Data Analysis after Record Linkage

Sources of Error, Consequences, and Possible Solutions

Martin Slawski Priyanjali Bukke Emanuel Ben-David

GEORGE MASON UNIVERSITY GEORGE MASON UNIVERSITY United States™ Census Bureau

June 4, 2023
IISA Conference
Golden, CO

Record Linkage

Generally, there are various ways of combining data from multiple sources. Often, this involves aggregated information, e.g., area-specific statistics.

Record Linkage (RL) is the most granular form of data integration, and refers to micro-level, i.e., record-by-record combination of two or more files.

RL or *entity resolution*, is formally the task of pairing records from multiple files and identifying which pairs belong to the same entity.

First_Name	Last_Name	Sex	BID	NH_Nights	
William	Smith	M	8LA6-RL1-LE17	1	1
Imari	Vasquez	F	NA	0	2
Morgan	Jones	F	8QP9-RD4-IP64	1	3
Roland	Matthews	M	NA	0	4
Sarah	Begum	F	9YZ3-RZ3-YC19	0	5 X

	First_Name	Last_Name	Sex	BID	ICD-9	NH_Nights
1	Bill	Smith	M	8LA6-RL1-LE17	29011	1
2	Imari	Vazquez	F	7OI6-LI1-WJ31	42840	0
2	Imani	Vasquez	F	5KR9-VF7-EI16	4401	0
3	Morgan	Jones	M	3QP9-RD4-IR55	40301	1
4	Roland	Matthews	M	6XM7-KA4-ZL20	86511	0
6	Donald	Miller	M	7OE2-HG2-EV16	00329	0
7	Agatha	Buckman	F	9WV8-WH4-MG19	5109	1
8	Betty	Wu	F	1SG8-EQ4-EN86	37173	1

Record Linkage: Examples

- IRS data to Census data,
- Hospital records and health insurance claims,
- Birth and death registries,
- Social media data (e.g., Twitter profiles) and surveys,
- Historical censuses
- ⋮

Uncertainty and Potential for Error in RL

- Uncertainty arises when there are no common unique identifiers in both files (such as `SSN`, `EIN`, `HPID`).
- Such identifiers are often removed to protect the privacy or confidentiality of the individuals.
- Unique identifiers can be different across multiple files. For example, education, health, and tax records use different personal identifiers.
- Linkage between such data files then relies on some quasi-identifiers (aka **comparison variables** or **matching variables**), such as names, DOB, addresses, etc.

Which records belong to the same individual?

f.name	m.name	l.name	m.o.b	lives in
Emanuel	Hyatt	Bendavid	Mar	New York, NY
Emmanuel	Ben	David	Dec	Washington, DC
Emanuel	NA	Ben-Dawid	Nov	Stanford, CA
Emanuel	NA	Ben-David	Mar	Ashland, OR
E.	NA	Ben-Davit	Nov	San Diego, CA

Variations can be due to misspellings, typographical errors, faulty record systems, or legal name changes.

Some Sources of Linkage Error

Field	Type	Examples
Names	Case	John Smith | JOHN SMITH
	Nicknames	Charles | Chuck
	Synonyms	William | Bill
	Prefixes	Dr. John Smith
	Suffixes	John Smith, III
	Punctuation	O'Malley | Smith-Taylor | Smith, Jr.
	Spaces	John Smith, Jr
	Digits	J2ohn Smith
	Initials	AM | A.M. | Anne Marie
	Transposition	John Smith | Smith John
Address	Abbreviations	RD | Road | DR | Drive
Dates	Format	01012013 | 01-01-2013 | 01JAN2013
	Invalid values	Month = 13 | Day = 32 | Birth year = 2025 | Date = 29FEB2013
Social Security Number	Format	999999999 | 999-99-9999 | 999 99 9999
Geographic locations	Abbreviations	NC | North Carolina
	FIPS codes	North Carolina = 37
	SSA codes	North Carolina = 34
	ZIP Codes	99999 | 99999-9999
	Concatenation of State and county codes	Mecklenburg County, NC | 37119
Sex	Format	Male / Female | M / F | 1 / 2

- Dusetzina SB, Tyree S, Meyer AM, et al. Linking Data for Health Services Research: A Framework and Instructional Guide [Internet]. An Overview of Record Linkage Methods. Available from: https://www.ncbi.nlm.nih.gov/books/NBK253312/

Probabilistic RL

Probabilistic RL aims to systematically address the uncertainty associated with the identification of matching pairs.

