# Implicates as Instrumental Variables: An Approach for Estimation and Inference with Probabilistically Matched Data*

Dhiren Patki† Matthew D. Shapiro‡

January 2023

**Abstract**

Linkage errors in probabilistically matched data sets can cause biases in the estimation of regression coefficients. This paper proposes an approach to obtain consistent estimates and valid inference that relies on instrumental variables. The novelty of the method is to show that instrumental variables arise naturally in the course of probabilistic record linkage thereby allowing for off-the-shelf implementation. Relative to existing approaches, the instrumental variable approach does not require integration of the record linkage and regression analysis steps, the estimation of complex models of linkage error, or computationally expensive methods to estimate standard errors. The instrumental variables approach performs well in Monte Carlo simulations of an environment highlighting a many-to-one linkage problem.

**Keywords:** Instrumental variables, regression analysis, record linkage.

*This research is supported by the Alfred P. Sloan Foundation through the CenHRS project at the University of Michigan. Any opinions and conclusions expressed herein are those of the authors and do not indicate concurrence by the Federal Reserve Bank of Boston, the principals of the Board of Governors, or the Federal Reserve System. All errors are our own. This is a pre-copyedited, author-produced version of an article accepted for publication in the *Journal of Survey Statistics and Methodology* following peer review. The version of record is available online at: https://doi.org/10.1093/jssam/smad005.

†Federal Reserve Bank of Boston. Email: Dhiren.Patki@bos.frb.org

‡Department of Economics and Institute for Social Research, University of Michigan. Email: shapiro@umich.edu

# 1 Introduction

Record linkage is an increasingly important data processing method for measurement and analysis. For example, statistical agencies that produce survey data can reduce respondent burden and lower costs by using pre-existing information on survey respondents that is held within administrative registers (see, e.g., National Academies of Sciences Engineering and Medicine (2017)). Record linkage can also broaden the analytic scope of survey or administrative data sets by adding variables that are not otherwise contained within a given data set (see, e.g., Warren et al. (2002) and Abowd et al. (2021)). Probabilistic record linkage arises in a setting where there are no unique identifiers to facilitate unambiguous matching of record pairs. In this context, linkage errors occur when a record in a given file is incorrectly paired to a candidate match in another file. If the linkage errors are random, estimating parameters of the joint relationship between variables, such as linear regression coefficients, can suffer from attenuation bias. Neter et al. (1965) provides an early overview of this concern.

There are two well-developed solutions that allow researchers to correct for linkage-error induced biases. The first approach exploits information from the linkage step to undo biases that arise in the regression analysis step. Examples of this approach include Scheuren and Winkler (1993) and Lahiri and Larsen (2005), who use match probabilities as weights to reduce or entirely eliminate biases in regression coefficients obtained using probabilistically matched data. The second approach does not require the linkage and the analysis steps to be integrated. Examples of this approach include Slawski and Ben-David (2019) and Slawski et al. (2021) who rely on sufficiently well matched data to model the unobserved permutations that give rise to linkage errors. Chambers and da Silva (2020) use information about linkage accuracy rates to correct for attenuation bias in regression coefficients.

In this paper we propose a new solution to obtain consistent estimates and valid inference for regression coefficients when analyzing data subject to linkage errors. Our approach builds on the well-known instrumental variables (IV) estimator that corrects for bias arising from measurement error in regressors (see, e.g., Bound et al. (2001)). The idea behind the IV approach is to first find a variable that has two features: it is uncorrelated with the measurement error, but is correlated with the mismeasured variable. Variables that meet these two criteria allow analysts to purge measurement error from the mismeasured variable and isolate variation arising from the unobserved true variable, thereby yielding consistent estimates of the regression coefficient of interest. Despite their theoretical appeal, instrumental variables are difficult to find in practice. The core contribution of our paper is to show that valid instruments arise naturally in the course of record linkage, which allows for off-the-shelf application of the IV approach. Relative to existing approaches, the IV approach is straightforward in that it does not require integration of the linkage and analysis steps or the estimation of complex models of linkage error. Furthermore, while the estimation of standard errors of regression coefficients with existing approaches is computationally expensive, the IV estimator obeys asymptotic normality thereby facilitating simple computation of standard errors and confidence intervals.

To see how instrumental variables arise in probabilistic record linkage, consider a generic example where researchers are interested in linking records in file A to records in file B. In the absence of unique identifiers, researchers estimate the likelihood that each candidate from file B is the right match for a record in file A, conditional on a set of available variables. It is common, in this type of setting, to enforce one-to-one linkage by selecting the most likely match. The key insight of our approach is to recognize that candidates eliminated when enforcing one-to-one matching share many characteristics in common with the unobserved true record. Each of these candidate matches can be thought of as providing an error-ridden estimate of the true variable of interest. The candidate-specific measures constitute valid instruments under the assumption that the linkage error for a given candidate is independent of the *other* candidates' error-ridden estimates of the true variable.

Applications that address record linkage as a missing data problem, such as Gutman et al. (2013) and Abowd et al. (2021), sample multiple match candidates from file B for each record in file A as implicates under a model that posits ignorability of linkage errors. We show that implicates can be used as valid instruments when the assumption of ignorability fails. That is, when the linkage error is correlated with the imputed variable. We illustrate the performance of our approach through Monte Carlo simulations in a variety of linkage-error environments and find that the performance of the two-stage least squares (TSLS) estimator, which exploits multiple implicates as instruments, is comparable to the benchmark unbiased estimator proposed by Lahiri and Larsen (2005).

In our simulations, we probabilistically link workers to employers and estimate a regression model on the matched data. There are several real-world examples of this class of applications. For instance, Jäckle et al. (2004) match European Community Household Panel (ECHP) Survey respondents in the UK to their employers, the US Census Bureau's Longitudinal Employer Household Dynamics (LEHD) data set matches workers to workplaces (see Abowd et al. (2009)), Abowd and Stinson (2013) match survey respondents in the Survey of Income and Program Participation (SIPP) to employers in the Census Business Register (BR), and Abowd et al. (2021) match survey respondents in the Health and Retirement Study (HRS) to employers in the BR. A distinct feature of worker-firm record linkage is that it can generate non-classical measurement error in employer size. We find that the IV approach performs well even in this context.

The rest of this paper is organized as follows. Section 2 provides an overview of probabilistic record linkage. Section 3 describes bias that can arise from imperfectly matched data. Section 4 explains the instrumental variables approach and describes the implementation of the TSLS estimator using multiple implicates. Section 5 provides an overview of the weighting-based estimator proposed by Lahiri and Larsen (2005). Section 6 describes how we structure our simulated linkage environment and compares the performance of a variety of estimators to recover regression coefficients. Section 7 concludes.

# 2 Record Linkage

Consider a case where two files containing overlapping sub-populations need to be linked. The records in file A are indexed by $i = 1, \ldots, N_A$ and the records in file B are indexed by $j = 1, \ldots, N_B$. Consequently, there are $N_A \times N_B$ record pairs to consider, which is potentially a very large number. To reduce the set of comparisons, researchers typically undertake an exercise known as "blocking," which groups record pairs that share certain characteristics. Pairs that share at least one characteristic are regarded as being potential matches whereas pairs that do not share any characteristics are considered non-matches. For example, in a setting where address information is available in both files, records that share Zip Codes can be regarded as potential matches while those that fail to share Zip Codes can be considered non-matches.

