Received: 2 March 2021 | Revised: 8 July 2021 | Accepted: 15 July 2021

DOI: 10.1002/wics.1570

FOCUS ARTICLE

WIREs COMPUTATIONAL STATISTICS WILEY

Regression with linked datasets subject to linkage error

Zhenbang Wang[1] | **Emanuel Ben-David**[2] | **Guoqing Diao**[3] | **Martin Slawski**[1]

[1]Department of Statistics, George Mason University, Fairfax, Virginia, USA

[2]Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, District of Columbia, USA

[3]Department of Biostatistics and Bioinformatics, The George Washington University, Washington, District of Columbia, USA

Correspondence
Emanuel Ben-David, Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, DC 20233, USA.
Email: emanuel.ben.david@census.gov

Funding information
CCF National Science Foundation, Grant/Award Number: 1849876

Edited by: Kimberly Sellers, Commissioning Editor and David W. Scott, Co-Editor-in-Chief

Abstract

Data are often collected from multiple heterogeneous sources and are combined subsequently. In combing data, record linkage is an essential task for linking records in datasets that refer to the same entity. Record linkage is generally not error-free; there is a possibility that records belonging to different entities are linked or that records belonging to the same entity are missed. It is not advisable to simply ignore such errors because they can lead to data contamination and introduce bias in sample selection or estimation, which, in return, can lead to misleading statistical results and conclusions. For a long while, this problem was not properly recognized, but in recent years a growing number of researchers have developed methodology for dealing with linkage errors in regression analysis with linked datasets. The main goal of this overview is to give an account of those developments, with an emphasis on recent approaches and their connection to the so-called "Broken Sample" problem. We also provide a short empirical study that illustrates the efficacy of corrective methods in different scenarios.

This article is categorized under:

Statistical Models > Model Selection

Statistical and Graphical Methods of Data Analysis > Robust Methods

Statistical and Graphical Methods of Data Analysis > Multivariate Analysis

KEYWORDS

Bayesian analysis, data integration, linkage error, mixture models, record linkage, regression

1 | INTRODUCTION

Record linkage or entity resolution has a long history of uses, for example in federal statistical agencies and healthcare organizations. To study and answer research questions or to make informed decisions, researchers and policymakers regularly rely on data and the results of statistical analysis. In many situations, data are combined from multiple sources because a single dataset with all necessary information is unavailable and collecting data on additional variables is burdensome, time consuming, and costly. In the absence of common unique identifiers across datasets, an important intermediate step in combing data is record linkage. Formally, record linkage is the task of finding records in a dataset that refer to the same entity across different datasets. Unfortunately, since record linkage relies on quasi-identifiers, linkage can result in two types of errors: false matches or synonymously *mismatches* (when records of two different entities are erroneously linked) and false non-matches or synonymously *missed matches* (when records belonging to

the same individual are not linked). In statistical analysis with linked data, linkage error can invalidate basic assumptions made about data generation, such as the joint distribution of observations or random sampling. Simply said, linkage error contaminates the data and introduces bias, which in return leads to misleading statistical results and conclusions. We note that linkage error tends to be more likely when large datasets are merged, because as the size of the datasets increases, there is a higher chance that quasi-identifiers such as names are non-unique.

For quite a while, this problem was not properly recognized, with the exception of few early studies by Neter et al. (1965) and Scheuren and Winkler (1993, 1997), which show that even a small linkage error rate can have a significantly adverse effect on subsequent statistical results and findings. In recent years, however, a variety of studies from different communities, such as Abowd et al. (2019), Massey et al. (2018), Rentsch et al. (2018), Di Consiglio and Tuoto (2018), and Hof and Zwinderman (2012), have emphasized the importance of the problem, and at the same time a growing number of research papers have proposed methodologies for dealing with linkage errors in regression analysis of linked datasets.

Overall, record linkage and adjustment of post-linkage analysis to account for potential linkage errors is a vast and active area of research in recent years; it is considered as one of the grand challenges in the Center for Statistical Research Methodology at the United States Census Bureau. In light of this, we herein do not make the attempt to provide a comprehensive review on various topics within this area, but instead present a concise account of a specific problem of great relevance to post-linkage analysis. To be precise, we give an overview on remedies for regression analysis contaminated by mismatch error. Without doubt, regression analysis in its many variants is the most essential and the most widely used statistical technique across a myriad of application domains. It is therefore natural to consider it as an anchor point in the development of a complete framework for post-linkage data analysis.

We emphasize that our exposition is limited to the impact of *mismatches*, that is, records that have been linked despite not corresponding to the same entity. *Missed matches* need to be addressed quite differently, for example, by using suitable missing data methodology as in Little and Rubin (1987), and will not be discussed in the sequel.

Moreover, our survey puts an emphasis on methods for *secondary analysis* as opposed to *primary analysis*. The latter describes the situation in which the data analyst has access to the individual files underlying the merged file as well as to the matching variables used during record linkage. By contrast, in the more challenging setting of *secondary analysis* only the merged file is given and knowledge about the linkage process is incomplete at best. Secondary analysis is rather common in order to limit the amount of data to be shared and to prevent access to sensitive information. For example, consider a survey to be complemented by variables contained in a database containing information about a much larger set of entities than those that participated in the survey. Common matching variables include demographics such as birthday, ZIP code of the residential address, and so on. It is well known, for example from Sweeney (2001) and Hundepool et al. (2012), that releasing such information is associated with considerable re-identification risks of individuals. Another important focus that is somewhat specific to this survey is the "Broken Sample" perspective in DeGroot et al. (1971), DeGroot and Goel (1976, 1980), Goel (1975), Goel and Ramalingam (1987) and recent developments in this regard, several of which have been put forth by the authors. The approach bears considerable potential given advances in computation that were not available at the time of its introduction in the 1970s, and specifically addresses underexplored settings in which no or only very little information about the data linkage process is available to the data analyst beyond the final output, that is, the linked file. Several advances on the "Broken Sample" problem were made in neighboring disciplines (machine learning and signal processing), and it is thus hoped that this survey will contribute to the formation of connections between those disciplines on the one hand and the substantial body of research on record linkage conducted in the statistics community on the other hand.

The outline of this paper is as follows. In Section 2, we provide a short summary of record linkage to provide more context for readers new to this area. Section 3 then introduces the central subject of this article and important notations. Sections 3.1–3.7 discuss the impact of mismatch error on linear regression analysis and inference as well as series of methods to adjust for or correct such error. Specific extensions of interest beyond the basic linear regression setting are discussed in Section 4. A brief empirical study and illustrations are contained in Section 5. Concluding remarks are provided in Section 6.

2 | RECORD LINKAGE IN A NUTSHELL

In an effort to keep this article self-contained, we here provide basic background on record linkage. Excellent introductions to this area can be found in the monographs (Christen, 2012; Herzog et al., 2007) and survey articles (Binette & Steorts, 2020; Brizan & Tansel, 2006; Enamorado et al., 2019; Winkler, 2006, 2014).

The main goal of combing data from multiple sources is to collect and gather more comprehensive information about a target population. In the absence of a common unique identifier, a main challenge is to identify which records in different data files refer to the same entity. Record linkage, entity resolution, or data matching is a task that specifically deals with this problem. A brief introduction is as follows. Two files F_1 and F_2 consisting of n and N records, respectively, pertaining to entities from a target population are to be combined. A record is a row in the data file that consists of the attributes (or characteristics) of an individual on a number of fields (or synonymously variables). The task of record linkage is to classify pairs of records $(i,j) \in F_1 \times F_2$ into two classes, the class of true matches M, that is, both records refer to the same entity, and the class of true non-matches U, that is, i and j refer to different entities. In the Fellegi–Sunter framework (Fellegi & Sunter, 1969), classification into M or U is based on the value of the ratio $R = p\left(\gamma_{ij}|M\right)/p\left(\gamma_{ij}|U\right)$, in which γ_{ij} is a vector that denotes the pattern of agreement/disagreement in each common field (or matching variable) of i and j, and $p\left(\gamma_{ij}|M\right)$ is the probability of observing γ_{ij} given that the pair is a match; $p\left(\gamma_{ij}|U\right)$ is defined analogously. Given upper and lower thresholds T_U and T_L, respectively, a given pair (i,j) is declared a match if $R \geq T_U$, and a non-match if $R \leq T_L$. If the value of R is between T_L and T_U, the pair is declared a possible match. This basic template has been developed further in various methods, including several recent Bayesian approaches (cf. Binette & Steorts, 2020).

A fundamental challenge in record linkage is the computational complexity associated with the comparison of a large number of pairs of records. Most approaches rely on blocking techniques which are computationally inexpensive operations to partition each data file into corresponding "blocks," that is, subsets of records so that it is certain that records from non-corresponding blocks are non-matches, and comparisons hence only need to be made for records that are in the same corresponding block. A basic approach is to select categorical fields, for example, gender, month of birth, or the zip code of an address, and then partition each data file by the levels of these fields. For a more comprehensive review of blocking techniques, we refer the reader to Steorts et al. (2014).

Regardless of what methodology is employed, record linkage is rarely error-free and two types of errors can occur when a pair is misclassified: a false match when a non-match is declared a match or a false non-match when a match is declared a non-match. In the subsequent sections, we discuss the impact of false matches (mismatches) on regression analysis of linked data files as illustrated in Section 3 (Figure 1).

3 | LINEAR REGRESSION WITH LINKED DATASETS

We start by fixing the setup and notation. We suppose that the covariates are contained in a file $F_{\mathbf{x}} = \{\mathbf{x}_i\}_{i=1}^N \subset \mathbb{R}^d$ and response variables are contained in another file $F_{\mathbf{y}} = \{\mathbf{y}_i\}_{i=1}^n \subset \mathbb{R}^m$, where $n \leq N$. The merged file is denoted by $F_{\mathbf{x}\cdot\mathbf{y}} = \{(\mathbf{x}_i,\mathbf{y}_i)\}_{i=1}^n$, to be distinguished from the latent quantity $F^*_{\mathbf{x}\cdot\mathbf{y}} = \left\{\left(\mathbf{x}_{\pi^*(i)},\mathbf{y}_i\right)\right\}_{i=1}^n$ containing $(\mathbf{x},\mathbf{y})$-pairs in their correct correspondence. Here, $\pi^* : \{1,\ldots,n\} \to \{1,\ldots,N\}$ is an unknown one-to-one map. A pair $(\mathbf{x}_i,\mathbf{y}_i)$ is a mismatch in $F_{\mathbf{x}\cdot\mathbf{y}}$ if $i \neq \pi^*(i)$, $1 \leq i \leq n$. We write $\mathbf{X}_N$, $\mathbf{X}$, and $\mathbf{Y}$ for the matrices whose rows are given by $\{\mathbf{x}_i^{\mathrm{T}}\}_{i=1}^N$, $\{\mathbf{x}_i^{\mathrm{T}}\}_{i=1}^n$, and $\{\mathbf{y}_i^{\mathrm{T}}\}_{i=1}^n$, respectively. Unless noted otherwise, the response variable is one-dimensional $(m = 1)$, in which case we drop

File A

ID	Age	Sex	ZIP	Edu	Salary($)
1	45	F	47134	Master	6030
2	36	M	31526	Doctorate	8427
3	25	M	63312	Bachelor	5616
4	30	M	17621	High School	3408
5	45	F	47134	Doctorate	7799
6	25	M	63312	Master	6500
7	55	F	17621	High School	3266
8	34	F	17621	Bachelor	4084

File B

ID	Age	Sex	ZIP	weeks_unemployed
5	45	F	47134	7
7	55	F	17621	21
3	25	M	63312	13
2	36	M	31526	5
1	45	F	47134	11
4	30	M	17621	19
6	25	M	63312	9
8	34	F	17621	15

