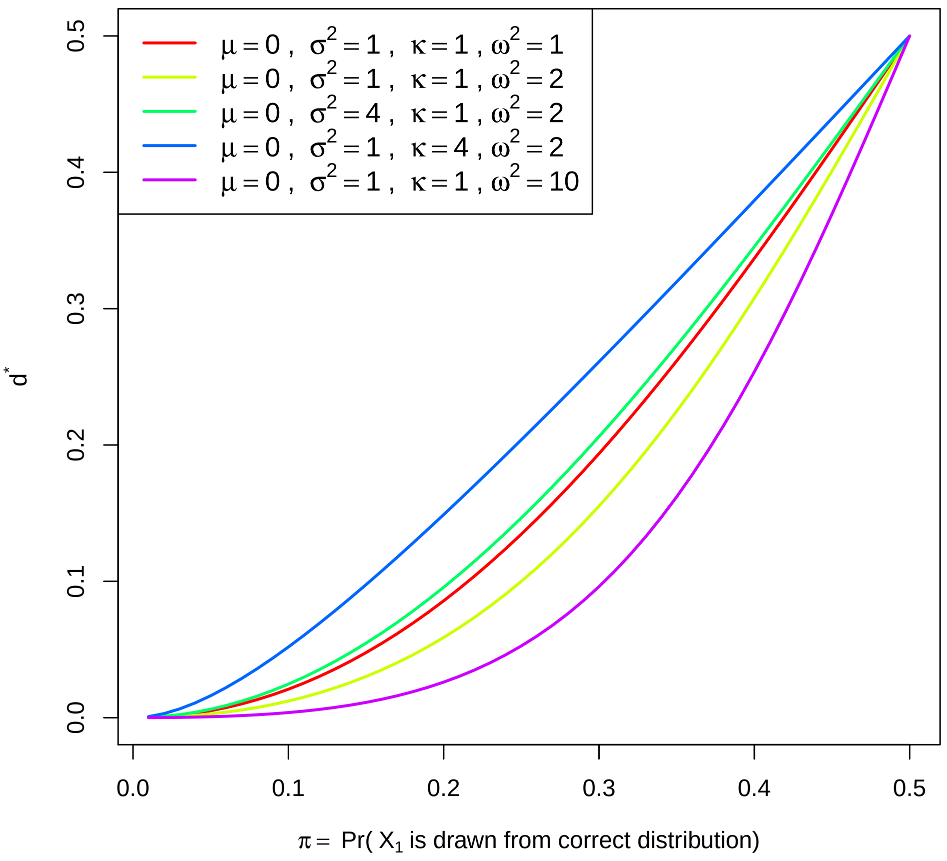
$$\hat{\mu}^* \equiv \hat{\mu}(d^*) = d^*\hat{\mu}_1 + (1-d^*)\hat{\mu}_2 \tag{22}$$

where d^* is defined as in (18). 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 $\pi = 0.5$, $\operatorname{Var}(\hat{\mu}_1) = \operatorname{Var}(\hat{\mu}_2)$ 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 $\operatorname{Var}(X_1) = \operatorname{Var}(X_2)$, 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 $(X_{i1}, X_{i2})_{i=1}^N$, where X_{i1} is drawn from F_X with probability π_i, and X_{i2} is drawn from F_X with probability

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



$1 - \pi_i$. In this case, d^* is calculated according to (18) 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[1].

As observed in Anderson, Honoré, and Lleras-Muney (2019), when π is unknown, it is possible to construct an unbiased linear estimator of $\hat{\mu}$ by weighting all observations equally,

$$\hat{\mu}^{AHL} = \frac{1}{N}\sum_{i=1}^{N} X_{i1} + \frac{1}{N}\sum_{i=1}^{N} X_{i2} - \kappa \tag{23}$$

[1] 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".

The variance of this estimator is

$$\text{Var}\left(\hat{\mu}^{AHL}\right) = \frac{\text{Var}\left(X_{1i} + X_{2i}\right)}{N} \tag{24}$$

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 (19)-(21) 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 $\text{Bias}(\hat{\pi})$. Alternatively, for a fixed value of $\text{Bias}(\hat{\pi})$, 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

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

$\hat{\pi}$ 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 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 $\widehat{\mu^*}$ 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 it is rarely the case that $N = 10$ 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. Whether this result holds more generally requires additional work to incorporate heterogenous π_i and $L > 2$ observations of X_ℓ.