Once blocking is complete, researchers estimate a model that predicts whether a given pair of records constitutes a match. Define match status, $s_{ij}$, as equal to 1 if a given pair of records is a match and 0 otherwise. Either by using data where $s_{ij}$ is known or by applying other statistical techniques, researchers estimate

$$p(\boldsymbol{r}_{ij}; \boldsymbol{\theta}) = P(s_{ij} = 1 \mid \boldsymbol{r}_{ij}), \tag{1}$$

where $\boldsymbol{r}_{ij}$ represents a vector of variables used to predict match status. For example, researchers may rely on comparison vectors that represent agreement status between variables in the two files, Jaro-Winkler scores that capture name or address similarity, or any other relevant variables. In practice, there are several approaches to estimating models that predict match status. These include the classic Fellegi and Sunter (1969) (FS) approach, Bayesian methods that incorporate parameter uncertainty (see, e.g., Tancredi and Liseo (2011), Gutman et al. (2013), and Steorts et al. (2016)), or machine learning (ML) methods that rely on classification algorithms (see, e.g., Christen (2008a) and Christen (2008b)).

Finally, researchers use the estimated model to assign matches. There are two distinct approaches to match assignment. The first is to enforce one-to-one matching and the second is to retain the uncertainty inherent in probabilistic record linkage by allowing one-to-many matches. For instance, FS enforces one-to-one matching by selecting the candidate associated with the highest estimated match probability, certain Bayesian applications use the posterior mode of the parameter vector $\boldsymbol{\theta}$ or minimize a loss function to enforce one-to-one matching, and most ML applications rely on classification algorithms that select the best match. In contrast, record linkage applications such as Gutman et al. (2013) and Abowd et al. (2021) propagate uncertainty in the linkage by multiply imputing matches. These multiple imputation (MI) approaches repeatedly sample from the set of potential matches to generate several distinct copies of the matched records between files A and B. Each copy is then treated as a completed data set. Inference about parameters of interest derived from the completed data sets are based on the combining formulas in Rubin (1987). A key assumption that underpins the MI approach is that unobserved determinants of match status are ignorable conditional on the predictors. That is

$$P(s_{ij}=1 \mid \boldsymbol{r}_{ij}) = P(s_{ij}=1 \mid \boldsymbol{r}_{ij}, \boldsymbol{t}_{ij}), \tag{2}$$

where $\boldsymbol{t}_{ij}$ represents all unobserved determinants of match status.

Although MI is traditionally used to impute variables, an important distinction in the context of record linkage is that MI is used to impute matches. All the variables associated with the matched record, as well as the underlying covariance structure of those variables, is therefore inherited through the record linkage.

# 3 Estimation of Regression Coefficients with Linked Data

Consider the following statistical model

$$y_i = \alpha + x_i^* \beta + \varepsilon_i, \tag{3}$$

where $\varepsilon_i$ is an i.i.d. error term and $E(\varepsilon_i \mid x_i^*) = 0$. Suppose that $y_i$ is observed only in file A, while $x_i^*$ is observed only in file B.

Analysts are interested in estimating $\beta$, which is contingent on record linkage between file A and file B. Because of linkage errors, analysts do not observe the true record pairs $(y_i, x_i^*)$, but instead observe the probabilistically matched record pairs $(y_i, x_{ij})$ where $x_{ij}$ represents the variable from record $j$ from file B that is linked to record $i$ of file A. The linkage error can be written as

$$u_{ij} = x_{ij} - x_i^*. \tag{4}$$

Assume throughout that $Cov(u_{ij}, \varepsilon_{ij}) = 0$. Denote the parameter vector $(\alpha, \beta)' = \boldsymbol{\gamma}$. The ordinary least squares (OLS) estimator of $\boldsymbol{\gamma}$ is

$$\hat{\boldsymbol{\gamma}}_{\texttt{OLS}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}, \tag{5}$$

where $\boldsymbol{X}$ is an $N \times 2$ matrix whose $ij$-th row is $(1, x_{ij})$ and $\boldsymbol{y}$ is an $N \times 1$ vector whose $i$-th element is $y_i$. The sample variance estimator of $\hat{\boldsymbol{\gamma}}_{\texttt{OLS}}$ is given by $N^{-1}\hat{\sigma}^2_{\texttt{OLS}}\hat{E}(\boldsymbol{X}'\boldsymbol{X})^{-1}$, where:

$$\hat{\sigma}^2_{\texttt{OLS}} = (N-2)^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\gamma}}_{\texttt{OLS}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\gamma}}_{\texttt{OLS}}) \tag{6}$$

$$\hat{E}(\boldsymbol{X}'\boldsymbol{X})^{-1} = N^{-1}(\boldsymbol{X}'\boldsymbol{X})^{-1} \tag{7}$$

The OLS estimate of the parameter of interest, $\hat{\beta}_{\texttt{OLS}}$, can be written as,

$$\begin{aligned}\hat{\beta}_{\texttt{OLS}} &= \frac{Cov(x_{ij}, y_i)}{V(x_{ij})} + o_p(1) \\ &= \beta\left(\frac{V(x_{ij}) - Cov(x_{ij}, u_{ij})}{V(x_{ij})}\right) + o_p(1).\end{aligned} \tag{8}$$

The extent to which the OLS estimator is biased depends on the sign of $Cov(x_{ij}, u_{ij})$. Conceptually, there are two distinct distinct ways in which linkage errors can generate bias in the estimation of $\beta$. We describe each of these frameworks and their implications next.

### 3.1 Classical Measurement Error Framework

Under the classical measurement error model, linkage errors reflect noise that is uncorrelated with the true variable. Thus, $Cov(x_i^*, u_{ij}) = 0$ and consequently $Cov(x_{ij}, u_{ij}) > 0$. In this framework, OLS estimation of $\beta$ suffers from attenuation bias. The magnitude of the bias is, of course, dependent on the quality of the record linkage. In situations where the majority of records are accurately matched, measurement error is trivial and the bias could be small. In cases where a large share of records are imperfectly matched, attenuation bias may severely limit inferences that can be drawn from the matched data set.

While it is common to posit classical measurement error and subsequent attenuation bias of regression coefficients in record linkage settings, this model of linkage errors is a special case that is not guaranteed to occur in practice. For instance, non-classical measurement error arises naturally in the many-to-one worker-to-employer linkage problem that that we simulate in Section 6. This is because, consistent with real-world data, the distribution of employer size is highly skewed. In particular, most firms are very small but a small number of large firms employs a disproportionately large share of the workforce. A consequence of the empirical distribution of firm size is that workers employed at the right tail of the firm size distribution are more likely to be matched with smaller firms and vice versa. This induces a negative correlation between the linkage error and the true size of the employer.

### 3.2 Non-Classical Measurement Error in the Imputation Error Framework

An alternative framework is one where the record linkage step is viewed as method of imputing missing data. In this case, $x_{ij}$ is the imputed value of $x_i^*$. If the imputation errors are ignorable, then $Cov(x_{ij}, u_{ij}) = 0$ and the estimation of $\beta$ by OLS will be consistent. If on the other hand, the errors are non-ignorable, then $Cov(x_{ij}, u_{ij}) \neq 0$ and OLS estimation can suffer either attenuation or amplification bias. The unrestricted case of arbitrary correlation between $x_{ij}$ and $u_{ij}$ arises in the simulated linkage and analysis that we conduct in Section 6.

## 4 Instrumental Variables Correct for Linkage Error-Induced Biases

IVs have long been understood as a method that can correct measurement-error induced biases in regression coefficients (see, e.g., Bound et al. (2001)). While the IV estimator provides a useful way of addressing biases, it is difficult in practice for researchers to find valid instruments. In one prominent example, Ashenfelter and Krueger (1994) address measurement error-induced attenuation bias in estimates of the economic return to education by collecting data on a special sample of twins and instrumenting for self-reported years of education with twin reported years of education. This creative example of “second measurements” illustrates the high burden of finding valid instruments to correct for measurement-error induced biases.