FIGURE 1 The setting of regression based on two linked files A and B. The response variable `weeks_unemployed` (duration of unemployment in weeks) is contained in File B, while the two potential predictor variables `Edu` (education level) and `Salary` (past monthly salary in US$) are contained in File A. The variables `Age` and `Sex` are potential predictor variables contained in both files. In combination with `ZIP` (zip code of home address), these three variables can thus be used as matching variables in record linkage. Possible sources of mismatch error corresponding to records with non-unique combinations of matching variables are highlighted via gray shading and framed boxes, respectively. Reproduced from Wang et al. (2020)

the boldface notation, that is, we write $(\mathbf{x},y)$, F_y, $\{y_i\}_{i=1}^n$, and so on. The restriction $n \le N$ is necessary to rule out the case of missed matches among the responses. The case $n < N$ is typically referred to as *sample-to-register linkage* (cf. Chambers & da Silva, 2020): here, the response is collected for a random subset of the population whose covariate information is stored in the register. The requirement that π^* be one-to-one is equivalent to not having duplicate entities in $F_{\mathbf{y}}$. Note that if $N = n$, the map π^* is a permutation of $\{1,\ldots,n\}$; even if $n \neq N$, π^* will be referred to as permutation for simplicity. The matrix representation of π^* is denoted by $\Pi^* = \left(\Pi^*_{i,j}\right)_{1 \le i \le n, 1 \le j \le N}$ where $\Pi^*_{i,j} = 1$ if $\pi^*(i) = j$ and $\Pi^*_{i,j} = 0$ otherwise, $1 \le i \le n$, $1 \le j \le N$. Note that Π^* is an element of the set of matching matrices $\mathcal{M}(n,N) = \left\{\Pi \in \{0,1\}^{n \times N} : \sum_{j=1}^N \Pi_{i,j} = 1, 1 \le i \le n, \sum_{i=1}^n \Pi_{i,j} \le 1, 1 \le j \le N\right\}$. Note that $\mathcal{M}(n,n) =: \mathcal{P}(n)$ is the set of permutation matrices, that is, n-by-n binary matrices with unit row and column sums. In the literature, π^* is predominantly treated as a latent random quantity, which makes sense intuitively given that the linkage process is associated with uncertainty arising, for example, from random errors, incompleteness, spelling or formatting variations in the matching variables.

We note that in several approaches to regression with linked data such as those reviewed in Sections 3.2 and 3.3, it is implicitly assumed that π^* is generated from a prior distribution on the set of permutations with known expectation. In essence, this unspecified distribution is determined by the linkage mechanism and the data generating process for the matching variables. However, in some situations, there may be little to no information about the linkage mechanism and the data generating process to justify such an assumption and to propagate uncertainty about the random permutation to subsequent statistical analysis. A remedy is to condition on the linkage mechanism and the matching variables, and to treat π^* as a parameter in the model; in fact, some works such as Pananjady et al. (2018), Hsu et al. (2017), Slawski and Ben-David (2019), and Zhang et al. (2019) have adopted a minimax paradigm that aims at safeguarding against *all possible* π^* from suitable subsets of $\mathcal{M}(n,N)$ as elaborated in Section 3.5.

3.1 | Impact of mismatch error on linear regression analysis and possible remedies

Suppose the data analyst intends to a fit a linear regression model of a single response variable y on $\mathbf{x}$ given the merged file $F_{\mathbf{x}\cdot y} = \{\mathbf{x}_i, y_i\}_{i=1}^n$. Building on the setting outlined in the introductory portion of this section, we assume the following model for all subsequent developments:

$$y_i = \mathbf{x}^{\mathrm{T}}_{\pi^*(i)}\beta^* + \sigma_* \varepsilon_{\pi^*(i)} \Leftrightarrow \mathbf{Y} = \Pi^*\left(\mathbf{X}_N\beta^* + \sigma_*\boldsymbol{\varepsilon}\right), \quad \mathbf{Y} = (y_i)_{i=1}^n, \ \boldsymbol{\varepsilon} = (\varepsilon_i)_{i=1}^N, \tag{1}$$

where the ε_i's are i.i.d. zero-mean, unit variance errors independent of Π^* and $\mathbf{X}_N$, σ_* is a non-negative number, and $\beta^* = \mathrm{argmin}_{\beta \in \mathbb{R}^d} \mathbf{E}_{\boldsymbol{\varepsilon}}\left[\left\|\mathbf{Y} - \Pi^*\mathbf{X}_N\beta\right\|_2^2\right]$. The solution β^* is assumed to be unique and independent[1] of Π^*. For our discussion, we assume that primary interest concerns inference for β^*, which includes point estimation, confidence intervals for individual coefficients, linear hypothesis tests, and so on. To reduce notation, we assume that a constant term (intercept) is included in the $\mathbf{x}_i$'s.

The *naive estimator* $\hat{\beta}_N$ refers to the least squares solution based on $F_{\mathbf{x}\cdot y}$ while ignoring the possibility of mismatch error as reflected in Equation (1). The mean squared estimation error (MSE) of this estimator can be quantified as

$$\mathbf{E}\left[\left\|\hat{\beta}_N - \beta^*\right\|_2^2\right] = \mathbf{E}\left[\left\|\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y} - \beta^*\right\|_2^2\right] = \mathbf{E}\left[\left\|\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\Pi^*\mathbf{X}_N\beta^* - \beta^*\right\|_2^2\right] + \sigma_*^2\frac{\mathrm{tr}\left(\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}/n\right)^{-1}\right)}{n}, \tag{2}$$

where the two terms on the right-hand side represent squared bias and variance, respectively. The latter is not affected by the presence of Π^*. Expanding the bias term further, we obtain that (cf. Appendix for a detailed derivation)

$$\mathbf{E}\left[\left\|\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\Pi^*\mathbf{X}_N\beta^* - \beta^*\right\|_2^2\right] \le \lambda_{\min}^{-1}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}/n\right)\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\left\{\left(\mathbf{x}_{\pi^*(i)} - \mathbf{x}_i\right)^{\mathrm{T}}\beta^*\right\}^2\right] \le \lambda_{\min}^{-1}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}/n\right)\frac{k}{n}\left\|\beta^*\right\|_2^2 D^2_{\mathbf{X}_N}, \tag{3}$$

where $\lambda_{\min}(\cdot)$ returns the minimum eigenvalue of a real symmetric matrix, $k = \left|\left\{i : \pi^*(i) \neq i\right\}\right|$ denotes the number of mismatches, and $D_{\mathbf{X}_N} = \max_{1 \leq i,j \leq N} \left\|\mathbf{x}_i - \mathbf{x}_j\right\|_2$. Treating the minimum eigenvalue term and $D_{\mathbf{X}_N}$ as constants, we observe that the squared bias roughly scales as $\frac{k}{n}\left\|\beta^*\right\|_2^2$. This indicates that the estimator $\hat{\beta}_N$ is not consistent unless $k/n \to 0$. Note that the second term in the rightmost expression in (2) represents the MSE of the *oracle estimator* based on an oracle that is equipped with knowledge of Π^*, or equivalently, the MSE of the usual least squares estimator if responses and predictors are matched correctly without error. The excess error of the order $k/n\left\|\beta^*\right\|_2^2$ relative to this oracle suggests that the performance of $\hat{\beta}_N$ is particularly poor if the so-called *signal-to-noise ratio* $\mathrm{SNR} = \left\|\beta^*\right\|_2^2/\sigma_*^2$ is large, that is, if the linear model achieves a good fit. In the same vein, (3) suggests that $\hat{\beta}_N$ can be a catastrophic estimator since even the trivial estimator $\hat{\beta}_0 = \mathbf{0}$ independent of the observed data may achieve a comparable MSE!

The case of a single predictor variable ($d = 1$). For simplicity, let us assume that the underlying intercept is zero, and can hence be omitted. In this case, $\left|\mathbf{E}_\epsilon\left[\hat{\beta}_N\right]\right| = \left|\sum_{i=1}^n x_i x_{\pi^*(i)}\right|\left|\beta^*\right| / \sum_{i=1}^n x_i^2 \leq \left|\beta^*\right|$, which indicates that the naive estimator suffers from an *attenuation bias*. An illustration is depicted in Figure 2, which also shows a massive degradation in model fit, with a drop in the coefficient of determination R^2 from 0.52 to 0.03.

Inference. The bias of $\hat{\beta}_N$ immediately implies that subsequent inference as for the usual least squares solutions does not yield valid conclusions. Using a similar reasoning as in that leading to (3), it can be shown that the error variance σ_*^2 will be over-estimated, which implies that hypothesis tests may suffer from substantially reduced power.

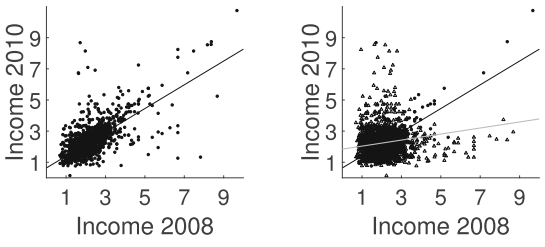
Remedies. The deficiencies of the naive solution $\hat{\beta}_N$ are apparent in light of the preceding discussion. Different mitigation strategies pursue the following goals. Adjustment methods aim at bias reduction or the MSE in general. The choice of the adjustment depends on multiple factors, including the goal of data analysis (e.g., valid inference vs. predictive performance), the fraction of mismatched data, and the SNR. Several recent papers consider the more ambitious goal of estimating Π^* (*permutation recovery*), which subsumes adjustment since being able to estimate Π^* with small error or even exactly naturally yields an improved estimator for β^*. *Mismatch recovery* aims at the identification of the set $\left\{i : \pi^*(i) \neq i\right\}$. Removing all mismatches yields an unbiased estimator, which, however, comes with a loss in statistical efficiency, since only a subset of the given data is used. In the sequel, we discuss a variety of approaches that pursue at least one of the aforementioned mitigation strategies under different assumptions.

3.2 | The Lahilri and Larsen estimator

Building on earlier work by Neter et al. (1965), Scheuren and Winkler (1993, 1997), and Winkler (1995), Lahiri and Larsen (2005) propose an unbiased estimator of β^* in Equation (1). Note that we have $\mathbf{Y} = \Pi^*\mathbf{Y}^*$ with $\mathbf{Y}^* = \mathbf{X}_N\beta^* + \sigma_*\boldsymbol{\varepsilon}$. Suppose that (i) $\mathbf{Y}^*$ and Π^* are conditionally independent given $\mathbf{X}_N$, and (ii) $\mathbf{E}\left[\Pi^* \middle| \mathbf{X}_N\right] = Q$. It follows that

$$\mathbf{E}[\mathbf{Y}|\mathbf{X}_N] = \mathbf{E}\left[\Pi^*\mathbf{Y}^* \middle| \mathbf{X}_N\right] = \mathbf{E}\left[\Pi^* \middle| \mathbf{X}_N\right]\mathbf{E}\left[\mathbf{Y}^* \middle| \mathbf{X}_N\right] = Q\mathbf{X}_N\beta^*, \tag{4}$$

where the first identity follows from assumption (i), which is typically, as in Han and Lahiri (2019), referred to as *non-informative linkage* or *linkage at random* assumption in the literature, in analogy to the missing at random mechanism



	oracle	naive
intercept	0.63	1.84
slope	0.76	0.19
residual variance	0.38	0.78
R^2	0.52	0.03

FIGURE 2 Fictitious example dataset based on the Italian household survey discussed in Tancredi and Liseo (2015). Here, the monthly household income (in 1000 Euros) in 2010 is regressed on the same quantity in 2008. Left: scatterplot and estimated regression line in the absence of mismatch error ("oracle"). Middle: scatterplot and regression line (gray) after linking two files containing the records for 2008 and 2010, respectively. Mismatches are represented by Δ, correct matches by •. Right: summary of the linear regression fits corresponding to the left and middle plot; R^2 denotes the coefficient of determination

in missing data analysis pioneered in Little and Rubin (1987). The relation in (4) immediately implies that least regression of $\mathbf{Y}$ on $Q\mathbf{X}_N$ yields the unbiased estimator

$$\hat{\beta}^{\mathrm{LL}} = \left(\mathbf{X}_N^{\mathrm{T}} Q^{\mathrm{T}} Q \mathbf{X}_N\right)^{-1} \mathbf{X}_N^{\mathrm{T}} Q^{\mathrm{T}} \mathbf{Y}. \tag{5}$$

Obtaining expressions for the asymptotic covariance of $\hat{\beta}^{\mathrm{LL}}$ is more intricate given the dependence of this quantity on the distribution of Π^*. As an intermediate result, Lahiri and Larsen (2005) show how to obtain a consistent estimator of σ_*^2. In Lahiri and Larsen (2005) as well as Han and Lahiri (2019), Jackknife and bootstrap estimators are proposed to assess the variability of $\hat{\beta}^{\mathrm{LL}}$.

