Journal of the Royal Statistical Society Series A: Statistics in Society, 2025, **188**, 896–919
https://doi.org/10.1093/jrsssa/qnae083
Advance access publication 26 August 2024
Original Article

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series A
Statistics in Society

A general framework for regression with mismatched data based on mixture modelling

Martin Slawski[1], Brady T. West[2], Priyanjali Bukke[1], Zhenbang Wang[1], Guoqing Diao[3] and Emanuel Ben-David[4]

[1]Department of Statistics, George Mason University, 4400 University Drive, Fairfax, VA 22030, USA
[2]Institute for Social Research, University of Michigan-Ann Arbor, 426 Thompson Street, Ann Arbor, MI 48106, USA
[3]Department of Biostatistics and Bioinformatics, George Washington University, 800 22nd Street, NW, Washington, DC 20052, USA
[4]Center for Statistical Research and Methodology, U.S. Census Bureau, 4600 Silver Hill Rd, Suitland, MD 20746, USA

Address for correspondence: Martin Slawski, Department of Statistics, George Mason University, 4400 University Drive, Fairfax, VA 22030, USA. Email: mslawsk3@gmu.edu

Abstract

The advent of the information age has revolutionized data collection and has led to a rapid expansion of available data sources. Methods of data integration are indispensable when a question of interest cannot be addressed using a single data source. Record linkage (RL) is at the forefront of such data integration efforts. Incentives for sharing linked data for secondary analysis have prompted the need for methodology accounting for possible errors at the RL stage. Mismatch error is a common consequence resulting from the use of nonunique or noisy identifiers at that stage. In this paper, we present a framework to enable valid postlinkage inference in the secondary analysis setting in which only the linked file is given. The proposed framework covers a variety of statistical models and can flexibly incorporate information about the underlying RL process. We propose a mixture model for linked records whose two components reflect distributions conditional on match status, i.e. correct or false match. Regarding inference, we develop a method based on composite likelihood and the expectation-maximization algorithm that is implemented in the R package `pldamixture`. Extensive simulations and case studies involving contemporary RL applications corroborate the effectiveness of our framework.

Keywords: composite likelihood, EM algorithm, mismatch error, mixture model, record linkage, secondary analysis

1 Introduction

The digital revolution has not only changed the sheer volume of data that is being generated but has also substantially impacted the way data are collected, disseminated, and analysed. As part of this ongoing development, there are increasing efforts to synthesize complementary pieces of information residing in multiple data sources that were gathered in isolation. Such 'data siloes' are commonly discussed as a barrier to leveraging the full potential inherent in the available data (e.g. Japec et al., 2015). Methods of data integration have thus become a crucial component in many contemporary data analysis pipelines. Record linkage (RL) (e.g. Binette & Steorts, 2022; Christen, 2012; Newcombe & Kennedy, 1962) combines individual records contained in multiple files and thus constitutes the most granular method of data integration (Lohr & Raghunathan, 2017). Record linkage comes with the promise of creating richer data sets from existing ones at virtually no extra cost, and has seen widespread use across many domains of applications.

Received: August 4, 2023. **Revised:** May 9, 2024. **Accepted:** July 23, 2024

Examples include the linkage of surveys and administrative records, insurance claims and hospital records, birth and death registers, historical censuses, etc.

The continuing importance of RL has catalysed efforts to improve the quality of linking procedures in the absence of unique identifiers that—if available—would render the identification of matching records an easy task. The use of quasi-identifiers such as names and addresses can be error-prone, possibly because of variations in spelling, mix-ups between middle names and double/hyphenated last names, change of residence over time, or data entry errors. Increasing awareness of data privacy has further limited the information available to RL procedures. As a result, the process can be highly ambiguous, with one record yielding many candidate matches in another file. Probabilistic RL techniques, e.g. those based on the Fellegi–Sunter method (Fellegi & Sunter, 1969), address such uncertainty systematically by assigning a matching score to each pair of records. Regardless of the increasing sophistication of the methods employed, by no means can it be guaranteed that linked files are free of errors. Mainstream implementations of probabilistic RL can be sensitive to the choice of the threshold for the matching score at which a pair is deemed a match. Proper choice of this threshold strikes a suitable balance between false matches (henceforth *mismatches*) and false nonmatches (*missed matches*). Both types of errors can negatively affect downstream statistical analyses (*postlinkage analysis*) performed on the linked file. As elaborated below, how to account for such errors and how to suitably propagate uncertainty from RL to subsequent analysis stages in a seamless fashion has remained a significant challenge.

While missed matches can induce sample selection bias similar to nonresponse in survey data (Little & Rubin, 2019), mismatches can cause data contamination and typically attenuated relationships when analysing associations, e.g. in regression analysis. This is a well-studied problem dating back to Neter et al. (1965), and important follow-up work was conducted by Scheuren and Winkler (1993, 1997) and Lahiri and Larsen (2005). Subsequently, a variety of approaches have been proposed to account for mismatches in postlinkage data analysis. This work can be roughly divided according to whether it addresses *primary analysis* or *secondary analysis*. The former refers to scenarios in which record linkage and downstream analysis are performed by the same individual, or the analyst at least has significant insights into the details of the underlying RL. In this situation, it is possible to directly propagate the uncertainty from RL; examples include Han and Lahiri (2019) and Hof and Zwinderman (2015); and hierarchical Bayes methods (e.g. Dalzell & Reiter, 2018; Gutman et al., 2013; Steorts et al., 2018; Tancredi & Liseo, 2015). By contrast, in the secondary analysis setting, the analyst only has access to the linked file rather than the individual files, and has limited knowledge about how RL was performed. For instance, the analyst may be given scores reflecting the likelihood of every linked record being a correct match, as in the recent study by Abowd et al. (2021), or indicators of the blocks within which linkage was performed as well as the mismatch rate within each block. A line of research pioneered by Chambers (2009) hinges on this information, typically in conjunction with the assumption of exchangeable linkage error in each block. Notable follow-up work along this line includes Chambers and da Silva (2020); Kim and Chambers (2012); Zhang and Tuoto (2021). We also refer to recent surveys of this literature (Chambers et al., 2023; Wang et al., 2022) and the references therein.

Driven by tendencies towards 'Open Science' and elevated requirements for reproducing scientific analyses, enabling data access and disseminating research results beyond a specific research group that has collected and prepared the underlying data, the significance of the secondary analysis setting is expected to grow considerably in the future. Since the primary data owners will often find it infeasible to share individual data sources, e.g. because of the aforementioned privacy considerations or the mere complexity of the RL task that would be too burdensome to replicate for prospective secondary data users, it has become a common practice to simply provide the final linked files. Secondary analysts, however, frequently analyse the data provided without any specific regard for possible linkage error arising during their creation. In the sequel, we provide three illustrative examples representing contemporary secondary analysis scenarios involving linked data.

(I) *Linkage of the Health and Retirement Study and Administrative Data*. The Health and Retirement Study (HRS, see hrs.isr.umich.edu) is the largest and most comprehensive nationally representative multidisciplinary panel study of Americans over the age of 50. The primary goal of the HRS is to explore the challenges and opportunities of aging and the associated transitions in lifestyle. Administrative Data such as Centers for Medicare and Medicaid Services (CMS) claims

data or Census data can be used to complement or corroborate survey responses. By linking information from the survey to such administrative data, researchers can obtain a better picture of the extent of inaccurate responses or self-assessment. In this context, it is common for RL and subsequent analysis to be performed by different teams. For instance, extracting the relevant information from CMS claims data requires considerable expertize with specific software systems designed primarily for internal use. The team performing RL may provide an overall assessment of the reliability of the generated links (e.g. the fraction of links deemed unreliable) or pairwise confidence scores for each candidate link that can be in the form of posterior match probabilities derived from a Fellegi–Sunter-type model or a classifier trained on clerically reviewed links (Abowd et al., 2019). Various types of statistical models could be employed at the analysis stage. In Abowd et al. (2021), a regression model is fitted to analyse the relationship between wages (from the HRS) and establishment size (from the Census Business register). In the case study presented in Section 7.2 below, a two-way contingency table analysis is performed to study the association between self-reported nursing home residence from the HRS and the presence of corresponding billing records in the CMS data.

(II) *Linkage of Historical Censuses*. The digitization of historical documents enables researchers to obtain a better quantitative understanding of past events and socio-economic evolutions. For example, the Longitudinal Intergenerational Family Electronic Micro-Database (LIFE-M) Project (Bailey et al., 2022, see also life-m.org) has created a public-use database obtained from linking historical decennial censuses and vital records covering millions of records from the late 19th and 20th centuries in the states of Ohio and North Carolina. The main purpose of the project is to study questions related to intergenerational mobility. The creators of LIFE-M acknowledge that a fraction of entries in this database might involve incorrect links of records in the underlying data sources. They report a suspected mismatch rate, quality indicators pertaining to the mode of linkage for each records (clerically reviewed vs. automatically linked), and name commonness scores indicative of the likelihood of a record being a correct match. Regression analysis is frequently used to study substantive questions, e.g. regression of the son's income on the father's income and other covariates (Bailey et al., 2020). Another example concerning longevity analysis is presented in Section 7.1 below.

(III) *Linkage of Social Media Data*. There is a growing interest in linking social media activity to a primary data source, such as survey data (Mneimeh, 2022; Stier et al., 2020). In this context, researchers have studied how social media activity aligns with behavioural patterns, political views, social status, etc. For instance, Liu et al. (2021) use predictive modelling to infer individual demographics, party affiliation, gun ownership, and other attributes (as provided in a corresponding survey) from tweets and Twitter biographies. In this setting, mismatch errors may arise when the names provided by the survey respondents do not match with the names used for social media accounts, the full names provided do not uniquely identify individuals, when social media platform handles associated with user accounts are provided with typos that prevent exact matching, or when platform handles change over time (Beuthner et al., 2021; Stier et al., 2020).

Contributions. In this paper, we develop a framework to account for mismatch errors in post-linkage analysis in the type of secondary analysis settings as illustrated above. This framework is *general* in the sense that (i) it covers various types of statistical analysis and models including regression modelling, curve fitting, covariance estimation, and contingency table analysis under one umbrella, and (ii) it can incorporate information of varying degrees about the preceding RL process that has generated the linked file to be analysed. As referenced in the above examples, such information may include estimates of mismatch rates, block indicators associated with the blocking variables that were used, and variables indicative of the correctness of linked records (e.g. clerically reviewed yes/no, scores from a probabilistic RL procedure or surrogates thereof). If no such information is available, the proposed approach will attempt to estimate the underlying mismatch rate from the data. This task can be accomplished under suitable conditions that include correct model specification.

In a nutshell, this framework relies on a two-component mixture model that ties together a model for the linked variables of interest and a model for a latent binary indicator of match status (correctly or incorrectly matched) for each record in the linked file. Estimation is based on composite likelihood (Lindsay, 1988; Varin et al., 2011), which provides a path towards valid (asymptotic) inference; we also sketch how the proposed method can be cast in a Bayesian framework. The

proposed approach extends prior work (Slawski et al., 2021) motivated by 'shuffled data problems' (DeGroot & Goel, 1980; Pananjady et al., 2018; Slawski & Ben-David, 2019) in multiple directions. In brief, the paper by Slawski et al. (2021) is limited to classical linear regression and a constant mismatch rate. The approach presented herein bears a close connection to the method in Hof and Zwinderman (2015). The main distinction is that the latter method is developed for the primary analysis setting and involves a pairwise composite likelihood, which renders the approach less scalable. Apart from that, we employ additional assumptions; while these assumptions may be considered strong, they render inference much more tractable. An R package (Bukke et al., 2024) implementing our approach based on a formula interface enabling straightforward model specification and inference is available on the Comprehensive R Archive Network (CRAN).

Organization. Formal descriptions of the setup, our approach, and its assumptions are provided in Section 2. We then outline the framework for inference in Section 3. Specific examples of interest are discussed in Section 4. Additional technical details and extensions are presented in Section 5. Simulation studies and real data analysis are presented in Sections 6 and 7, respectively. We conclude with a summary of the main findings and discuss directions for future work in Section 8.