In this section, we explain how an off-the-shelf set of valid instruments arises automatically in the course of probabilistic record linkage. In particular, when the record linkage is multiply imputed, each implicate provides information about the underlying variable of interest. We show that the two-stage least squares (TSLS) estimator, which uses implicates as instrumental variables corrects for linkage error induced biases. The value of our approach is threefold. First, because the implicates are part of the matched data, researchers can rely on an off-the-shelf set of instruments to address linkage error-induced biases instead of needing to look for valid instruments elsewhere. Second, unlike existing weighting-based approaches to regression estimation with linked data, the TSLS estimator is easy to implement because it allows researchers to obtain consistent estimates of regression coefficients without requiring access to the data used in the linkage step or the need to obtain or estimate match probabilities or other paradata from the linkage process. Finally, unlike weighting-based approaches where standard errors are computationally expensive to compute, the TSLS estimator obeys asymptotic normality and its variance is easily computed with standard software.

Consider a setting where researchers multiply impute missing variables for each record in file A using plausible candidates from file B. Let the implicates be indexed by $m = 1, \dots, M$. Extending Equation (4) to apply across all $M$ implicates

$$u_{ij}^{(m)} = x_{ij}^{(m)} - x_i^*. \tag{9}$$

Suppose that ignorability does not hold. I.e., suppose that $Cov(x_{ij}^{(m)}, u_{ij}^{(m)}) \neq 0$. Assume instead that the variable associated with implicate $m$ is uncorrelated with the imputation error for any *other* implicate $m'$. That is,

$$Cov(x_{ij}^{(m)}, u_{ij}^{(m')}) = 0 \,\forall\, m \neq m'. \tag{10}$$

The extent to which this condition holds is closely related to the properties of the probabilistic record linkage. In particular, Equation (10) is likely to hold in two cases:

1. The first case is where match probabilities are accurately concentrated such that repeated sampling of a match candidate as an implicate only occurs when the candidate is in fact the true match. In all instances where this happens, the linkage error will be zero.

2. The second case is where match probabilities are sufficiently diffuse such that false matches are never repeatedly sampled as implicates.

Related to the second case, note that Equation (10) does not hold when the match probabilities are concentrated in a biased way such that a false match is repeatedly sampled as an implicate. When this happens, the linkage errors for implicates $m$ and $m'$ will be identical and non-zero. In this case, Equation (10) turns into the non-ignorable error condition:

$$Cov(x_{ij}^{(m)}, u_{ij}^{(m')}) = Cov(x_{ij}^{(m)}, u_{ij}^{(m)}) \neq 0. \tag{11}$$

A key point of the Monte Carlo exercise in Section 6 is to evaluate the success of the IV strategy, which relies on validity of Equation (10).

Without loss of generality, we choose the first implicate as the mismeasured variable and the remaining implicates as instruments. Define the $N \times M$ matrix of instruments $\boldsymbol{Z}$ whose $ij$-th row is $\boldsymbol{z}_{ij} = (1, x_{ij}^{(2)}, \ldots, x_{ij}^{(M)})$. Define the $N \times 2$ matrix $\boldsymbol{X}^{(1)}$ whose $ij$-th element is $(1, x_{ij}^{(1)})$. Write the linear projection of $\boldsymbol{X}^{(1)}$ onto $\boldsymbol{Z}$ as

$$\hat{\boldsymbol{X}} = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X}^{(1)}. \tag{12}$$

Note here that while $x_{ij}^{(1)}$ is modeled as a scalar to facilitate exposition, it can also be a vector. The procedure generalizes straightforwardly to the case where there are multiple regressors. Since each implicate provides a complete record of variables, for each value that needs to be imputed there is an associated value in the other implicates that provides an additional instrumental variable.

Denote the $ij$-th row of $\hat{\boldsymbol{X}}$ by $(1, \hat{x}_{ij})$. Under the assumption implied by Equation (10), $\hat{x}_{ij}$ isolates variation in $x_{ij}^{(1)}$ that is uncorrelated with $u_{ij}$. Thus, $Cov(\hat{x}_{ij}, u_{ij}^{(1)}) = 0$. The TSLS estimator of $\boldsymbol{\gamma}$ is then written as

$$\hat{\boldsymbol{\gamma}}_{\texttt{TSLS}} = (\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}})^{-1}\hat{\boldsymbol{X}}'\boldsymbol{y}. \tag{13}$$

Consistency of the parameter of interest, $\hat{\beta}_{\texttt{TSLS}}$, follows because:

$$\begin{aligned}\hat{\beta}_{\texttt{TSLS}} &= \left(\frac{Cov(\hat{x}_{ij}, y_i)}{V(\hat{x}_{ij})}\right) + o_p(1)\\ &= \beta\left(\frac{Cov(\hat{x}_{ij}, (x_{ij}^{(1)} - u_{ij}^{(1)}))}{V(\hat{x}_{ij})}\right) + o_p(1)\\ &= \beta + o_p(1).\end{aligned} \tag{14}$$

Consistency of the TSLS estimator implies that $\sqrt{N}(\hat{\boldsymbol{\gamma}}_{\texttt{TSLS}} - \boldsymbol{\gamma})$ is asymptotically normally distributed with mean zero and variance

$$\sigma^2(E(\boldsymbol{X}^{(1)'}\boldsymbol{Z})E(\boldsymbol{Z}'\boldsymbol{Z})^{-1}E(\boldsymbol{Z}'\boldsymbol{X}^{(1)}))^{-1}, \tag{15}$$

where $\sigma^2$ is the residual variance. The asymptotic variance is estimated by plugging in sample analogs of the terms in Equation (15) as follows:

$$\hat{\sigma}^2_{\texttt{TSLS}} = (N-2)^{-1}(\boldsymbol{y} - \boldsymbol{X}^{(1)}\hat{\gamma}_{\texttt{TSLS}})'(\boldsymbol{y} - \boldsymbol{X}^{(1)}\hat{\gamma}_{\texttt{TSLS}}) \tag{16}$$

$$\hat{E}(\boldsymbol{X}^{(1)'}\boldsymbol{Z}) = N^{-1}(\boldsymbol{X}^{(1)'}\boldsymbol{Z}) \tag{17}$$

$$\hat{E}(\boldsymbol{Z}'\boldsymbol{Z}) = N^{-1}(\boldsymbol{Z}'\boldsymbol{Z}) \tag{18}$$

$$\hat{E}(\boldsymbol{Z}'\boldsymbol{X}^{(1)}) = N^{-1}(\boldsymbol{Z}'\boldsymbol{X}^{(1)}). \tag{19}$$

So the sample variance estimator of $\hat{\boldsymbol{\gamma}}_{\texttt{TSLS}}$ is

$$N^{-1}\hat{\sigma}^2_{\texttt{TSLS}}(\hat{E}(\boldsymbol{X}^{(1)'}\boldsymbol{Z})\hat{E}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\hat{E}(\boldsymbol{Z}'\boldsymbol{X}^{(1)}))^{-1}. \tag{20}$$

Note that the estimation of $\hat{\sigma}^2_{\texttt{TSLS}}$ in Equation (16) relies on $\boldsymbol{X}^{(1)}$ not $\hat{\boldsymbol{X}}$. Hausman (1983) provides additional details on the specification, estimation, and properties of the TSLS estimator. TSLS is the generalization of the IV estimator when, as is the case with our application, there are more instrumental variables than parameters.

## 5 Weighting-Based Approach

Lahiri and Larsen (2005) (LL) provide a weighting-based estimator that corrects for bias arising from linkage-induced errors. We briefly describe the implementation of the LL estimator in this section because it is well known and therefore provides a useful benchmark relative to which we can evaluate the performance of the TSLS approach.