An obvious shortcoming of the estimator (5) is that it assumes that Q is known or at least estimable. If Q is estimated, the additional uncertainty needs to be accounted for, and in the complete absence of information about the linkage process with only the linked file being given, the approach cannot be applied. At the same time, the improvement over the naive least squares solution can be good to excellent as long as blocking variables are given that partition the observations into a reasonably number of disjoint bocks, and Π^* is assumed to match uniformly at random within each block. This is confirmed in our empirical study in Section 5 and has been also verified analytically in Wang et al. (2020).

3.3 | Chambers' estimator

Kim and Chambers (2012) and similarly Chambers and da Silva (2020) propose a method that generalizes the approach of Lahiri and Larsen (2005). Suppose the records in $F_{\mathbf{x}}$ can be partitioned into B blocks, that is, distinct and non-overlapping sets such that linkage errors only occur within these blocks. The estimation of the regression parameters in Chambers and da Silva (2020) and Kim and Chambers (2012) is then based on the following so-called adjusted estimating equation

$$H_{adj}^*(\beta) := \sum_{b=1}^{B} G_b(\mathbf{X}_b, \beta)(\mathbf{Y}_b - Q_b \mathbf{X}_b \beta) = \mathbf{0}, \tag{6}$$

where $\mathbf{X}_b$ and $\mathbf{Y}_b$ are the row submatrix and subvector associated with the bth block, and accordingly Q_b is the corresponding principal submatrix (bth diagonal block) of $Q = \mathbf{E}\left[\Pi^* \middle| \mathbf{X}_N\right]$ as in the previous subsection, $b = 1, \ldots, B$. Moreover, $\{G_b(\mathbf{X}_b, \beta)\}_{b=1}^B$ are weighting matrices that may depend on the respective submatrices $\{\mathbf{X}_b\}_{b=1}^B$ and β.

It is easy to see that the solution to Equation (6), henceforth denoted by $\hat{\beta}^C$, is an unbiased estimator of β^*. We note that in the absence of blocking, Equation (6) reduces to $G(\mathbf{X}_N, \beta)(\mathbf{Y} - Q\mathbf{X}_N\beta) = \mathbf{0}$; for the particular choice $G(\mathbf{X}_N, \beta) = \mathbf{X}_N^{\mathrm{T}} Q^{\mathrm{T}}$, we obtain the LL-estimator (5). The choice $G(\mathbf{X}_N, \beta) = \mathbf{X}_N^{\mathrm{T}}$ yields the Chambers' original proposal in Chambers (2009). Both choices yield linear unbiased, but generally not efficient estimators. An optimal choice of the weighting matrix would minimize the asymptotic covariance,[2] that is, would generate a BLUE (best linear unbiased) estimator. This yields $G_b(\mathbf{X}_b, \beta) = \mathbf{X}_b^{\mathrm{T}} Q_b^{\mathrm{T}} \{\mathrm{Cov}(\mathbf{Y}_b | \mathbf{X}_b)\}^{-1}$, $b = 1, \ldots, B$; since the $\{\mathrm{Cov}(\mathbf{Y}_b | \mathbf{X}_b)\}_{b=1}^B$ are unknown, they need to be estimated, cf. Chambers and da Silva (2020) and Chambers (2009).

A setting specifically considered in Chambers and da Silva (2020) and Kim and Chambers (2012) is termed *exchangeable linkage error* (*ELE*), in which the off-diagonal elements of Q_b are constant (and hence all diagonal elements are equal to a complementary constant). Novel insights into the ELE setting are recently presented in Zhang and Tuoto (2020).

Han and Lahiri (2019) extend the approach pioneered by Chambers to a primary data analysis setting in which regression and record linkage are carried out simultaneously. In order to incorporate uncertainty resulting from the linkage process, a jackknife method is employed to estimate bias, variance, and mean squared error of the proposed estimators.

3.4 | Bayesian approaches

In addition to the frequentist methods considered herein, there is also a significant line of research concerning Bayesian approaches to the linear regression problem with linked data files, a selection of which is reviewed below.

Tancredi and Liseo (2015) propose an approach in which the observed matching variables in two files, denoted by matrices $\mathbf{V_y}$ and $\mathbf{V}_x$ are possibly faulty versions of their true but latent matrices $\mathbf{V}^*_{\mathbf{y}}$ and $\mathbf{V}^*_x$ respectively. The main parameter that determines the true linkage is the $n \times N$ matching matrix Π^* with $\Pi^*_{ij} = 1$ if (i,j) is a match and 0 if (i,j) is a non-match. The main ingredients of this Bayesian approach are specifications for the distributions $p\left(\mathbf{V_x}|\mathbf{V}^*_{\mathbf{x}}\right)$, $p\left(\mathbf{V}_y|\mathbf{V}^*_y\right)$, $p\left(\mathbf{V}^*_{\mathbf{x}},\mathbf{V}^*_y\middle|\Pi^*\right)$, $p\left(\mathbf{X}^*_N|\Pi^*\right)$ for the distribution of the true but unobserved matrix of covariates $\mathbf{X}^*_N$, and finally a prior $p\left(\Pi^*\right)$ for the matching matrix Π^*. The approach then proceeds via posterior simulation by means of a Metropolis–Hastings algorithm that updates Π^* one-match at a time at each move. Whenever a move is accepted, the values of the variables corresponding to that match are updated by drawing from the specified distributions. In the end, the regression parameters β and σ^2 are updated using Equation (1). Steorts et al. (2018) propose an extension of this approach in which records are linked based on a random partition model.

Another influential Bayesian method is that of Gutman et al. (2013) (cf. also Section 3.7). In Gutman et al. (2013), the matching variables are assumed to be error-free, and the resulting blocking serves as a key element in order to reduce the number of paired comparisons. Dalzell and Reiter (2018) extend the approach of Gutman et al. (2013) to allow faulty matching variables in the file containing the covariates. A brief summary of their approach is as follows. For each block b, $F_{\mathbf{x}}$ and F_y are partitioned into corresponding sub-files $F_{\mathbf{x},b}$ and $F_{y,b}$, $b = 1,...,B$. By adding dummy records to either $F_{\mathbf{x},b}$ or $F_{y,b}$, we can always assume that both are of the same size. When a record in $F_{\mathbf{x},b}$ or $F_{y,b}$ is linked to a dummy record, we assume that the corresponding values for the response or the covariates are missing at random. The matching matrix Π^*_b associated with $F_{\mathbf{x},b}$ and $F_{y,b}$ is hence a permutation matrix, and the matching matrix Π^* for $F_{\mathbf{x}}$ and F_y is determined by $\left(\Pi^*_1,...,\Pi^*_B\right)$. The approach proceeds by specifying distributions $p\left(\mathbf{V_x}|\mathbf{V}^*_{\mathbf{x}}\right)$, $p\left(\mathbf{V}^*_{\mathbf{x}},\mathbf{V}_y|\Pi^*\right)$, and $p\left(\mathbf{X}^*_N|\Pi^*\right)$; note that $\mathbf{V}_y$ is assumed not to be faulty, hence inclusion of $\mathbf{V}^*_y$ is not needed. Each of the $\left\{\Pi^*_b\right\}_{b=1}^B$ follows a uniform prior. MCMC sampling is used to generate draws from the posterior distribution of Π^*, and the results are then used to estimate the regression parameter.

3.5 | "Broken Sample" perspective

The "Broken Sample" problem roots in a series of works by DeGroot and Goel such as DeGroot et al. (1971), DeGroot and Goel (1976, 1980), Goel (1975), and Goel and Ramalingam (1987). Let $P_{\mathbf{x},\mathbf{y}}$ denote the joint distribution of the random variables $\mathbf{x}$ and $\mathbf{y}$, and suppose interest concerns a functional θ^* of $P_{\mathbf{x},\mathbf{y}}$ such as, for example, the covariance $\Sigma_{\mathbf{xy}} = \mathrm{Cov}(\mathbf{x},\mathbf{y})$. However, in order to perform this task, the data analyst is only given two separate samples $F_{\mathbf{x}} = \{\mathbf{x}_i\}_{i=1}^n$ and $F_{\mathbf{y}} = \{\mathbf{y}_i\}_{i=1}^n$ such that $\left\{\left(\mathbf{x}_{\pi^*(i)},\mathbf{y}_i\right)\right\}_{i=1}^n$ are i.i.d. from $P_{\mathbf{x},\mathbf{y}}$, with π^* being an unknown permutation of $\{1,...,n\}$. This setting is obviously closely related to that outlined at the beginning of Section 3, essentially corresponding to the special case $N = n$ and the complete absence of information about which of the pairs $\left(\mathbf{x}_i,\mathbf{y}_j\right)_{i\le j}$ are likely to be matches. In fact, in the "Broken Sample" the corresponding merged file $F_{\mathbf{x}\cdot\mathbf{y}} = \{(\mathbf{x}_i,\mathbf{y}_i)\}_{i=1}^n$ obtained by linking matching indices may not contain a single correct match since the order of the samples in $F_{\mathbf{x}}$ may be arbitrarily shuffled relative to the order in $F_{\mathbf{y}}$.

Perhaps somewhat surprisingly, the "Broken Sample" problem has recently experienced a revival in the engineering and machine learning literature under terms such as "Unlabeled Sensing" in Pananjady et al. (2018), Haghighatshoar and Caire (2017), Tsakiris (2018) or "Learning from Shuffled Data" in Hsu et al. (2017), Abid et al. (2017), Pananjady et al. (2017), Slawski et al. (2019), Rigollet and Weed (2019), and Carpentier and Schlüter (2016), driven by specific applications as well as as by computational advances. The hope is that some of the developments in those fields turn out to be fruitful for the problem surveyed in this overview article.

Inference. In the original formulation of the "Broken Sample" problem, inference concerns θ^* while π^* is treated as a nuisance parameter. Unsurprisingly, even basic estimation problems become rather challenging in comparison to their conventional counterparts. In addition, the unknown permutation π^* may impose substantial computational challenges given the combinatorial nature of this quantity. The linear regression problem under the "Broken Sample" setting, which has been dubbed "Unlabeled Sensing" in Haghighatshoar and Caire (2017) in the area of signal processing, has been subject to extensive study during recent years, such as in Pananjady et al. (2018), Haghighatshoar and Caire (2017), Hsu et al. (2017), and Abid et al. (2017). The associated setting offers immediate insights into the principal challenges. Along this line, consider the modified least squares problem

$$\min_{\beta \in \mathbb{R}^d, \Pi \in \mathcal{P}(n)} \|\mathbf{Y} - \Pi \mathbf{X}\beta\|_2^2 = \min_{\Pi \in \mathcal{P}(n)} \left\|P_{\mathbf{X}}^{\perp} \Pi \mathbf{Y}\right\|_2^2, \tag{7}$$

where $\mathbf{Y} = (y_i)_{i=1}^n$, the rows of $\mathbf{X}$ are given by $\{\mathbf{x}_i^{\mathrm{T}}\}_{i=1}^n$, $\mathcal{P}(n)$ denotes the set of n-by-n permutation matrices, and $P_{\mathbf{X}}^{\perp}$ denotes the projection on the orthogonal complement of the subspace spanned by the columns of $\mathbf{X}$. For ease of exposition, we suppose for (the remainder of this section) that both $\mathbf{Y}$ and $\mathbf{X}$ are centered so that the intercept can be dropped.