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

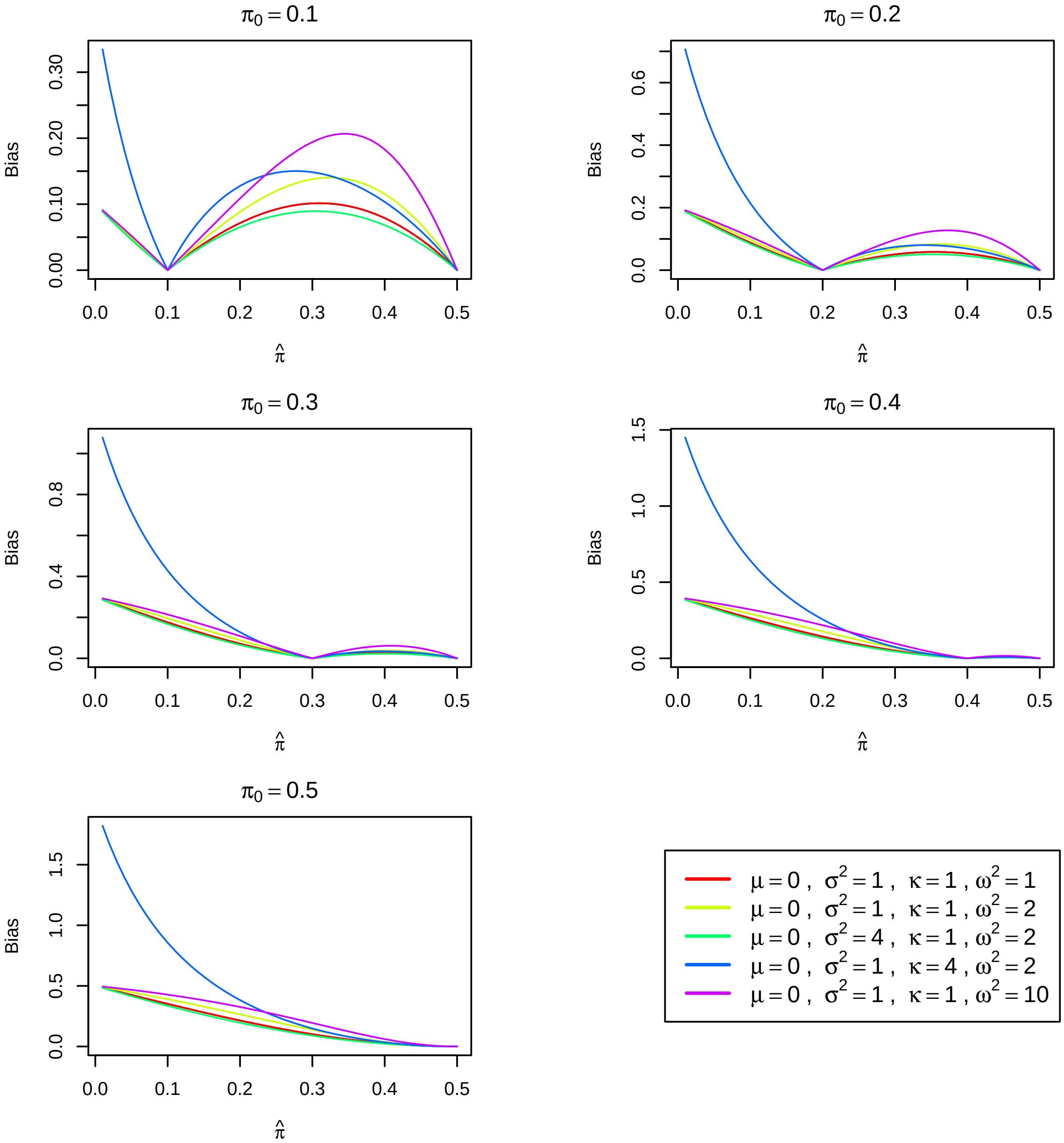


Figure 3: Variance of $\hat{\mu}^*$ as a function of $\hat{\pi}$ with $N = 1$

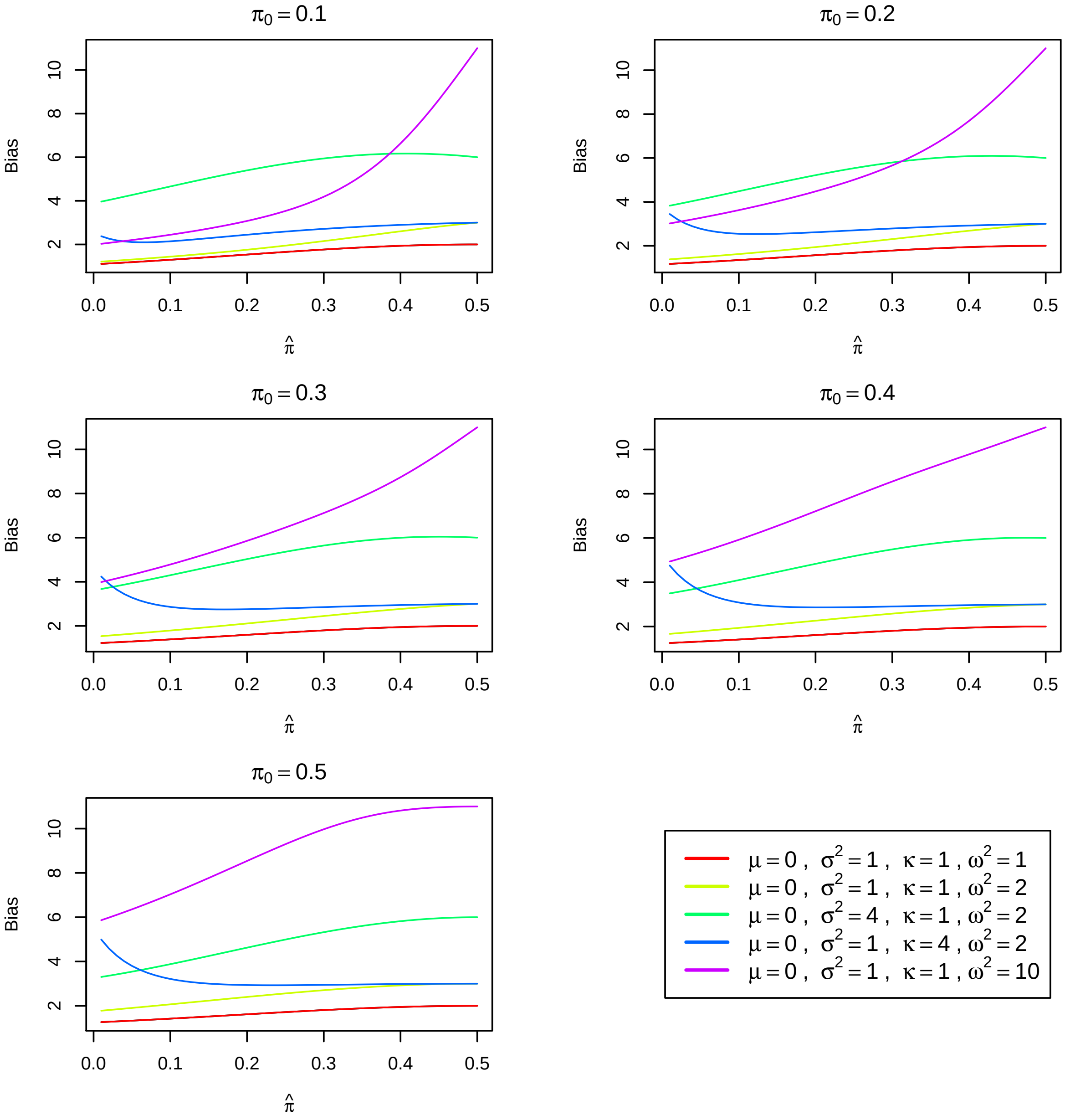


Table 2: MSE ratio for $\hat{\mu}^*$ and $\hat{\mu}^{AHL}$ for $N = 100$ and $(\mu, \sigma^2, \kappa, \omega^2) = (0,1,1,2)$

	$\hat{\pi}$									
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

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 the y's, as in the model:

$$a_1 y_{i1} + a_2 y_{i2} - \kappa = x_i'\beta + \varepsilon_i, \qquad \text{Var}\left(\varepsilon | x_i\right) = \sigma^2 \tag{25}$$