The LL estimator exploits the fact that

$$P(s_{ij} = 1 \mid \boldsymbol{r}_{ij}) = P(x^*_i = x_{ij} \mid \boldsymbol{r}_{ij}). \tag{21}$$

Define $\mathcal{S}(i)$ as the set of match candidates from file B that can be paired to unit $i$ in file A and let $k = 1, \ldots, N_{\mathcal{S}(i)}$ index those candidates. Note that Equation (21) requires that the true match is included in the set of potential candidates; i.e., that there are no false negative errors in blocking. Using Equation (21), the conditional expectation of the $i$-th unobserved true value, $E(x^*_i \mid x_{i1} \ldots, x_{iN_{\mathcal{S}(i)}}, \boldsymbol{r}_{i1}, \ldots, \boldsymbol{r}_{iN_{\mathcal{S}(i)}})$, can be written as

$$\tilde{x}_i = \sum_{k \in \mathcal{S}(i)} p(\boldsymbol{r}_k; \hat{\boldsymbol{\theta}}) \times x_k, \tag{22}$$

where $p(\boldsymbol{r}_k; \hat{\boldsymbol{\theta}})$ is an estimate of the candidate-specific match probability conditional on a set of predictors, $\boldsymbol{r}_k$. The LL estimator is given by

$$\hat{\gamma}_{\text{LL}} = (\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}})\tilde{\boldsymbol{X}}'\boldsymbol{y}, \tag{23}$$

where $\tilde{\boldsymbol{X}}$ is an $N \times 2$ matrix whose $i$-th row is $(1, \tilde{x}_i)$. Consistency of the parameter of interest, $\hat{\beta}_{\text{LL}}$, follows because:

$$\begin{aligned}\hat{\beta}_{\texttt{LL}} &= \frac{Cov(\tilde{x}_i, y_i)}{V(\tilde{x}_i)} + o_p(1) \\ &= \beta\left(\frac{Cov(\tilde{x}_i, x_i^*)}{V(\tilde{x}_i)}\right) + o_p(1) \\ &= \beta + o_p(1). \end{aligned} \tag{24}$$

The variance of $\hat{\boldsymbol{\gamma}}_{\text{LL}}$ is a function of the uncertainty in $\hat{\boldsymbol{\theta}}$ due to Equation (22). To account for the explicit two-step nature of the estimation, LL approximate the second-step estimation variance with the bootstrap as

$$V_{\texttt{BOOT}}(\hat{\boldsymbol{\gamma}}_{\text{LL}}) = E_*(V(\hat{\boldsymbol{\gamma}}_{\text{LL}})) + V_*(\hat{\boldsymbol{\gamma}}_{\text{LL}}), \tag{25}$$

where $E_*$ and $V_*$ represent expectation and variance over the bootstrap distribution. Within each bootstrap repetition, $V(\hat{\boldsymbol{\gamma}}_{\text{LL}})$ in the first term of Equation (25) is estimated as $N^{-1}\hat{\sigma}^2_{\text{LL}}\hat{E}(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}})^{-1}$, where:

$$\hat{\sigma}^2_{\text{LL}} = (N-2)^{-1}(\boldsymbol{y} - \tilde{\boldsymbol{X}}\hat{\boldsymbol{\gamma}}_{\text{LL}})'(\boldsymbol{y} - \tilde{\boldsymbol{X}}\hat{\boldsymbol{\gamma}}_{\text{LL}}) \tag{26}$$

$$\hat{E}(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}}) = N^{-1}(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}}). \tag{27}$$

Denoting bootstrap repetitions by $b = 1, \ldots, B$, the first term of Equation (25) is approximated as

$$E_*(V(\hat{\boldsymbol{\gamma}}_{\text{LL}})) \approx B^{-1}\sum_{b=1}^{B} V(\hat{\boldsymbol{\gamma}}_{\text{LL}}(\hat{\boldsymbol{\theta}}^{(b)})), \tag{28}$$

and the second term is approximated as

$$V_*(\hat{\boldsymbol{\gamma}}_{\text{LL}}) \approx B^{-1}\sum_{b=1}^{B}\left(\hat{\boldsymbol{\gamma}}_{\text{LL}}(\hat{\boldsymbol{\theta}}^{(b)}) - \hat{\boldsymbol{\gamma}}_{\text{LL}}(\hat{\boldsymbol{\theta}})\right)\left(\hat{\boldsymbol{\gamma}}_{\text{LL}}(\hat{\boldsymbol{\theta}}^{(b)}) - \hat{\boldsymbol{\gamma}}_{\text{LL}}(\hat{\boldsymbol{\theta}})\right)', \tag{29}$$

where $\hat{\boldsymbol{\theta}}^{(b)}$ and $\hat{\boldsymbol{\gamma}}_{\text{LL}}(\hat{\boldsymbol{\theta}}^{(b)})$ are estimates from the b-th bootstrap repetition.

# 6 Monte Carlo Simulation

In this section we simulate data sets that are constructed by probabilistically matching workers to employers. After constructing the matched data sets, we estimate the relationship between a worker-level outcome and an employer-level characteristic using a variety of regression-based estimators. A real-world example of such a relationship is the observed positive association between the logarithm of worker-level hourly wages and the logarithm of employer size (i.e., the number of employees), which reflects, among other factors, the transmission of employer-level productivity into workplace size and worker-level compensation (see, e.g., Brown and Medoff (1989) and Bloom et al. (2018)).

To facilitate exposition, we use variables from this example to name our simulated variables.

## 6.1 Setup

In each Monte Carlo simulation we create a register of $j = 1, \ldots, 500$ firms whose size is distributed $\mathcal{LN}(3, 1)$. To assure that we have a discrete number of workers per firm, we round the simulated firm size. This parameterization allows us to create an environment where each of the firms employs many workers and a small number of firms are very large. We record information on the population of workers employed at each firm, which represents the universe of true worker-firm links. The location of each employer is given by $l_j^* \sim \mathcal{U}(0, 2\pi)$. We simulate the logarithm of hourly wage for each worker $i$ as

$$\log(\text{wage}_i) = \alpha + \log(\text{size}_i^*)\beta + \varepsilon_i, \tag{30}$$

where $\text{size}_i^*$ refers to the size of the firm that employs worker $i$ and $\varepsilon_i \sim \mathcal{N}(0, 1)$ is an i.i.d. disturbance term. From these data, we draw a random sample of $1,000$ workers. Note that we sample $1,000$ workers from the population of worker-firm links, which is much larger than the population of firms because each firm employs many workers. Each of the sampled workers provides a noisy report of their employer's location, which is given by

$$l_i = l_i^* + e_i, \tag{31}$$

where $e_i \sim \mathcal{U}(-\overline{e}, \overline{e})$. We simulate record linkage under two different reporting error environments: $\overline{e} = \frac{\pi}{600}$ in the low-error environment and $\overline{e} = \frac{\pi}{100}$ in the high-error environment. Table 1 provides details on all of the parameters used in the simulations.

Define file A as the random sample of workers containing two variables: $(\text{wage}_i, l_i)$. Define file B as the register of all employers also containing two variables: $(\text{size}_j^*, l_j^*)$. Record linkage in this setting seeks to match workers in file A to their employers in file B. We do this in three distinct steps.

### Step 1: Blocking

Let $k = 1, \ldots, N_A \times N_B$ index observations formed by pairing records from file A to records in file B. Retain pairs where $l_k^* \in [l_k - 2\overline{e}, l_k + 2\overline{e}]$. This blocking screen treats candidate employers that are within the margin of location-specific reporting error for each worker as potential matches and considers all other pairs as non-matches. Denote the set of blocked pairs for worker $i$ by $\mathcal{S}(i)$.