The identity in (7) shows that given the permutation $\hat{\Pi}$ that minimizes the right-hand side, the corresponding solution $\hat{\beta} = \hat{\beta}(\hat{\Pi})$ can be obtained by least squares regression of $\mathbf{Y}$ on $\hat{\Pi}\mathbf{X}$. At the same time, the problem on the right-hand side of (7) is a specific instance of a *quadratic assignment problem*—a class of optimization problems notorious for their computational hardness that includes, among others, the famous traveling salesman and graph isomorphism problems (see Burkard et al., 2009). Indeed, it was established recently in Pananjady et al. (2018) that (7) is in general NP-hard for $d > 1$.

For $d = 1$, problem (7) can be solved in closed form. It is shown in Pananjady et al. (2018) that the minimizing permutation matrix is characterized by the condition

$$\left|\left\langle \mathbf{Y}, \hat{\Pi}\mathbf{X} \right\rangle\right| = \max\left\{ \sum_{i=1}^n x_{(i)} y_{(i)}, \left| \sum_{i=1}^n x_{(i)} y_{(n-i)} \right| \right\}, \tag{8}$$

where the subscript (i) refers to the ith order statistic, $1 \le i \le n$, for example, $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$. Equation (8) implies that $\hat{\Pi}$ can be obtained by sorting $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$. The corresponding solution for the regression coefficient is given by $\hat{\beta} = \langle \mathbf{Y}, \hat{\Pi}\mathbf{X} \rangle / \|\mathbf{X}\|_2^2$. It is shown in Abid et al. (2017) that $\hat{\beta}$ suffers from an *amplification bias*: under model (7), it holds that $\hat{\beta} \to \mathrm{sign}(\beta^*)\sqrt{(\beta^*)^2 + \sigma_*^2}$, and the corresponding estimator for the error variance σ_*^2 converges to zero as $n \to \infty$. This can be understood as a manifestation of a broader over-fitting phenomenon noted in Hsu et al. (2017), and Slawski and Ben-David (2019) that affects the statistical properties of the solution of (7), independent of the computational challenge. Subsequent discussion hence focuses on strategies addressing at least one of the two major limitations of the "Broken Sample" formulation (7).

3.6 | Approaches for sparse and partial shuffling

Solving "Broken Sample" problems is somewhat too ambitious since this amounts to performing record linkage and inference for an unknown parameter simultaneously, without any additional information on potential matches. In other words, the "Broken Sample" problem can be seen as an extreme instance of post-linkage data analysis in which (almost) all of the $(\mathbf{x}_i, \mathbf{y}_i)$ are mismatched; in fact, it is well known that when matching pairs at random, the number of correct matches approximately follows a Poisson distribution with expectation one (see Diaconis, 1988), that is, the fraction of correct matches scales as $1/n$. By contrast, linked files encountered in practice typically exhibit low to moderate mismatch rates given that record linkage is performed in an informed fashion (i.e., based on matching variables that pinpoint likely matches) rather than "blindly." This suggests that a relaxed version of the "Broken Sample" problem in which the underlying permutation π^* is constrained to move a bounded fraction of indices $0 \le \alpha_* < 1$, that is, $\left|\{i : \pi^*(i) \ne i\}\right| \le \alpha_* n$. In the sequel, we distinguish between two regimes for α_*, which we refer to *sparse* and *partial* shuffling, respectively. The former refers to the case in which α_* is small, say, less than 0.2, and certainly less than 0.5, that is, mismatches are the exception rather than the rule. *Partial shuffling* is less restrictive, requiring only that α_* is bounded away from one. Intuitively, consistent estimation should still be possible in this regime as long as the correct matches can be separated from mismatches, and the latter can be eliminated when performing the statistical analysis of interest.

Sparse shuffling. Consider $\xi^* = (I_n - \Pi^*)\mathbf{X}\beta^*$. The linear regression model (7) can then be rewritten as $\mathbf{Y} = \mathbf{X}\beta^* + \xi^* + \sigma_* \boldsymbol{\varepsilon}$, which in turn suggests the penalized least squares optimization problem

$$\min_{\beta \in \mathbb{R}^d, \xi \in \mathbb{R}^n} \|\mathbf{Y} - \mathbf{X}\beta - \xi\|_2^2 + \lambda \|\xi\|_1, \tag{9}$$

where $\lambda > 0$ is a tuning parameter. Adopting the "lasso trick" of Tibshirani (1996), the ℓ_1-penalty is used to promote sparsity in ξ in order to account for the fact that ξ^* is sparse with at most $\alpha^* \cdot n$ nonzero entries. It is shown in She and Owen (2012) that (9) is equivalent robust M-estimation with the Huber loss in Huber (1964). This connection can be used, as in She and Owen (2012), Antoniadis (2007), and Loh (2018) for example, for the proper calibration of λ via concomitant scale estimation. Let $\hat{\beta}^{\text{rob}}$ denote the minimizer of (9). It is shown in Slawski and Ben-David (2019) that under regularity conditions, $\left\|\hat{\beta}^{\text{rob}} - \beta^*\right\|_2^2 \leq C\sigma_*^2(\alpha_* + d/n)$ for some constant $C > 0$ with high probability[3] provided α_* is sufficiently small. Comparison with the corresponding MSE bound (2) and (3) for the naive estimator indicates that the amount fof improvement achieved by $\hat{\beta}^{\text{rob}}$ increases with the SNR $= \left\|\beta^*\right\|_2^2/\sigma_*^2$.

Partial shuffling. Consider model (7) under the additional assumption of Gaussian errors, that is, $\varepsilon_i \sim N(0,1)$, $1 \leq i \leq n$, and introduce latent mismatch indicators $z_i = I(\pi^*(i) \neq i)$, $1 \leq i \leq n$.

Assuming

- Independence between mismatched response and predictors, that is, i.e., $\mathbf{x}_i \perp\!\!\!\perp y_i \,|z_i = 1$,
- Homogeneous mismatch probability, that is, $\mathbf{P}(z_i = 1) = \alpha_*$ independent of i, the conditional distribution of y_i given $\mathbf{x}_i$ is given by the two-component mixture model

$$y_i \,|\, \mathbf{x}_i \sim (1-\alpha_*)N\left(\mathbf{x}_i^{\mathrm{T}}\beta^*, \sigma_*^2\right) + \alpha_* F_y,\ 1 \leq i \leq n, \tag{10}$$

where F_y represents the marginal distribution of the $\{y_i\}_{i=1}^n$. The mixture representation (10) suggests the "likelihood"

$$L(\beta, \alpha, \sigma^2) = \prod_{i=1}^n \left[(1-\alpha)\frac{1}{\sigma}\phi\left(\frac{y_i - \mathbf{x}_i^{\mathrm{T}}\beta}{\sigma}\right) + \alpha f_y(y_i)\right], \tag{11}$$

where ϕ denotes the standard normal density and f_y denotes the density associated with F_y. Since f_y is unknown, one may either follow a plug-in approach in which f_y is estimated in a parametric or non-parametric fashion from the $\{\mathbf{y}_i\}_{i=1}^n$, or estimate it within (11) noting that marginally, the $\{y_i\}_{i=1}^n$ are generated by the Gaussian location mixture

$$f_y(y) = \sum_{i=1}^n \frac{1}{n}\frac{1}{\sigma_*}\phi\left(\frac{y-\mu_i}{\sigma_*}\right),\ \ \mu_i = \mathbf{x}_i^{\mathrm{T}}\beta^*,\ 1 \leq i \leq n.$$

It is important to realize that (11) does not represent a proper likelihood since the joint distribution of the $y_i|\mathbf{x}_i$, $1 \leq i \leq n$, is *not* given by product of the marginal distributions (10). Nevertheless, the "likelihood" (11) falls within the framework of *composite-* or *pseudo*-likelihood of Lindsay (1988) and Varin et al. (2011). Theory for this framework asserts that the minimizer of (11) is $\sqrt{n}$-consistent and asymptotic normal (albeit generally not efficient), and that the asymptotic covariance matrix can be estimated consistently as well.

Compared to (9), the pseudo-likelihood approach (10) has several advantages. First, it achieves faster estimation rates (see below for details). Second, it yields consistent estimators of the error variance and the mismatch rate α_*, and inference for the parameters can be performed based on the expression for the asymptotic covariance matrix. Third, the fraction of mismatches that can be tolerated is significantly higher; at least in theory, it can be any constant fraction less than one. At the same time, the approach (10) has the following limitations: (1) unlike (9), optimization of the pseudo-likelihood is a non-convex optimization problem, and heuristics such as the EM algorithm as proposed in Slawski, Diao, and Ben-David (2020) are not guaranteed to deliver the global optimum, but may return spurious local optima depending on the initialization; (2) the assumptions preceding (10) are somewhat restrictive, in particular the first bullet, since mismatches typically involve $\mathbf{x}_i$'s that are correlated, where the correlation is induced by close agreement on the matching variables used for record linkage.

Refinements. The approaches (9) and (10) can be built on to perform *mismatch recovery* and *permutation recovery* (cf. Section 3.5). The former refers to identification of the set $S_* = \{i : \pi^*(i) \neq i\}$. After the identification of S_*, the data analyst can simply obtain a linear regression fit based on the remaining data $\{(\mathbf{x}_i, y_i)\}_{i \notin S_*}$. The MSE of this estimator $\widetilde{\beta}$

is given by $\left\|\widetilde{\beta}-\beta^*\right\|_2^2 \le (1-\alpha_*)^{-1}\sigma_*^2 d/n$, which is the limit of what can be achieved by the pseudo-likelihood approach (10) and which can be significantly lower than the MSE bound for $\hat{\beta}^{\mathrm{rob}}$. Note that in order to achieve a comparable MSE rate of $\sigma_*^2 d/n$, it suffices to identify a *superset* of S_*. Under suitable conditions, this can be achieved by post-processing the output of (9) and (10). Specifically, for (9) we consider the minimizer $\hat{\xi}$ and compare its entries to σ_* respectively an estimate thereof: if $|\hat{\xi}_i|$ exceeds a certain multiple of σ_*, consider the corresponding pair $(\mathbf{x}_i, y_i)$ a potential mismatch. Alternatively, in order to circumvent the explicit selection of a threshold, one may sort the $\{|\hat{\xi}_i|\}_{i=1}^n$ and remove the pairs corresponding to the $\overline{\alpha}\cdot n$ largest values, where $\overline{\alpha}$ is a (guessed) upper bound on α_*. Similar procedures can be used in conjunction with (10) by considering plug-in estimates of the posterior mismatch probabilities

$$\mathbf{P}(z=1|(\mathbf{x}_i,y_i)) = \frac{\alpha_* f_y(y_i)}{(1-\alpha)\frac{1}{\sigma_*}\phi\left(\frac{y_i-\mathbf{x}_i^{\mathrm{T}}\beta^*}{\sigma_*}\right)+\alpha_* f_y(y_i)}, \quad 1\le i\le n.$$

For the purpose of mismatch recovery, all unknown quantities are substituted by their estimates.