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

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

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},

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

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

$$d^* = \frac{\text{Var}\left(\hat{\beta}_2\middle|x_i\right) - \text{Cov}\left(\hat{\beta}_1, \hat{\beta}_2\middle|x_i\right)}{\text{Var}\left(\hat{\beta}_1\middle|x_i\right) + \text{Var}\left(\hat{\beta}_2\middle|x_i\right) - 2\text{Cov}\left(\hat{\beta}_1, \hat{\beta}_2\middle|x_i\right)} \tag{26}$$

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 $\left(\hat{\beta}_i\middle|x_i\right)$ are the same as in (19)-(21), but replacing μ, σ^2, κ, and ω^2 with (under conditional homoskedasticity),

$$\begin{aligned}
\tilde{\mu} &= \beta \\
\tilde{\sigma}^2 &= \sigma^2 E[x_i x_i]^{-1} \\
\tilde{\kappa} &= E[x_i x_i']^{-1} E[x_i]\kappa \\
\tilde{\omega}^2 &= (\omega^2 + \kappa^2) E[x_i x_i']^{-1}
\end{aligned}$$

and the bias and variance should behave as in Figures 2 and 3 for misspecified beliefs $\hat{\pi} \neq \pi$.

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

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i, \quad \varepsilon_i \mid x_{1i}, x_{2i} \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2) \tag{27}$$

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, Boustan, Eriksson, Feigenbaum, and Perez (2019)[2] 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"). 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

[2] For example, 7% of observations have misreported first names and 17% of observations have misreported last names.

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

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.

5.2 Linking the x- and y-datafiles

Taking the x- and y-datafiles as given, I implement four record linkage procedures to obtain matched datasets with the general structure in (2). This section offers a brief overview of these methods, however the interested reader should refer to Harron, Goldstein, and Dibben (2015); Christen (2012) and Herzog, Scheuren, and Winkler (2007) to learn more about modern record linkage methods.

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 how to measure (partial) agreement between (w_i, w_j), and (iv) designating (i, j) pairs as matches.

Step (i) accounts for differences in w_i and w_j that arise as a result of transcription error or misreporting, even when observations i and j refer to the same individual. In practice, the researcher may define binary matching variables that correspond with events such as i and j were born in the same month or i and j have last names ending with the same three characters. If the matching variables include full strings, then the researcher may standardize them by removing spaces and non-alphabetic characters, replacing common nicknames with full names, or pre-processing names with phonetic algorithms to account for possible misspellings.

Example 2. The identifiers in the simulated data do not include non-alphabetic characters, but do include misspelled names, so I pre-process all of the first and last names using the New York State Identification and Intelligence (NYSIIS) phonetic algorithm, and include these phonetically-spelled names among the matching variables. The other matching variable is birth year, which does not require standardization. Other popular phonetic algorithms include Soundex (Odell and Russell, 1918) and Metafone (Philips, 1990); 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 assigned at birth in the present-day 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. Only pairs assigned to the same block are considered as potential matches, and pairs that belong to different blocks are automatically designated as non-matches. Thus, blocking variables should be recorded with minimal error, otherwise blocking can increase the Type II error rate.

Example 2 (cont'd). Given the size of my simulated datafiles, I do not need to impose any blocking rule for computational feasibility; however, the age-band threshold imposed by my deterministic matching methods acts as one. These methods require that potential matches have recorded birth years that lie within a 2-year band. Since I generate errors in the recorded birth year by rounding independent, random draws from a Normal(0, 2.5) distribution to the nearest integer, about 12 percent of these errors should result in true matches outside the threshold. Furthermore, only half of the observations in the x-datafile contain errors in their recorded birth year, so this blocking structure should imply a Type II error rate of about 6 percent. Out of the 500 observations in the x-datafile, this corresponds to approximately 28 false non-matches.

Step (iii) defines a metric for quantifying the similarity between non-numeric variables, such as the Jaro-Winkler distance or Levenshtein "edit" distance for strings. This allows the researcher to declare as a "name match", or "partial name match" any observation pair whose string distance is lower than a pre-specified threshold. These metrics are related to, but different from the standardization methods described in Step (ii). This is pointed out by Abramitzky, Mill, and Perez (2018), who observe that "Abramtziky" is coded differently than "Abramitzky" using the NYSIIS algorithm, but the Jaro-Winkler distance between these two names is very low (0.02, where the minimum distance is 0, and the maximum distance is 1). Also, "James Tennes" and "James Thomas" have the same NYSIIS code, but the Jaro-Winkler distance between "Tennes" and "Thomas" is 0.4

Example 2 (cont'd). The probabilistic record linkage methods in this paper use Jaro-Winkler distances to calculate the similarity between strings. This metric gives higher weight to discrepancies in the first part of the string, as this is where errors are less likely to be made when recording names (Jaro, 1989; Winkler, 2006). The Jaro-Winkler distance is also the default string metric for many record linkage packages, so this choice best reflects how researchers often match datasets in practice.

Whereas Steps (i)-(iii) involve decisions about pre-processing data that can be incorporated in any linkage procedure, Step (iv) is where the most meaningful differences among procedures arise. The decision to match an (i, j) pair requires trading off the possibility of introducing Type I or Type II error. Probabilistic methods developed by Fellegi and Sunter (1969) show how to construct the optimal linkage rule subject to pre-specified tolerances for Type I and Type II errors. Deterministic methods offer a proxy for these methods. The rest of this section is devoted to discussing each method in detail.

5.2.1 Deterministic Methods

The deterministic matching methods used in this paper are based upon those used by Abramitzky, Boustan, and Eriksson (2012, 2014, 2019), and Ferrie (1996). I implement two versions of the same method. The first matches each observation in the x-datafile to at most one observation in the y-datafile, and discards any observations that result in multiple matches. The second version designates as a match *any* pair of observations that have the same phonetically spelled first and last names, and whose birth years fall within a 2-year band, which can result in linking a single observation to multiple matches.

The basic algorithm that I use 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_i 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 (i, j) as a match, and stop searching for additional possible matches.

   (b) If there are multiple possible matches in the y-datafile, discard the observation i.

   (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 an exact name match, then discard the observation.

4. Repeat Steps 2 and 3 for each record in the y-datafile, searching for matches in the x-datafile.

5. Return the matched dataset equal to the intersection of the two sets of matches pro-

duced by Steps 3 and 4.