### Step 2: Model Estimation

Create a training data set by drawing a sample of 100 workers from the population of workers. Applying the blocking screen defined above, let $t = 1, \ldots, N_T$ index worker-employer pairs associated with 100 sampled workers. Reveal the true match status for each pair, $s_t$. Estimate the probability that a given pair is a match conditional on a set of predictors $\boldsymbol{r}_t$ using logistic regression:

$$P(s_t = 1 \mid \boldsymbol{r}_t) = \frac{\exp(\boldsymbol{r}_t\boldsymbol{\theta})}{1 + \exp(\boldsymbol{r}_t\boldsymbol{\theta})}. \tag{32}$$

Note that the training data set is a standalone sample of workers rather than a sub-sample of file A. Consequently, predictions about match status obtained by fitting Equation (32) using the training data are out-of-sample relative to file A.

Define the absolute distance between worker-reported location and the true location of the candidate employer as

$$d_t = |l_t - l_t^*|. \tag{33}$$

Further define the within-block share of employment accounted for by a candidate employer as

$$b_t = \frac{\text{size}_t}{\sum_{t \in \mathcal{S}(i)} \text{size}_t}. \tag{34}$$

We estimate match probabilities using two different groups of predictors, which we refer to as Model 1 and Model 2 respectively. Model 1 reflects an environment where record linkage is based only on variables that contain pair-level identifying information. It also reflects an environment where other relevant predictors of match status may not available. The set of predictors in Model 1 is

$$\boldsymbol{r}_t = (1, d_t, d_t^2, d_t^3).$$

While $d_t$ reflects physical distance in the simulation exercise, this variable is a stand in for measures of proximity observed in practice such as the Jaro-Winkler metric of record-pair similarity between name or address strings.

In Model 2, we enrich the set of predictors to include variables that appear in subsequent analysis steps such as log(wage) and log(size) along with higher order terms in these variables. The inclusion of variables are relevant for predicting match status makes the ignorability assumption of Equation (2) more tenable, thereby facilitating MI inference. The full set of predictors in Model 2 is

$$\boldsymbol{r}_t = (1, d_t, d_t^2, d_t^3, \log(\text{wage}_t), \log(\text{wage}_t)^2, \log(\text{wage}_t)^3, \log(\text{size}_t), \log(\text{size}_t)^2, \log(\text{size}_t)^3, b_t, b_t^2, b_t^3).$$

To multiply impute the record linkage, we re-sample with replacement from the training data set $m = 1, \ldots, 10$ times and estimate $\hat{\boldsymbol{\theta}}^{(m)}$ separately for each repetition.

### Step 3: Assignment of Matches

For each candidate pair formed by blocking records in file A with those in file B, we obtain $p(\boldsymbol{r}_k; \hat{\boldsymbol{\theta}}^{(m)})$ using the m-th estimate of $\boldsymbol{\theta}$. We then normalize the $p(\boldsymbol{r}_k; \hat{\boldsymbol{\theta}}^{(m)})$ to sum to one across the different candidate employers for each worker, $i$. Denote the normalized match probability by

$$\check{p}(\boldsymbol{r}_k; \hat{\boldsymbol{\theta}}^{(m)}) = \frac{p(\boldsymbol{r}_k; \hat{\boldsymbol{\theta}}^{(m)})}{\sum_{k \in \mathcal{S}(i)} p(\boldsymbol{r}_k; \hat{\boldsymbol{\theta}}^{(m)})}. \tag{35}$$

For each of the $m = 1, \ldots, 10$ estimates of $\boldsymbol{\theta}$, we sample one candidate employer with probability proportional to the normalized match probability. This step yields 10 multiply imputed employer matches from file B for each worker in file A.

## 6.2 Estimators Used in Regression Analysis

Once the record linkage is complete, we estimate the regression in Equation (30) using a variety of estimators.

1. **Oracle estimator:** This infeasible estimator uses the true value of log employer size as the explanatory variable. The oracle estimator provides a benchmark for conducting inference in an error-free setting.

2. **TSLS estimator:** This estimator uses implicates 2-10 as instruments for implicate 1 as described in Section 4.

3. **IV estimator:** This estimator is a special case of the TSLS estimator that relies on implicate 2 as an instrument for implicate 1; i.e, $\boldsymbol{z}_{ij} = (1, x_{ij}^{(2)})$.

4. **Lahiri-Larsen (LL) estimator:** We implement the LL estimator by first computing the expected value of true employer size for each worker $i$ using normalized match probabilities as weights

$$\widetilde{\text{size}}_i = \sum_{k \in \mathcal{S}(i)} \check{p}(\boldsymbol{r}_k; \hat{\boldsymbol{\theta}}) \times \text{size}_k. \tag{36}$$

   Note that $\mathcal{S}(i)$ represents the set of all employer candidate matches that survive the blocking threshold. In Equation (36), $\hat{\boldsymbol{\theta}}$ represents the parameter estimate for the first implicate; i.e., $m = 1$. We then use $\log(\widetilde{\text{size}}_i)$ as the explanatory variable in Equation (30) to estimate $\hat{\boldsymbol{\gamma}}_{\text{LL}}$. The variance of the LL estimator is computed using 500 bootstrap repetitions. Within each bootstrap repetition, we repeat Step 2 by re-sampling with replacement from the training data set and estimate $\hat{\boldsymbol{\theta}}^{(b)}$ for each repetition. We then re-construct the record linkage (Step 3) and estimate $\hat{\boldsymbol{\gamma}}_{\text{LL}}(\hat{\boldsymbol{\theta}}^{(b)})$ for each repetition. With these estimates in hand, we apply Equation (25) to obtain the bootstrapped variance of the LL estimator. Note that we replace $\hat{\boldsymbol{\gamma}}_{\text{LL}}(\hat{\boldsymbol{\theta}})$ in Equation (29) by the average across bootstrap repetitions, $B^{-1} \sum_{b=1}^{B} \hat{\boldsymbol{\gamma}}_{\text{LL}}(\hat{\boldsymbol{\theta}}^{(b)})$.

5. **OLS-MI estimator:** This estimator uses OLS to estimate $\boldsymbol{\gamma}$ with 10 multiply imputed data sets. We obtain point and variance estimates using the combining formulas in Rubin (1987)

as follows

$$\hat{\gamma}_{\texttt{OLS-MI}} = 10^{-1} \sum_{m=1}^{10} \hat{\gamma}^{(m)}, \tag{37}$$

where $\hat{\gamma}^{(m)}$ is the OLS estimate of $\gamma$ using the m-th completed data set. The within-implicate variance is

$$V_W(\hat{\gamma}_{\texttt{OLS-MI}}) = 10^{-1} \sum_{m=1}^{10} V(\hat{\gamma}^{(m)}). \tag{38}$$

The between-implicate variance is

$$V_B(\hat{\gamma}_{\texttt{OLS-MI}}) = (10-1)^{-1} \sum_{m=1}^{10} \left(\hat{\gamma}^{(m)} - \hat{\gamma}_{\texttt{OLS-MI}}\right)\left(\hat{\gamma}^{(m)} - \hat{\gamma}_{\texttt{OLS-MI}}\right)'. \tag{39}$$

The total variance associated with $\hat{\gamma}_{\texttt{OLS-MI}}$ is

$$V(\hat{\gamma}_{\texttt{OLS-MI}}) = V_W(\hat{\gamma}_{\texttt{OLS-MI}}) + (1 + 10^{-1}) V_B(\hat{\gamma}_{\texttt{OLS-MI}}). \tag{40}$$

Under the assumption of ignorable linkage errors, this procedure propagates uncertainty in $\hat{\boldsymbol{\theta}}$ as well as uncertainty in latent match status.