- The Fellegi–Sunter (FS) model (1969) formalizes the probabilistic record linkage theory.
- Consider two data files A and B.
- The set of all pairs of records $\{(a, b) : a \in A,\ b \in B\}$ can be partitioned as

$$M = \{(a, b) \mid a \text{ and } b \text{ refer to the same entity}\}$$
$$U = \{(a, b) \mid a \text{ and } b \text{ refer to two different entities}\}$$

- These sets are called **m**atched and **un**matched classes, respectively.

Basic Ideas

- In the FL formulation, the main goal of RL is to make a decision about each pair as to whether it is a match (i.e., it belongs to M) or a non-match (i.e., it belongs to U).
- In a nutshell, the basic ideas are as follows.
  1. Compare records a and b based on their values on variables common to both files.
  2. Assign a score to each comparison.
  3. Make a decision based on the score to declare (a, b) a **match** or **non-match**.
- In this formulation, it is clear that RL is essentially a supervised (classification) or unsupervised (clustering) machine-learning problem.

Comparing Records

- To compare records in a pair, we use variables common to both files A, B.
- These are called **matching variables**.
- We measure the level of agreement/disagreement on each matching variable using a desirable metric.
- For example, below shows 5 matching variables:
- Below shows an example with 5 matching variables.

	`firstname`	`lasttname`	`DOB`	`SEX`	`zipcode`
a	Gillian	Browser	1978	Female	47301
b	Jilliam	Browse	1987	Female	47301

Using Similarity Measures for Comparisons

- The comparison vector may yield $\gamma_{ab} = (.7, .9, 0, 1, 1)$.
- $.7, .9$ indicate partial agreement between `firstname`, `lastname` (using a string distance)
- 0 indicates disagreement (using the Hamming distance)
- 1 indicates agreement (using the Hamming distance).
- When the matching variable is a string, several string distance metrics are available to choose from, such as:
- Levenshtein distance (or edit distance), Jaro distance, Jaro-Winkler distance, and token-based distance functions.
- The Jaro-Winkler string distance is usually the default metric.

Comparison Vector in General Setting

- Comparisons on K matching variables result in a comparison vector

$$\gamma_{ab} = (\gamma_{ab}^1, \ldots, \gamma_{ab}^K)^\top \in \mathbb{R}^K$$

- The k-th component γ_{ab}^k is formed based on the level of agreement between the k-th matching variable.

$$\gamma_{ab}^k = \begin{cases} 0 & \text{disagreement} \\ \vdots & \text{some level of agreement} \\ L_k & \text{agreement} \end{cases}$$

- For example, with two levels: $\gamma_{ab}^k = 1$ for agreement and $\gamma_{ab}^k = 0$ for disagreement on `Gender`.

The Fellegi-Sunter Linkage Rule

- Given the comparison vector $\gamma = \gamma_{ab}$, a (decision) linkage rule designates (a, b) as a match (decision A_1), a possible match (decision A_2), or a non-match (decision A_3).

- The linkage rule is a mapping $d : \gamma \mapsto \{A_1, A_2, A_3\}$.
- There are two types of errors associated with the linkage rule:
  - **Mismatch error** occurs when a true non-match is declared a match.
  - **Missed-match error** occurs when a true match is declared a non-match.