Regarding *permutation recovery*, a crucial observation from a computational viewpoint is the following:

$$\min_{\Pi\in\mathcal{P}(n)}\|\mathbf{Y}-\Pi\mathbf{X}\beta\|_2^2 = \min_{\Pi\in\mathcal{P}(n)}\langle\mathbf{Y},\Pi\mathbf{X}\beta\rangle + c = \min_{\Pi\in\mathcal{P}(n)}\mathrm{tr}\left(\Pi\mathbf{X}\beta\mathbf{Y}^{\mathrm{T}}\right)+c = \min_{\Pi\in\mathscr{B}(n)}\mathrm{tr}(\Pi\mathbf{C})+c, \tag{12}$$

for any *fixed* β, where $c=\|\mathbf{Y}\|_2^2+\|\Pi\mathbf{X}\beta\|_2^2=\|\mathbf{Y}\|_2^2+\|\mathbf{X}\beta\|_2^2$ does not depend on Π, $\mathbf{C}=\mathbf{X}\beta\mathbf{Y}^{\mathrm{T}}$, and $\mathscr{B}(n)$ is the convex hull of $\mathcal{P}(n)$, known as the set of doubly stochastic matrices or the Birkhoff polytope. The right-hand side of (12) is an instance of a *linear* assignment problem (LAP) that are specific linear programs, and hence can be solved efficiently via various algorithms in Burkard et al. (2009). In particular, in virtue the rightmost identity in (12), there are at least one integral minimizer (i.e., in this case a permutation matrix) of the linear program. That minimizer can be obtained by specialized algorithms such as the Hungarian algorithm in Kuhn (1955) or the Auction algorithm in Bertsekas and Castanon (1992). The principal implication of (12) is a computationally tractable *two-stage* estimation approach: in the first stage, an estimator for the regression parameter is obtained, for example, from (9) or (10); in the second stage, that estimator is substituted for β in (12) to obtain an estimate of Π^*. As mentioned in Section 3.5 that estimate generally suffers from an underlying overfitting phenomenon, which can be curbed by imposing suitable restrictions on Π such as a bound on the number of indices $\{i:\pi^*(i)\neq i\}$ that can be moved, or block structure resulting from blocking variables used during record linkage. We refer to section 2.4 in Slawski and Ben-David (2019) for more details on two-stage estimation.

3.7 | EM algorithm and data augmentation

One natural and hence also historically early take on the "Broken Sample problem" is the treatment as a missing data problem, with the unknown permutation Π^* representing missing data. This perspective prompts the use of a traditional tool for handling missing data: the expectation–maximization (EM) algorithm. Under the assumption of normal errors, that is, $\varepsilon_i\sim N(0,1)$, $1\le i\le n$, the complete data negative log-likelihood is given by

$$\ell_C(\beta,\sigma^2|\mathbf{X},\mathbf{Y},\Pi^*) \propto \left\|\mathbf{Y}-\Pi^*\mathbf{X}\beta\right\|_2^2/\sigma^2 + n\log(\sigma^2) = \left\{-2\langle\mathbf{Y},\Pi^*\mathbf{X}\beta\rangle+\|\mathbf{X}\beta\|_2^2\right\}/\sigma^2+n\log(\sigma^2)+\|\mathbf{Y}\|_2^2. \tag{13}$$

Given initial iterates $\hat{\beta}_{(0)}$ and $\hat{\sigma}^2_{(0)}$, the **E**-step with respect to Π^* takes the form

$$\begin{aligned}
&\Pi_{(0)} := \mathbf{E}\left[\Pi^*\middle|\mathbf{X},\mathbf{Y},\hat{\beta}_{(0)},\hat{\sigma}^2_{(0)}\right] = \sum_{\Pi\in\mathcal{P}(n)}\Pi\cdot\mathbf{P}\left(\Pi^*=\Pi|\mathbf{X},\mathbf{Y},\hat{\beta}_{(0)},\hat{\sigma}^2_{(0)}\right),\\
&\text{where } \mathbf{P}\left(\Pi^*=\Pi|\mathbf{X},\mathbf{Y},\hat{\beta}_{(0)},\hat{\sigma}^2_{(0)}\right) = \exp\left(-\frac{\left\|\mathbf{Y}-\Pi\mathbf{X}\hat{\beta}_{(0)}\right\|_2^2}{2\hat{\sigma}^2_{(0)}}\right)\Big/\left[\sum_{\Pi'\in\mathcal{P}(n)}\exp\left(-\frac{\left\|\mathbf{Y}-\Pi'\mathbf{X}\hat{\beta}_{(0)}\right\|_2^2}{2\hat{\sigma}^2_{(0)}}\right)\right].
\end{aligned} \tag{14}$$

Substituting Π^* in (13) by $\Pi_{(0)}$ gives rise to the expected complete data negative log-likelihood whose minimization, that is, the **M**-step, amounts to least squares regression of $\Pi_{(0)}^{\mathrm{T}}\mathbf{Y}$ on $\mathbf{X}$, which yields new iterates $\hat{\beta}_{(1)}$ and $\hat{\sigma}^2_{(1)}$. Clearly, the **E**-step (14) cannot be performed in practice unless n is tiny. Wu (1998) proposes to approximate the expectation via a specific MCMC sampling scheme known as the Fisher–Yates shuffling algorithm. This approach has been rediscovered recently in Abid and Zou (2018), and has been developed further in Gutman et al. (2013) (cf. Section 3.4). That work builds on the connection between the EM algorithm and data augmentation in Tanner and Wong (1987), which is closely rated to Gibbs sampling. The scheme can be summarized as follows.

$$\text{Augmentation step}: \text{ sample } \Pi^{(j)} \sim p\left(\Pi^* \middle| \mathbf{Y}, \mathbf{X}, \beta_{(k)}, \sigma^2_{(k)}\right), \quad j = 1, \ldots, M, \tag{15}$$

$$\text{Posterior step}: \text{ sample } \beta_{(k+1)} \sim \frac{1}{M}\sum_{j=1}^{M} p\left(\beta \middle| \Pi^{(j)}, \mathbf{Y}, \mathbf{X}, \sigma^2{}_{(k)}\right), \text{ and sample } \sigma^2_{(k+1)} \sim \frac{1}{M}\sum_{j=1}^{M} p\left(\sigma^2 \middle| \Pi^{(j)}, \mathbf{Y}, \mathbf{X}, \beta_{(k)}\right), \tag{16}$$

where $p(\cdot\,|\,\cdot)$ is used generically to denote conditional distributions, and the subscript (k) refers to the kth iterate (sample). The augmentation step can be realized by adopting the same sampling algorithm that is used for approximating the **E**-step (14), and the posterior step is realized by noting that $p(\beta|\cdot)$ and $p(\sigma^2|\cdot)$ are Gaussian and inverse-χ^2, respectively.

The framework in Gutman et al. (2013) assumes that Π^* is block-structured given blocking variables used for record linkage, which yields considerable reductions in complexity. In this situation, the **E**-step in (14) as well as the above augmentation step can be performed independently block-by-block. Moreover, the posterior distribution $p(\Pi^*|\cdot)$ can be sampled from exactly without the need to resort to MCMC sampling. As mentioned in Section 3.4, Dalzell and Reiter (2018), in a follow-up work, suggest an important refinement that addresses the possibility of errors in the blocking variables and hence incorrect block partitioning.

4 | EXTENSIONS

4.1 | Multivariate linear regression with linked datasets

There is specific motivation to consider the "Broken Sample" problem in regression with *multiple* response variables. Intuitively, the hope is that multidimensional response will facilitate the identification of mismatches and permutation recovery. The idea is that for a single response variable, the discrepancy between expected response given the response and the observed must be quite substantial in order to be reliably detected, whereas for multiple responses, even small discrepancies might add up and are thus more easily to detect. This intuition is confirmed and made precise in Zhang et al. (2019), Pananjady et al. (2017), and Slawski, Ben-David, and Li (2020). While all the aforementioned practical approaches to (simplified versions) of the "Broken Sample problem" can potentially benefit considerably from the availability of multiple response variables, we will herein focus on a generalization of the penalized least squares approach (9) to showcase the central insights. The "Broken Sample" setting in the context of multivariate linear regression in matrix form can be written as

$$\mathbf{Y} = \Pi^*\mathbf{X}B^* + E\Sigma_*^{1/2}, \tag{17}$$

where $\mathbf{Y}$ is an n-by-m matrix whose rows are given by $\{\mathbf{y}_i^{\mathrm{T}}\}_{i=1}^n$, $B^* \in \mathbb{R}^{d\times m}$ is the matrix of regression coefficients, the rows of E are independent zero-mean, isotropic errors, and $\Sigma_* \in \mathbb{R}^{m\times m}$ is positive definite. To keep our exposition simple, we henceforth assume that $\Sigma = \sigma_*^2 I_m$, for $\sigma_*^2 > 0$. The approach (9) can be generalized seamlessly to the multivariate response setting. Defining $\Xi^* = (I_n - \Pi^*)\mathbf{X}B^*$, and re-writing (17) accordingly, prompts the penalized least squares problem

$$\min_{B\in\mathbb{R}^{d\times m},\Xi\in\mathbb{R}^{n\times m}} \|\mathbf{Y}-\mathbf{X}B-\Xi\|_{\mathrm{F}}^2+\lambda\sum_{i=1}^{n}\left\|\Xi^{\mathrm{T}}e_i\right\|_2, \tag{18}$$

where $\|\cdot\|_F$ denotes the Frobenius norm, and $\{e_i\}_{i=1}^n$ denotes the canonical basis of $\mathbb{R}^n$. The penalty in (18) is known as group lasso- or $\ell_1-\ell_2$ penalty in Yuan and Lin (2006). It promotes row-wise sparsity of Ξ rather than entry-wise sparsity by leveraging information across multiple columns (responses). It is shown in Slawski, Ben-David, and Li (2020) that (18) indeed improves over m separate applications of the formulation (9) for a single response variable. The improvements that can be achieved via multiple responses can be even more dramatic as far as mismatch recovery and permutation recovery based on the solution of (18) in a two-stage fashion (cf. Equation (12) are concerned): it is shown in Pananjady et al. (2017), Zhang et al. (2019), and Slawski, Ben-David, and Li (2020) that exact permutation recovery is achievable under considerably relaxed conditions on the SNR in comparison to the univariate response case in Pananjady et al. (2018) and Slawski and Ben-David (2019). This result hinges crucially on the so-called *stable rank* of the matrix B^*, a measure that quantifies the effective number of linearly independent columns in B^*, which makes sense since permutation recovery in the presence of noise cannot be expected to become easier if the different columns in $\mathbf{Y}$ represent redundant information.

4.2 | Generalized linear models with linked datasets

Most of the methodology discussed herein is also applicable to the broader class of generalized linear models (GLMs, McCullagh & Nelder, 1989) in which the distribution of the response belongs to an exponential family. The methods of Lahiri & Larsen (cf. Section 3.2) and Chambers (cf. Section 3.3) can be used by defining suitable unbiased estimating equations in alignment with the underlying GLM (Chambers, 2009; Chambers & da Silva, 2020; Han & Lahiri, 2019). The approaches for sparse and partial shuffling outlined in Section 3.6 can be modified: for the penalized least squares formulation (9), the $\{\xi_i\}_{i=1}^n$ can be absorbed into the linear predictor, and the least squares objective is replaced by the negative log-likelihood of the GLM under consideration in Wang et al. (2020); for the pseudo-likelihood method (11), the Gaussian density needs to be replaced by the corresponding density of the exponential family response. Finally, the approaches hinging on the missing data perspective in Section 3.7 can be adapted similarly: the EM algorithm can still be applied since the complete data likelihood associated with GLMs can be expressed as a linear function in Π^*.

4.3 | Other extensions

In this survey, we have focused on linear regression or closely related settings. A recent paper Shi et al. (2018) considers the "Broken Sample" setting for spherical regression, a specific instance of multivariate regression in which both the predictors and responses are related by a unitary transformation on the sphere in $\mathbb{R}^m$; this setting has interesting applications in natural language processing. Thus far, not much work has been done beyond (generalized) linear regression models. There is an evolving line of research on isotonic regression (Balabdoui et al., 2020; Carpentier & Schlüter, 2016; Rigollet & Weed, 2019), even though the setting in those references concerns the case of complete shuffling of predictors and responses, which, as mentioned before, is predominantly of theoretical interest rather than being applicable in typical record linkage applications. The scarcity of work on more flexible (semi- or nonparametric) regression models naturally suggests directions of future research, with the hope that some of the approaches discussed herein can be built on.

5 | REAL DATA ANALYSIS

We consider three benchmark datasets for model (1) given in Table 1. The regression model used in here could be found in Tancredi and Liseo (2015) for ISD and Slawski, Diao, and Ben-David (2020) for END and CPS. All of them contain matching variables which could be used to generate blocks of records that have agreements on matching variables. We consider the following three settings for real data analysis.