6. **OLS-Best estimator:** This estimator enforces one-to-one linkage as is common in many record linkage applications. We implement this estimator by assigning each worker in file A the candidate employer with the highest normalized match probability in file B. Since this estimator relies on a single match per record, we use $\hat{\boldsymbol{\theta}}$ constructed from the first implicate, $m = 1$, when computing the normalized match probability estimates. Note that, unlike the other estimators, the OLS-Best estimator does not capture uncertainty in $\hat{\boldsymbol{\theta}}$ or uncertainty regarding latent match status.

## 6.3 Comparative Performance of Estimators

We simulate 500 low- and high-error environments and conduct record linkage using Model 1 and Model 2 in each error environment. For each simulation, this produces four conditions under which we can estimate the parameter of interest, $\beta$. The panels of Figure 1 show an example of a single simulation of the respective error environments and record linkage models. Log true employer size is on the horizontal axis whereas log matched employer size associated with implicate 1 is on the vertical axis. Concentration along the 45-degree axis is indicative of linkage accuracy whereas dispersion away from the 45-degree axis is indicative of linkage error. The top left panel highlights the atypical nature of measurement error in the context of the worker-firm record linkage problem. At the left tail of true employer size, workers tend to be matched with larger firms while the opposite is true at the right tail. This happens because most employers are small but most workers are employed at large firms. Bunching along a given value of the horizontal axis happens when multiple workers employed at the same firm are incorrectly matched. In the top left panel, for example, workers employed at the largest firm in this particular simulation (log(size) $\approx$ 6) are incorrectly matched with smaller firms. As the figure makes clear, each error environment and modeling specification generates substantial variation in linkage error.

Table 2 shows sample characteristics of the matched files in different record linkage environments. Columns (1) and (2) show estimates under the high-error environment, while columns (3) and (4) show estimates under the low-error environment. All statistics represent averages over the 500 simulations. The first row shows the linkage precision rate (i.e., the proportion of workers in file A that are correctly matched to employers in file B), which characterizes the overall degree of linkage error. Within the high-error environment, moving from the simpler specification implied by Model 1 to the more complex specification implied by Model 2 increases linkage precision from about 20 percent to 32 percent. In the low-error environment, the precision rates for the two models rise to 66 percent and 75 percent respectively. The second row shows the average block size (i.e., number of candidate matches from file B for each record in file A), which is between 10-11 in the high-error environment and about 2.5 in the low-error environment. Record linkage computation time, which is inversely proportional to block size but proportional to the complexity of the linkage model is shown in the following row. The next set of rows show correlations between linkage error and different measures of log size. Without loss of generality, we use the linkage error associated with implicate 1 (i.e., $u_{ij}^{(1)}$) in each correlation. The correlation between log true employer size and the linkage error is negative across all the columns, which rules out classical measurement error. This negative association is a by-product of the skewness of the firm size distribution: workers employed at the largest firms are more likely to be mis-matched with smaller firms and vice versa. Similarly, the correlation between log matched employer size associated with implicate 1 and the linkage error for implicate 1 is non-zero across all the columns, which rules out ignorablility. The final row shows the correlation between the instrument constructed using Equation (12) and the linkage error. Validity of the IV strategy requires zero correlation between these variables, which is approximately true in all four columns.

Table 3 presents point and variance estimates for $\hat{\beta}$ constructed from 500 Monte Carlo simulations. Panel A shows estimates in the high-error environment while Panel B shows estimates in the low-error environment. For each estimator, we show the Monte Carlo mean of the point estimates ($E_{\texttt{MC}}(\hat{\beta})$), the Monte Carlo mean of the variance estimates ($E_{\texttt{MC}}(V(\hat{\beta}))$), the Monte Carlo variance of the point estimates $V_{\texttt{MC}}(\hat{\beta})$, and the average time required to compute the point and variance estimates in each simulation. The first row shows the infeasible oracle estimator, which represents a situation where linkage is completely error free. The second and third rows show the TSLS and IV estimators, which both use implicates as instruments to correct for linkage-error induced bias. As one would expect, TSLS outperforms the IV estimator in terms of efficiency as it exploits information from 9 variables rather than just 1. While the estimated variance for the TSLS estimator is about a quarter as large as the IV estimator in the high-error environment, this advantage diminishes substantially in the low-error environment. The fourth row shows the LL estimator, which performs well in all linkage environments.

While the TSLS estimator does exhibit a small but non-trivial bias under Model 1, it is substantially easier to implement than the LL estimator because variance estimation does not rely on the bootstrap. This advantage is reflected in the substantial difference in computation times between the TSLS and LL estimators. Comparing variances, one sees that the TSLS estimator is less efficient than the LL estimator in the high-error environment but comparably efficient in the low-error environment. Contrasting the Monte Carlo means of the variance estimates to the Monte Carlo variances reveals that the IV variance estimator provides a reasonable approximation to the variance under all configurations, except for Model 2 in the low-error environment. The TSLS variance estimator is nearly unbiased under Model 1. For Model 2, the TSLS variance estimator under-estimates the variance. We find that the LL variance estimator performs better in the low-error environment than it does in the high-error environment.

The last two rows of of the table show the performance of the two alternative estimators. OLS-MI depends on ignorability of the linkage error while OLS-Best assumes away linkage error-related concerns by using the best match. In both low- and high-error environments with Model 1, these estimators exhibit severe bias. The performance of both estimators improves when we use Model 2, which conditions on a much richer set of predictors. The most evident improvement occurs in the low-error environment with Model 2, which highlights that these estimators can perform well if the linkage is of sufficiently high precision.

Taken together, the estimates shown in Table 3 highlight the value of using implicates as instruments in the context of probabilistic record linkage. The key takeaway from these simulations is that the assumption implied by Equation (10) holds in a variety of linkage environments and, consequently, the TSLS estimator performs very well. In results not reported here, we conducted the same simulation exercise with regression disturbance terms distributed $\mathcal{N}(0,2)$ instead of $\mathcal{N}(0,1)$ as a robustness check. We find qualitatively very similar results in the presence of higher error variance to those reported in Table 3.

# 7 Conclusion

In this paper, we study the problem of conducting regression analysis using probabilistically matched data, which is subject to biases that arise from linkage errors. We propose an instrumental variable-based approach to address this concern that is based on multiply imputed record linkage. In particular, when ignorability assumptions required to obtain valid inference with MI fail, the implicates can be used as instrumental variables to obtain consistent estimates of regression coefficients. This approach has several advantages relative to existing methods. It does not require analysts to estimate match probabilities or complex models of linkage error, or obtain access to information used in the linkage step such as matching variables that may be confidential. Finally, relative to alternative methods, the standard errors for the IV-based approach are easy to compute using conventional software since the estimator obeys asymptotic normality. While the setting we consider in this paper is one with a single mis-measured regressor, the approach we outline applies equally in settings with multiple mis-measured regressors. We conduct simulated record linkage of workers to employers and subsequent regression analysis in low- and high-reporting error environments as well as with simple and complex models of record linkage. Across these simulated settings, we find that the underlying assumptions that deliver consistency of the IV-based approach hold and that the TSLS estimator performs well even in the presence of non-ignorable linkage error.

# References

**Abowd, John M. and Martha H. Stinson**, "Estimating Measurement Error in Annual Job Earnings: A Comparison of Survey and Administrative Data," *Review of Economics and Statistics*, 2013, *95* (5), 1451–1467.