Notation. We use the following conventions regarding probability density functions (PDFs): instead of writing $f_{\mathbf{x}}(\mathbf{x}_0)$ for the density of a random vector $\mathbf{x}$ evaluated at a point $\mathbf{x}_0$, we drop the symbol in the subscript and simply write $f(\mathbf{x}_0)$ with the convention that the corresponding random variable is inferred from the symbol in the argument. Similar conventions are adopted for joint and conditional PDFs, i.e. we use $f(\mathbf{a}_0, \ldots, \mathbf{z}_0)$ instead of $f_{\mathbf{a}\ldots\mathbf{z}}(\mathbf{a}_0, \ldots, \mathbf{z}_0)$ and $f(\mathbf{x}_0 \,|\, \mathbf{y}_0)$ instead of $f_{\mathbf{x}\,|\,\mathbf{y}=\mathbf{y}_0}(\mathbf{x}_0)$, etc. Subscripts in f will be present in case there is no argument. By default, symbols will be boldfaced to indicate vector-valued quantities, with the understanding that boldfaced quantities may also represent scalars as special case; occasionally, normal instead of bold font is used to highlight a scalar quantity. The dependence of PDFs on parameters is expressed via $f(\cdot; \ldots)$, where … represents a list of parameters. A table summarizing frequently used symbols is given below.

$\mathbb{I}(\cdot)$	indicator function	$\mathbf{u} \perp\!\!\!\perp \mathbf{v}$	random variables $\mathbf{u}$ and $\mathbf{v}$ are independent
m	mismatch indicator	$\phi(\mathbf{y} \,|\, \mathbf{x})$	conditional PDF of $\mathbf{y}$ given $\mathbf{x}$ (regression setup)
$\mathbf{P}(\ldots)$	probability	$\boldsymbol{\theta}$	parameter describing the $(\mathbf{x}, \mathbf{y})$-relationship
$\mathbf{E}[\ldots]$	expectation	$\mathbf{z}$	covariates informative of mismatch indicator
$[\ldots]^{(t)}$	iteration counter	$h(\mathbf{z})$	$\mathbf{P}(m = 0 \,|\, \mathbf{z})$
$\mathrm{logit}(x)$	$\log(x/(1-x))$	$\boldsymbol{\gamma}$	parameter associated with h
		$\boldsymbol{\theta}^*, \boldsymbol{\gamma}^*$ etc.	'ground truth' parameter values

2 Methods

The goal of record linkage is to merge two individual files $F^\star_{\mathbf{x}} = \{\mathbf{x}^\star_j\}_{j=1}^M$ and $F^\star_{\mathbf{y}} = \{\mathbf{y}^\star_k\}_{k=1}^N$ into a new file $F^\star_{\mathbf{x}\bowtie\mathbf{y}} = \{(\mathbf{x}^\star_{\ell_i}, \mathbf{y}^\star_{\ell_i})\}_{i=1}^v$ of pairs corresponding to identical statistical units. For simplicity, we assume that every $\mathbf{y}^\star_i$, $1 \le i \le N$, has one and only one match in $F^\star_{\mathbf{x}}$ (and hence $M \ge N = v$). We also assume that the missing links in the larger file $F^\star_{\mathbf{x}}$ are ignorable.[1] Data linkage is assumed to produce an imperfectly combined file $F_{\mathbf{x}\bowtie\mathbf{y}} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$ with $\mathbf{x}_i \in F^\star_{\mathbf{x}}$ and $\mathbf{y}_i \in F^\star_{\mathbf{y}}$, $1 \le i \le n \le N$, containing *mismatched pairs* $(\mathbf{x}_i, \mathbf{y}_i) \notin F^\star_{\mathbf{x}\bowtie\mathbf{y}}$ and lacking correct matches $\mathbf{F}^\star_{\mathbf{x}\bowtie\mathbf{y}} \setminus \mathbf{F}_{\mathbf{x}\bowtie\mathbf{y}}$ (*missed matches*). Throughout this paper, we focus on mismatches and assume that missed matches are ignorable.

With each linked pair in $\mathbf{F}_{\mathbf{x}\bowtie\mathbf{y}}$, we may additionally observe variables $\mathbf{z}_i$ pertaining to the confidence in the correctness of the link, $1 \le i \le n$, which yields triplets $\{(\mathbf{x}_i, \mathbf{y}_i, \mathbf{z}_i)\}_{i=1}^n$. Accordingly, we define latent mismatch indicators $m_i = \mathbb{I}((\mathbf{x}_i, \mathbf{y}_i) \notin \mathbf{F}^\star_{\mathbf{x}\bowtie\mathbf{y}})$, $1 \le i \le n$.

[1] Missing at random in regression settings with the $\mathbf{y}$ variable as the response, missing completely at random in unsupervised settings. See Section 4.3 for a definition of 'unsupervised settings'.

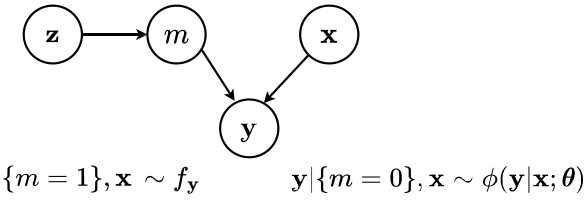


Figure 1. Directed acyclic graph representation of the model associated with the likelihood (3) for regression settings. Note that the covariates for modelling the latent match indicator and the covariates for modelling the response variable are assumed to be independent.

Assumptions **(A1)** The $\{(m_i, \mathbf{z}_i)\}_{i=1}^n$ are independent of both $\{\mathbf{x}_i\}_{i=1}^n$ and $\{\mathbf{y}_i\}_{i=1}^n$.

(A2) The following two-component mixture model is assumed for each pair $(\mathbf{x}_i, \mathbf{y}_i)$:

$$(\mathbf{x}_i, \mathbf{y}_i) \,|\, \mathbf{z}_i, \{m_i = 0\} \sim \phi_i(\cdot; \boldsymbol{\theta}^*), \quad \text{(IND): } \mathbf{y}_i \perp\!\!\!\perp \mathbf{x}_i \,|\, \mathbf{z}_i, \{m_i = 1\}, \tag{1}$$

where the $\phi_i(\cdot; \boldsymbol{\theta}^*)$ are PDFs depending on an unknown parameter of interest $\boldsymbol{\theta}^*$ but neither on m_i nor $\mathbf{z}_i$, $1 \le i \le n$. The second item in (1) will be referred to via the abbreviation (IND) in the sequel.

(A3) $\mathbf{P}(m_i = 0 \,|\, \mathbf{z}_i) = h(\mathbf{z}_i; \boldsymbol{\gamma}^*)$ for some known function h and unknown parameter $\boldsymbol{\gamma}^*$ (of secondary interest), $1 \le i \le n$.

Given **(A1)**–**(A3)**, the likelihood in the parameters $(\boldsymbol{\theta}, \boldsymbol{\gamma})$ of a single triplet $(\mathbf{x}_i, \mathbf{y}_i, \mathbf{z}_i)$ can be shown to be of the form (cf. Section A in the online supplementary material for a derivation)

$$L_i(\boldsymbol{\theta}, \boldsymbol{\gamma}) \propto f(\mathbf{x}_i; \boldsymbol{\theta}) \times f(\mathbf{y}_i; \boldsymbol{\theta}) \times \left\{1 - h(\mathbf{z}_i; \boldsymbol{\gamma})\right\} + \phi_i(\mathbf{x}_i, \mathbf{y}_i; \boldsymbol{\theta}) \times h(\mathbf{z}_i; \boldsymbol{\gamma}), \tag{2}$$

where $\propto$ here means equality up to multiplicative constants not involving $\boldsymbol{\theta}$ or $\boldsymbol{\gamma}$.

In a regression setup, $\phi_i(\cdot; \boldsymbol{\theta})$ depends on $\boldsymbol{\theta}$ only via the conditional PDF of the response variable given covariates, here denoted by $\phi_i(\cdot \,|\, \cdot; \boldsymbol{\theta})$, $1 \le i \le n$. Moreover, $f(\mathbf{x}_i; \boldsymbol{\theta}) = f(\mathbf{x}_i)$, $1 \le i \le n$, typically does not depend on the regression parameter. The likelihood (2) accordingly can be decomposed as

$$\begin{aligned} L_i(\boldsymbol{\theta}, \boldsymbol{\gamma}) &= f(\mathbf{x}_i; \boldsymbol{\theta}) \times f(\mathbf{y}_i; \boldsymbol{\theta}) \times \left\{1 - h(\mathbf{z}_i; \boldsymbol{\gamma})\right\} + \phi_i(\mathbf{y}_i \,|\, \mathbf{x}_i; \boldsymbol{\theta}) \times f(\mathbf{x}_i; \boldsymbol{\theta}) \times h(\mathbf{z}_i; \boldsymbol{\gamma}) \\ &\propto f(\mathbf{y}_i; \boldsymbol{\theta}) \times \left\{1 - h(\mathbf{z}_i; \boldsymbol{\gamma})\right\} + \phi_i(\mathbf{y}_i \,|\, \mathbf{x}_i; \boldsymbol{\theta}) \times h(\mathbf{z}_i; \boldsymbol{\gamma}), \quad 1 \le i \le n, \end{aligned} \tag{3}$$

The corresponding representation as a directed acyclic graph (DAG) is shown in Figure 1. Observe that as a consequence of **(A1)**, the match indicator only depends on the $\{\mathbf{z}_i\}_{i=1}^n$ but not on the covariates $\{\mathbf{x}_i\}_{i=1}^n$.

We here briefly note that while the marginal densities $f(\mathbf{x})$ and $f(\mathbf{y})$ are typically not known, their estimation is not affected by mismatch error and is thus straightforward; we refer to Section 5.1 for more details.*(Composite) likelihood*. Multiplication of the individual terms $L_i(\boldsymbol{\theta}, \boldsymbol{\gamma})$, $1 \le i \le n$, yields the likelihood

$$L(\boldsymbol{\theta}, \boldsymbol{\gamma}) = \prod_{i=1}^n L_i(\boldsymbol{\theta}, \boldsymbol{\gamma}), \tag{4}$$

to be maximized with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$. We note that the product over the observation-specific L_i's (referred to as marginal likelihoods) in general may not qualify as a 'proper' (joint) likelihood of the set of pairs $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$. Dependencies among pairs may arise from the mismatch indicators $\{m_i\}_{i=1}^n$ since (in)correct linkage of one pair may affect (in)correct linkage of another pair, specifically if one-to-one matchings are enforced. For example, consider the situation of two records

$\{a, b\}$ in the first file with correct matches $\{a', b'\}$ in the second file. Incorrectly matching a to b' will cause b to be mismatched as well since the requirement of a one-to-match entails that (i) at least one match has to be found for b and (ii) b cannot match to b' since none of the records can appear in more than one linked pair. In light of these considerations, (4) is more adequately referred to as a *composite likelihood*, which is a general term used for likelihoods constructed from factorizing terms that individually represent proper likelihoods, whereas their product may not. Regardless, maximizers of composite likelihoods can be studied within the framework of M-estimation and enjoy properties similar to (proper) maximum likelihood estimators such as $\sqrt{n}$-consistency and asymptotic normality with an asymptotic covariance matrix that exhibits the familiar sandwich form (e.g. Lindsay, 1988; Varin et al., 2011); cf. Section 3.2 below.

We also note that the above composite likelihood can be extended to accommodate survey data arising from a complex probability sample in finite population settings, following Binder (1983). The factors in (4) can be raised to powers of survey weights $\{w_i\}_{i=1}^n$ available for each survey respondent to form a consistent estimator of the population pseudo-likelihood. While we do not consider applications involving survey weights in the remainder of this paper, future applications could certainly consider this possibility.

Secondary analysis setting. We here emphasize that the proposed approach is motivated by secondary analysis in which no additional information beyond the imperfectly linked file $F_{\mathbf{x}\bowtie\mathbf{y}}$ may be available. In particular, none of the two individual files $F_{\mathbf{x}}^\star$, $F_{\mathbf{y}}^\star$ may be given. Additional information from the linkage process such as match probabilities can be incorporated in terms of the variables $\{\mathbf{z}_i\}_{i=1}^n$ in a model for the mismatch indicators (cf. Section 4.4 below). Note that the $\{\mathbf{z}_i\}_{i=1}^n$ are allowed to be empty, in which case the $\{m_i\}_{i=1}^n$ are treated as identically distributed Bernoulli random variables.

A related approach addressing the *primary analysis* setting (in which linkage and subsequent data analysis are considered in an integrated fashion) is developed in Hof and Zwinderman (2015). Their formulation is based on a *pairwise* composite likelihood over all pairs $F_{\mathbf{x}}^\star \times F_{\mathbf{y}}^\star$ and associated comparison vectors $\{\mathbf{c}_{jk}\}$.