Sparse shuffling (estimation of β^*). For each dataset, we consider multiple independent random permutations for each value of $\alpha_* = k/n$. Note that we are not considering any given blocking information in the datasets. The random permutation is generated by shuffling α_* fraction of indices in the data. We assume that the fraction of mismatches α_* varies from 0.15 to 0.4 in steps of 0.05. The following approaches are compared.

naive, oracle. Plain least squares and estimation of β^* with knowledge of Π^*, respectively.
robust. β^* is estimated according to (9) based on the `robustfit` function in `MATLAB` with the argument `wfun` = 'huber' in light of the connection between (9) and Huber regression made in She and Owen (2012).
mixture. β^* is given by the pseudo-likelihood approach (10) with initial solution of β^* given by *robust* method.
EM. The permutation is initialized as the identity, and the number of MCMC iterations per EM iteration is set to 4000 for ISD, CPS and 80,000 for END given a "burn-in period" of half of MCMC iterations of their counterparts. The EM iteration is set to be 200.
Lahiri and Larsen (LL), Chambers (C). The approach in Sections 3.2 and 3.3 with $Q = \left(1 - \alpha_* - \frac{\alpha_*}{n-1}\right) I_n + \frac{\alpha_*}{n-1} \mathbf{1}_n \mathbf{1}_n^{\mathrm{T}}$.

For evaluation of the approaches across experimental configurations, we mainly look at the relative estimation error (REE) and R^2. REE is defined by $\left\| \hat{\beta}^{\text{est}} - \hat{\beta}^{\text{oracle}} \right\|_2 / \left\| \hat{\beta}^{\text{oracle}} \right\|_2$, where $\hat{\beta}^{\text{est}}$ denotes corresponding estimator given by specific approaches. Note that REE equals zero for $\hat{\beta}^{\text{est}} = \hat{\beta}^{\text{oracle}}$, thus REE can be interpreted as the excess error relative to the oracle estimator. The coefficient of determination R^2 is computed as $1 - \left\| \mathbf{Y}^* - \mathbf{X}\hat{\beta}^{\text{est}} \right\|_2^2 / \left\| \mathbf{Y}^* - \overline{\mathbf{Y}}^* \right\|_2^2$ where $\overline{\mathbf{Y}}^* = \sum_{i=1}^n y_i^*/n$. Higher values of R^2 correspond to better performance.

As shown in Figure 3, some of results are not sensitive to the fraction of mismatches k/n. For the ISD and END data, the *EM* and *mixture* approaches do not perform worse as the fraction of mismatches increasess. As discussed in Sections 3.1 and 3.6, the estimation errors of *naive* and *robust* are proportional to $\alpha_* = k/n$ which is confirmed by the study presented here.

Blocking (estimation of β^*). Depending on the application, matching variables or additional information about the linkage process are provided. In this case, the analyst will create blocks of records to reduce the burden for record linkage for the whole dataset. Mismatches or permutation of records will only occur within blocks of records defined by agreement on the matching variables, that is, mismatches will not involve pairs from different blocks. It is thus natural to consider the case in which Π^* is a "block permutation." Formally, let $\mathcal{P}(\mathcal{G})$ be the set of block permutations induced by the resulting index subsets $\mathcal{G} = \{G_b\}_{b=1}^B$ and $\cup_{b=1}^B G_b = \{1, \ldots, n\}$. Denote Π_b^* the permutation on the index subset G_b, $b = 1, \ldots, B$, and accordingly $\Pi^* = \mathrm{diag}\{\Pi_b^*\}_{b=1}^B$. The latter is generated by randomly permuting records within each block sharing the same quasi-identifier given in each datasets. For example, for the ISD data, we only consider permutations that permute indices within blocks defined by entities having the same value on the `Sex`, `Year of birth`, `Marital status`, `Employment status`, and `Working sector`. We consider 20 independent replications of Π^*. While metrics for evaluating the estimation of β^* is the same as the one given in *sparse shuffling*, the approaches discussed here differ from from those described under *sparse shuffling* in the sense that they are equipped with the knowledge about the underlying blocks.

EM. EM algorithm approach with MCMC sampler for (14) in block-wise fashion. Each block permutation is initialized as the identity, and the number of MCMC iterations per EM iteration and per block is set to 2000 given a "burn-in period" of 500. The number of EM iterations is set as 200.

DA. Data augmentation approach given in Section 3.7 with MCMC sampler for the *augmentation step* (15) in block-wise fashion. The regression parameter is estimated by the mean of posterior distribution of β^*. Each block permutation

TABLE 1 Overview of datasets considered in this section

Dataset	Abbreviation	n	d	R^2	σ_*^2	B	$\alpha_* = k/n$
Italian Survey data	ISD	2211	2	0.52	0.38	1301	0.41
El Nino data	END	93,935	5	0.91	0.26	5029	0.95
CPS wage data	CPS	534	11	0.70	0.04	465	0.12

Note: R^2 and σ_*^2 here refer to the case in the absence of mismatches. B denotes number of blocks. σ_* represents the average fraction of mismatches given random block permutation.

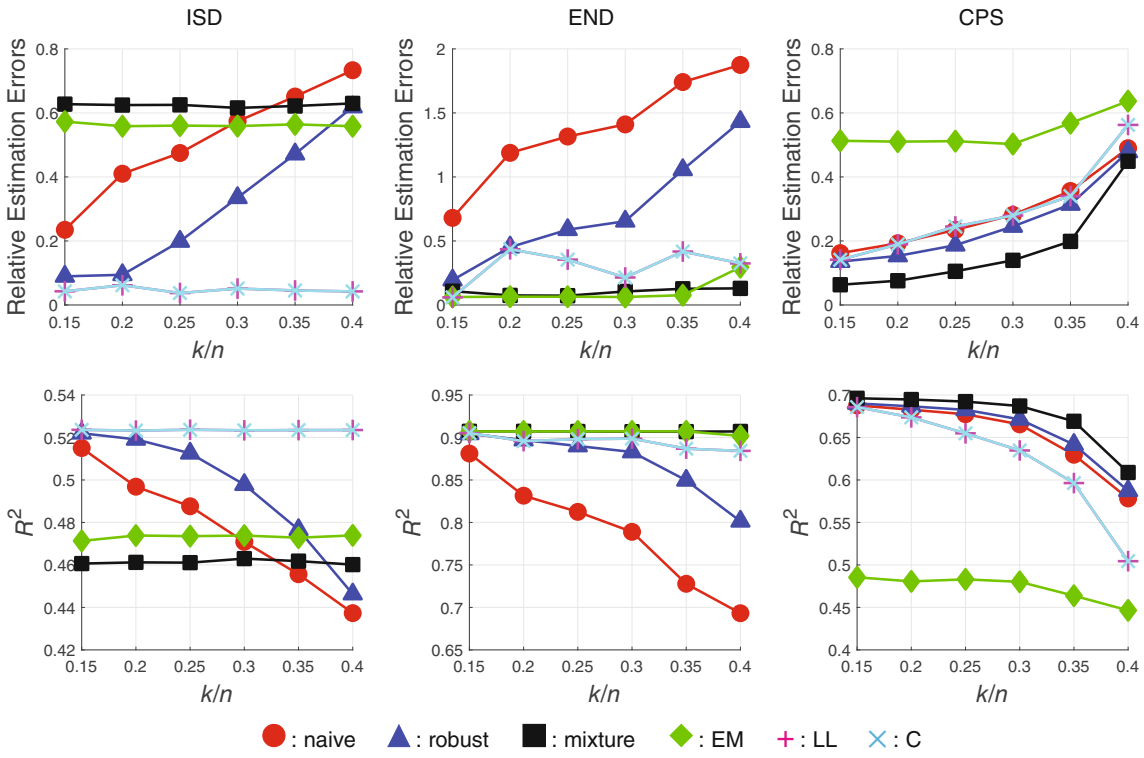


FIGURE 3 Comparison of approaches without knowledge of blocks. Each maker corresponds to different approaches given above. Each point in the graphs represents an average of the given metric over independent replications. The number of replications equals 100 for both ISD and CPS and 20 for END given the large size of the dataset

TABLE 2 Comparison of approaches with and without knowledge of blocks

		Methods without knowledge of blocking					Methods with knowledge of blocking			
Data	**Metric**	**Naive**	**Robust**	**Mixture**	**EM**	**DA**	**EM**	**DA**	**LL**	**C**
ISD	REE	0.193	**0.028**	0.538	0.566	0.260	**0.016**	0.029	0.062	0.101
	R^2	0.518	**0.522**	0.477	0.473	0.513	**0.524**	**0.524**	0.523	0.522
END	REE	0.985	0.800	0.278	0.761	**0.171**	0.433	0.272	**0.029**	0.303
	R^2	0.856	0.864	**0.890**	0.872	0.871	0.904	0.903	**0.906**	0.904
CPS	REE	0.124	0.106	**0.036**	0.516	0.323	0.023	0.036	**0.016**	0.057
	R^2	0.687	0.691	**0.698**	0.474	0.612	**0.699**	**0.699**	**0.699**	0.697

Note: The bold numbers are either the lowest REE or highest R^2 in each rows for corresponding estimator. Each cell is the average metric over 20 replications. REE is defined by $\left\|\hat{\beta}^{\text{est}}-\hat{\beta}^{\text{oracle}}\right\|_2 / \left\|\hat{\beta}^{\text{oracle}}\right\|_2$ and R^2 is defined by $1-\left\|\mathbf{Y}^*-\mathbf{X}\hat{\beta}^{\text{est}}\right\|_2^2 / \left\|\mathbf{Y}^*-\bar{\mathbf{Y}}^*\right\|_2^2$ where $\bar{\mathbf{Y}}^*=\sum_{i=1}^{n} y_i^*/n$.

is initialized as the identity, and the number of MCMC iterations per DA iteration is set to 1000. The overall number of iterations is set to 200, and the number of samples for *augmentation step M* is set to be 25.

Lahiri and Larsen (*LL*), *Chambers* (*C*). The approach in Sections 3.2 and 3.3 with $Q=\text{diag}\{Q_b\}_{b=1}^{B}$, $Q_b=\frac{1}{n_b}\mathbf{1}_{n_b}\mathbf{1}_{n_b}^{\mathrm{T}}$, and $\sum_{b=1}^{B} n_b = n$.

As can be seen from Table 2, the results are much different between with and without knowledge of blocks. The approaches equipped with knowledge of blocks yield significant improvement over those without knowledge of blocks as indicated by lower REE and higher R^2. Among the approaches without knowledge of blocks, *robust*, *DA*, and *mixture* perform the best. Among the methods with knowledge of blocks, *Lahiri and Larsen* (*LL*) is consistently better than the other approaches, which make sense since it yields an unbiased estimator of β^* given the matrix Q.

Blocking (estimation of Π^*). As discussed in Section 3.6 under "refinements", estimation of Π^* can be done in a two-stage fashion. After obtaining an estimator of β^*, we can estimate Π^* by solving a specific instance of a LAP. In this paragraph, we evaluate the performance of this approach. Formally, we consider the following optimization problem:

$$\min_{\Pi \in \mathcal{P}(\mathcal{G})} \left\| \Pi^{\mathrm{T}} \mathbf{Y} - \mathbf{X} \hat{\beta}^{\mathrm{est}} \right\|_2^2, \tag{19}$$

where we recall that $\mathcal{P}(\mathcal{G})$ is defined as the set of block permutations induced by the collection of index subsets $\mathcal{G}$. For the $\hat{\beta}^{\mathrm{est}}$ in (19), we use the estimator with knowledge of blocks achieving the smallest REE for each of the datasets according to Table 2. As shown in Figure 4, a substantial reduction in mismatch error is achieved, that is, $\hat{\Pi}\mathbf{Y}$ is visibly closer to $\mathbf{Y}^*$ than $\mathbf{Y}$ (top row vs. middle row). The refitting of the regression model with corrected response $\hat{\Pi}\mathbf{Y}$ indicates that the resulting fitted values align well with the fitted values obtained from a mismatch-free dataset.