_ **, Bryce E. Stephens, Lars Vilhuber, Fredrik Anderson, Kevin L. McKinney, Marc Roemer, and Simon Woodcock**, "The LEHD Infrastructure Files and the Creation of the Quarterly Workforce Indicators," in Timothy Dunne, J. Bradford Jensen, and Mark J. Roberts, eds., *Producer Dynamics: New Evidence from Micro Data*, University of Chicago Press, 2009, pp. 149–230.

_ **, Joelle Abramowitz, Margaret C. Levenstein, Kristin McCue, Dhiren Patki, Trivellore Raghunathan, Ann M. Rodgers, Matthew D. Shapiro, Nada Wasi, and Dawn Zinsser**, "Finding Needles in Haystacks: Multiple Imputation Record Linkage Using Machine Learning," 2021. U.S. Census Bureau Working Paper No. CES-21-35.

**Ashenfelter, Orley and Alan Krueger**, "Estimates of the Economic Return to Schooling from a New Sample of Twins," *American Economic Review*, 1994, *84* (5), 1157–1173.

**Bloom, Nicholas, Fatih Guvenen, Benjamin S. Smith, Jae Song, and Till von Wachter**, "Inequality and the Disappearing Large Firm Wage Premium," *American Economic Association Papers and Proceedings*, 2018, *108*, 317–322.

**Bound, John, Charles Brown, and Nancy Mathiowetz**, "Measurement Error in Survey Data," in James J. Heckman and Edward Leamer, eds., *Handbook of Econometrics*, North-Holland, 2001, pp. 3705–3843.

**Brown, Charles and James Medoff**, "The Employer Size-Wage Effect," *Journal of Political Economy*, 1989, *97*, 1027–1059.

**Chambers, Ray and Andrea Diniz da Silva**, "Improved Secondary Analysis of Linked Data: A Framework and an Illustration," *Journal of the Royal Statistical Society, Series A*, 2020, *183* (1), 37–59.

**Christen, Peter**, "Automatic Record Linkage Using Seeded Nearest Neighbour and Support Vector Machine Classification," in "Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining" Association for Computing Machinery 2008, pp. 151–159.

_ , "Automatic Training Example Selection for Scalable Unsupervised Record Linkage," in "Proceedings of the 12th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining" Springer-Verlag 2008, pp. 511–518.

**Fellegi, Ivan P. and Alan B. Sunter**, "A Theory for Record Linkage," *Journal of the American Statistical Association*, 1969, *64* (328), 1183–1210.

**Gutman, Roee, Christopher C. Afendulis, and Alan M. Zaslavsky**, "A Bayesian Procedure for File Linking to Analyze End-of-Life Medical Costs," *Journal of the American Statistical Association*, 2013, *108* (501), 34–47.

**Hausman, Jerry A.**, "Specification and Estimation of Simultaneous Equation Models," in Zvi Griliches and Michael D. Intriligator, eds., *Handbook of Econometrics*, North-Holland, 1983, pp. 392–445.

**Jäckle, Anette, Emanuela Sala, Stephen P. Jenkins, and Peter Lynn**, "Validating Survey Data: Experiences Using Employer Records and Government Benefit (Transfer) Data in the UK," *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 2004, pp. 4802–4809.

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

**National Academies of Sciences Engineering and Medicine**, *Innovations in Federal Statistics: Combining Data Sources While Protecting Privacy*, Washington, DC: National Academies Press, 2017.

**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*, 1005–1027.

**Rubin, Donald B.**, *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley, 1987.

**Scheuren, Fritz and William Winkler**, "Regression Analysis of Data Files that are Computer Matched I," *Survey Methodology*, 1993, *19*, 39–58.

**Slawski, Martin and Emanuel Ben-David**, "Linear Regression with Sparsely Permuted Data," *Electronic Journal of Statistics*, 2019, *13*, 1–36.

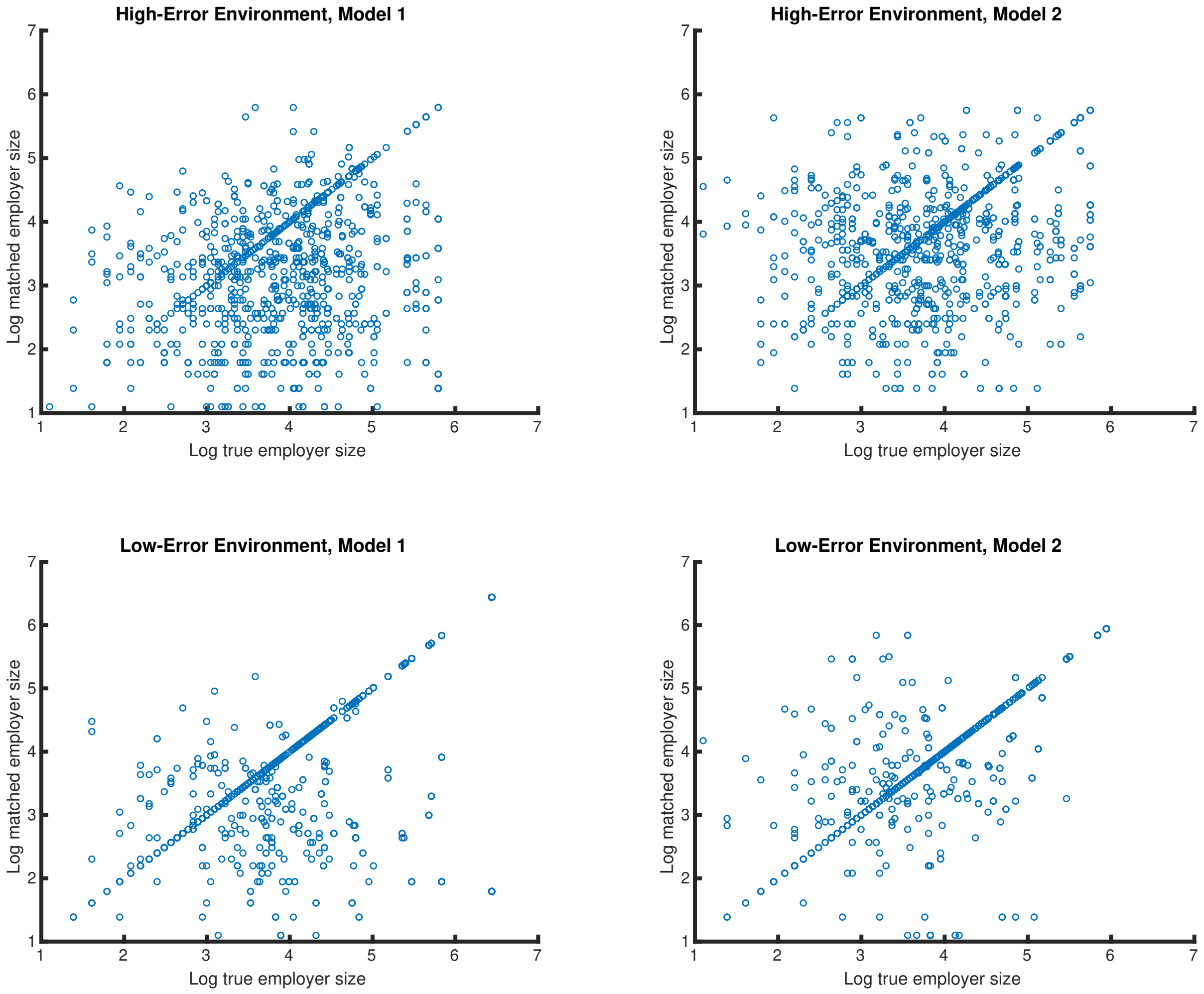
**_ , Guoqing Diao, and Emanuel Ben-David**, "A Pseudo-Likelihood Approach to Linear Regression With Partially Shuffled Data," *Journal of Computational and Graphical Statistics*, 2021, *30* (4), 991–1003.

