Regression analysis with linked data

Rachel Anderson*

This Version: October 16, 2019

Abstract

This paper compares different methods for estimating parametric models with linked data, i.e. when x and y are observed in distinct datasets with imperfect identifiers. This setup requires that the researcher must attempt to identify which observations in the x- and y-datafiles refer to the same individual, prior to performing inference about the joint or conditional distributions of x and y. At a minimum, random errors in the matching step introduce measurement error that must be accounted for in subsequent inference; however, additional concerns about sample selection arise when these errors are correlated with unobservables that affect x or y.

1 Introduction

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

*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).

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

Although frequently encountered in economics, problems related to merging also appear in many other fields, including statistics, computer science, operations research, and epidemiology, and under many names. such as record linkage, data linkage, entity resolution, instance identification, de-duplication, merge/purge processing, and entity reconciliation. In economics, interest in matching has emerged in response to the increasing use of large administrative datasets across the profession, and because errors in matching may unintentionally introduce sample selection and bias results. While there are a number of recent papers that compare the performance of popular matching methods on the representativeness and accuracy of the datasets that they produce, the literature on how to perform estimation using linked data is notably lacking (Abramitzky et al., 2019, 2018; Bailey et al., 2017).

In the past, authors have handled multiple matches by generating a "composite match" equal to the average of the linked observations (Bleakley and Ferrie, 2016), constructing bounds on the parameter of interest using different configurations of matched data (Nix and Qian, 2015), or using methods that allow for multiple outcomes (Anderson et al., 2019). In statistics, authors have suggested using probabilistic record linkage and multiple imputation to correct for bias introduced by errors in the matching process Scheuren and Winkler (1993); Lahiri and Larsen (2005); Goldstein et al. (2012).

This paper contributes to the literature by developing a unified approach for inference using linked data. Building upon a well-established literature that compares the performance of different record linkage methods, this paper studies how the *outputs* of these procedures – such as whether the linked dataset contains multiple matches per observation, or probabilities that any given record pair refers to a true match – can be used to correct for bias and improve efficiency during estimation.

My preliminary results 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

This paper considers the problem of 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. We 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. Also, since $N_y \geq N_x$, 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.

This matching step consists of constructing a 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 the obvious choice is to set $\varphi(i) = j$ if and only if $w_i = w_j$, and $\varphi(i) = \varnothing$ otherwise. In practice, however, w_i and w_j may contain variables that are not unique and prone to typographical error, so that it is difficult to distinguish false links from true links, or discern among multiple links with identical w_j.

To fix ideas, consider again the example of Aizer et al. (2016), who seek to estimate the effect 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 observations of y_j, person j's age at death. Both of the x- and y-datafiles also contain w_i and w_j, which include the observation's first and last names and year of birth, so that common names and typographical error are likely to complicate even the most conservative assignment of φ.

In light of these challenges, researchers have developed a variety of automated record linkage methods for constructing φ, some of which are described in Section 4. Such procedures result in a dataset $\{x_i, \{y_{i\ell}\}_{\ell=1}^{L_i}\}_{i=1}^{N}$, such that a single observation x_i may be linked to L_i possible values of y. Note that $N_x \geq N$, which holds with equality if and only if every observation x_i is assigned a match. Additionally, some record linkage procedures output values $\pi_{i\ell}$ equal to the estimated probability that $y_{i\ell}$ is the true match, and $\sum_{\ell=1}^{L_i} \pi_{i\ell} = 1$ for all i.

As there is already a well-established literature that compares the performance of different record linkage methods, this paper studies how the *outputs* of these procedures – such as whether multiple matches are allowed or estimates of $\pi_{i\ell}$ are available – can be used to

improve the estimation step and correct for bias introduced by errors in the matching step.

3 Numerical Example

The purpose of this section is to introduce a numerical example that will be used to illustrate the different matching techniques discussed in this paper. The benefits of using synthetic data are that I can control the degree of similarity among identifying variables, and overlap between datasets, all while knowing the true match status of each (x_i, y_j) record pair. As a result, I can compare how sensitive my results are to data quality, and compare how the various matching and estimation procedures perform relative to the correctly specified model applied to a dataset containing only correct links.

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, $x_{1i} \overset{i.i.d}{\sim}$ Bernoulli(0.5), and $x_{2i} \overset{i.i.d}{\sim} \mathcal{N}(0, 2)$. The y_i values are generated according to the linear relationship,

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

with $(\beta_0, \beta_1, \beta_2, \sigma^2) = (2, 0.5, 1, 2)$, so that estimating the correctly specified linear regression model yields an R^2 value of approximately 0.50. Each observation is assigned a vector of identifying variables, w_i, which includes a first and last name drawn at random from a list of first and last names[1], and a random birthday between January 1, 1900, and December 31, 1925. The resulting dataset looks like the observations in the top panel of Figure 1.