I implement the version that allows for multiple matches in exactly the same way, except that 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.

Although there are many ways to alter the above procedure, such as using a ±5-year age band, replacing NYSIIS with another phonetic algorithm, or allowing for partial string agreement by incorporating Jaro-Winkler string distances, I design my methods to mimic the algorithm from Abramitzky, Boustan, Eriksson, Feigenbaum, and Perez (2019). As discussed above, I predict that the choice of the 2-year age band, combined with my data generation process, will result in a Type II error rate of about 5.7 percent for the deterministic methods. However, the focus of this paper is to study estimation methods for linked data, so I favor simplicity over choosing the optimal variant, and defer to Bailey, Cole, Henderson, and Massey (2017) and Abramitzky, Boustan, Eriksson, Feigenbaum, and Perez (2019) for a discussion of these matters.

5.3 Probabilistic Method

The probabilistic matching methods implemented in this paper are implemented using the fastLink package in `R` created by Enamorado, Fifield, and Imai (2019). Their methods are based upon the canonical work of Fellegi and Sunter (1969), which I now review.

Fellegi and Sunter (1969) present the record linkage task as a classification problem, where each (i, j) record pair belongs either to the set of matches (M), or non-matches (U). If the pairs are evaluated according to K, comparison criteria, represented as a *comparison vector*,

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

then the probability of observing a particular configuration of $\boldsymbol{\gamma}_{\mathbf{ij}}$ can be represented by the mixture distribution:

$$P(\boldsymbol{\gamma}_{\mathbf{ij}}) = P(\boldsymbol{\gamma}_{\mathbf{ij}}|M)p_M + P(\boldsymbol{\gamma}_{\mathbf{ij}}|U)p_U \tag{28}$$

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, and p_M and $p_U = 1 - p_M$ are the marginal probabilities of observing a matched or unmatched pair.

Using Bayes' Rule, we can write the probability of (i, j) belonging to M or U, conditional on observing $\boldsymbol{\gamma}_{\mathbf{ij}}$, as

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

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

If we can estimate the parameters of the mixture distribution in (28), then we can estimate the probability that any two records refer to the same entity using the formulas above. These probabilities can in turn be used to designate pairs as matches, or to quantify uncertainty about the matched dataset.

In the context of this paper, the comparison vector $\boldsymbol{\gamma}_{\mathbf{ij}}$ reflects agreements between w_i and w_j, such as "i and j have the same birth year" or "i and j have the same phonetically spelled last name." Yet even if all of the γ_{ij}^k are binary, $\boldsymbol{\gamma}_{\mathbf{ij}}$ has $2^K - 1$ possible configurations of $\boldsymbol{\gamma}_{\mathbf{ij}}$, and so it is convenient to assume the comparison fields γ_{ij}^k are independent across k conditional on match status. This reduces the number of parameters necessary to describe each mixture class, since we can factor

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

In principle, this assumption can be relaxed using log-linear models, as in Larsen and Rubin (2001); however, the conditional independence assumption is appropriate for my application, because I generate the errors for different categories of w_i independently.

Since membership to M or U is not observed, Larsen and Rubin (2001) suggest using the expectation-maximization (EM) algorithm from Dempster, Laird, and Rubin (1977) to simultaneously estimate the parameters in (28) and classify record pairs as matches or non-matches. Crucially, there is no restriction that the posterior match probabilities in (30) sum to 1 for a fixed observation i. Obtaining unique matches requires using a linear sum assignment program that maximizes the sum of the posterior matching probabilities subject to the constraint that no observations in either datafile can be matched multiple times.

The fastLink package by Enamorado, Fifield, and Imai (2019) estimates the posterior match probability for each pair of observations in the sample, as outputted by the EM algorithm. Their method allows the user to specify a lower bound for the posterior probability of a match that will be accepted; I set this equal to 0.7 because the default value of 0.85 does not result in any matches. The fastLink algorithm also allows the user to specify whether unique matches are desired, and, if so, returns the set of matches that solves the linear sum assignment program assigned above. I use this option to obtain two versions of linked data, one that enforces unique matches, and one that accepts all matches with posterior match probabilities greater than 0.7, which I then normalize to obtain estimates of $\pi_{i\ell}$.

6 Monte Carlo Results

Following the data generating process described in Section 5.1, I generate 1,000 x- and y-dataset pairs, such that each of the 500 observations in the x-datafile has a unique, true match in the y-dataset, but the identifiers in the y-dataset may repeat, and the identifiers in the x-dataset contain random errors. I then match each dataset pair a total of four times, using two deterministic matching methods and two probabilistic record linkage methods that produce either single or multiple matches per observation, as summarized in Table 4.

Each linkage method produces a distinct matched dataset, and so the matching step produces a total of 4,000 matched datasets. For each dataset, I compute and compare the following estimators:

Table 4: Overview of Matching Methods

Method	Unique Match	Matched Dataset
Deterministic	Yes	$(x_i, y_i)_{i=1}^N$
Deterministic	No	$\big(x_i, \{y_{i\ell}\}_{\ell=1}^{L_i}\big)_{i=1}^N$
Probabilistic	Yes	$(x_i, y_i, \pi_i)_{i=1}^N$
Probabilistic	No	$\big(x_i, \{y_{i\ell}\}_{\ell=1}^{L_i}, \{\pi_{i\ell}\}_{\ell=1}^{L_i}\big)_{i=1}^N$

1. the OLS estimator that treats each link as a distinct observation
2. the OLS estimator that uses only observations assigned a unique match
3. the SW estimator
4. the AHL estimator
5. the OLS estimator that uses all 500 correctly linked record pairs

6.1 Matching results

The first column of Table 5 reports the proportion of observations in the x-datafile that are linked to at least one observation in the y-datafile, averaged across the Monte Carlo replications. These rates range between 71 and 79 percent across methods, however Figure 5 shows that the deterministic method with multiple matches consistently matches more distinct observations than any other procedure. Additionally, the probabilistic methods have the same match rate on average, which suggests that allowing for multiple matches adds additional matches per observations, as opposed to matching new individuals. This is likely an artifact of the fastLink algorithm, because the posterior probability threshold for designating a match is the same for both methods. By contrast, the match rate increases for the deterministic algorithm when multiple matches are allowed, because it matches observations that were discarded otherwise.