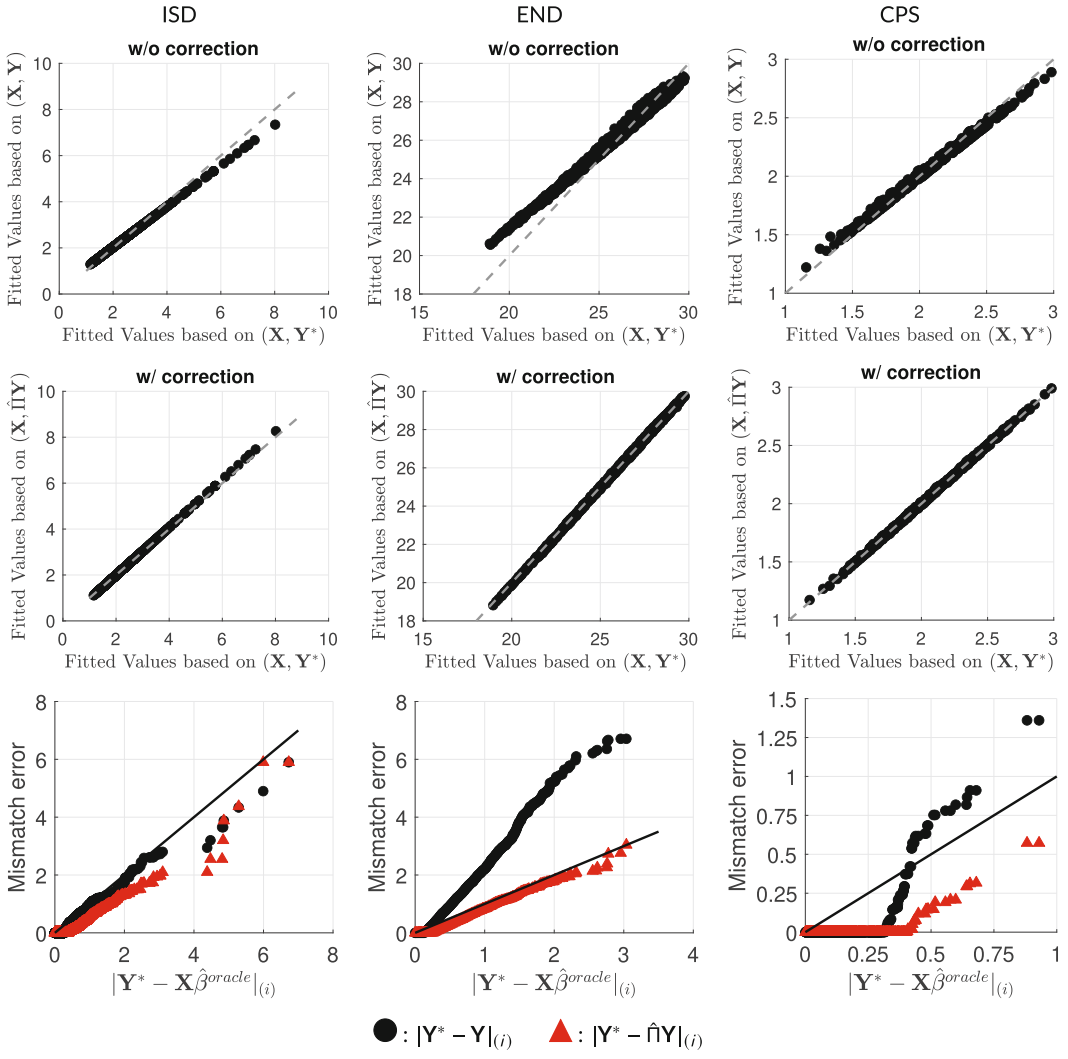


FIGURE 4 Comparison of two-stage approach for estimation of Π^* for corresponding data in Table 1. Top row: fitted values based on "mismatch-free" data $(\mathbf{X},\mathbf{Y}^*)$ vs. fitted values based on data with mismatches $(\mathbf{X},\mathbf{Y})$. Middle row: fitted values based on "mismatch-free" data $(\mathbf{X},\mathbf{Y}^*)$ vs. fitted values based on the corrected data $(\mathbf{X},\hat{\Pi}\mathbf{Y})$ with $\hat{\Pi}$ denoting the solution of optimization problem (19). Bottom row: *Q*–*Q* plots of the absolute differences between the true responses and their fitted values based on the *oracle* estimator $\left\{\left|y_i^* - \mathbf{x}_i^T \hat{\beta}^{\mathrm{oracle}}\right|\right\}_{i=1}^n$ vs. the absolute mismatch errors in the merged file $\left\{\left|y_i^* - y_i\right|\right\}_{i=1}^n$ (dots) and their counterparts $\left\{\left|y_i^* - y_{\hat{\pi}(i)}\right|\right\}_{i=1}^n$ after correction (triangles) based on (19). The idea underlying the bottom plots is that after correction, the remaining mismatch error is supposed to exhibit a similar distribution as the noise of the regression model, here approximated by the distribution of the residuals associated with the oracle estimator

6 | CLOSING REMARKS

In this survey article, we reviewed the impact of mismatch error on regression analysis when the predictor and response variables are originally contained in two separate files that are combined via record linkage. Several types of approaches that aim to mitigate the effect of potential errors in that process were discussed, with an emphasis on recent advances concerning the "Broken Sample" paradigm. In our opinion, the latter has an important benefit: by casting the problem explicitly in terms of an unknown permutation, the problem at hand is addressed more directly compared to approaches that primarily aim at unbiasedness or at least bias correction when averaging with respect to the randomness of the underlying permutation. In addition, recent computational and statistical advances have rendered the "Broken Sample" paradigm applicable in a much broader range of situations, and there is considerably more research under way, not only in statistics, but also in signal processing and machine learning.

Regarding the current state of research, it is worth pointing out that at this point, there is not a single approach that performs universally well independent of key parameters such as the fraction of mismatches, SNR, the amount of available information about the linkage process, and the goals of the data analyst. It is therefore important to keep those factors in mind when deciding on the type of mitigation method to apply.

Lastly, despite its importance and scope, the topic discussed herein can be seen as part of a significantly larger area evolving around *post-linkage data analysis*, that is, the analysis of linked files with an awareness for potential linkage errors. So far, corrective methods were mainly developed for regression, but needless to say, there is a whole array of multivariate data analysis techniques (PCA, cluster analysis, etc.) that are vulnerable to linkage error and thus might benefit from the use of suitable adjustment methods.

ACKNOWLEDGMENT

The first and the last author were partially supported by the NSF grant CCF-1849876.

CONFLICT OF INTEREST

The authors have declared no conflicts of interest for this article.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

AUTHOR CONTRIBUTIONS

Zhenbang Wang: Conceptualization (equal); data curation (equal); formal analysis (lead); investigation (equal); methodology (equal); resources (equal); software (equal); validation (equal); visualization (equal); writing – original draft (supporting); writing – review and editing (supporting). **Emanuel Ben-David:** Conceptualization (equal); investigation (supporting); methodology (supporting); project administration (equal); resources (supporting); supervision (supporting); writing – original draft (supporting); writing – review and editing (supporting). **Diao Guoqing:** Conceptualization (equal); methodology (equal); supervision (supporting); visualization (equal); writing – review and editing (supporting). **Martin Slawski:** Conceptualization (equal); data curation (equal); formal analysis (equal); investigation (equal); methodology (lead); supervision (lead); validation (supporting); visualization (supporting); writing – original draft (supporting); writing – review and editing (supporting).

ORCID

Emanuel Ben-David https://orcid.org/0000-0002-4230-4497

ENDNOTES

[1] This assumptions is mild since it is fulfilled as long as any row sub-matrix of $\mathbf{X}_N$ with n rows has rank at least d.

[2] With respect to the following ordering on the positive semidefinite cone: $A \geq B$ if $A - B$ is positive semidefinite.

[3] With probability tending to one as $n \to \infty$.

RELATED WIREs ARTICLES

Matching and record linkage

Secondary analysis of linked data

Statistical Analysis with Linked Data
Statistical Inference, Learning and Models in Big Data

REFERENCES

Abid, A., Poon, A. & Zou, J. (2017). Linear regression with shuffled labels. arXiv:1705.01342.

Abid, A., & Zou, J. (2018). Stochastic EM for shuffled linear regression. In *Allerton conference on communication, control, and computing* (pp. 470–477).

Abowd, J., Abramowitz, J., Levenstein, M., McCue, K., Patki, D., Raghunathan, T., Rodgers, A. M., Shapiro, M., & Wasi, N. (2019) *Optimal probabilistic record linkage: Best practice for linking employers in survey and administrative data* [Working papers]. U.S. Census Bureau, Center for Economic Studies. Retrieved from https://EconPapers.repec.org/RePEc:cen:wpaper:19-08

Antoniadis, A. (2007). Wavelet methods in statistics: Some recent developments and their applications. *Statistics Surveys*, *1*, 16–55.

Balabdoui, F., Doss, C., & Durot, C. (2020). *Unlinked monotone regression*. arXiv:2007.00830.

Bertsekas, D., & Castanon, D. (1992). A forward/reverse auction algorihtm for asymmetric assignment problems. *Computational Optimization and Applications*, *1*, 277–297.

Binette, O., & Steorts, R. (2020). *(Almost) All of entity resolution*. arXiv preprint arXiv:2008.04443.

Brizan, D. G., & Tansel, A. (2006). A survey of entity resolution and record linkage methodologies. *Communications of the IIMA*, *6*, 41–50.

Burkard, R., Dell'Amico, M., & Martello, S. (2009). *Assignment problems: Revised reprint*. SIAM.

Carpentier, A., & Schlüter, T. (2016). Learning relationships between data obtained independently. In *Proceedings of the international conference on artifical intelligence and statistics (AISTATS)* (Vol. 51, pp. 658–666). JMLR.org.

Chambers, R. (2009). *Regression analysis of probability-linked data*. Tech. rep., Statistics New Zealand.

Chambers, R., & da Silva, A. D. (2020). Improved secondary analysis of linked data: A framework and an illustration. *Journal of the Royal Statistical Society Series A*, *183*, 37–59.

Christen, P. (2012). *Data matching: Concepts and techniques for record linkage, entity resolution, and duplicate detection*. Springer.

Dalzell, N., & Reiter, J. (2018). Regression modeling and file matching using possibly erroneous matching variables. *Journal of Computational and Graphical Statistics*, *27*, 728–738.

DeGroot, M., Feder, P., & Goel, P. (1971). Matchmaking. *The Annals of Mathematical Statistics*, *42*, 578–593.

DeGroot, M., & Goel, P. (1976). The matching problem for multivariate normal data. *Sankhya, Series B*, *38*, 14–29.

DeGroot, M., & Goel, P. (1980). Estimation of the correlation coefficient from a broken random sample. *The Annals of Statistics*, *8*, 264–278.

Di Consiglio, L., & Tuoto, T. (2018). When adjusting for the bias due to linkage errors: A sensitivity analysis. *Statistical Journal of the IAOS*, *34*, 589–597.

Diaconis, P. (1988). Group representations in probability and statistics. In *Lecture notes—Monograph series* (Vol. 11). Institute of Mathematical Statistic.

Enamorado, T., Fifield, B., & Imai, K. (2019). Using a probabilistic model to assist merging of large-scale administrative records. *American Political Science Review*, *113*, 353–371.

Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, *64*, 1183–1210.

Goel, P. (1975). On re-pairing observations in a broken sample. *The Annals of Statistics*, *3*, 1364–1369.

Goel, P., & Ramalingam, T. (1987). Some properties of the maximum likelihood strategy for re-pairing a broken random sample. *Journal of Statistical Planning and Inference*, *16*, 237–248.

Gutman, R., Afendulis, C., & Zaslavsky, A. (2013). A Bayesian procedure for file linking to analyze end-of-life medical costs. *Journal of the American Statistical Association*, *108*, 34–47.