Next, I split the ground truth dataset into the x- and y-datafile, which contain the (x_1, x_2, w) and (y, w) values respectively. To construct the x-datafile, I select 500 observations

[1]The first and last name lists contain 41 and 24 names, respectively, and can be found in the replication files. Note that the number of possible names is smaller than the number of observations to ensure that there are multiple observations with the same name.

at random from the ground truth dataset, and introduce random errors in their corresponding identifiers. To make the synthetic data resemble a real application, I set the probabilities of introducing transcription errors equal to those reported for the 1940 Census data in Abramitzky et al. (2019). These errors include deleting characters (e.g., "Anderson" becomes "Andersn"), exchanging vowels (e.g., "Rachel" becomes "Rachal"), and swapping English phonetic equivalents (e.g. "Ellie" becomes "Elie"). I also add normally distributed errors to the birth day, month, and year. The probabilities of introducing an error are set to match the transcription error rates reported in the 1940 Census by Abramitzky et al. (2019); for example, 7% of observations have misreported first names and 17% of observations have misreported last names. The bottom panel of Figure 1 illustrates how the x- and y-datafiles are split visually.

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

4 Record Linkage Methods

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

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

constitute (partial) agreements, and (iv) designating (i, j) pairs as matches.[2]

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

Step (ii) reduces the computational burden of a matching procedure when $N_x \times N_y$ is large by partitioning $X \times Y$ into "blocks." Only records within the same block are attempted to be matched, while records in different blocks are assumed to be non-matches. Blocking variables should be recorded with minimal error, otherwise blocking may adversely affect the Type II error rate.

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

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

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

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

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

(iv) Designate as matches all (i, j) pairs with Jaro-Winkler scores exceeding a pre-determined cut-off; and, if a record i has multiple possible matches that exceed the cut-off, then

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

choose the corresponding j with the highest score (or pick one match at random if there is a tie).

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

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

Except in rare cases, the estimated matching functions obtained by switching (iv) and (iv*) will differ, if only because the former method matches each x with at most one y, the latter potentially matches the same x with multiple y. This example also illustrates the difference between deterministic and probabilistic record linkage methods: while (iv) uses pre-determined rules to designate pairs as matches, (iv*) uses statistical theory to inform the selection of the decision rule. Probabilistic record linkage also involves the estimation of match weights, which can be incorporated in subsequent estimation steps.

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

4.1 Deterministic

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

1. Clean names in the x- and y- datafiles to remove any non-alphabetic characters and account for common mis-spellings and nicknames (e.g., so that Ben and Benjamin would be considered the same name).

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

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

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

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

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

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

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

fail.

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

4.2 Probabilistic Record Linkage

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

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

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

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

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

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

$$P(\boldsymbol{\gamma_{ij}}) = P(\boldsymbol{\gamma_{ij}}|M)p_M + P(\boldsymbol{\gamma_{ij}}|U)p_U \tag{3}$$

where $P(\boldsymbol{\gamma_{ij}}|M)$ and $P(\boldsymbol{\gamma_{ij}}|U)$ are the probabilities of observing the pattern $\boldsymbol{\gamma_{ij}}$ conditional on the record pair (i, j) belonging to M or U, respectively. The proportions p_M and $p_U = 1 - p_M$ are the marginal probabilities of observing a matched or unmatched pair. Applying Bayes' Rule, we obtain the probability of $(i, j) \in M$ conditional on observing $\boldsymbol{\gamma_{ij}}$,

$$P(M|\boldsymbol{\gamma_{ij}}) = \frac{p_M P(\boldsymbol{\gamma_{ij}}|M)}{P(\boldsymbol{\gamma_{ij}})} \tag{4}$$

Thus, if we can estimate p_M, $P(\boldsymbol{\gamma_{ij}}|M)$ and $P(\boldsymbol{\gamma_{ij}}|U)$, then we can estimate the probability that any two records refer to the same entity using (4). These probabilities can then be used to designate pairs as matches, or to estimate the false positive rate associated with a particular match configuration using the formulas in Fellegi and Sunter (1969).

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

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

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

$$P(\gamma_{ij}|C) = \prod_{k=1}^{K} P(\gamma_{ij}^k|C)^{\gamma_{ij}^k}(1 - Pr(\gamma_{ij}^k|C))^{1-\gamma_{ij}^k} \qquad C \in \{M, U\} \tag{5}$$

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.

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

Since membership to M or U is not actually observed, a convenient way of simultaneously estimating p_M, p_U and classifying record pairs as matches or non-matches is via mixture modeling, with mixture distributions $P(\boldsymbol{\gamma_{ij}}|M)$ and $P(\boldsymbol{\gamma_{ij}}|U)$. The parameters can be estimated using the expectation-maximization (EM), first applied to record linakge by Larsen and Rubin (2001). For this paper, I use the `fastLink` algorithm developed by Enamorado et al. (2019).