The second column of Table 5 displays the average number of links assigned by each matching procedure. For the methods that produce multiple matches, each (i, j) combination counts as a distinct link, so that an observation with L linked outcomes counts as L matches. To explore whether these numbers are driven primarily by linking the observations to many matches, or by linking more observations to possibly multiple, but fewer, matches, I report the average number of links per observation in Table 6. On average, the deterministic method seems to assign larger L_i per observation. Whereas the probabilistic method assigns a unique match to 59 percent of observations, and two matches to 33 percent of observations, the deterministic method assigns, on average, one, two, and three matches to 52, 35, and 11 percent of observations, respectively. Both algorithms tend not to assign more than four matches for many observations. Note that these probabilities do not sum to one, because they reflect the average of $P(L_i = \ell)$ across 1,000 observations, and not the average *distribution* of L_i across datasets.

Of the methods that match pairs uniquely, the deterministic matching algorithm seems to produce higher quality matches, as measured by its low Type I error (0.03) relative to that of the probabilistic method (0.11). It also produces a lower Type II error – the deterministic method fails to match 31 percent of true links, and the probabilistic method misses 35 percent. As discussed in Section 5, about 6 percent of the Type II error can be attributed to the blocking structure imposed by the deterministic algorithm, yet it still outperforms the probabilistic method.

The last column in Table 5 reports the average probability that the set of matches $\{y_{i\ell}\}_{\ell=1}^{L_i}$ contains the true match. This is an important metric to study, because the estimation methods used in this paper assume that this probability is equal to 1. Notably, this assumption is most likely to fail for the probabilistic method with unique matches (it finds the correct match only 89 percent of the time), but allowing for multiple matches in improves this probability significantly. Both of the deterministic methods perform very well on this metric, as they include the correct match 97 and 99 percent of the time. Breaking down this probability by the number of matches L_i in Table 6 shows that allowing for multiple matches increases the probability that the true match is included in the sample.

Table 5: Summary of matching algorithm performance

Method	Match Rate	# Matches	Type I	Type II	P(Contains True)
Deterministic (Multiple)	0.79 (0.02)	505.08 (17.30)	0.23 (0.02)	0.22 (0.02)	0.99 (0.01)
Deterministic (Single)	0.71 (0.02)	356.50 (10.60)	0.03 (0.01)	0.31 (0.02)	0.97 (0.01)
Probabilistic (Multiple)	0.74 (0.02)	435.65 (14.94)	0.18 (0.02)	0.28 (0.02)	0.97 (0.01)
Probabilistic (Single)	0.74 (0.02)	369.15 (9.65)	0.11 (0.02)	0.35 (0.02)	0.89 (0.02)

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

Table 6: Performance of multiple match methods by value of L_i

L	1	2	3	4	5	6+
Deterministic						
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)
Probabilistic						
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)

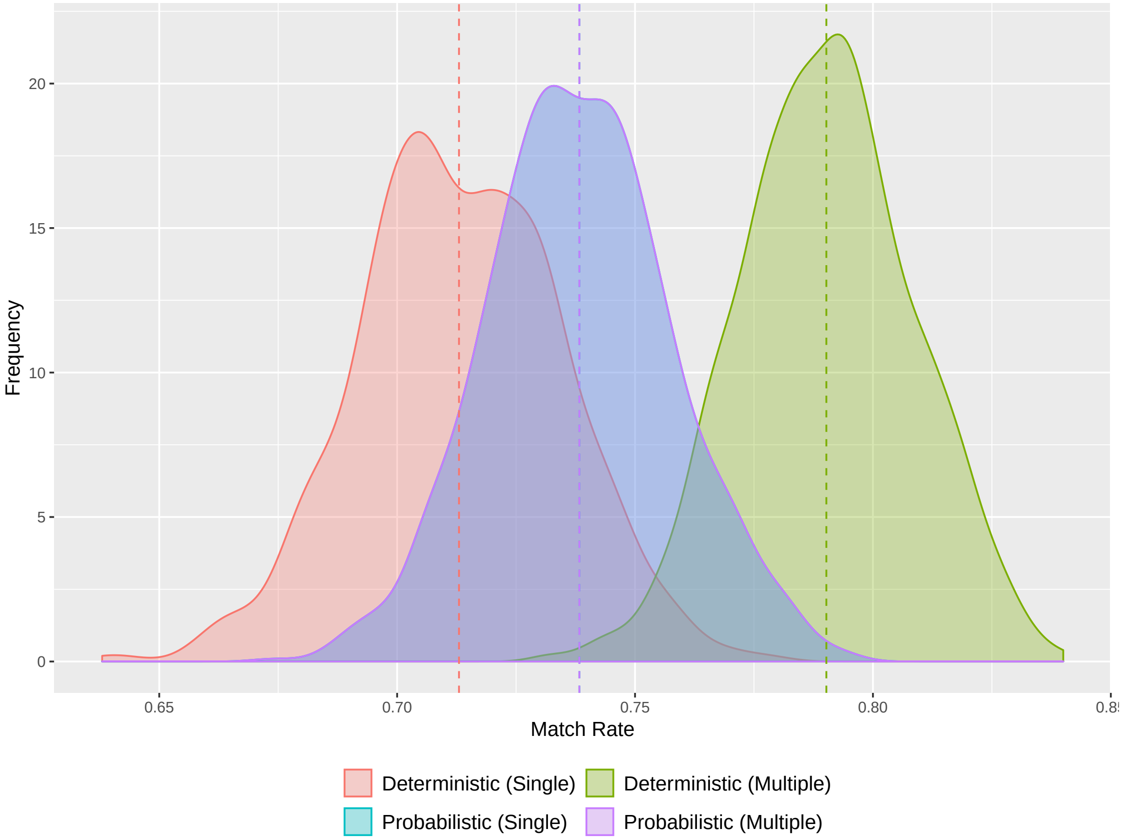
Note: Based on 1,000 simulations. Standard deviations are reported in parentheses.

6.2 Estimation Results

Based on Figure 5, the AHL estimator appears to perform better than the SW estimator for both datasets, because the SW estimator for β_0 is not centered around the true value, and the distribution of the SW estimator has fatter tails than that of the AHL estimator for all parameter values. The AHL estimator also performs better than the SW estimator as measured by the median absolute deviation across the 1,000 replications, and for both datasets (see Table7). Furthermore, the AHL estimator performs better than the OLS estimator that uses only those observations assigned $L_i = 1$ by the methods that allow L_i to vary freely, whereas the SW estimator performs worse. The AHL and SW estimators cannot be compared using the datasets matched by ABE Single or PRL Single, because they are both equal to the Naive OLS estimator when $L_i = 1$.