The Fellegi–Sunter Model

- In the FL model, the linkage rule is defined based on the likelihood ratio

$$d(\gamma) = \frac{L(\gamma \mid M)}{L(\gamma \mid U)}.$$

- The FL model specifies:
- An upper threshold T_μ and a lower threshold T_λ.

$$\text{A pair } (a,b) \text{ is declared a } \begin{cases} \text{match} & \text{if } d(\gamma_{ab}) > T_\mu \\ \text{possible match} & \text{if } T_\lambda \leq d(\gamma_{ab}) \leq T_\mu \\ \text{non-match} & \text{if } d(\gamma_{ab}) < T_\lambda \end{cases}$$

- The main intuition is that

$$\text{if } (a,b) \in M, \text{ then } L(\gamma_{ab} \mid M) > L(\gamma_{ab} \mid U).$$

Reducing the Computational Cost of RL

- In the linkage of two data files A, B, the number of comparisons grows quadratically.
- However the number of possible matches only increases linearly.
- For example, if file A has $2k$ records and file B has $3k$ records, ordinarily $6M$ comparisons are required.
- However there are at most 2000 matches, assuming no duplicates in files.
- Blocking is a technique for reducing the computational cost of RL by avoiding unnecessary comparisons of a large number of pairs.
- Blocking filter outs record pairs that are very unlikely to be matched.

Blocking

- For blocking, each data file is partitioned into smaller blocks using some **blocking key variables**.

File **A**

	firstname	lastname	DOB	SEX	zipcode
a_1	Maggie	Paynne	1965	1	24044
a_2	Adrian	Murphy	1995	0	35033
a_3	Johnn	Pitts	1997	0	24044
a_4	Kelly	Vogt	1997	1	17670
a_5	Raymon	Nobble	1986	0	17540
a_6	Yinan	Chen	2000	0	35033

File **B**

	firstname	lastname	DOB	SEX	zipcode
b_1	Kevin	Black	1983	0	35033
b_2	John	Pitts	1997	0	24044
b_3	Kelly	Gross	1997	1	24044
b_4	Ray	Nobble	1968	0	17540
b_5	Yin	Chen	2000	0	35033

- In the example above, `zipecode` is used as a blocking key variable.

An Example of Blocking

- A pair (a, b) is a candidate to be a match only if a, b are in the blocks formed with the same blocking key value.
- Other pairs are declared non-matches.

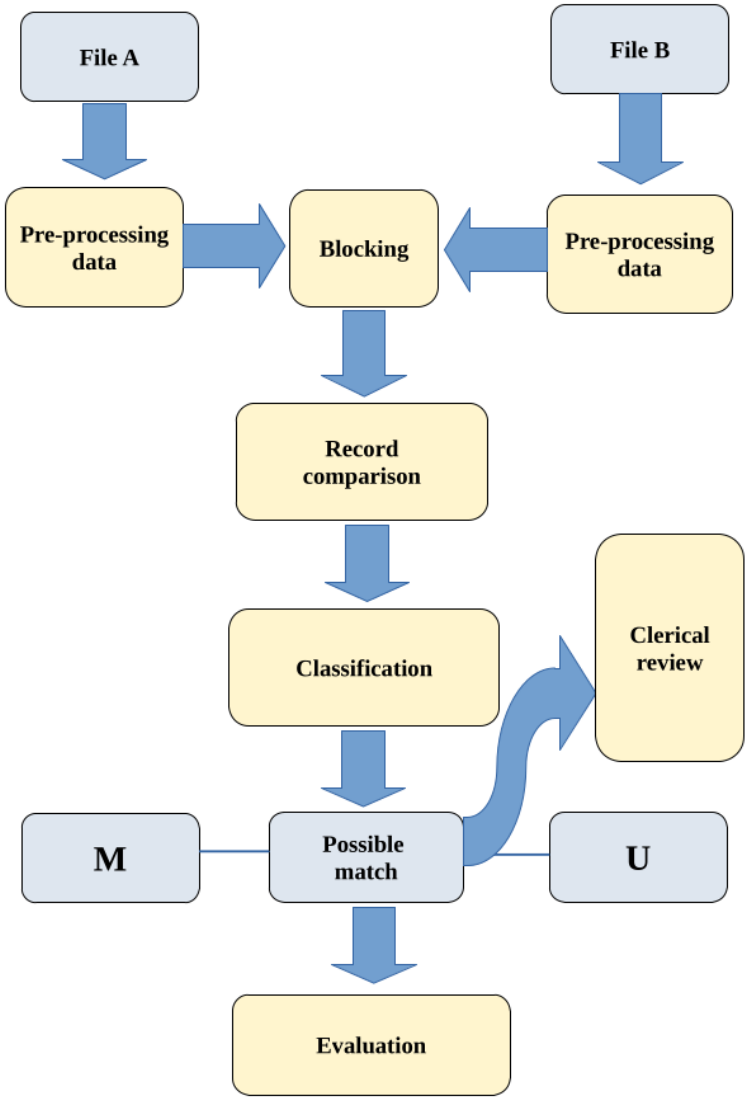
Candidate pairs generated from zipcode blocking