5 Estimation with linked data

Consider first the case where each x observation is linked to a single value of y, i.e. $L_i = 1$ for all i. The data consists of (x_i, z_i) for $i = 1, \ldots, N$, where z_i may or may not correspond to 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,

$$\hat{\beta}_N = (X'X)^{-1}X'z \tag{6}$$

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 $q_i = (q_{i1}, \ldots, q_{iN_y})'$, we can write the bias of $\hat{\beta}_N$ conditional on the observed values of y as,

$$\text{bias}(\hat{\beta}_N|y) = E[(\hat{\beta}_N - \beta)|y] = (X'X)^{-1}X'B \tag{7}$$

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 (7), Scheuren and Winkler (1993) proposed estimating $\hat{B}$ to correct for the bias of $\hat{\beta}_N$. To reduce the computational burden of constructing $\hat{B}$, they suggest using the first and second highest elements of the vector q_{ij_1} and q_{ij_2} and their corresponding values y_{ij_1} and y_{ij_2} to compute $\hat{B}_i^{TR} = (q_{ij_1} - 1)y_{ij_1} + q_{ij_2}y_{ij_2}$, and then calculating

$$\hat{\beta}_{SW} = \hat{\beta}_N - (X'X)^{-1}X'\hat{B}^{TR} \tag{8}$$

Although $\hat{B}^{TR}$ can incorporate an arbitrary number of elements of q_i, Scheuren and Winkler (1993) note that if the probability is high that the best candidate link is the true link, then the truncation with two links results in a very small bias.

The downside of the Scheuren and Winkler (1993) method is that it requires knowledge of q_{ij}, as well as the second most likely value of y for each observation of x. This information is not typically available when using deterministic matching procedures such as those developed by Abramitzky et al. (2012). However, even if estimates of q_{ij} are available (as may be the case when using probabilistic record linkage), the bias may persist if the estimates $\widehat{q_{ij}}$ are correlated with x or y, as this would introduce endogeneity.

The endogeneity problem arises because $\widehat{q_{ij}}$ are typically calculated by plugging in estimates of the parameters $\psi \equiv \{p_M, P(\boldsymbol{\gamma_{ij}}|M),\ P(\boldsymbol{\gamma_{ij}}|U)\}$ into equation (4). Thus, $\widehat{q_{ij}}$ will be correlated with x or y if errors in the matching variables, which determine the distribution of $\hat{\psi}$, are correlated with x or y. This is likely to be a problem in economics applications, such as in Nix and Qian (2015), where y measures whether a person's recorded ethnicity changes between Census years, but changes in names (the matching variables) are also strongly correlated with y.

One possible solution to the challenges described above is to use the estimator from Anderson et al. (2019), which is unbiased if the errors in estimating the conditional expectation of y be independent of the x_i, which holds if x_i and y_i are random samples conditional on

the matching variables[4]. Specifically, for the model in (1), the AHL estimator is 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 $\sigma(X_i) = \sigma_\varepsilon^2 + (L_i - 1)\sigma_\nu^2$.

In practice, the AHL is estimated in three steps: (i) estimating $\hat{g}(w_i, L_i)$ using nonparametric methods such as k-Nearest Neighbors, local polynomial regression, or kernel density estimators; (ii) estimating $\hat{\beta}$ by applying OLS to (9) to construct $\hat{\sigma}(X_i)$, and (iii) computing $\hat{\beta}^{WLS}$ using the formula in (10). The resulting estimator is consistent and asymptotically normal under the regularity conditions described in Anderson et al. (2019).

Like the Scheuren and Winkler (1993) estimator, the AHL estimator requires that the true match for x_i is included among the matches $\{y_{i\ell}\}_{\ell=1}^{L_i}$ for all ℓ. The simulations in Section 7 suggest that this is a reasonable assumption when multiple matches are allowed. When $L_i = 1$ for all i, the AHL estimator reduces to the OLS estimator $\hat{\beta}_N$, so it is only meaningful to compute for linked datasets with some $L_i > 1$.

The AHL estimator also requires that x and y are random samples conditional on the matching variables. Practically speaking, this means thats all individuals with the same

[4]Technically, the result in Anderson et al. (2019) is proven for GMM, so the necessary condition is that errors in estimating a conditional moment condition are independent of x_i

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.

Other approaches for regression analysis using linked data have been proposed by Lahiri and Larsen (2005) and Neter et al. (1965), however neither is appropriate for the setup described in Section 2. The methods in Lahiri and Larsen (2005) assume that each observation appearing y-datafile is generated according to the DGP in (1), and that its corresponding value of x appears in the x-datafile. This is a problem both conceptually and for implementation when entries in the x-datafile represent a strict subset of observations in the y-datafile. The methods in Neter et al. (1965) are simplified versions of those in Scheuren and Winkler (1993), and hence face the same implementation issues described above.

By contrast, the AHL estimator is agnostic about whether the incorrectly linked y_j are generated by (1) or by some other data generating process. The AHL estimator has additional robustness properties due to the fact that it weights multiple matches equally and does not require estimating match probabilities, which are explored in the following section.