I calculate the AHL estimator by setting $\hat{g}(w_i, L_i)$ equal to the unconditional mean of the y_j, which, despite being the correct choice for my data generating process, means that the AHL estimator may perform better in scenarios where w or L is correlated with y. By

Figure 5: Match Rates by Linking Procedure



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

contrast, the SW estimator is biased when the matching variables w are correlated with y, so the gains in accuracy achieved by the AHL estimator relative to the SW estimator reported here can be interpreted as a lower bound.

Table 7 also compares the AHL and SW estimators to the OLS estimator that uses only the true matches appearing in the matched datasets. As expected, no estimator performs as this theoretical estimator, however neither AHL nor SW performs significantly worse in terms of median absolute deviations.

Figure 6.2 serves as both a benchmark for comparing the AHL and SW methods, as well as an additional metric for evaluating the performance of the matching methods. Note that the OLS estimator that incorporates all matches (counting multiple links as distinct observations) suffers from attenuation bias in its estimate of $\hat{\beta}_2$. This is because the additional matches, by necessity, introduce measurement error.

Figure 6: Monte Carlo Distribution of the SW and AHL Estimators

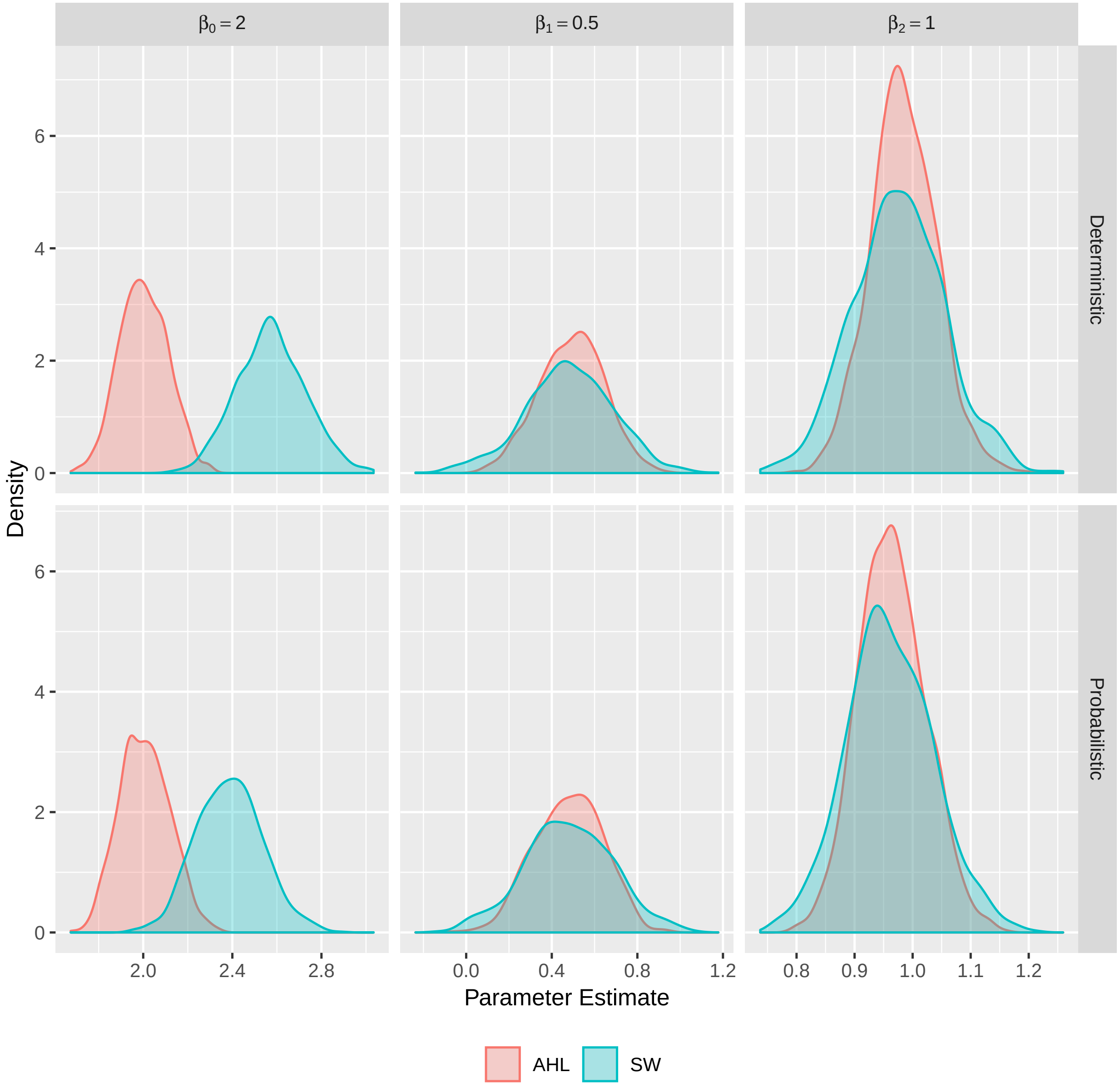


Figure 7: Monte Carlo Distribution of Benchmark OLS Estimators

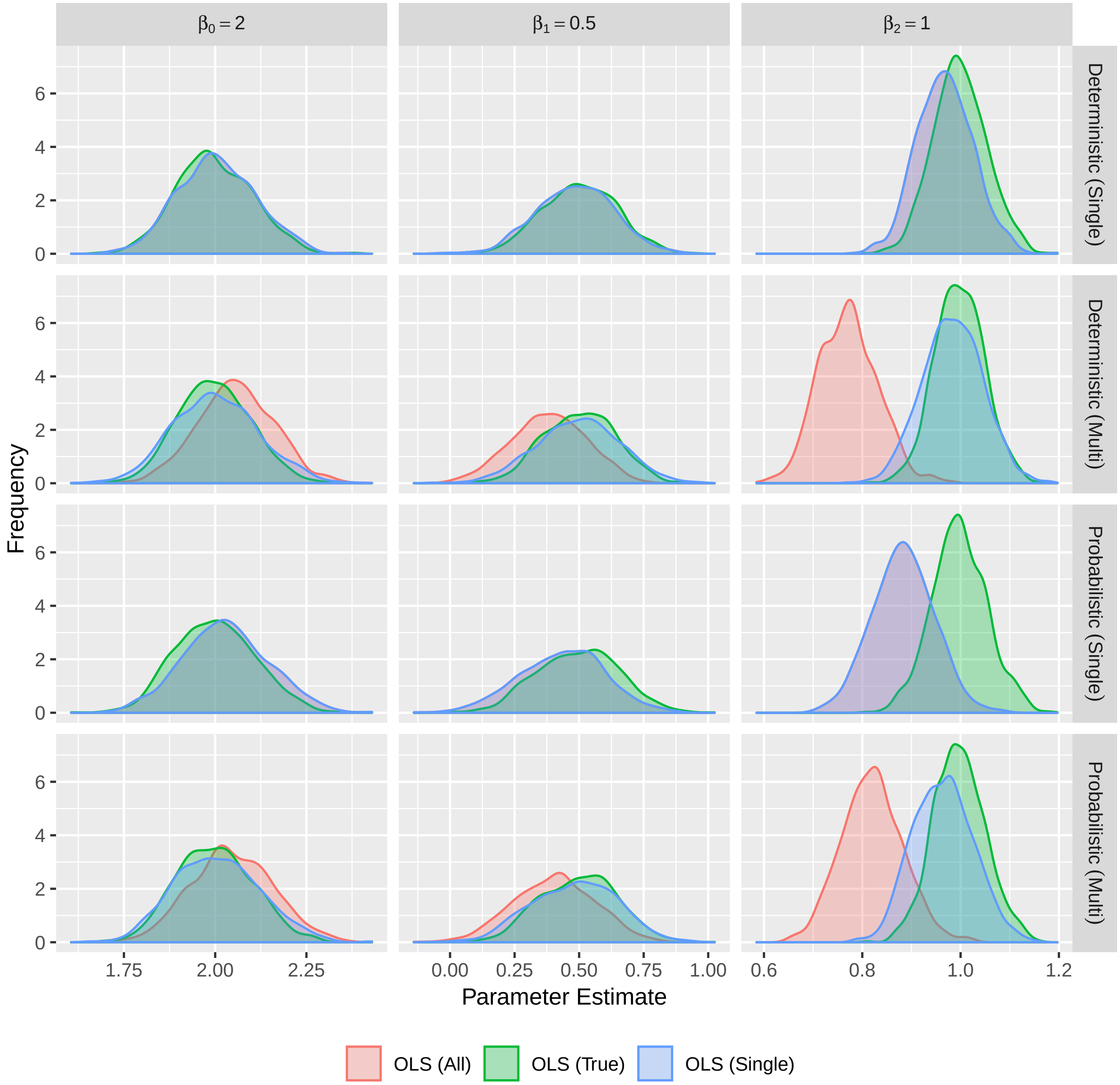


Table 7: Median Absolute Deviations for Estimators

Parameter	AHL	SW	OLS (All)	OLS (True)	OLS (Single)
Deterministic (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
Deterministic (Multiple)					
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
Probabilistic (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
Probabilistic (Multiple)					
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 Conclusion

The Monte Carlo study in this paper involved arbitrary choices about the types of identifiers and the errors added to them that may influence the results in important ways, and so further work is necessary to determine whether the patterns described in Sections 4 and 6 generalize to other settings. For now, both the theoretical and numerical analysis in this paper suggest that allowing for multiple matches may offer a solution for analyzing data that are linked with imperfect and non-unique identifiers. Furthermore, my theoretical results suggest that multiple matches should be weighted equally for asymptotically linear estimators such as OLS and GMM, unless the match probability can be estimated exactly or the sample size is very small.