3 Inference

In the following, we describe the main ingredients of our inferential framework, with specific details and extensions postponed to Section 5. Selected examples are reviewed in Section 4.

3.1 EM algorithm

Direct maximization of the composite likelihood tends to be challenging. Treating the mismatch indicators $\{m_i\}_{i=1}^n$ as missing data naturally prompts the use of the EM algorithm. The resulting E-step involves simple closed-form updates akin to those in conventional mixture models (cf. Section B in the online supplementary material for their derivations), and the M-step updates for $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ decouple into separate optimization problems. Moreover, the update for $\boldsymbol{\theta}$ typically reduces to an optimization problem that would be encountered in the *absence of mismatches* with additional observation weights. As a result, existing software can be used as long as these weights can be incorporated. The general template is presented below, assuming for now that $f_{\mathbf{x}}$ and $f_{\mathbf{y}}$ are known; we refer to Section 5.1 for details on this aspect.

The complete data (composite)likelihood is given by

$$\begin{aligned} L^{\mathrm{c}}(\boldsymbol{\theta}, \boldsymbol{\gamma}) &= \prod_{i=1}^n f(\mathbf{x}_i, \mathbf{y}_i, \mathbf{z}_i, m_i; \boldsymbol{\theta}, \boldsymbol{\gamma}) \\ &\propto \prod_{i=1}^n f(\mathbf{x}_i, \mathbf{y}_i \mid \mathbf{z}_i, m_i; \boldsymbol{\theta}, \boldsymbol{\gamma}) f(m_i \mid \mathbf{z}_i; \boldsymbol{\gamma}) \\ &= \prod_{i=1}^n \Big\{ [f(\mathbf{x}_i) \times f(\mathbf{y}_i) \times (1 - h(\mathbf{z}_i; \boldsymbol{\gamma}))]^{m_i} \times [\phi_i(\mathbf{x}_i, \mathbf{y}_i; \boldsymbol{\theta}) \times h(\mathbf{z}_i; \boldsymbol{\gamma})]^{1-m_i} \Big\}, \end{aligned}$$

where for the last line we use the assumption that f_{x} and f_{y} known. Taking logarithms, the complete data negative (composite) log-likelihood is given by (modulo additive constants):

$$\begin{aligned}\ell^{\mathrm{c}}(\boldsymbol{\theta}, \boldsymbol{\gamma}) = &- \sum_{i=1}^{n} \big\{ m_i \log\left(1 - h(\mathbf{z}_i; \boldsymbol{\gamma})\right) + (1 - m_i) \log\left(h(\mathbf{z}_i; \boldsymbol{\gamma})\right) \big\} - \\ &- \sum_{i=1}^{n} (1 - m_i) \log\left(\phi_i(\mathbf{x}_i, \mathbf{y}_i; \boldsymbol{\theta})\right)\end{aligned} \tag{5}$$

E-step. In the E-step, we evaluate $\widehat{m}_i^{(t)} = \mathbf{P}(m_i = 1 \mid (\mathbf{x}_i, \mathbf{y}_i, \mathbf{z}_i); \boldsymbol{\theta}^{(t)}, \boldsymbol{\gamma}^{(t)})$, $1 \leq i \leq n$, given the observed data $\{(\mathbf{x}_i, \mathbf{y}_i, \mathbf{z}_i)\}_{i=1}^n$ and current iterates $(\boldsymbol{\theta}^{(t)}, \boldsymbol{\gamma}^{(t)})$ for the parameters.

M-step. The expected complete data negative (composite) log-likelihood then results as

$$\begin{aligned}\ell^{(t)}(\boldsymbol{\theta}, \boldsymbol{\gamma}) = &- \sum_{i=1}^{n} \Big\{ \widehat{m}_i^{(t)} \log\left(1 - h(\mathbf{z}_i; \boldsymbol{\gamma})\right) + (1 - \widehat{m}_i^{(t)}) \log\left(h(\mathbf{z}_i; \boldsymbol{\gamma})\right) \Big\} - \\ &- \sum_{i=1}^{n} (1 - \widehat{m}_i^{(t)}) \log\left(\phi_i(\mathbf{x}_i, \mathbf{y}_i; \boldsymbol{\theta})\right)\end{aligned} \tag{6}$$

Note that minimization over $\boldsymbol{\gamma}$ involves only the first term, which is seen to be the log-likelihood of a binary regression model with 'responses' $\{\widehat{m}_i^{(t)}\}_{i=1}^n$, covariates $\{\mathbf{z}_i\}_{i=1}^n$, and link function h. Minimization over $\boldsymbol{\theta}$ involves the log-likelihood encountered in the absence of mismatches with additional observation-specific weights.

3.2 Standard errors

For fully parametric models, asymptotic standard errors can be obtained from well-known properties of composite maximum likelihood estimators. Specifically, letting $(\widehat{\boldsymbol{\theta}}_n, \widehat{\boldsymbol{\gamma}}_n)$ denote the maximizer of the composite likelihood (4) and $(\boldsymbol{\theta}^*, \boldsymbol{\gamma}^*)$ the corresponding population parameters, we have (under suitable regularity conditions) that

$$\sqrt{n}\left(\begin{pmatrix}\widehat{\boldsymbol{\theta}}_n \\ \widehat{\boldsymbol{\gamma}}_n\end{pmatrix}\right) \to N\left(\begin{pmatrix}\boldsymbol{\theta}^* \\ \boldsymbol{\gamma}^*\end{pmatrix}, \mathbf{E}[\nabla^2 \ell(\boldsymbol{\theta}^*, \boldsymbol{\gamma}^*)]^{-1} \mathbf{E}[\nabla \ell(\boldsymbol{\theta}^*, \boldsymbol{\gamma}^*) \nabla \ell(\boldsymbol{\theta}^*, \boldsymbol{\gamma}^*)^\top] \mathbf{E}[\nabla^2 \ell(\boldsymbol{\theta}^*, \boldsymbol{\gamma}^*)]^{-1}\right),$$

in distribution, where $\nabla \ell$ and $\nabla^2 \ell$ denote the gradient and Hessian of $\ell = -\log L$. Moreover, the above covariance can be estimated consistently by substituting the expectations with their empirical counterparts.

4 Specific examples

In this section, we work out the specifics of the general template in the previous section for several popular regression setups. We also present applications to covariance estimation and contingency table analysis. Modelling of the latent mismatch indicators is discussed in a dedicated subsection.

4.1 Generalized linear models

We start by considering linear regression with Gaussian errors, reproducing results in earlier work (Slawski et al., 2021, Section 3). In this case, we have

$$-\log \phi_i(y_i, \mathbf{x}_i; \boldsymbol{\theta}) = -\log \phi(y_i \mid \mathbf{x}_i; \boldsymbol{\theta}) = \frac{1}{2} \log(\sigma^2) + \frac{1}{2\sigma^2}(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}), \quad 1 \leq i \leq n,$$

with $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$. As a result, the M-step for $\boldsymbol{\theta}$ based on (6) reduces to the following:

$$\widehat{\boldsymbol{\beta}}^{(t+1)} \leftarrow \operatorname*{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{n} \{(1 - \widehat{m}_i^{(t)})(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2\}, \quad \widehat{\sigma}^{2\,(t+1)} \leftarrow \frac{\sum_{i=1}^n (1 - \widehat{m}_i^{(t)})(y_i - \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}^{(t+1)})}{\sum_{i=1}^n (1 - \widehat{m}_i^{(t)})}.$$

We note that unless stated otherwise, the intercept is included in the $\{\mathbf{x}_i\}_{i=1}^n$.

An extension to the class of generalized linear regression models (GLMs, McCullagh & Nelder, 1989) is obtained via the specification

$$-\log \phi_i(\mathbf{x}_i, y_i; \boldsymbol{\theta}) = -\log \phi(y_i \mid \mathbf{x}_i; \boldsymbol{\theta}) = \frac{\psi(\vartheta(\mathbf{x}_i^\top \boldsymbol{\beta})) - y_i \vartheta(\mathbf{x}_i^\top \boldsymbol{\beta})}{\sigma} + c(y_i, \sigma), \quad 1 \le i \le n,$$

for a link function ϑ, cumulant ψ, scale parameter σ, and partition function c. It is customary to use the canonical link in which case ϑ equals the identity map. Popular examples include (i) logistic regression with $\psi(\cdot) = \log(1 + \exp(\cdot))$ and $\sigma = 1$, and (ii) Poisson regression with $\psi(\cdot) = \exp(\cdot)$ and $\sigma = 1$. A popular example with a noncanonical link is (iii) Gamma regression with log-link with $\vartheta = -\exp(-\cdot)$, $\psi(\cdot) = -\log(-\cdot)$, and $c(y, \sigma) = \frac{\sigma - 1}{\sigma}\log(y) + \frac{\log(\sigma)}{\sigma} + \log(\Gamma(1/\sigma))$.

In all three cases, the M-step for $\boldsymbol{\theta}$ based on (5) is performed by first obtaining $\widehat{\boldsymbol{\beta}}^{(t+1)}$ via a (regular) GLM fit with data $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ and observation weights $\{1 - \widehat{m}_i^{(t)}\}_{i=1}^n$, and then (if necessary) updating the scale parameter σ by minimizing the M-step objective (5) over σ with $\boldsymbol{\beta}$ fixed to $\boldsymbol{\beta}^{(t+1)}$. The latter is a *one-dimensional* optimization problem and hence easy to solve via appropriate routines.

4.2 Cox proportional hazards regression

For the (semiparametric) Cox proportional hazards (PH) model, the response variable is given by a right-censored survival time. Accordingly, the data set is of the form $\{((y_i, \delta_i), \mathbf{x}_i)\}_{i=1}^n$, where $\delta_i = 1$ if y_i is observed without right-censoring and $\delta_i = 0$ otherwise, $1 \le i \le n$. The Cox PH model postulates that

$$-\log \phi(y_i, \delta_i \mid \mathbf{x}_i, ; \boldsymbol{\theta}) = -\delta_i \log \lambda(y_i \mid \mathbf{x}_i; \boldsymbol{\theta}) + \Lambda(y_i \mid \mathbf{x}_i; \boldsymbol{\theta}), \quad 1 \le i \le n,$$

with $\lambda(y_i \mid \mathbf{x}_i; \boldsymbol{\theta}) = \lambda_0(y_i) \cdot \exp(\mathbf{x}_i^\top \boldsymbol{\beta})$ and $\Lambda(y_i \mid \mathbf{x}_i; \boldsymbol{\theta}) = \exp(\mathbf{x}_i^\top \boldsymbol{\beta})\Lambda_0(y_i)$, $1 \le i \le n$, where λ and Λ denote the (conditional) hazard and cumulative hazard functions, respectively, depending on baseline hazard and cumulative hazard functions λ_0 and Λ_0, respectively. Here, $\boldsymbol{\theta} = (\boldsymbol{\beta}, \lambda_0)$ contains the nuisance parameter λ_0. The M-step for $\boldsymbol{\theta}$ is given by

$$\min_{\boldsymbol{\beta}, \lambda_0} \left\{ -\sum_{i=1}^n (1 - \widehat{m}_i^{(t)}) \left\{ \delta_i \left[\log(\lambda_0(y_i)) + \mathbf{x}_i^\top \boldsymbol{\beta} \right] + \exp(\mathbf{x}_i^\top \boldsymbol{\beta}) \Lambda_0(y_i) \right\} \right\},$$

where the term inside the curly brackets equals the (full) negative log-likelihood of the Cox model with observation weights $\{(1 - \widehat{m}_i^{(t)})\}_{i=1}^n$. In a nutshell, the M-step can be performed based on an accordingly weighted partial negative log-likelihood in $\boldsymbol{\beta}$ and a weighted Breslow estimator for updating the baseline hazard; cf. Section D in the online supplementary material for details.

4.3 Unsupervised problems

To illustrate the unsupervised setting, we consider (i) estimation of the covariance matrix of a multivariate normal random vector $(\mathbf{x}^\top \ \mathbf{y}^\top)^\top$, and (ii) parameter estimation for a two-way contingency table for categorical variables. Here, the term 'unsupervised' refers to the fact that the roles of $\mathbf{x}$ and $\mathbf{y}$ are symmetric in the sense that there is no distinction between predictor and response variables.