6 Optimality of the AHL Estimator

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

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

$$\hat{\mu}_1 = \frac{X_1}{\pi} - \frac{1-\pi}{\pi}\kappa \tag{11}$$

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

$$\hat{\mu_2} = \frac{X_2}{1-\pi} - \frac{\pi}{1-\pi}\kappa \tag{12}$$

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

$$\hat{\mu} = a_1 X_1 + a_2 X_2 - a_3\kappa \tag{13}$$

which has the following expectation,

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

so that unbiased, requires

$$a_1\pi + a_2(1-\pi) = 1 \implies a_2(a_1) = \frac{1}{1-\pi} - \frac{a_1\pi}{1-\pi} \tag{14}$$

$$a_1(1-\pi) + a_2\pi = a_3 \implies a_3(a_1) = \frac{\pi}{1-\pi} + \frac{a_1 - 2a_1\pi}{1-\pi} \tag{15}$$

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

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

When $a_1 = \frac{1}{\pi}$, then $\hat{\mu} = \hat{\mu}_1$; and if $a_1 = 0$ then $\hat{\mu} = \hat{\mu_2}$.

We can write:

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

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

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

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

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

$$\begin{aligned}\text{Var}\left(\hat{\mu}_1\right) &= \frac{\text{Var}(X_1)}{\pi^2} = \frac{1}{\pi^2}\left(\pi\sigma^2 + (1-\pi)\omega^2 + \pi(1-\pi)(\mu-\kappa)^2\right) \\ \text{Var}\left(\hat{\mu}_2\right) &= \frac{\text{Var}\left(X_2\right)}{(1-\pi)^2} = \frac{1}{(1-\pi)^2}\left((1-\pi)\sigma^2 + \pi\omega^2 + \pi(1-\pi)(\mu-\kappa)^2\right)\end{aligned}$$

This follows from the law of total variance, with the random variable $D = 1$ if X_1 is drawn from the correct distribution (and X_2 is drawn from the incorrect distribution), and $D = 0$ otherwise.

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

Similarly,

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

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

$$\begin{aligned} g(\sigma^2, \omega^2, (\mu-\kappa)^2, x) &\equiv \frac{1}{x^2}\left(x\sigma^2 + (1-x)\omega^2 + x(1-x)(\mu-\kappa)^2\right) \\ \text{Var}\left(\hat{\mu}_1\right) &= g(\sigma^2, \omega^2, (\mu-\kappa)^2, \pi) \\ \text{Var}\left(\hat{\mu}_2\right) &= g(\sigma^2, \omega^2, (\mu-\kappa)^2, 1-\pi) \end{aligned}$$

Importantly,

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

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

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

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

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

are known for all i, then the optimal estimator is

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

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

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

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

7 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 1: 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 2: 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.

7.1 Matching results

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

- the proportion of observations in the x-datafile that are linked to at least one observation in the y-datafile (match rate),
- the total number of links made by the matching algorithm,
- the proportion of links that are incorrect (Type I error rate),

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

For the linked datasets that contain multiple matches per observation, I report also the average number of links per observation, and how often those links include the true match (Table 2).

As seen in Table 1, the average match rates range between 71 and 79 percent across the various matching procedures, but plotting the distribution of match rates in Figure 2shows 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 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 2, which shows that allowing multiple matches improves the accuracy of the ABE algorithm. Table 2 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 1, 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.

7.2 Estimation Results

I compare the estimators according to median absolute deviation, and plot histograms of the estimated values in Figures 3-9. 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.

8 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 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.

Table 3: Median Absolute Deviations for Estimators

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

9 Appendix: 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.

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

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.*

Figure 1: Creation of Synthetic Datasets

ID	y	x_1	x_2	First Name	Last Name	Birthday
1	y_1	$x_{1,1}$	$x_{2,1}$	Tyler	Ashenfelter	1915-05-13
2	y_2	$x_{1,2}$	$x_{2,2}$	Brandon	Christensen	1904-06-27
				$\vdots$		
195	y_{195}	$x_{1,195}$	$x_{2,195}$	Samantha	Andersen	1914-08-18
196	y_{196}	$x_{1,196}$	$x_{2,196}$	Victoria	Andersen	1918-11-25
				$\vdots$		
1000	y_{500}	$x_{1,500}$	$x_{2,500}$	Vicky	Anderson	1915-04-14