References

Abramitzky, R., L. Boustan, K. Eriksson, J. Feigenbaum, and S. Perez (2019): "Automated Linking of Historical Data," *NBER Working Paper.*

Abramitzky, R., L. P. Boustan, and K. Eriksson (2012): "Europe's Tired, Poor, Huddled Masses: Self-Selection and Economic Outcomes in the Age of Mass Migration," *American Economic Review*, 102(5), 1832–56.

——— (2014): "A Nation of Immigrants: Assimilation and Economic Outcomes in the Age of Mass Migration," *Journal of Political Economy*, 122(3), 467–506.

——— (2019): "To the New World and Back Again: Return Migrants in the Age of Mass Migration," *ILR Review*, 72(2), 300–322.

Abramitzky, R., R. Mill, and S. Perez (2018): "Linking Individuals Across Historical Sources: a Fully Automated Approach," Working Paper 24324, National Bureau of Economic Research.

Aizer, A., S. Eli, J. Ferrie, and A. Lleras-Muney (2016): "The Long-Run Impact of Cash Transfers to Poor Families," *American Economic Review*, 106(4), 935–71.

Anderson, R., B. Honoré, and A. Lleras-Muney (2019): "Estimation and inference using imperfectly matched data," *Working paper.*

Bailey, M., C. Cole, M. Henderson, and C. Massey (2017): "How Well Do Automated Linking Methods Perform? Lessons from U.S. Historical Data," Working Paper 24019, National Bureau of Economic Research.

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

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977): "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1), 1–38.

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

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

Ferrie, J. P. (1996): "A New Sample of Males Linked from the Public Use Microdata Sample of the 1850 U.S. Federal Census of Population to the 1860 U.S. Federal Census Manuscript Schedules," *Historical Methods*, 29(4), 141–156.

GOLDSTEIN, H., K. L. HARRON, AND A. M. WADE (2012): "The analysis of record-linked data using multiple imputation with data value priors.," *Statistics in medicine*, 31 28, 3481–93.

HARRON, K., H. GOLDSTEIN, AND C. DIBBEN (2015): *Methodological Developments in Data Linkage*. John Wiley Sons Inc., United States.

HERZOG, T. N., F. J. SCHEUREN, AND W. E. WINKLER (2007): *Data Quality and Record Linkage Techniques*. Springer Publishing Company, Incorporated, 1st edn.

JARO, M. A. (1989): "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida," *Journal of the American Statistical Association*, 84(406), 414–420.

LAHIRI, P., AND M. D. LARSEN (2005): "Regression Analysis with Linked Data," *Journal of the American Statistical Association*, 100(469), 222–230.