(i) **Multivariate normal data**. The parameter is given $\boldsymbol{\theta} = \boldsymbol{\Sigma}$, structured according to blocks $\boldsymbol{\Sigma}_{\mathbf{xx}}$, $\boldsymbol{\Sigma}_{\mathbf{xy}}$, $\boldsymbol{\Sigma}_{\mathbf{yy}}$ (and $\boldsymbol{\Sigma}_{\mathbf{yx}} = \boldsymbol{\Sigma}_{\mathbf{xy}}^\top$), with $\boldsymbol{\Sigma}_{\mathbf{xx}} = \mathrm{Cov}(\mathbf{x})$, $\boldsymbol{\Sigma}_{\mathbf{yy}} = \mathrm{Cov}(\mathbf{y})$, and $\boldsymbol{\Sigma}_{\mathbf{xy}} = \mathrm{Cov}(\mathbf{x}, \mathbf{y})$. For simplicity, we assume that $\mathbf{E}[\mathbf{x}]$ and $\mathbf{E}[\mathbf{y}]$ are both zero; in fact, the estimation of these quantities is not affected by mismatch error in the linked file $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$, nor is the estimation of the (marginal) covariances $\boldsymbol{\Sigma}_{\mathbf{xx}}$ and $\boldsymbol{\Sigma}_{\mathbf{yy}}$. We here slightly depart from the principle according to which the marginals $f_{\mathbf{x}}$ and $f_{\mathbf{y}}$ are

considered fixed (known or substituted by a plug-in estimator), and instead jointly estimate $\mathbf{f}\Sigma_{\mathbf{xx}}$, $\mathbf{f}\Sigma_{\mathbf{yy}}$, and $\mathbf{f}\Sigma_{\mathbf{xy}}$ in a way that is computationally most convenient. Specifically, noting that

$$f(\mathbf{x}_i;\boldsymbol{\theta})\times f(\mathbf{y}_i;\boldsymbol{\theta})\propto|\boldsymbol{\Gamma}|^{-1/2}\exp\left(-\frac{1}{2}\begin{pmatrix}\mathbf{x}_i\\ \mathbf{y}_i\end{pmatrix}^{\top}\boldsymbol{\Gamma}^{-1}\begin{pmatrix}\mathbf{x}_i\\ \mathbf{y}_i\end{pmatrix}\right),\quad \boldsymbol{\Gamma}:=\begin{pmatrix}\boldsymbol{\Sigma}_{\mathbf{xx}} & 0\\ 0 & \boldsymbol{\Sigma}_{\mathbf{yy}}\end{pmatrix},\quad 1\le i\le n,$$

we can consider a modification of the objective in the M-step (6) by not dropping the terms depending on the marginals $f_{\mathbf{x}}$ and $f_{\mathbf{y}}$. The resulting modified expected complete data negative (composite) log-likelihood then takes the form

$$\begin{aligned}&-\sum_{i=1}^{n}\left\{(1-\widehat{m}_i^{(t)})\log\left(\phi_i(\mathbf{x}_i,\mathbf{y}_i;\boldsymbol{\theta})\right)+\widehat{m}_i^{(t)}\cdot f(\mathbf{x}_i;\boldsymbol{\theta})\cdot f(\mathbf{y}_i;\boldsymbol{\theta})\right\}\\ &\quad\propto\left\{-\log|\boldsymbol{\Omega}|\left(\sum_{i=1}^{n}(1-\widehat{m}_i^{(t)})\right)+\operatorname{tr}(\boldsymbol{\Omega}\mathbf{S}^{(t)})-\log|\boldsymbol{\Psi}|\left(\sum_{i=1}^{n}\widehat{m}_i^{(t)}\right)+\operatorname{tr}(\boldsymbol{\Psi}\mathbf{S}_{\mathrm{ind}}^{(t)})\right\},\end{aligned}\tag{7}$$

$$\boldsymbol{\Omega}=\boldsymbol{\Sigma}^{-1},\quad \boldsymbol{\Psi}=\boldsymbol{\Gamma}^{-1},\quad \mathbf{S}^{(t)}=\sum_{i=1}^{n}(1-\widehat{m}_i^{(t)})\begin{pmatrix}\mathbf{x}_i\\ \mathbf{y}_i\end{pmatrix}\begin{pmatrix}\mathbf{x}_i\\ \mathbf{y}_i\end{pmatrix}^{\top},\quad \mathbf{S}_{\mathrm{ind}}^{(t)}=\sum_{i=1}^{n}\widehat{m}_i^{(t)}\begin{pmatrix}\mathbf{x}_i\mathbf{x}_i^{\top} & 0\\ 0 & \mathbf{y}_i\mathbf{y}_i^{\top}\end{pmatrix}.$$

Minimization with respect to $\boldsymbol{\Omega}$ and $\boldsymbol{\Psi}$ yields the following closed-form updates:

$$\boldsymbol{\Omega}^{(t+1)}=\left(\mathbf{S}^{(t)}\Big/\left(\sum_{i=1}^{n}(1-\widehat{m}_i^{(t)})\right)\right)^{-1},\quad \boldsymbol{\Psi}^{(t+1)}=\left(\mathbf{S}_{\mathrm{ind}}^{(t)}\Big/\left(\sum_{i=1}^{n}\widehat{m}_i^{(t)}\right)\right)^{-1}.$$

(ii) Two-way contingency tables. Consider two categorical random variables $\mathbf{x}$ and $\mathbf{y}$ taking values in categories numbered $\{1,\dots,K\}$ and $\{1,\dots,L\}$, respectively. Let $\theta_{kl}=\mathbf{P}(\mathbf{x}=k,\mathbf{y}=l)$, $1\le k\le K$, $1\le l\le L$, denote the corresponding joint probabilities, and accordingly let $\boldsymbol{\theta}=(\theta_{kl})_{k,l}$. Given a linked file $\{(x_i,y_i)\}_{i=1}^{n}$ whose correctly matched pairs are distributed as $(\mathbf{x},\mathbf{y})$, we note that the independence assumption in (1) implies that for mismatched pairs the resulting contribution to the likelihood is given by $f(x_i;\boldsymbol{\theta})\times f(y_i;\boldsymbol{\theta})=\theta_{x_i+}\cdot\theta_{+y_i}$, where the subscript + indicates summation over the corresponding index. As a notable difference from models discussed above, we note that the parameter $\boldsymbol{\gamma}$ of the model $h(\cdot;\boldsymbol{\gamma})$ for the mismatch indicators can no longer be inferred from the data. In fact, consider the case in which $h(\cdot;\gamma)=1-\gamma$, $\gamma\in(0,1)$, is a constant: it is easy to see that the resulting composite likelihood (4) is always maximized by setting $\gamma=0$ since the parameters (θ_{kl}) correspond to a saturated model achieving perfect fit regardless of the specific $\{(x_i,y_i)\}_{i=1}^{n}$. This issue can be addressed by fixing γ. Similar to the approach taken for the multivariate Gaussian model in (7), we propose to work with two separate sets of parameters representing a saturated and an independence model, respectively, and to drop the associated (linear) constraints that would couple these two sets of parameters. Specifically, the expected complete data negative (composite) log-likelihood takes the form

$$\begin{aligned}&-\sum_{i=1}^{n}\left\{(1-\widehat{m}_i^{(t)})\log\left(\phi_i(x_i,y_i;\boldsymbol{\theta})\right)+\widehat{m}_i^{(t)}\cdot f(x_i;\boldsymbol{\theta})\times f(y_i;\boldsymbol{\theta})\right\}\\ &\quad=-\sum_{k,l}(1-\widehat{m}_{kl}^{(t)})\log(\pi_{kl})+\sum_{k,l}\widehat{m}_{kl}^{(t)}\log(\psi_{k+}\cdot\psi_{+l}),\end{aligned}$$

where we note that the $\{\widehat{m}_i\}_{i=1}^{n}$ are constant across observations falling into the same cell (k,l) of the associated contingency table, $1\le k\le K$, $1\le l\le L$. In the above display, $\psi_{k+}=\sum_l\pi_{kl}$, $1\le k\le K$, and $\psi_{+l}=\sum_k\pi_{kl}$, $1\le l\le L$, but for computational simplicity this constraint is dropped when performing the minimization with respect to $\{\pi_{kl}\}$, $\{\psi_{k+}\}$, and $\{\psi_{+l}\}$. This minimization amounts to fitting separate saturated and independence models to reweighted samples with (effective) sample sizes of $\sum_{k,l}(1-\widehat{m}_{kl}^{(t)})$ and $\sum_{k,l}\widehat{m}_{kl}^{(t)}$, respectively, and can be implemented via weighted

Poisson regressions in light of connections between log-linear models and Poisson regression (Agresti, 2012).

4.4 Modelling the latent mismatch indicator

In the preceding sections, we have elaborated on the specifics of various models concerning the relationship between $\mathbf{x}$ and $\mathbf{y}$. The second major aspect of modelling concerns the latent mismatch indicators. Since these are binary, the use of a logistic regression model can be considered the standard choice, i.e. in the context of (2) and (3)

$$\mathbf{P}(m_i = 0 \mid \mathbf{z}_i) = h(\mathbf{z}_i; \boldsymbol{\gamma}) = \frac{\exp(\mathbf{z}_i^\top \boldsymbol{\gamma})}{1 + \exp(\mathbf{z}_i^\top \boldsymbol{\gamma})}, \quad 1 \leq i \leq n.$$

Note that the above specifies a marginal model and does not entail independence of the $\{m_i\}_{i=1}^n$. If no auxiliary covariates $\{\mathbf{z}_i\}_{i=1}^n$ informative of the match status are available, an intercept-only model can be employed which is equivalent to assuming a constant mismatch rate (regardless of the choice of the link function). It is worth stressing that despite the similarities in modelling, estimation of the parameters is more challenging than in (plain) binary regression since the $\{m_i\}_{i=1}^n$ are not observed. To facilitate parameter estimation, it can be helpful to integrate prior knowledge about the underlying mismatch rate by imposing a linear constraint on the average linear predictor of the form $(\frac{1}{n}\sum_{i=1}^n \mathbf{z}_i)^\top(-\boldsymbol{\gamma}) \leq b$, where $b \in \mathbb{R}$ corresponds to the logit of the assumed mismatch rate. Such a constraint can be incorporated in a straightforward manner within the approach to inference presented in Section 3.

5 Miscellaneous details and extensions

This section complements Sections 3 and 4, filling in additional details on the estimation of marginal PDFs and outlining an extension to a Bayesian setup.

5.1 Estimation of marginal PDFs

In Section 3, the marginal densities $f_{\mathbf{x}}$ and $f_{\mathbf{y}}$ were treated as known quantities. In practice, this is not the case even though the estimation is considered less of a challenge given that mismatch error affects the estimation (of parameters) of the joint distribution but not of the marginals. We distinguish between two approaches: (i) *plug-in estimation* and (ii) *integrated estimation*. In the first approach, the marginal PDFs are estimated beforehand and substituted in place of the corresponding population quantities; in the second approach, the marginal PDFs are updated along with the parameter $\boldsymbol{\theta}$ of primary interest. (i) The plug-in approach reduces to plain density estimation of $f_{\mathbf{y}}$ (and also of $f_{\mathbf{x}}$ outside regression setups), and various methods ranging from fully nonparametric to parametric are available to perform this task. Particular examples include kernel density estimation or the use of empirical probability mass functions if the range of the associated random variable is discrete and small in size. Note that while in the plug-in approach, the marginal PDFs are not updated during the EM iterations, they enter in the E-step as well as in the evaluation of the composite likelihood at the iterates $\{(\widehat{\boldsymbol{\theta}}^{(t)}, \widehat{\boldsymbol{\gamma}}^{(t)})\}_{t\geq 1}$. (ii) In the integrated approach, $f_{\mathbf{x}}$ and $f_{\mathbf{y}}$ are updated with $\boldsymbol{\theta}$. If correctly paired observations are i.i.d. with joint PDF $f_{\mathbf{x},\mathbf{y}}(\cdot, \cdot; \boldsymbol{\theta})$, the relationships $f_{\mathbf{x}}(\cdot; \boldsymbol{\theta}) = \int f_{\mathbf{x},\mathbf{y}}(\cdot, \mathbf{y}; \boldsymbol{\theta})\, \mathrm{d}\mathbf{y}$ and $f_{\mathbf{y}}(\cdot) = \int f_{\mathbf{x},\mathbf{y}}(\mathbf{x}, \cdot; \boldsymbol{\theta})\, \mathrm{d}\mathbf{x}$ prompt updates along with $\boldsymbol{\theta}$. The possibility of such updates typically arises in unsupervised settings, and results in additional constraints on $\boldsymbol{\theta}$ that may not be easy to implement(cf. Section 4.3). In standard fixed design regression setups, $f_{\mathbf{y}}$ can be expressed as the finite mixture