$\downarrow$

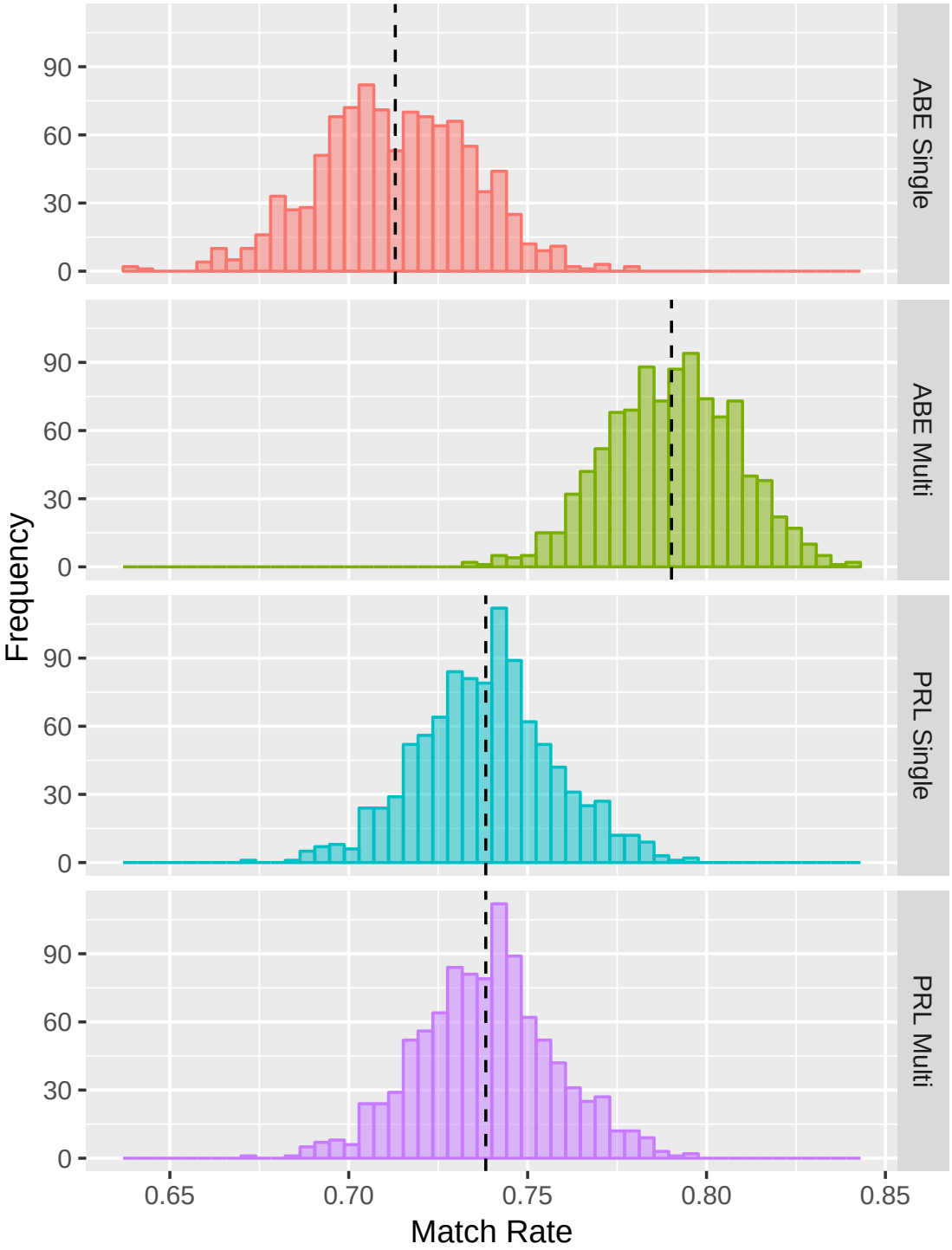
x-Datafile

ID	x	Name	Birthday
2	$(x_{1,2}, x_{2,2})$	Branden Christenson	1905-06-27
		...	
195	$(x_{1,195}, x_{2,195})$	Samantha Anderson	1914-08-21
198	$(x_{1,198}, x_{2,198})$	Jon Smyth	1918-12-20
		...	
1000	$(x_{1,1000}, x_{2,1000})$	Vic Andersn	1915-04-14