Haghighatshoar, S., & Caire, G. (2017). Signal recovery from unlabeled samples. In *International symposium on information theory (ISIT)* (pp. 451–455). IEEE.

Han, Y., & Lahiri, P. (2019). Statistical analysis with linked data. *International Statistical Review*, *87*, 139–157.

Herzog, T., Scheuren, F., & Winkler, W. (2007). *Data quality and record linkage techniques*. Springer.

Hof, M. H. P., & Zwinderman, A. H. (2012). Methods for analyzing data from probabilistic linkage strategies based on partially identifying variables. *Statistics in Medicine*, *31*, 4231–4242.

Hsu, D., Shi, K., & Sun, X. (2017). Linear regression without correspondence. In *Advances in neural information processing systems (NIPS)* (pp. 1531–1540). nips.cc.

Huber, P. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, *53*, 73–101.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E., Spicer, K., & Wolf, P.-P. D. (2012). *Statistical disclosure control*. John Wiley & Sons.

Kim, G., & Chambers, R. (2012). Regression analysis under incomplete linkage. *Computational Statistics and Data Analysis*, *56*, 2756–2770.

Kuhn, H. (1955). The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, *2*, 83–97.

Lahiri, P., & Larsen, M. D. (2005). Regression analysis with linked data. *Journal of the American Statistical Association*, *100*, 222–230.

Lindsay, B. (1988). Composite likelihood methods. *Contemporary Mathematics*, *80*, 221–239.

Little, R., & Rubin, D. (1987). *Statistical analysis with missing data*. John Wiley & Sons.

Loh, P. (2018). *Scale calibration for high-dimensional robust regression*. arXiv:1811.02906.

Massey, C. G., Genadek, K. R., Alexander, J. T., Gardner, T. K., & O'Hara, A. (2018). Linking the 1940 U.S. census with modern data. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, *51*, 246–257.

McCullagh, P., & Nelder, J. (1989). *Generalized Linear Models*. Chapman and Hall.

Neter, J., Maynes, S., & Ramanathan, R. (1965). The effect of mismatching on the measurement of response error. *Journal of the American Statistical Association*, *60*, 1005–1027.

Pananjady, A., Wainwright, M., & Cortade, T. (2017). *Denoising linear models with permuted data*. arXiv:1704.07461.

Pananjady, A., Wainwright, M., & Cortade, T. (2018). Linear regression with shuffled data: Statistical and computational limits of permutation recovery. *IEEE Transactions on Information Theory*, *64*, 3826–3300.

Rentsch, C., Harron, K., Urassa, M., Todd, J., Reniers, G., & Zaba, B. (2018). Impact of linkage quality on inferences drawn from analyses using data with high rates of linkage errors in rural Tanzania. *BMC Medical Research Methodology*, *18*, 1–9.

Rigollet, P., & Weed, J. (2019). Uncoupled isotonic regression via minimum Wasserstein deconvolution. *Information & Inference*. arXiv: 1806.10648.

Scheuren, F., & Winkler, W. (1993). Regression analysis of data files that are computer matched I. *Survey Methodology*, *19*, 39–58.

Scheuren, F., & Winkler, W. (1997). Regression analysis of data files that are computer matched II. *Survey Methodology*, *23*, 157–165.

She, Y., & Owen, A. (2012). Outlier detection using nonconvex penalized regression. *Journal of the American Statistical Association*, *106*, 626–639.

Shi, X., Lu, X., & Cai, T. (2018). *Spherical regresion under mismatch corruption with application to automated knowledge translation*. arXiv: 1810.05679.

Slawski, M., & Ben-David, E. (2019). Linear regression with sparsely permuted data. *Electronic Journal of Statistics*, *13*(1). https://doi.org/10.1214/18-ejs1498

Slawski, M., Ben-David, E., & Li, P. (2020). A two-stage approach to multivariate linear regression with sparsely mismatched data. *Journal of Machine Learning Research*, *21*, 1–42.

Slawski, M., Diao, G. & Ben-David, E. (2020). A pseudo-likelihood approach to linear regression with partially shuffled data. Journal of Computational and Graphical Statistics. arXiv:1910.01623.

Slawski, M., Rahmani, M., & Li, P. (2019). A sparse representation-based approach to linear regression with partially shuffled labels. In *Proceedings of the thirty-fifth conference on uncertainty in artificial intelligence (UAI)*.

Steorts, R. C., Tancredi, A., & Liseo, B. (2018). Generalized Bayesian record linkage and regression with exact error propagation. In *International conference on privacy in statistical databases* (pp. 279–313). Springer.

Steorts, R. C., Ventura, S. L., Sadinle, M., & Fienberg, S. E. (2014). A comparison of blocking methods for record linkage. In J. Domingo-Ferrer (Ed.), *Privacy in statistical databases* (pp. 253–268). Springer International Publishing.

Sweeney, L. (2001). *Computational disclosure control: A primer on data privacy protection* [Ph.D. thesis]. Massachusetts Institute of Technology.

Tancredi, A., & Liseo, B. (2015). Regression analysis with linked data: Problems and possible solutions. *Statistica*, *75*, 19–35.

Tanner, M., & Wong, W. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, *82*, 528–540.

Tibshirani, R. (1996). Regression shrinkage and variable selection via the lasso. *Journal of the Royal Statistical Society Series B*, *58*, 671–686.

Tsakiris, M. (2018). *Eigenspace conditions for homomorphic sensing*. arXiv:1812.07966.

Varin, C., Reid, N., & Firth, D. (2011). An overview of composite likelihood estimation. *Statistica Sinica*, *21*, 5–42.

Wang, Z., Ben-David, E., & Slawski, M. (2020). *Estimation in exponential family regression based on linked data contaminated by mismatch error*. arXiv:2010.00181.

Winkler, W. (1995). Matching and record linkage. *Business Survey Methods*, *1*, 355–384.

Winkler, W. (2006). *Overview of record linkage and current research directions*. Tech. rep., Statistical Research Division U.S. Census Bureau.

Winkler, W. E. (2014). Matching and record linkage. *WIREs Computational Statistics*, *6*, 313–325.

Wu, Y. N. (1998). *A note on broken sample problem*, Tech. rep., Department of Statistics, University of Michigan.

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B*, *68*, 49–67.

Zhang, H., Slawski, M., & Li, P. (2019). Permutation recovery from multiple measurement vectors in unlabeled sensing. In *IEEE international symposium on information theory (ISIT)*. IEEE.

Zhang, L.-C., & Tuoto, T. (2020). Linkage-data linear regression. *Journal of the Royal Statistical Society Series A*, *184*, 522–547.

How to cite this article: Wang, Z., Ben-David, E., Diao, G., & Slawski, M. (2022). Regression with linked datasets subject to linkage error. *Wiley Interdisciplinary Reviews: Computational Statistics*, *14*(4), e1570. https://doi.org/10.1002/wics.1570

APPENDIX A.: DERIVATION OF EQUATION (2) AND EQUATION (3)

Consider Equation (2). We have

$$\begin{aligned}
\mathbf{E}\left[\left\|\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}-\beta^{*}\right\|_2^2\right] &\overset{\text{(i)}}{=} \mathbf{E}\left[\left\|\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\Pi^{*}\left(\mathbf{X}_N\beta^{*}+\sigma_{*}\boldsymbol{\varepsilon}\right)-\beta^{*}\right\|_2^2\right] \\
&\overset{\text{(ii)}}{=} \mathbf{E}\left[\left\|\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\Pi^{*}\mathbf{X}_N\beta^{*}-\beta^{*}\right\|_2^2\right]+\sigma_{*}^2\mathbf{E}\left[\left\|\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\Pi^{*}\boldsymbol{\varepsilon}\right\|_2^2\right] \\
&\overset{\text{(iii)}}{=} \mathbf{E}\left[\left\|\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\Pi^{*}\mathbf{X}_N\beta^{*}-\beta^{*}\right\|_2^2\right]+\sigma_{*}^2\mathbf{E}\left[\mathrm{tr}\left(\Pi^{*}\varepsilon\left(\Pi^{*}\varepsilon\right)^{\mathrm{T}}\mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-2}\mathbf{X}^{\mathrm{T}}\right)\right] \\
&\overset{\text{(iv)}}{=} \mathbf{E}\left[\left\|\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\Pi^{*}\mathbf{X}_N\beta^{*}-\beta^{*}\right\|_2^2\right]+\sigma_{*}^2\mathrm{tr}\left(\mathbf{E}\left[\Pi^{*}\varepsilon\left(\Pi^{*}\varepsilon\right)^{\mathrm{T}}\right]\mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-2}\mathbf{X}^{\mathrm{T}}\right) \\
&\overset{\text{(v)}}{=} \mathbf{E}\left[\left\|\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\Pi^{*}\mathbf{X}_N\beta^{*}-\beta^{*}\right\|_2^2\right]+\sigma_{*}^2\mathrm{tr}\left(\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\right) \\
&= \mathbf{E}\left[\left\|\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\Pi^{*}\mathbf{X}_N\beta^{*}-\beta^{*}\right\|_2^2\right]+\sigma_{*}^2\frac{\mathrm{tr}\left(\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}/n\right)^{-1}\right)}{n},
\end{aligned}$$

where in (i), we substitute the model equation (1); in (ii), we expand the square and use the fact that the resulting cross-term equals to zero since $\boldsymbol{\varepsilon}$ is zero-mean and is independent of $\mathbf{X}$ and Π^{*}; in (iii), we expand the quadratic form in terms of a trace; in (iv), we use linearity of expectation; in (v), we use the fact that $\Pi^{*}\boldsymbol{\varepsilon}$ has i.i.d. unit variance entries, and the shuffling property of the trace.

Regarding Equation (3), we have that

$$\begin{aligned}
\mathbf{E}\left[\left\|\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\Pi^{*}\mathbf{X}_N\beta^{*}-\beta^{*}\right\|_2^2\right] &= \mathbf{E}\left[\left\|\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\Pi^{*}\mathbf{X}_N\beta^{*}-\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{X}\beta^{*}\right\|_2^2\right] \\
&\overset{\text{(i)}}{\leq} \left\|\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\right\|_2^2\mathbf{E}\left[\left\|\Pi^{*}\mathbf{X}_N\beta^{*}-\mathbf{X}\beta^{*}\right\|_2^2\right] \\
&\overset{\text{(ii)}}{\leq} \lambda_{\min}^{-1}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}/n\right)\frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\left[\left\{\left(\mathbf{x}_{\pi^{*}(i)}-\mathbf{x}_i\right)^{\mathrm{T}}\beta^{*}\right\}^2\right] \\
&\overset{\text{(iii)}}{\leq} \lambda_{\min}^{-1}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}/n\right)\frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\left[\left\|\mathbf{x}_{\pi^{*}(i)}-\mathbf{x}_i\right\|_2^2\right]\left\|\beta^{*}\right\|_2^2 \\
&\leq \lambda_{\min}^{-1}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}/n\right)\frac{k}{n}\left\|\beta^{*}\right\|_2^2D_{\mathbf{X}_N}^2, \quad D_{\mathbf{X}_N} \coloneqq \max_{1\leq i,j\leq N}\left\|\mathbf{x}_i-\mathbf{x}_j\right\|_2.
\end{aligned}$$

In (i), we use that $\|\mathbf{A}\mathbf{v}\|_2^2 \leq \|\mathbf{A}\|_2^2\|\mathbf{v}\|_2^2$ for a matrix $\mathbf{A}$ and a vector $\mathbf{v}$, where $\|\mathbf{A}\|_2$ denotes the spectral norm of $\mathbf{A}$; in (ii), we use that $\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}/n\right)^{-1}\mathbf{X}^{\mathrm{T}}/\sqrt{n}$ is the pseudo-inverse of $\mathbf{X}/\sqrt{n}$; in (iii), we use the Cauchy-Schwarz inequality. The last inequality follows from the fact that the number of mismatches $\left|\left\{1\leq i\leq n:\pi^{*}(i)\neq i\right\}\right|$ is bounded by k.