$f_{\mathbf{y}}(\cdot; \boldsymbol{\theta}) = \int \phi(\cdot \mid \mathbf{x}; \boldsymbol{\theta})\, \mathrm{d}P(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n \phi(\cdot \mid \mathbf{x}_i; \boldsymbol{\theta})$, where P denotes the atomic measure with atoms $\{\mathbf{x}_i\}_{i=1}^n$ each having mass $1/n$. For example, in classical linear regression with i.i.d. Gaussian errors (cf. Section 4.1), the above mixture density becomes the Gaussian location mixture

$$f_{\mathbf{y}}(\cdot; \boldsymbol{\beta}, \sigma) = \frac{1}{n}\sum_{i=1}^n \frac{1}{\sigma}\varphi\left(\frac{\cdot - \mathbf{x}_i^\top \boldsymbol{\beta}}{\sigma}\right), \tag{8}$$

where φ denotes the PDF of the $N(0, 1)$-distribution. Incorporating this into the EM approach in Section 3.1 would break the simplicity of the updates, which appears too much of a price to pay given only minor gains in statistical efficiency over the plug-in approach in which f_y could simply be replaced by a kernel density estimator based on the $\{y_i\}_{i=1}^n$. As a compromise, an initial kernel density estimator can be replaced by the representation in (8) with $(\boldsymbol{\beta}, \sigma)$ substituted by estimates $(\widehat{\boldsymbol{\beta}}, \widehat{\sigma})$ obtained from a first round of EM iterations.

A simplified approach for GLMs is to model f_y in terms of an intercept-only GLM (and potentially a scale parameter). In particular, this is relevant to binary GLMs in which case the intercept is simply a one-to-one transformation of $\mathbf{P}(y_i = 1)$, $1 \le i \le n$. In Normal GLMs, if the predictor variables follow a Normal distribution, then f_y is also a Normal distribution with unknown mean (intercept) and standard deviation (scale parameter). It is justifiable to adopt the latter model at least as a simple approximation outside the setting of Normal predictors (cf. Slawski et al., 2021).

5.2 Bayesian inference

There are situations where it can be useful to recast the proposed approach in a Bayesian framework to facilitate inference. In particular, this is the case in which regularization is imperative to deal with a large number of parameters. For example, RL is often performed after blocking, and each (of the potentially many) blocks may be associated with its own mismatch rate. We do not pursue this case further and instead consider another scenario of interest, namely smooth curve fitting via penalized splines. Specifically, the 'roughness penalty' (Green & Silverman, 1993) is realized via an (improper) Gaussian prior on the spline coefficients. This connection facilitates the data-driven choice of the level of smoothing, which is particularly helpful when other criteria such as Generalized Cross-Validation (Craven & Wahba, 1978) are not easily applicable. Moreover, subsequent inference (e.g. point-wise standard errors for the regression curve) becomes rather straightforward within a Bayesian framework.

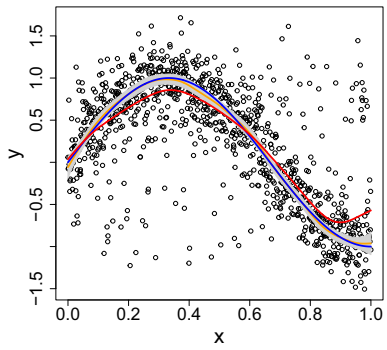
Specifically, we consider the following setup expressed in a hierarchical Bayes fashion:

$$\begin{aligned} &f(\alpha) \propto 1, \qquad f(\sigma^2) \propto (\sigma^2)^{-1}, && f(\tau^2) \propto (\tau^2)^{-1} \\ &\{m_i\}_{i=1}^n \mid \alpha \overset{\text{i.i.d}}{\sim} \text{Bernoulli}(\alpha), && f(\boldsymbol{\beta} \mid \tau^2) \propto (\tau^2)^{-r/2} (\det \mathbf{S}^+)^{1/2} \exp\left(-\frac{1}{2\tau^2} \boldsymbol{\beta}^\top \mathbf{S} \boldsymbol{\beta}\right) \\ &y_i \mid x_i, \{m_i = 0\}, \boldsymbol{\beta}, \sigma^2 \sim N(s_{\boldsymbol{\beta}}(x_i), \sigma^2), && y_i \mid \{m_i = 1\} \sim f_y, \quad i = 1, \ldots, n, \end{aligned} \tag{9}$$

where for $\boldsymbol{\beta} \in \mathbb{R}^d$, the function $s_{\boldsymbol{\beta}}(x) = \sum_{j=1}^d \beta_j B_j(x)$ is a cubic spline expansion with coefficients $\boldsymbol{\beta} = (\beta_j)_{j=1}^d$ and basis functions $\{B_j\}_{j=1}^d$ on some interval $[a, b]$ covering the range of the predictor variable. To keep the setup simple, the mismatch indicators are assumed to be i.i.d. Bernoulli random variables.[2] The prior $f(\boldsymbol{\beta} \mid \tau^2)$ is an established construct in the spline literature, (cf., e.g. Ruppert et al., 2003); in (9), $^+$ denotes the Moore–Penrose pseudo-inverse, and r equals the rank of the roughness penalty matrix $\mathbf{S}$. Note that improper Gamma priors are placed on σ^2 and τ^2, with σ^2/τ^2 corresponding to the effective smoothing parameter. Conveniently, under (9) posterior inference can be performed via Gibbs sampling with standard distributions for the full conditionals (cf. Section E.3 in the online supplementary material).

As an illustration, we simulate data $y_i = \sin(\frac{3}{2}\pi x_i) + 0.25 \cdot \varepsilon_i$, $\{\varepsilon_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} N(0, 1)$, $x_i = (i-1)/(n-1)$, $1 \le i \le n = 1,000$, and randomly shuffle 20% of the (x_i, y_i)-pairs. We then fit a cubic spline (25 equispaced knots) to the resulting data, without any adjustment, using the R package `mgcv` (Wood, 2017) as well as with the approach outlined above (a kernel density estimator based on the $\{y_i\}$ is used for f_y). The results are shown in Figure 2. It can be seen that the proposed approach successfully remediates the effect of mismatches and that all model parameters are estimated accurately. Moreover, the level of smoothing with the proposed approach aligns closely with the level of smoothing for correctly matched data, whereas ignoring mismatches yields an oversmoothed fit. In this regard, the Bayesian perspective is particularly helpful since it is

[2] More accurately, one could assume that the $\{m_i\}_{i=1}^n$ arise from a permutation that moves a fraction of α indices. Under this assumption, the mismatch indicators are pairwise independent asymptotically as $n \to \infty$, cf. Section E.2 in the online supplementary material.



underlying sine function
unadjusted spline fit
adjusted spline fit (point-wise posterior means)
95% point-wise credible bands

	naive	oracle	posterior mean	95% credible interval
α	–	–	.1966 (0.2†)	[.155, .235]
σ^2	.222	.061	.0630 (0.0625†)	[.055, .072]
σ^2/τ^2	1088.6*	413.1*	383.9	[89.4, 911.5]

*: smoothing parameter selected by `mgcv`.
†: true parameter values.

Figure 2. Left: Realizations from a noisy sine function with 20% random mismatches, unadjusted spline fit, and adjusted spline fit (point-wise posterior mean and credible bands).

unclear how to select the smoothing parameter in a penalized likelihood framework and conduct subsequent statistical inference. For more discussion and additional numerical studies regarding the impact of mismatch error on nonlinear regression, the interested reader is referred to Section E.1 of the online supplementary material of this article.

6 Simulations

We here present the results of a set of simulation studies to investigate the empirical performance of our approach in a series of different scenarios, including the case of (partial) model misspecification and gentle violations of some of the underlying assumptions listed at the beginning of Section 2. We consider a single predictor variable x and an outcome variable y following a Poisson distribution with $\mathbf{E}[y \mid x] = \exp(\beta_0^* + \beta_1^* x)$. Specifically, we consider a fixed design with $\{x_i\}_{i=1}^n$, $n = 1,000$ uniformly spaced between 1 and 5 and $\beta_0^* = 0.5$, $\beta_1^* = 2$. Several settings are considered for the mismatch indicator:

Constant. The $\{m_i\}_{i=1}^n$ are sampled i.i.d. from a Bernoulli distribution with probability of success $\alpha^* \in \{0, 0.05, \ldots, 0.3\}$. Given the $\{m_i\}_{i=1}^n$ the corresponding subset of the $\{y_i\}_{i=1}^n$ is permuted according to a right circular shift.[3]

Blockwise. The data set is subdivided into four subsets (blocks) of equal size. Within each of the blocks, the mismatch rates are constant (equal to 0, 0.1, 0.4, 0.6, respectively), and the procedure described under *Constant* is applied.

Logistic. In an attempt to mimic the situation in which output from probabilistic record linkage is available to the data analyst operating on linked data, we consider auxiliary data $z_i = \text{logit}(p_i)$, where the p_i's take the place of match 'probabilities' assigned to the ith linked pair (as potentially supplied by a record linkage procedure), $1 \le i \le n$. Here, the $\{p_i\}_{i=1}^n$ are drawn i.i.d. from a Beta distribution with parameters 4.5 and 0.5. Subsequently, the $\{m_i\}_{i=1}^n$ are generated according to the logistic model

$$\text{logit}\{\mathbf{P}(m_i = 0 \mid z_i)\} = \gamma_0^* + \gamma_1^* z_i, \quad 1 \le i \le n, \tag{10}$$

where $\gamma_0^* = -0.5$ and $\gamma_1^* = 1$. Given the $\{m_i\}_{i=1}^n$, the y_i's are permuted as described under *Constant*.

For all three settings, the model for the mismatch indicator is specified accordingly when applying our approach. The marginal density f_y is estimated via a kernel density estimator with a rectangular kernel and bandwidth fixed to 100 throughout all simulations.

In addition to the above settings, we consider three further settings associated with model misspecification and/or violation of assumptions. We conduct 10k replications per setting.

Mis-y. The linear predictor in the Poisson model is misspecified in that a quadratic model in x is used to generate the y's. Specifically, $\mathbf{E}[y \mid x] = \beta_0^* + \beta_1^* x + \beta_2^* x^2$ with $\beta_2^* = 0.05$. This model for y is combined with the constant mismatch rate scenario described above.

[3] The corresponding index permutation π on is of the form $\pi(i_1) = i_2, \ldots, \pi(i_{k-1}) = i_k, \pi(i_k) = i_1$.

Mis-m. As a modification of the setting 'logistic' above, model (10) is changed as follows:

$$\text{logit}\{\mathbf{P}(m_i = 0 \mid z_i)\} = 0.5(\gamma_0^* + \gamma_1^* z_i \mathbb{I}(2 \leq x_i \leq 4), \quad 1 \leq i \leq n, \tag{11}$$

When applying our approach, we instead fit a logistic model linear in z in accordance with (10). Note that in addition to using a misspecified model for the mismatch indicator, the fact that the mismatch indicator depends on x also constitutes a violation of the independence assumption (**A1**).

Mis-ind. In this setting the x_i's are partitioned into 50 blocks of size 20. Within each block, the x_i's are simulated according to a Gaussian copula inducing dependence between each set of 20 x_i's whose marginal distribution is uniform on [1, 5]. The associated covariance matrix of the Gaussian copula is taken as the equi-correlation matrix with unit diagonal elements and off-diagonal elements equal to 0.5. A constant mismatch rate is assumed within each block and the y_i's for which $m_i = 1$ are permuted as described under *Constant*. Note that this simulation design violates assumption (**A2**), part (IND).

Results. Under correct model specifications, the proposed approach largely performs as expected. Confidence interval coverage levels achieve the nominal 95% for all model parameters, with slight undercoverage for the parameter γ^* (the logit of the correct match rate $1 - \alpha^*$) only under the *Blockwise* setting; we suspect that this might be attributable to the reduced sample size in each block. Table 1 also shows that the impact of mismatches becomes noticeable once 10% of the observations are incorrectly matched. Plain GLM estimates for the regression parameters follow a typical pattern of attenuation characterized by an inflated intercept and a reduced slope. By contrast, with the proposed adjustment, the estimation of the regression parameters is not visibly affected. In addition, substantial losses in statistical efficiency in the absence of mismatches ($\alpha^* = 0$) are not observed either.