y-Datafile

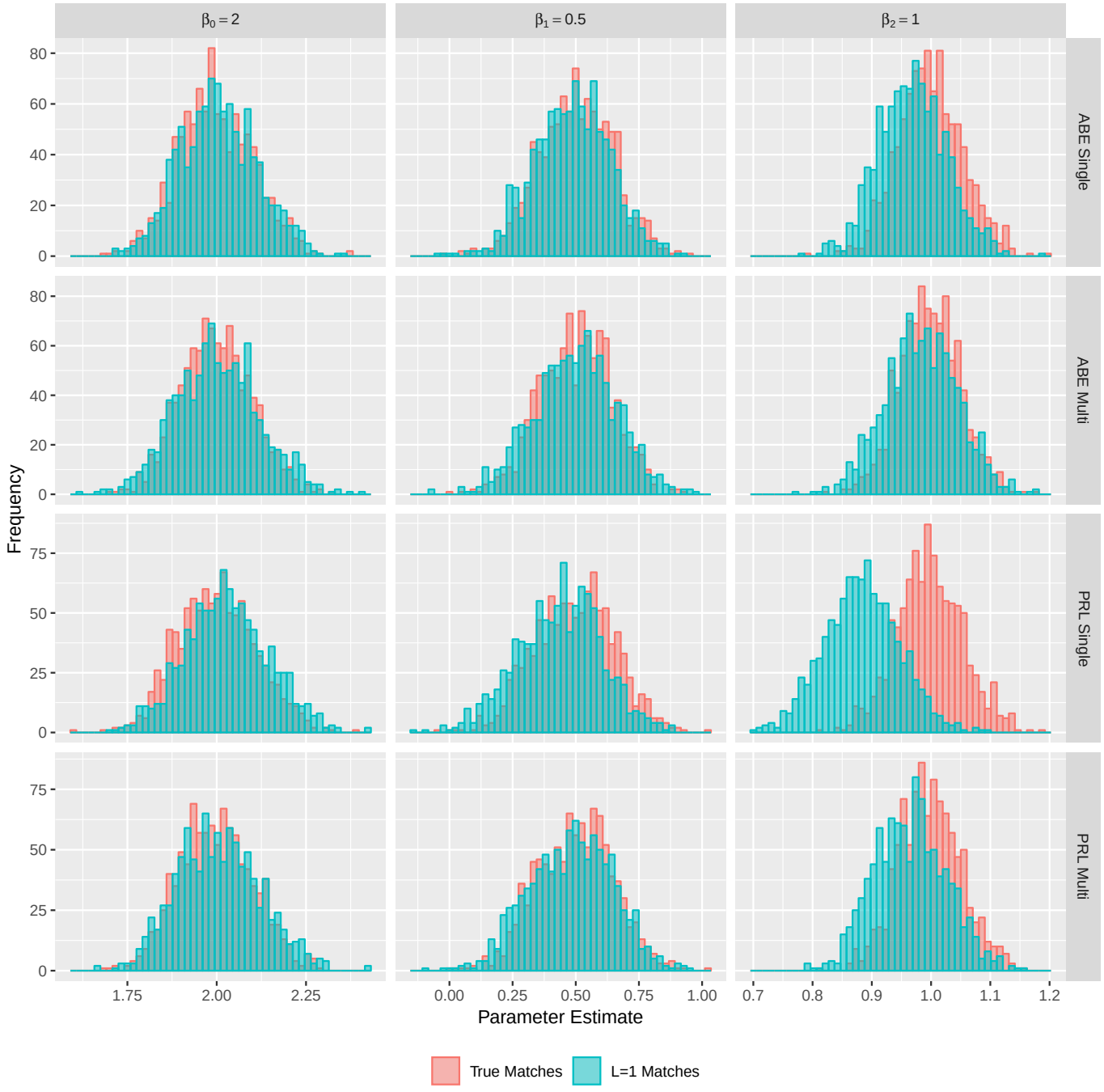
ID	y	Name	Birthday
1	y_1	Tyler Ashenfelter	1915-05-13
2	y_2	Brandon Christensen	1904-06-27
		...	
195	$y_{1,195}$	Samantha Anderson	1914-08-18
		...	
1000	y_{1000}	Vicky Anderson	1915-04-14

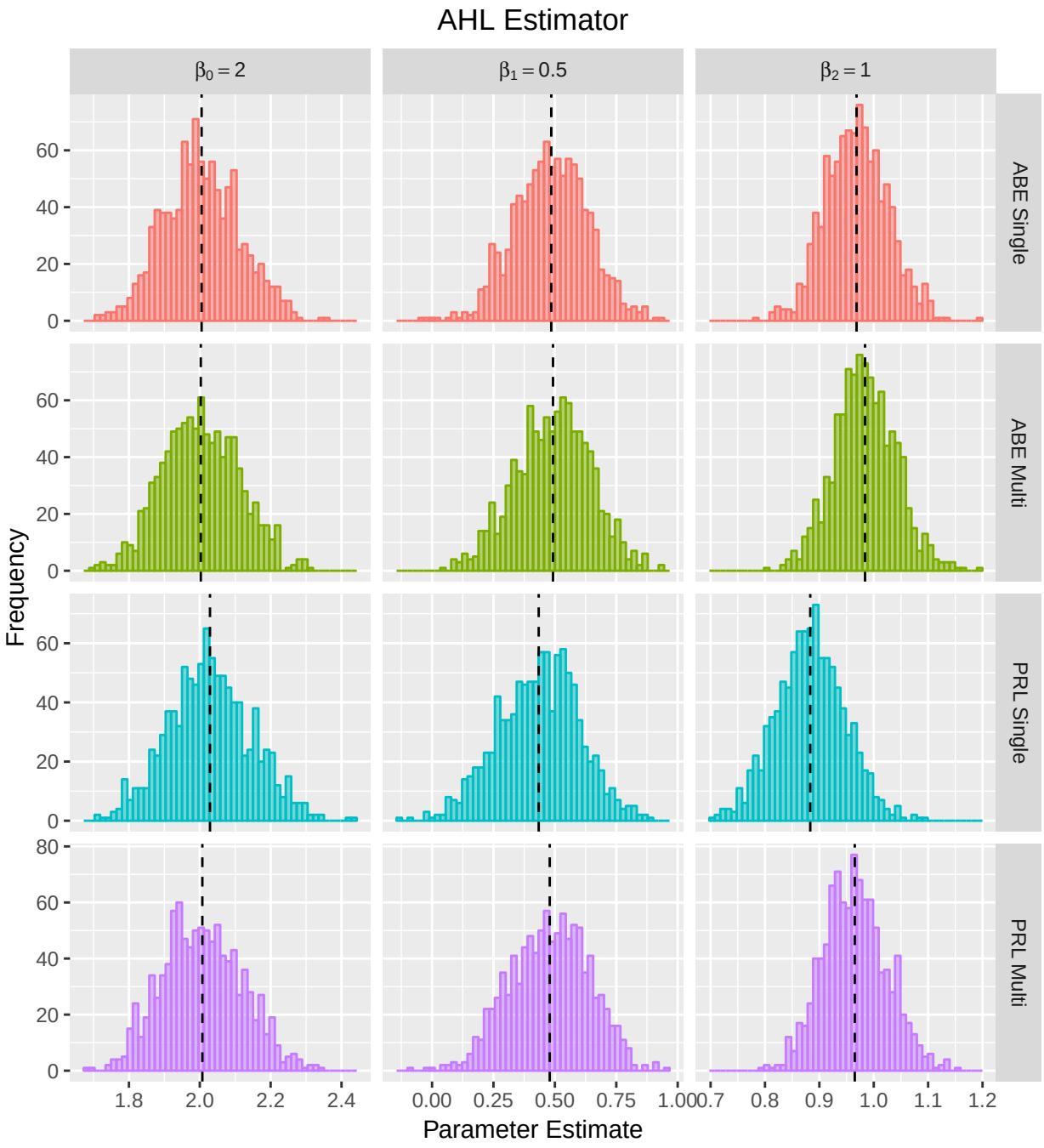
Figure 2: Match Rates by Linking Procedure