BKV `zipcode`	Candidate pairs
24044	(a_1, b_2), (a_3, b_2), (a_1, b_3), (a_3, b_3)
35033	(a_2, b_1), (a_6, b_1), (a_2, b_5), (a_6, b_5)
17540	(a_5, b_4)

	b_2	b_3	b_1	b_5	b_4
a_1	(a_1, b_2)	(a_1, b_3)	(a_3, b_1)	(a_1, b_5)	(a_1, b_4)
a_3	(a_3, b_2)	(a_3, b_3)	(a_3, b_1)	(a_3, b_5)	(a_3, b_4)
a_2	(a_2, b_2)	(a_2, b_3)	(a_2, b_1)	(a_2, b_5)	(a_2, b_4)
a_6	(a_6, b_2)	(a_6, b_3)	(a_6, b_1)	(a_6, b_5)	(a_6, b_4)
a_5	(a_5, b_2)	(a_5, b_3)	(a_5, b_1)	(a_5, b_5)	(a_5, b_4)
a_4	(a_4, b_2)	(a_4, b_3)	(a_4, b_1)	(a_4, b_5)	(a_4, b_4)

- In this example the number of comparisons, from 30, is reduced to 9.

Summary Chart for RL



Health Care Credential Data Example

Studied in Slawski, West, Bukke, Wang, Diao, Ben-David (2023)

Individual Files to Link: Nurse Records

File A – Nurse License Records

	LastName	FirstName	MiddleName	BirthYear	FirstIssueDate	LastInitial
1	bayhon	mark lowell	sorolla	1987	20190418	b
2	wellsly	paula	ann	1968	20000502	w
3	patel	krupa	missing	1992	20211103	p
4	records	colleen	meehan	1977	20050120	r
5	childress	michaela	lynn	1999	20210818	c
6	foster	leah	marie	1990	20160701	f
...						

File B – Nurse Temporary Permit Records

	LastName	FirstName	MiddleName	BirthYear	FirstIssueDate	LastInitial
1	strange	danielle	nicole	1981	20191212	s
2	bright	sarah	nicole	1989	20210309	b
3	morrow	jennifer	elizabeth	1979	20210603	m
4	lisk	loree	missing	1977	20200526	l
5	barker	benjamin	stephen hong	1989	20210505	b
6	cahhal	andrea	missing	1995	20210708	c
...						

After Blocking Using ”LastInitial” and ”BirthYear”

File A – Nurse License Records

	LastName	FirstName	MiddleName	BirthYear	FirstIssueDate	LastInitial
14	theard	nadia	missing	1779	20210920	t
101152	brinton	patricia	cram	1935	19570103	b
197470	bussman	barbara	giuntoli	1935	19570702	b
215538	bass	patricia	ann	1935	19890306	b
198045	bowman	donna	braymiller	1935	19821119	b
16555	berges	mary	missing	1935	19900202	b
...						

File B – Nurse Temporary Permit Records

	LastName	FirstName	MiddleName	BirthYear	FirstIssueDate	LastInitial
78110	theard	nadia	missing	1779	20210722	t
20045	borcherding	jerome	missing	1935	20110304	b
71371	davis	alice	jane	1935	20120814	d
55717	myers	anne	randall	1935	20120323	m
49000	derstine	jill	b	1936	20180807	d
64233	clark	sandra	l	1937	20090319	c
...						