In the presence of model misspecification and/or violation of assumptions, the regression coefficients are still estimated accurately and confidence level coverage is maintained, with the exception of setting *Mis-y* in which the linear predictor is misspecified. For the latter setting, performance is evaluated in terms of the Kullback–Leibler divergence (KLD) between the $\{\mu_i^* = \mathbf{E}[y_i \mid x_i]\}_{i=1}^n$ and the corresponding estimates $\{\widehat{\mu}_i\}_{i=1}^n$ given an incorrectly specified linear predictor; the KLD with adjustment range between 6.0 ($\alpha^* = 0$) and 7.8 ($\alpha^* = 0.3$) after adjustment, whereas without adjustment the KLD equals only 1.9 for $\alpha^* = 0$ but then jumps to 208 for $\alpha^* = 0.05$ and increases to almost 3.3k for $\alpha^* = 0.3$. While it is found that the different forms of mis-specifications studied here do not have a noticeable impact concerning the estimation of the x–y relationship, estimation of the model parameters pertaining to the latent mismatch indicators $\{m_i\}_{i=1}^n$ is affected more noticeably. For instance, in the setting Mis-y the mismatch rate is consistently overestimated by about 10% , and in the setting Mis-ind the mismatch rate is slightly underestimated. This would be expected since substantial correlations within blocks of observations reduce the impact of mismatch error. Similarly, for the setting *Mis-m* in which the generation of the mismatch indicators departs from the assumed logistic regression model, the impact of mismatch error is still clearly noticeable but reduced. This is because the overall mismatch rate is lower and mismatch error affects a narrower range of the predictor variable (only observations with x taking values in [2, 4] instead of the full range [1, 5]), which in turn limits the range of error in the response resulting from mismatches.

Section F of the online supplementary material contains the corresponding simulation results for a smaller sample size ($n = 100$). While the coverage rates tend to be slightly below the nominal coverage level, the results generally agree with what is reported here.

7 Applications

In this section, we illustrate our methodology in three case studies involving real data sets obtained from record linkage, including (i) a longevity analysis based on historical linkage, (ii) an analysis of two-way contingency tables obtained from linking Medicare claims and survey responses, and (iii)

Table 1. Simulation results based on 10k replications

Constant																			
		$\widehat{\beta}_0$			$\widehat{\beta}_1$			$\widehat{\gamma}$			$\widehat{\beta}_0$ (mis-ind)			$\widehat{\beta}_1$ (mis-ind)			$\widehat{\gamma}$ (mis-ind)		
α^*		RB	SD	CG	RB	SD	CG	RB	SD	CG	RB	SD	CG	RB	SD	CG	RB	SD	CG
0	w/	$2e^{-5}$	$4e^{-3}$	0.95	$8e^{-6}$	$1e^{-3}$	0.95				$6e^{-5}$	$4e^{-3}$	0.95	$4e^{-6}$	$1e^{-3}$	0.95			
	w/o	$2e^{-5}$	$2e^{-3}$	0.95	$8e^{-6}$	$1e^{-3}$	0.95				$6e^{-5}$	$4e^{-3}$	0.95	$4e^{-6}$	$1e^{-3}$	0.95			
0.05	w/	$4e^{-5}$	$4e^{-3}$	0.95	$2e^{-6}$	$1e^{-3}$	0.95	−0.01	0.15	0.95	$9e^{-5}$	$4e^{-3}$	0.95	$5e^{-6}$	$9e^{-4}$	0.95	−0.02	0.16	0.94
	w/o	2.4	0.34	0.06	−0.13	0.08	0.07				1.1	0.20	0.27	−0.06	0.05	0.27			
0.10	w/	$1e^{-5}$	$5e^{-3}$	0.95	$1e^{-6}$	$1e^{-3}$	0.95	−0.01	0.11	0.95	$1e^{-4}$	$5e^{-3}$	0.95	$5e^{-6}$	$1e^{-3}$	0.95	−0.03	0.15	0.92
	w/o	4.2	0.38	0	−0.23	0.09	0				2.1	0.30	0.03	−0.12	0.07	0.03			
0.15	w/	$1e^{-4}$	$5e^{-3}$	0.95	$4e^{-6}$	$1e^{-3}$	0.95	−0.01	0.09	0.95	$2e^{-4}$	$5e^{-3}$	0.95	$1e^{-5}$	$1e^{-3}$	0.95	−0.04	0.11	0.90
	w/o	5.7	0.37	0	−0.32	0.09	0				3.0	0.28	0	−0.17	0.06	0			
0.20	w/	$6e^{-5}$	$5e^{-3}$	0.95	$3e^{-6}$	$1e^{-3}$	0.95	−0.01	0.08	0.95	$2e^{-4}$	$5e^{-3}$	0.95	$1e^{-5}$	$1e^{-3}$	0.95	−0.04	0.08	0.90
	w/o	7.0	0.35	0	−0.40	0.08	0				3.8	0.27	0	−0.21	0.06	0			
0.25	w/	$3e^{-4}$	$5e^{-3}$	0.95	$2e^{-5}$	$1e^{-3}$	0.95	−0.02	0.07	0.95	$2e^{-4}$	$5e^{-3}$	0.95	$1e^{-5}$	$1e^{-3}$	0.95	−0.05	0.08	0.88
	w/o	8.0	0.33	0	−0.46	0.08	0				4.5	0.27	0	−0.26	0.06	0			
0.3	w/	$1e^{-4}$	$5e^{-3}$	0.95	$8e^{-6}$	$1e^{-3}$	0.95	−0.02	0.07	0.95	$5e^{-4}$	$5e^{-3}$	0.95	$3e^{-5}$	$1e^{-3}$	0.95	−0.07	0.07	0.87
	w/o	8.9	0.31	0	−0.51	0.08	0				5.2	0.27	0	−0.30	0.06	0			

Blockwise																		
	$\widehat{\beta}_0$			$\widehat{\beta}_1$			$\widehat{\gamma}_1$			$\widehat{\gamma}_2$			$\widehat{\gamma}_3$			$\widehat{\gamma}_4$		
	RB	SD	CG	RB	SD	CG	RB	SD	CG	RB	SD	CG	RB	SD	CG	RB	SD	CG
w/	$3e^{-4}$	$6e^{-3}$	0.95	$2e^{-5}$	$1e^{-3}$	0.95	–	[a]	–	0.21	0.25	0.91	0.09	0.13	0.90	0.04	0.13	0.94
w/o	2.8	0.12	0	−0.16	0.03	0												

(*continued*)

Table 1. Continued

Logistic													
		$\widehat{\beta}_0$			$\widehat{\beta}_1$			$\widehat{\gamma}_0$			$\widehat{\gamma}_1$		
		RB	**SD**	**CG**	**RB**	**SD**	**CG**	**RB**	**SD**	**CG**	**RB**	**SD**	**CG**
(10)	**w/**	$1e^{-4}$	$5e^{-3}$	0.95	$8e^{-6}$	$1e^{-3}$	0.95	0.02	0.22	0.95	$5e^{-3}$	0.10	0.95
	w/o	4.9	0.33	0	−0.28	0.08	0						
mis-m (11)	**w/**	$3e^{-5}$	$5e^{-3}$	0.95	$1e^{-6}$	$1e^{-3}$	0.95	–	[b]	–	–	[b]	–
	w/o	0.58	0.05	0	−0.03	0.01	0						

Note. Instead of the mismatch rate α^*, we estimate $\gamma^* = \log((1-\alpha^*)/\alpha^*)$.
RB = relative bias; SD = standard deviation; CG = coverage rate of confidence intervals. **w/** = with adjustment (proposed approach); **w/o** = without adjustment, i.e. plain GLM estimation.
[a] Not reported because the mismatch rate in block one is zero.
[b]Not reported because the corresponding part of the model is misspecified.

the investigation of time trends in the issuance of nurse licenses. The data sets for (i) and (iii) are open access.[4]

7.1 Longevity analysis

As mentioned in the Introduction, the Life-M project provides multigenerational data from the 20th century that was gathered from various data sources including birth certificates, death certificates, marriage certificates, and decennial censuses. In our case study, we study the relationship between the age of death and the year of birth obtained from linking birth and death certificates. Longitudinal Intergenerational Family Electronic Micro-Database used a hybrid of two linkage procedures: a fraction of the records were selected for manual linkage by trained research assistants ('hand-linked' records); the remaining records were linked based on probabilistic record linkage without clerical review ('machine-linked' records). The latter records are more inclined to have mismatch errors (Bailey et al., 2022).

Initial analyses of these data suggest that the death record sources and collection periods influence the trend in the age at death as a function of the year of birth. Therefore, we focus on birth cohorts where longevity tends to increase overall as expected. After visual exploration of the entire data available ($n \approx 155$k), we decided to use a cubic polynomial to model the relationship between year of birth and age at death (dependent variable); a cubic fit is used to capture the nonlinear relationship between the two variables during a specific time period (1883–1906).

Our approach is applied as follows. We assume a Gaussian regression model for the predictor–response relationship. Regarding the latent mismatch indicators $\{m_i\}_{i=1}^n$, we assume that all 2,159 hand-linked records are correctly matched (i.e. for the corresponding records it holds that $m_i = 0$). For machine-linked records, the mismatch indicator is considered unknown and is modelled via a logistic regression model whose predictors are given by the commonness of the first name (`commf`) and the last name (`comml`) of the associated individuals. Since these variables are readily available and probabilistic record linkage was primarily based on names, they are considered suitable surrogates in lieu of more specific information about the correctness of matches as would be output by a probabilistic record linkage procedure. The marginal distribution of the response variable is assumed to follow a Gaussian distribution whose parameters are estimated from the entire data and subsequently treated as fixed. In summary, the inference is based on the specifications

$$\begin{aligned} &y_i \,|\, x_i, \{m_i = 0\} \sim N(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3, \sigma^2), \quad y_i \,|\, \{m_i = 1\} \sim N(\mu, \tau^2), \\ &m_i \,|\, \texttt{commf}_i,\ \texttt{comml}_i \sim \text{Bernoulli}\left(\frac{\exp(\gamma_0 + \gamma_1 \cdot \texttt{commf}_i + \gamma_2 \cdot \texttt{comml}_i)}{1 + \exp(\gamma_0 + \gamma_1 \cdot \texttt{commf}_i + \gamma_2 \cdot \texttt{comml}_i)}\right), \quad 1 \le i \le n. \end{aligned} \tag{12}$$

Additionally, the Life-M team expects the mismatch rate among the machine-linked rates to be around 5% (Bailey et al., 2022). This information is incorporated by imposing corresponding constraints on the average of the linear predictors as described in Section 4.4. Specifically, we consider the upper bounds -3 and -2.5 on the logit scale, corresponding to about 5% and 7.5%, respectively, on the probability scale; the latter bound allows for a slightly higher fraction of mismatches as expected. While date of birth is used as a matching variable during linkage, there are no indications that mismatch rates depend (substantially) on the year of birth, the predictor variable in (12). Therefore, the assumption of no overlap between the predictors and variables informative of match status as mandated by (**A1**) in Section 2 is justifiable here.

Results are summarized in Figure 3 and Table 2. The estimated coefficients and predictions of the cubic fit generated by approach (12) are well within the realm of the naive analysis without adjustment for mismatches and an analysis confined to the much smaller subset of hand-linked records only. Figure 3 indicates that predictions under the naive and adjusted approaches start diverging from birth cohort 1897, with predictions under the naive approach falling below those under the adjusted approach and those under the hand-linked only analysis. Adjustment yields small reductions of the estimated residual standard error (about 2.5% and 4%, respectively) and the standard errors of the coefficients of the cubic polynomial tend to be slightly smaller as well. First-name commonness and last-name commonness are both predictive of the latent match

[4] (i) https://tinyurl.com/4wwv4uzf and (iii) https://tinyurl.com/5xs8pwsf.