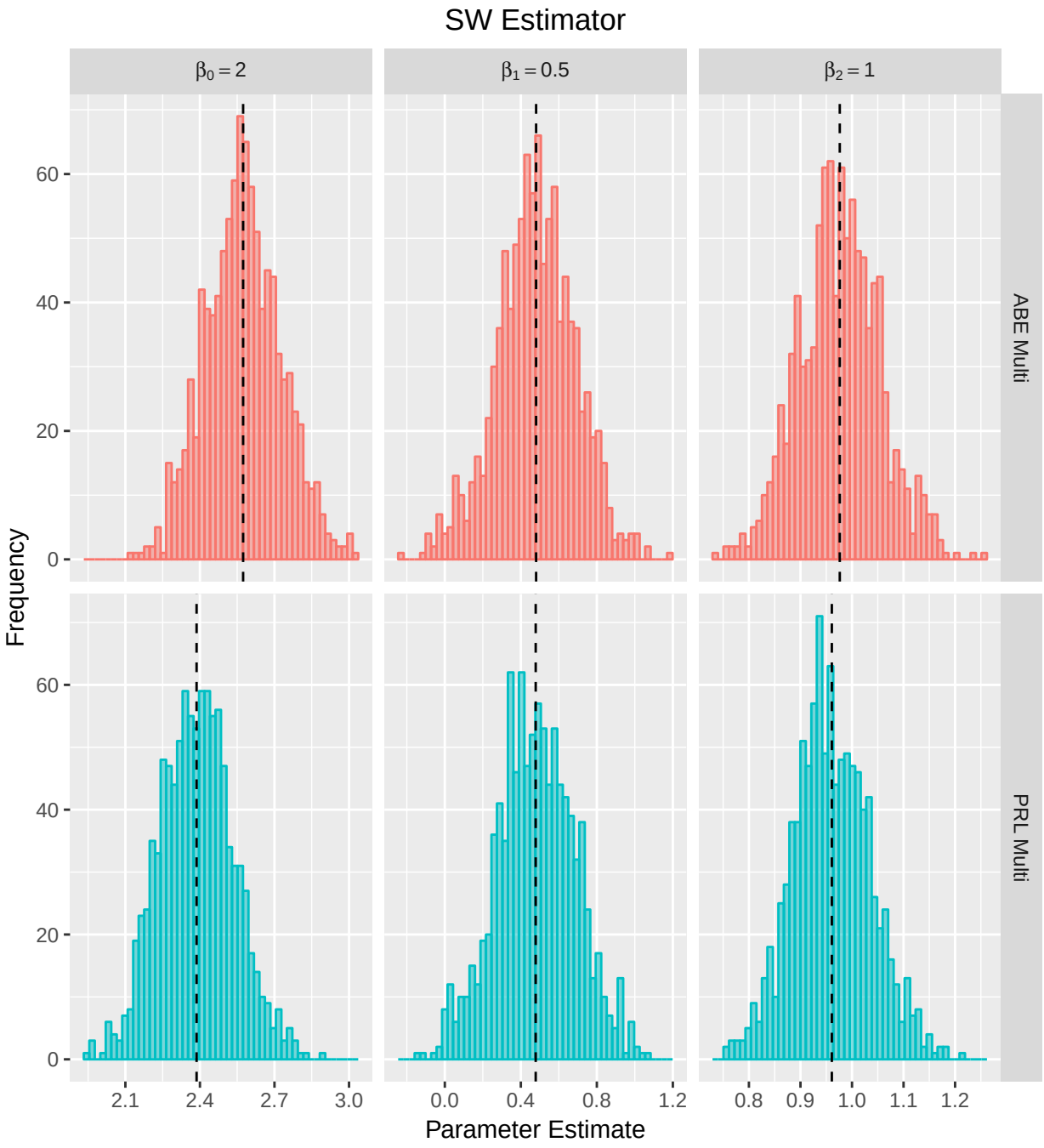
*Based on 1,000 simulations. Vertical line indicates the sample mean.