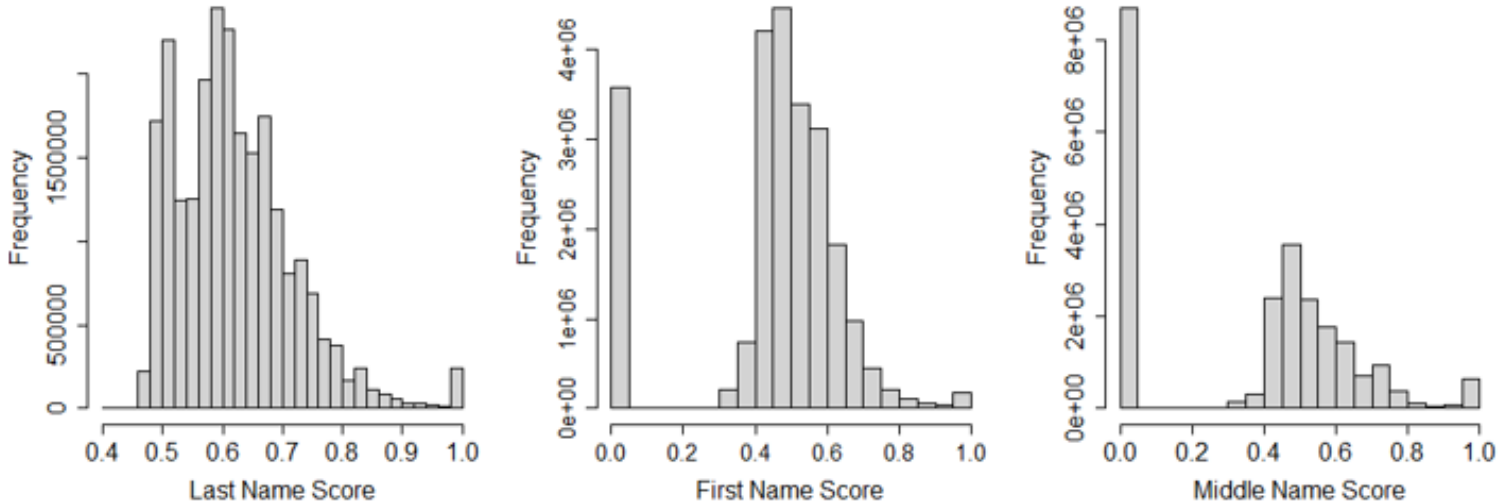
Compare Records Using First, Last, and Middle Names

Comparison Vectors

id1	id2	LastName	FirstName	MiddleName
14	78110	1.0000000	1.0000000	1.0000000
101152	20045	0.6055195	0.4305556	0.0000000
197470	20045	0.5402597	0.4365079	0.4900794
215538	20045	0.5022727	0.4305556	0.0000000
198045	20045	0.6727273	0.4555556	0.4952381
16555	20045	0.5318182	0.4722222	1.0000000

Classify Links Based on a Threshold

Choose Threshold = 0.85



Linked Data – Nurse License and Temporary Permit Records

id.a	LastName	FirstName	MiddleName	BirthYear	FirstIssueDate	id.b	LastName_TP	FirstName_TP	MiddleName_TP	BirthYear_TP	FirstIssueDate_TP
14	theard	nadia	missing	1779	20210920	78110	theard	nadia	missing	1779	20210722
83872	borcherding	jerome	missing	1935	20110909	20045	borcherding	jerome	missing	1935	20110304
37872	davis	alice	jane	1935	20120904	71371	davis	alice	jane	1935	20120814
188813	myers	anne	randall	1935	20121228	55717	myers	anne	randall	1935	20120323
281242	derstine	jill	b	1936	20180815	49000	derstine	jill	b	1936	20180807
62162	clark	sandra	l	1937	20010607	64233	clark	sandra	l	1937	20090319

Demo in R:

1. Install "RecordLinkage" package and load all packages used in this demo

2. Read in and prepare demo data

File A – Nurse License Records ("`sub_a`")

	LastName	FirstName	MiddleName	BirthYear	FirstIssueDate	LastInitial
1	spear	kelsey	thurman	1987	20100218	s
2	savella	jona marie	fernandez	1987	20211121	s
3	streeter	kelsey	marie	1987	20210219	s
...						
155	xavier	anjaly	missing	1987	20160914	x
...						