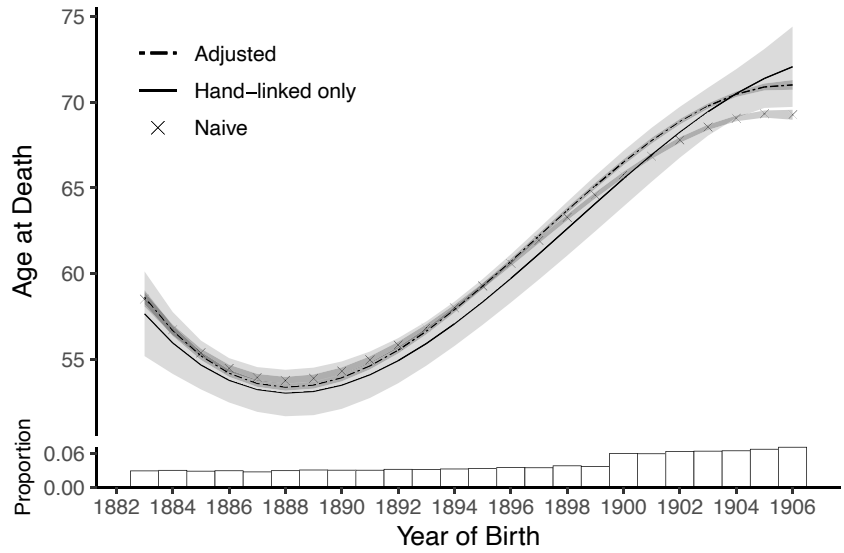


Figure 3. Predicted survival times for birth cohorts 1883–1906 based on the Life-M data with point-wise confidence intervals (grey-shaded areas; the wide light grey area corresponds to the results based on the 'hand-linked' records only). 'Adjusted' refers to the results under model (12) with the average linear predictor in the model for the $\{m_i\}$ upper bounded by −3 (∼5% mismatch rate). The proportion of the birth cohorts is shown at the bottom of the plot.

status. The sign of the coefficient for first name commonness is unexpected though (since intuitively the more common a name, the more likely mismatches tend to occur). We hence also explored the use of an interaction model with the same two predictor variables but since this change neither improved interpretability nor model fit we decided to retain the main effect model.

7.2 Agreement of Medicare claims and survey responses

The second case study presents an application of the two-way contingency table methodology outlined in Section 4.3. This case study is based on a linkage between a survey conducted as part of the HRS and Medicare claims data (see the Introduction). Such linkages to administrative data are routinely performed as part of the HRS, e.g. to online supplementary material or validate data reported by survey respondents given their consent to link.

In 2020, the HRS reevaluated the Medicare record linkage performed in 2018, and identified 59 cases in the 2018 linkage that were likely mismatches, either because these cases were linked to *different* claims records in the new linkage in 2020, or these cases could not be linked to a claims record in 2020. In this case study, we focus on the effects of including these likely mismatched cases in a contingency table analysis of the 2018 HRS data. Specifically, we look at the bivariate association between self-reports of nursing home attendance in the past 2 years and administrative records of nursing home attendance in that same time frame. Of specific interest to the HRS is the level of agreement between these two measures, for the purpose of investigating potential measurement error.

We note that the overall rate of likely mismatches in 2018 is rather small, given that there were 8,665 consenting respondents in total. For the sake of this illustration, we, therefore, selected a simple random sample of 300 HRS respondents who were not deemed to be mismatches in 2018, effectively simulating a mismatch rate of $59/359 = 0.164$. In this scenario, the mismatched cases may have an effect on the contingency table analysis.

In the analysis, we used the four proportions defining the two-by-two contingency table based on the 300 exact matches as the benchmark proportions for evaluation. For reasons explained in Section 4.3, the mismatch rate is assumed to be known. We computed the mean relative absolute error (MRAE) of the four proportions defining the contingency table, the KLD as a measure of distance between the proportions in the contingency table and the proportions based on the exact

Table 2. Parameter estimates for the longevity analysis (standard error in parentheses)

	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\sigma}$	$\widehat{\gamma}_0$	$\widehat{\gamma}_1$	$\widehat{\gamma}_2$
Naive	58.5 (0.2)	−46.7 (1.8)	130.4 (4.0)	−72.9 (2.5)	21.2 (0.04)			
Adjusted[a]	58.6 (0.1)	−51.0 (1.5)	140.3 (3.9)	−76.8 (2.6)	20.7 (0.06)	−5.98 (0.5)	−1.45 (0.6)	7.2 (0.3)
Adjusted[b]	58.7 (0.2)	−52.5 (1.6)	143.2 (3.9)	−77.7 (2.7)	20.4 (0.06)	−4.86 (0.4)	−1.38 (0.4)	6.1 (0.3)
Hand-linked	57.7 (1.3)	−44.2 (11.6)	118.6 (27.9)	−59.9 (18.5)	19.0 (0.29)			

Note. The letters [a] and [b] refer to the bounds −3.0 and −2.5 on the average of the linear predictors in the model for the mismatch indicators (cf. Section 4.4)

Table 3. Results of the Health and Retirement Study–Medicare claims contingency table analysis

	MRAE	KLD	GOF	Association	Kappa
Naive	0.1685	0.0020	1.4780	125.12	0.6130
Adjusted	0.1334	0.0017	1.0995	147.06	0.6625
Exact	0.0000	0.0000	0.0000	146.80	0.6569

Note. 'Exact' refers to the analysis based on the correct matches only and is used as a benchmark.

matches, the one-sample χ^2 goodness of fit (GOF) measure for the four proportions (again using the benchmark proportions), the χ^2 measure of association between the two variables, and Cohen's κ statistic.

Table 3 presents the results of our analysis. Compared to an analysis of the 300 known exact matches, the naive analysis of the 359 cases would result in a higher MRAE of the four proportions defining the table, a larger KLD based on the four proportions, a larger χ^2 GOF measure, an attenuated χ^2 measure of association, and an attenuated kappa statistic (understating the level of agreement between the two variables). The adjustment approach described in Section 4.3 assuming the aforementioned mismatch rate of 0.164 would reduce the errors and yield measures of agreement that are more consistent with the known true values.

7.3 Investigation of trends in nurse license processing times

In this section, we evaluate the utility of the proposed approach on curve fitting via penalized splines (cf. Section 5.2). Specifically, we study an application to a nurse credential database from the state of Washington between 1 January 2009 and 31 December 2021. Each entry in this database corresponds to one specific nurse practice license issued to one specific nurse, containing the following information: full name of the nurse and their year of birth, credential number, issue, and expiration dates, status (active, closed, expired), and type of license (e.g. 'registered nurse license', 'medical assistant certification', 'registered nurse temporary practice permit'). Nurses are commonly issued temporary permits prior to receiving a regular license. In our study, we investigate the average duration of the associated transitional period (in #days), which is of interest to researchers in health metrics (Flaxman, 2022, Personal communication). For this purpose, the two data subsets corresponding to temporary permits and regular licenses, respectively, are extracted and subsequently linked.

Data linkage is performed by first blocking on the year of birth and first initial of the last name of the nurse, and then string matching of first, middle, and last names within each block using the Jaro–Winkler metric (Winkler, 1990). We consider both exact name matching (restrictive linkage) and inexact name matching (generous linkage); in the latter case, two records have declared a match as long as the Jaro–Winkler match scores for each name variable exceeds the threshold 0.85 (chosen ad-hoc via visual inspection of the histograms of the scores). The resulting restrictively linked and generously linked files consist of about 61k and 78k records, respectively, after removing obvious mismatches whose waiting periods between permits were negative.

Ranges for the underlying mismatch rates in these two files were determined as follows: the first estimate assumes that the number of mismatches is about the same as the number of obvious mismatches associated with negative durations; the second estimate is based on excessively large durations ($\geq$1.5 years). This yields the range [3.7%, 8.1%] for the generously linked file and [0.4%, 1.0%] for the restrictively linked file. The still noticeable fraction in the latter file despite exact name matching can be attributed mostly to multiple instances of the license issue process for the same nurse. Given the available information, it is unclear how to determine the true match status with certainty even with a clerical review: in case of multiple issuances, only the earliest and latest dates of issuance are recorded, i.e. any intermediate dates are not given.

The goal of the analysis is to identify trends/variations over time in the average duration of the aforementioned transitional period from the time a temporary permit is issued until it gets substituted by a regular nurse license. We let $\{x_i\}_{i=1}^n$ denote the temporary permit issue dates (scaled to [0, 1] such that 1 January 2009 and 31 December 2021 correspond to 0 and 1, respectively) and

consider the duration until the regular license issue date as the dependent variables $\{y_i\}_{i=1}^n$. The latter is obtained from the linked files and hence in part incorrect as a consequence of mismatch error. Since observations with negative durations are dropped, the impact of mismatches producing smaller durations is significantly reduced (as a result of a truncation at zero) in comparison to mismatches producing inflated durations. In order to flexibly capture trends in average duration over time, the corresponding mean function for correctly matched observations is modelled via a cubic spline, i.e.

$$\mathrm{E}[y_i \mid x_i, m_i = 0; \boldsymbol{\beta}] = s_{\boldsymbol{\beta}}(x_i), \quad 1 \leq i \leq n, \quad s_{\boldsymbol{\beta}}(x) = \sum_{j=1}^{d} \beta_j B_j(x),$$

where the $\{B_j\}_{j=1}^d$ represent the associated B-spline basis functions given 1,000 knots placed evenly in [0, 1]. Conditional on $\{m_i = 1\}$, we assume an intercept-only model for the dependent variable y_i, $1 \leq i \leq n$. For simplicity, we assume Gaussian models (with different variances) for each of these specifications in order to apply the proposed approach, but alternative models (e.g. Poisson) could be used as well.

The Bayesian inference approach outlined in Section 5.2 is applied to both the generously and the restrictively linked data set. Unlike the case study in Section 7.1 neither of these linked files contains any information indicative of the match status of the individual records. The number of MCMC iterations is set to 10,000 after a burn-in period of length 100, out of which every tenth MCMC sample is retained for posterior inference. In addition to an 'out-of-the-box' application, we also run the approach with the residual standard deviation σ of the spline regression model fixed to a range of fractions $\{0.1, 0.15, \ldots, 0.95, 1\}$ of the residual standard deviation $\widehat{\sigma}_0$ from a 'naive' spline fit without accounting for mismatches. While the resulting mean functions do not change substantially, the additional (varying) constraint on σ allows us to explore a range of plausible solutions and associated estimates of the mismatch rate in the absence of specific information about the underlying record linkage. The ratio $\sigma/\widehat{\sigma}_0$ can be interpreted as the relative reduction in root mean squared error after accounting for mismatch error.

Figure 4 shows that without adjustment for mismatches, the estimated mean functions fluctuate strongly at the beginning of the time line; even when the restrictively linked file is used, the average duration exhibits fluctuations of ∼50 days. The impact of mismatch error is indeed expected to be more pronounced at the beginning of the time period than towards the end since excess durations resulting from incorrect linkage can be more drastic. Interestingly, a second window of rapid fluctuations is observed between 0.75 and 0.85 (scaled time scale); however, these fluctuations are present before and after adjustment for mismatch error and are hence more likely to be genuine.

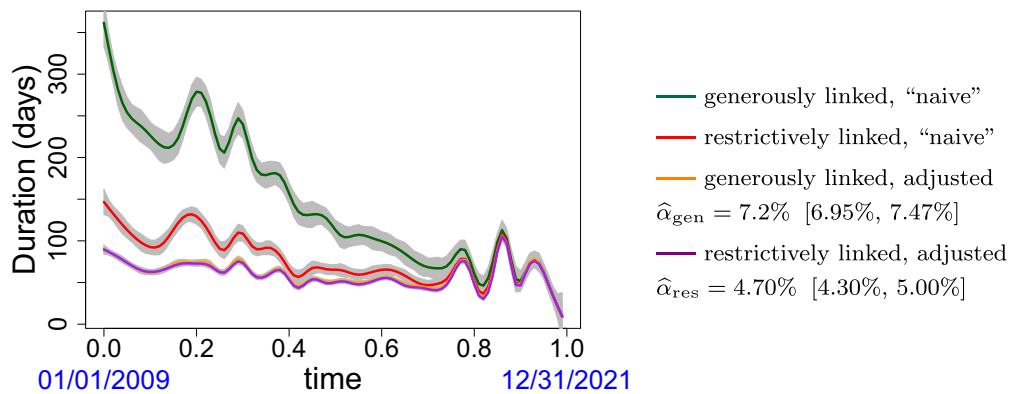


Figure 4. Estimated mean functions for the duration of regular nurse license issuance with and without adjustment for mismatch error based on the generously and restrictively linked files. As explained in the text, for the results after adjustment we report results for which the mismatch rate hits a plateau among the range of solutions under consideration as given in Table 4.