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

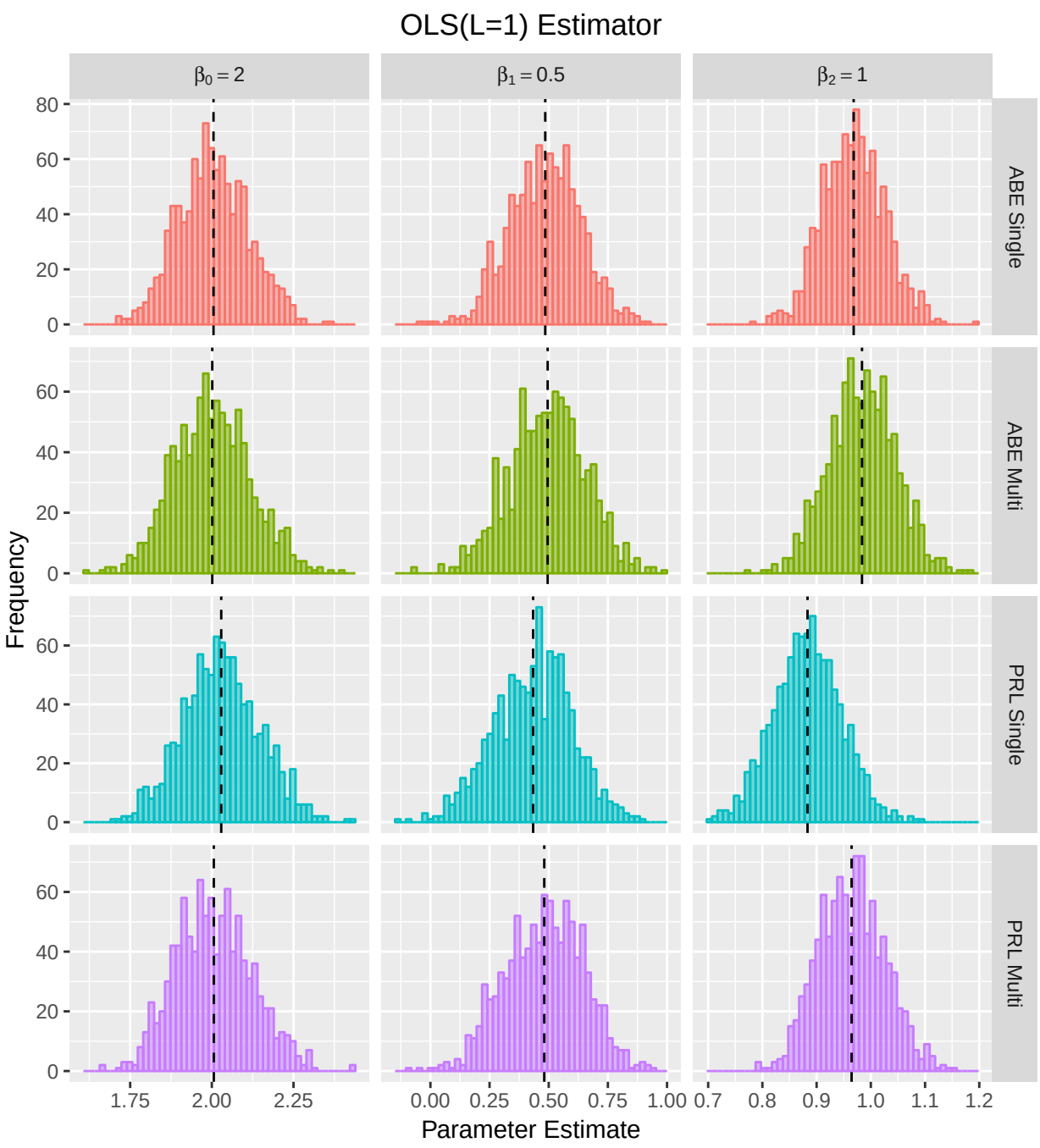




*Based on 1,000 simulations. Vertical line indicates the sample mean.



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



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