LARSEN, M. D., AND D. B. RUBIN (2001): "Iterative Automated Record Linkage Using Mixture Models," *Journal of the American Statistical Association*, 96(453), 32?41.

NETER, J., E. S. MAYNES, AND R. RAMANATHAN (1965): "The Effect of Mismatching on the Measurement of Response Errors," *Journal of the American Statistical Association*, 60(312), 1005–1027.

NIX, E., AND N. QIAN (2015): "The Fluidity of Race: Passing in the United States, 1880-1940," Working Paper 20828, National Bureau of Economic Research.

ODELL, M., AND R. RUSSELL (1918): "The soundex coding system," .

PHILIPS, L. (1990): "Hanging on the Metaphone," *Computer Language Magazine*, 7(12), 39–44, Accessible at `http://www.cuj.com/documents/s=8038/cuj0006philips/`.

SAS (2019): "SAS 9.4 Functions and CALL Routines: Reference, Fifth Edition," Report, SAS Institute Inc.

SCHEUREN, F., AND W. WINKLER (1993): "Regression analysis of data files that are computer matched," *Survey Methodology*, 19.

WINKLER, W. E. (2006): "Overview of record linkage and current research directions," Discussion paper, U.S. Census Bureau.

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.

Table 8: MSE ratio for $\hat{\mu}^*$ and $\hat{\mu}^{AHL}$ for $N = 1,000$ and $(\mu, \sigma^2, \kappa, \omega^2) = (0,1,1,1)$

	$\hat{\pi}$									
π	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
0.1	0.585	1.542	0.661	0.301	0.197	0.166	0.177	0.245	0.482	1
0.2	0.094	0.220	0.652	1.273	0.761	0.460	0.380	0.424	0.647	1
0.3	0.035	0.062	0.125	0.299	0.755	1.113	0.886	0.758	0.839	1
0.4	0.018	0.028	0.048	0.091	0.195	0.458	0.884	1.027	0.985	1
0.5	0.011	0.016	0.025	0.042	0.079	0.166	0.379	0.757	0.985	1
0.6	0.007	0.010	0.015	0.024	0.042	0.080	0.177	0.424	0.839	1
0.7	0.005	0.007	0.010	0.015	0.026	0.047	0.098	0.245	0.647	1
0.8	0.004	0.005	0.007	0.011	0.017	0.030	0.062	0.154	0.482	1
0.9	0.003	0.004	0.006	0.008	0.012	0.021	0.042	0.104	0.360	1

8.2 Variance formulas

$\mathrm{Var}(X_1)$ and $\mathrm{Var}(X_2)$ 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:

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

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

8.3 Additional MSE Tables

Table 9: MSE ratio for $\hat{\mu}^*$ and $\hat{\mu}^{AHL}$ for $N = 1,000$ 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

Table 10: MSE ratio for $\hat{\mu}^*$ and $\hat{\mu}^{AHL}$ for $N = 1,000$ and $(\mu, \sigma^2, \kappa, \omega^2) =$ (0,4,1,2)

	$\hat{\pi}$									
π	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
0.1	0.938	1.288	0.932	0.622	0.478	0.431	0.452	0.555	0.776	1
0.2	0.258	0.522	0.959	1.150	0.953	0.770	0.701	0.736	0.869	1
0.3	0.105	0.185	0.348	0.646	0.980	1.063	0.976	0.921	0.947	1
0.4	0.056	0.089	0.152	0.275	0.499	0.802	0.997	1.015	0.995	1
0.5	0.034	0.052	0.083	0.140	0.250	0.451	0.733	0.947	1.002	1
0.6	0.023	0.034	0.051	0.083	0.143	0.259	0.475	0.769	0.967	1
0.7	0.017	0.023	0.035	0.054	0.091	0.162	0.310	0.583	0.899	1
0.8	0.012	0.017	0.025	0.038	0.062	0.110	0.211	0.434	0.809	1
0.9	0.010	0.013	0.019	0.028	0.045	0.078	0.151	0.327	0.714	1

Table 11: MSE ratio for $\hat{\mu}^*$ and $\hat{\mu}^{AHL}$ for $N = 1,000$ and $(\mu, \sigma^2, \kappa, \omega^2) = $ (0,1,4,2)

	$\hat{\pi}$									
π	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
0.1	0.133	1.399	0.331	0.160	0.123	0.121	0.146	0.221	0.463	1
0.2	0.016	0.062	0.322	1.146	0.585	0.361	0.324	0.391	0.630	1
0.3	0.006	0.016	0.047	0.157	0.574	1.052	0.830	0.726	0.828	1
0.4	0.003	0.007	0.017	0.044	0.122	0.356	0.823	1.012	0.981	1
0.5	0.002	0.004	0.009	0.020	0.047	0.120	0.321	0.719	0.978	1
0.6	0.001	0.003	0.005	0.011	0.025	0.057	0.145	0.387	0.822	1
0.7	0.001	0.002	0.004	0.007	0.015	0.033	0.079	0.219	0.624	1
0.8	0.001	0.001	0.003	0.005	0.010	0.021	0.050	0.136	0.459	1
0.9	0.001	0.001	0.002	0.004	0.007	0.015	0.034	0.092	0.340	1

Table 12: MSE ratio for $\hat{\mu}^*$ and $\hat{\mu}^{AHL}$ for $N = 1,000$ and $(\mu, \sigma^2, \kappa, \omega^2) = $ (0,1,1,10)

	$\hat{\pi}$									
π	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
0.1	2.247	4.501	1.933	0.739	0.386	0.263	0.230	0.274	0.508	1
0.2	0.400	0.730	1.574	2.456	1.414	0.729	0.503	0.488	0.694	1
0.3	0.153	0.218	0.344	0.623	1.192	1.546	1.150	0.883	0.902	1
0.4	0.080	0.101	0.136	0.202	0.335	0.612	1.010	1.125	1.033	1
0.5	0.049	0.058	0.072	0.096	0.140	0.231	0.425	0.757	0.980	1
0.6	0.033	0.037	0.044	0.055	0.075	0.114	0.202	0.412	0.790	1
0.7	0.024	0.026	0.030	0.036	0.047	0.067	0.114	0.237	0.587	1
0.8	0.018	0.019	0.021	0.025	0.032	0.044	0.072	0.150	0.428	1
0.9	0.014	0.015	0.016	0.019	0.023	0.031	0.050	0.102	0.317	1