Table 4. Estimated mismatch rates (posterior means) and 95% credible intervals (bracketed) for the nurse credential data depending on the ratio of residual standard deviation $\widehat{\sigma}/\widehat{\sigma}_0$ before and after adjustment; in the leftmost column (asterisked) $\widehat{\sigma}$ is unrestricted, while for the remaining columns the ratio $\widehat{\sigma}/\widehat{\sigma}_0$ is fixed to the value in the column header

$\widehat{\sigma}_{\text{gen}}/\widehat{\sigma}_0$	0.08⋆	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
$\widehat{\alpha}_{\text{gen}}$	0.161 [0.158, 0.164]	0.135 [0.132, 0.138]	0.081 [0.079, 0.084]	0.073 [0.070, 0.075]	0.072 [0.070, 0.075]	0.076 [0.073, 0.078]	0.084 [0.081, 0.088]	0.098 [0.095, 0.103]	0.124 [0.118, 0.130]	0.175 [0.166, 0.184]	0.96 [0.961, 0.965]
$\widehat{\sigma}_{\text{res}}/\widehat{\sigma}_0$	0.14⋆	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
$\widehat{\alpha}_{\text{res}}$	0.188 [0.183, 0.193]	0.274 [0.269, 0.278]	0.132 [0.128, 0.135]	0.081 [0.078, 0.084]	0.058 [0.055, 0.061]	0.049 [0.047, 0.052]	0.046 [0.043, 0.049]	0.047 [0.043, 0.050]	0.053 [0.049, 0.058]	0.073 [0.066, 0.081]	0.051 [0.017, 0.144]

Note. The top half of the table contains the results for the generously linked data set (gen), and the bottom half to the restrictively linked data set (res).

Moreover, after adjustment, the estimated mean functions are significantly closer to the estimated mean functions (unadjusted) based on the restrictively linked file, and the estimated mean functions after adjustment are essentially identical *regardless of whether the adjustment was based on the generously or the restrictively linked file*.

Table 4 shows that for the generously linked file, the estimated mismatch rate plateaus for $\widehat{\sigma}/\widehat{\sigma}_0 = 0.4$ yielding an estimate of 7.2% of mismatches, which is well within the anticipated range between 3.7% and 8.1%. For the restrictively linked file, the estimated mismatch rate plateaus for $\widehat{\sigma}/\widehat{\sigma}_0 = 0.65$ at the value 4.7%, which is still within the realm of the anticipated range and significantly lower than the estimate based on the generously linked file.

8 Conclusion

In this paper, we have developed a general framework to enable valid postlinkage inference in the presence of mismatch error in the challenging secondary analysis setting. The proposed framework is flexible in the sense that limited information about the linkage process can be incorporated, and that the same machinery can be applied to handle various models for the mismatch indicator and the linked substantive variables. The approach is scalable with a run time that is linear in terms of the size of the sample and convenient from the perspective of implementation. A corresponding R package `pldamixture` is available on CRAN. Results from simulations and case studies with real data consolidate the usefulness for postlinkage analysis.

At the same time, the work presented here prompts various avenues of future research. First, the contingency table example presented in Section 4.3 prompts the question of model identifiability, which is not studied in depth in this paper. Significant additional research is needed to characterize the class of identifiable models. Second, it is of interest to further investigate the sensitivity of our approach vis-à-vis violations of the main assumptions even though the simulations shown here indicate at least a moderate degree of robustness. Third, it is worthwhile to consider extensions covering the linkage of more than two files. Fourth, while mismatch error has undoubtedly received much more attention, false nonmatches (missed matches) are similarly important; handling both types of error in an integrated fashion is a desirable goal. Finally, our approach for contingency table analysis highlights a connection to synthetic data methods such as postrandomization (Gouweleeuw et al., 1998) for disclosure control, and it would appear to be worth elaborating on that connection in more detail.

Acknowledgments

We would like to thank Jessica Faul for providing the data used in Section 7.2 and Abraham Flaxman for suggestions and discussions leading to the analysis in Section 7.3. We would like to thank two reviewers, an associate editor, and the editor-in-chief Bianca DeStavola for their careful reading and valuable comments that have led to substantial improvements of the paper.

Conflicts of interest: None declared.

Funding

M.S., B.T.W., P.B., and Z.W. were partially supported by National Science Foundation Grant #2120318.

Data availability

Data and R code for reproducing the simulation studies and analysis in Section 7.3 are available from the George Mason University Dataverse repository (https://doi.org/10.13021/orc2020/D2YJXX), which also contains the R code for reproducing the analysis in Section 7.1. The associated data can be obtained from the platform Open ICPSR (https://tinyurl.com/4wwv4uzf). The data used for the analysis in Section 7.2 cannot be publicly shared since they involve person-level data associated with restricted access. The data for Section 7.3 were retrieved from the official Washington state open data portal (https://tinyurl.com/5xs8pwsf).

Supplementary material

Supplementary material is available online at *Journal of the Royal Statistical Society: Series A*.

References

Abowd J., Abramowitz J., Levenstein M., McCue K., Patki D., Raghunathan T., Rodgers A., Shapiro M., Wasi N., & Zinsser D. (2021). Finding needles in haystacks: Multiple-imputation record linkage using machine learning. Federal Reserve Bank of Boston Research Department. Working Papers No. 22-11.

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. https://EconPapers.repec.org/RePEc:cen:wpaper:19-08.

Agresti A. (2012). *Categorical data analysis*. John Wiley & Sons.

Bailey M., Cole C., Henderson M., & Massey C. (2020). How well do automated linking methods perform? Lessons from US historical data. *Journal of Economic Literature*, *58*(4), 997–1044. https://doi.org/10.1257/jel.20191526

Bailey M., Lin P., Mohammed A. S., Mohnen P., Murray J., Zhang M., & Prettyman A. (2022, December). LIFE-M: The longitudinal, intergenerational family electronic micro-database. Inter-university consortium for political and social research (ICPSR).

Beuthner C., Breuer J., & Jünger S. (2021). *Data linking-linking survey data with geospatial, social media, and sensor data* (Technical Report). GESIS Leibniz Institute for the Social Sciences. https://doi.org/10.15465/gesis-sg_en_039

Binder D. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review/Revue Internationale de Statistique*, *51*(3), 279–292. https://doi.org/10.2307/1402588

Binette O., & Steorts R. (2022). (Almost) all of entity resolution. *Science Advances*, *8*(12), eabi8021. https://doi.org/10.1126/sciadv.abi8021

Bukke P., Wang Z., Slawski M., West B., Ben-David E., & Diao G. (2024). *pldamixture: Post-linkage data analysis based on mixture modelling*. R package version 0.1.0. https://CRAN.R-project.org/package=pldamixture.

Chambers R. (2009). *Regression analysis of probability-linked data* (Technical Report). Official Research Series, Vol. 4. 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*(1), 37–59. https://doi.org/10.1111/rssa.12477

Chambers R., Fabrizi E., Ranalli M., Salvati N., & Wang S. (2023). Robust regression using probabilistically linked data. *WIREs Computational Statistics*, *15*(2), e1596. https://doi.org/10.1002/wics.1596

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

Craven P., & Wahba G. (1978). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, *31*(4), 377–403. https://doi.org/10.1007/BF01404567

Dalzell N., & Reiter J. (2018). Regression modeling and file matching using possibly erroneous matching variables. *Journal of Computational and Graphical Statistics*, *27*(4), 728–738. https://doi.org/10.1080/10618600.2018.1458624

DeGroot M., & Goel P. (1980). Estimation of the correlation coefficient from a broken random sample. *The Annals of Statistics*, *8*(2), 264–278. https://doi.org/10.1214/aos/1176344952

Fellegi I. P., & Sunter A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, *64*(328), 1183–1210. https://doi.org/10.1080/01621459.1969.10501049

Gouweleeuw J., Kooiman P., & Wolf P. D. (1998). Post-randomisation for statistical disclosure control: Theory and implementation. *Journal of Official Statistics*, *14*(4), 463.

Green P., & Silverman B. (1993). *Nonparametric regression and generalized linear models: A roughness penalty approach*. CRC Press.

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*(501), 34–47. https://doi.org/10.1080/01621459.2012.726889

Han Y., & Lahiri P. (2019). Statistical analysis with linked data. *International Statistical Review*, *87*(S1), 139–157. https://doi.org/10.1111/insr.12295

Hof M., & Zwinderman A. (2015). A mixture model for the analysis of data derived from record linkage. *Statistics in Medicine*, *34*(1), 74–92. https://doi.org/10.1002/sim.v34.1

Japec L., Kreuter F., Berg M., Biemer P., Decker P., Lampe C., Lane J., O'Neil C., & Usher A. (2015). *AAPOR report on big data*. (Technical Report). American Association for Public Opinion Research.

Kim G., & Chambers R. (2012). Regression analysis under incomplete linkage. *Computational Statistics and Data Analysis*, *56*, 2756–2770. https://doi.org/10.1016/j.csda.2012.02.026
Lahiri P., & Larsen M. D. (2005). Regression analysis with linked data. *Journal of the American Statistical Association*, *100*(469), 222–230. https://doi.org/10.1198/016214504000001277
Lindsay B. (1988). Composite likelihood methods. *Contemporary Mathematics*, *80*, 221–239. https://doi.org/10.1090/conm/080/999014
Little R., & Rubin D. (2019). *Statistical analysis with missing data*. John Wiley & Sons.
Liu Y., Singh L., & Mneimneh Z. (2021). A comparative analysis of classic and deep learning models for inferring gender and age of Twitter users. In *Proceedings of the 2nd International Conference on Deep Learning Theory and Applications (DeLTA)*. SCITEPRESS Digital Library.
Lohr S., & Raghunathan T. (2017). Combining survey data with other data sources. *Statistical Science*, *32*(2), 293–312. https://doi.org/10.1214/16-STS584
McCullagh P., & Nelder J. (1989). *Generalized linear models*. Chapman and Hall, London.
Mneimeh Z. (2022). Evaluation of consent to link Twitter data to survey data. *Journal of the Royal Statistical Society Series A*, *185*, 364–386. https://doi.org/10.1111/rssa.12949
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.
Newcombe H., & Kennedy J. (1962). Record linkage: Making maximum use of the discriminating power of identifying information. *Communications of the ACM*, *5*(11), 563–566. https://doi.org/10.1145/368996.369026
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*(5), 3826–3300. https://doi.org/10.1109/TIT.2017.2776217
Ruppert D., Wand M., & Carroll R. (2003). *Semiparametric regression*. Cambridge University Press.
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, 12). Regression analysis of data files that are computer matched II. *Survey Methodology*, *23*, 157–165.
Slawski M., & Ben-David E. (2019). Linear regression with sparsely permuted data. *Electronic Journal of Statistics*, *13*(1), 1–36. https://doi.org/10.1214/18-EJS1498
Slawski M., Diao G., & Ben-David E. (2021). A pseudo-likelihood approach to linear regression with partially shuffled data. *Journal of Computational and Graphical Statistics*, *30*, 991–1003. https://doi.org/10.1080/10618600.2020.1870482
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.
Stier S., Breuer J., Siegers P., & Thorson K. (2020). Integrating survey data and digital trace data: Key issues in developing an emerging field. *Social Science Computer Review*, 38(5), 503–516.
Tancredi A., & Liseo B. (2015). Regression analysis with linked data: Problems and possible solutions. *Statistica*, *75*(1), 19–35. https://doi.org/10.6092/issn.1973-2201/5821
Varin C., Reid N., & Firth D. (2011). An overview of composite likelihood estimation. *Statistica Sinica*, *21*, 5–42.
Wang Z., Ben-David E., Diao G., & Slawski M. (2022). Regression with linked datasets subject to linkage error. *WIREs Computational Statistics*, *14*(4), e1570. https://doi.org/10.1002/wics.1570
Winkler B. (1990). *String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage* (Technical Report). U.S. Census Bureau, Statistical Research Division.
Wood S. (2017). *Generalized additive models: An introduction with R*. Chapman & Hall/CRC.
Zhang L.-C., & Tuoto T. (2021). Linkage-data linear regression. *Journal of the Royal Statistical Society Series A*, *184*, 522–547. https://doi.org/10.1111/rssa.12630