**Steorts, Rebecca C., Rob Hall, and Stephen E. Feinberg**, "A Bayesian Approach to Graphical Record Linkage and De-duplication," *Journal of the American Statistical Association*, 2016, *111* (516), 1660–1672.

**Tancredi, Andrea and Brunero Liseo**, "A Hierarchical Bayesian Approach to Record Linkage and Population Size Problems," *Annals of Applied Statistics*, 2011, *5* (2B), 1553–1585.

**Warren, Joan L., Carrie N. Klabunde, Deborah Schrag, Peter B. Bach, and Gerald F. Riley**, "Overview of the SEER-Medicare Data: Content, Research Applications, and Generalizability to the United States Elderly Population," *Medical Care*, 2002, *40* (8), IV3–IV18.

**Figure 1:** Examples of Linkage Errors in Different Reporting Error and Record Linkage Modeling Environments



**Notes:** This figure shows the relationship between actual log employer size and matched log employer size in four different linkage environments. Each graph is based on a single simulation with 1,000 observations. Differences between the true value and the matched value arise from variation in reporting errors and the relative complexity of the record linkage model. Model 1 only uses distance to predict match status, whereas Model 2 uses distance, wage, size, and within-block share of employment to predict match status.

**Table 1:** Parameters Used in Monte Carlo Simulations

| Simulation Parameter | Value or Distribution |
|---|---|
| Number of workers sampled in file A | 1000 |
| Number of employers in file B | 500 |
| Number of workers sampled in training data | 100 |
| Number of imputations | 10 |
| Number of bootstrap repetitions | 500 |
| Employer size distribution | $\mathcal{LN}(3, 1)$ |
| Employer location distribution | $\mathcal{U}(0, 2\pi)$ |
| log(wage)-log(size) equation $\alpha$ | 1 |
| log(wage)-log(size) equation $\beta$ | 0.25 |
| log(wage) disturbance term distribution | $\mathcal{N}(0, 1)$ |
| $\bar{e}$ in high-error environment | $\frac{\pi}{100}$ |
| $\bar{e}$ in low-error environment | $\frac{\pi}{600}$ |

**Notes:** This table shows parameter values and distributions used in the Monte Carlo simulations.

**Table 2:** Characteristics of Matched Files in Different Reporting Error and Record Linkage Modeling Environments

| | High-error environment | | Low-error environment | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| | Linkage Model 1 | Linkage Model 2 | Linkage Model 1 | Linkage Model 2 |
| Linkage precision rate | 0.197 | 0.321 | 0.658 | 0.736 |
| Average block size | 10.817 | 11.229 | 2.524 | 2.657 |
| Record linkage computation time (seconds) | 3.753 | 4.993 | 2.307 | 3.601 |
| Correlation between true log size and linkage error | -0.586 | -0.599 | -0.431 | -0.328 |
| Correlation between matched log size and linkage error | 0.683 | 0.585 | 0.561 | 0.339 |
| Correlation between instrument and linkage error | 0.030 | -0.015 | 0.079 | 0.009 |

**Notes:** Each linkage environment is simulated 500 times. Statistics shown in the table represent averages across 500 simulations. The linkage precision rate is defined as the fraction of observations in file A that are correctly matched to records in file B. Record linkage computation time includes MI and bootstrap repetitions for estimation of $\hat{\theta}^{(m)}$ and $\hat{\theta}^{(b)}$. Without loss of generality, linkage error is defined as $u_{ij}^{(1)}$ and log matched size is defined as $\log(\text{size}_{ij}^{(1)})$. The instrument is constructed using implicates 2-10, as in Equation (12). Model 1 only uses a cubic polynomial in distance to predict match status, whereas Model 2 uses a cubic polynomial in distance, cubic polynomial in log wage, cubic polynomial in log size, and cubic polynomial in within-block share of employment to predict match status.

**Table 3:** Performance of Estimators in Different Reporting Error and Record Linkage Modeling Environments

| | **A: High-Error Environment** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Linkage Model 1 | | | | Linkage Model 2 | | | |
| | | | | Estimation | | | | Estimation |
| | $E_{\texttt{MC}}(\hat{\beta})$ | $E_{\texttt{MC}}(V(\hat{\beta}))$ | $V_{\texttt{MC}}(\hat{\beta})$ | time (seconds) | $E_{\texttt{MC}}(\hat{\beta})$ | $E_{\texttt{MC}}(V(\hat{\beta}))$ | $V_{\texttt{MC}}(\hat{\beta})$ | time (seconds) |
| Oracle | 0.251 | 0.0013 | 0.0012 | 0.0007 | 0.260 | 0.0011 | 0.0010 | 0.0005 |
| TSLS | 0.214 | 0.0089 | 0.0090 | 0.0006 | 0.278 | 0.0061 | 0.0084 | 0.0006 |
| IV | 0.233 | 0.0496 | 0.0447 | 0.0003 | 0.295 | 0.0213 | 0.0287 | 0.0002 |
| Lahiri-Larsen | 0.286 | 0.0020 | 0.0045 | 17.3530 | 0.231 | 0.0025 | 0.0046 | 17.0770 |
| OLS-MI | 0.041 | 0.0020 | 0.0003 | 0.3306 | 0.080 | 0.0027 | 0.0006 | 0.3269 |
| OLS-Best | 0.038 | 0.0011 | 0.0010 | 0.0069 | 0.170 | 0.0020 | 0.0021 | 0.0074 |
| | **B: Low-Error Environment** | | | | | | | |
| | Linkage Model 1 | | | | Linkage Model 2 | | | |
| | | | | Estimation | | | | Estimation |
| | $E_{\texttt{MC}}(\hat{\beta})$ | $E_{\texttt{MC}}(V(\hat{\beta}))$ | $V_{\texttt{MC}}(\hat{\beta})$ | time (seconds) | $E_{\texttt{MC}}(\hat{\beta})$ | $E_{\texttt{MC}}(V(\hat{\beta}))$ | $V_{\texttt{MC}}(\hat{\beta})$ | time (seconds) |
| Oracle | 0.254 | 0.0009 | 0.0008 | 0.0006 | 0.257 | 0.0012 | 0.0011 | 0.0005 |
| TSLS | 0.224 | 0.0017 | 0.0018 | 0.0007 | 0.261 | 0.0016 | 0.0028 | 0.0006 |
| IV | 0.233 | 0.0033 | 0.0035 | 0.0003 | 0.263 | 0.0021 | 0.0036 | 0.0003 |
| Lahiri-Larsen | 0.269 | 0.0012 | 0.0013 | 17.1910 | 0.257 | 0.0019 | 0.0026 | 15.8080 |
| OLS-MI | 0.116 | 0.0013 | 0.0005 | 0.3344 | 0.203 | 0.0020 | 0.0012 | 0.3216 |
| OLS-Best | 0.117 | 0.0008 | 0.0008 | 0.0044 | 0.239 | 0.0013 | 0.0016 | 0.0044 |

**Notes:** Each linkage environment is simulated 500 times. Parameter estimates of $\beta$ (true value = 0.25) represent the Monte Carlo mean over 500 simulations. $E_{\texttt{MC}}(V(\hat{\beta}))$ is the Monte Carlo mean of the variance estimate of $\hat{\beta}$ over 500 simulations. $V_{\texttt{MC}}(\hat{\beta})$ is the Monte Carlo variance of $\hat{\beta}$. Computation time represents the Monte Carlo mean of the time taken to compute the point estimate and variance for each type of estimator.