File B – Nurse Temporary Permit Records ("`sub_b`")

	LastName	FirstName	MiddleName	BirthYear	FirstIssueDate	LastInitial
1	sloope	dena	annette	1987	20210528	s
2	shifflett	allison	whitney	1987	20210113	s
3	schmitt	shannon	danielle	1987	20150624	s
4	sandner	jessica	eleanor	1987	20140731	s
...						

Demo in R:

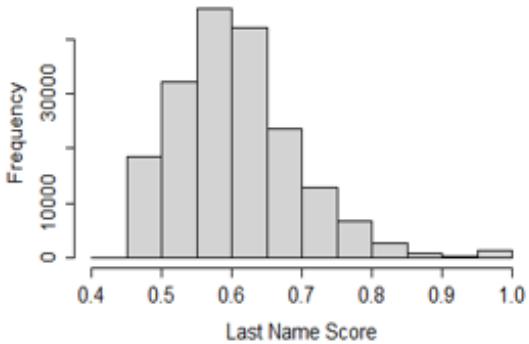
3. Compare Records After Blocking

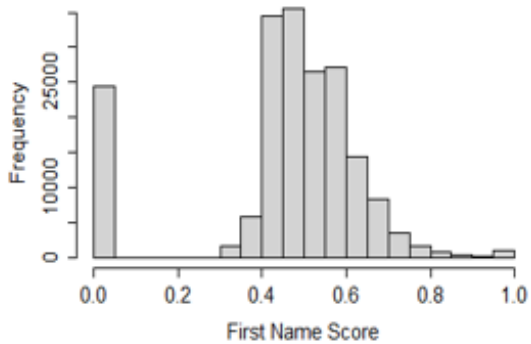
```
head(comparison_vectors_demo, 4)
```

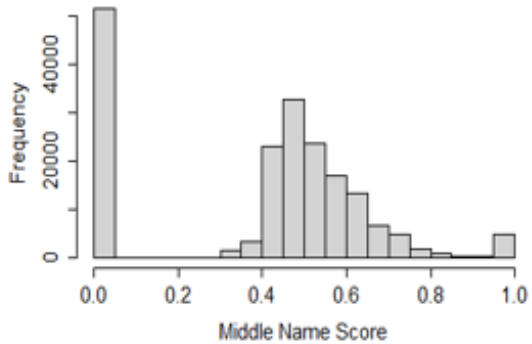
Comparison Vectors -

id1	id2	LastName	FirstName	MiddleName
1	1	0.5100000	0.4722222	0.0000000
1	2	0.4933333	0.5396825	0.5238095
1	3	0.5028571	0.0000000	0.0000000
1	4	0.7085714	0.5396825	0.5238095
...				

4. Classify Links







Choose Threshold = 0.85

Demo in R:

5. Obtain Linked Data Set

```
head(lds_demo, 4)
```

id.a	LastName	FirstName	MiddleName	BirthYear	FirstIssueDate	id.b	LastName_TP	FirstName_TP	MiddleName_TP	BirthYear_TP	FirstIssueDate_TP
3	streeter	kelsey	marie	1987	20210219	323	streeter	kelsey	marie	1987	20200512
5	staudinger	katherine	jane	1987	20180520	224	staudinger	katherine	jane	1987	20180428
7	shifflett	allison	whitney	1987	20210406	2	shifflett	allison	whitney	1987	20210113
4	swain	carrie	elizabeth	1987	20200825	172	swain	carrie	elizabeth	1987	20200413
...											

Consequences of Linkage Error

Linkage error and Post-Linkage Data Analysis

There are two types of errors associated with RL:

1) False matches (**Mismatches**):

- Non-matches erroneously designated as matches.

2) False non-matches (**Missed-matches**):

- Matches that were not identified as such.

- What are the sequences of such errors on subsequent data analysis (**Post-Linkage Data Analysis)**, PLDA)?
- Can PLDA be adjusted to alleviate adverse effects of linkage error?

Consequences of false non-matches

Missed matches can be conceptualized as follows:

File A	File B
a_1	b_1
a_2	b_2
⋮	⋮
⋮	⋮
a_M	b_M

Ideal file w/o missing any matches

File A	File B
a_1	b_1
a_2	?
⋮	⋮
⋮	⋮
a_M	b_M

File missing match no. 2

Ignoring missing matches is thus comparable to running a complete-case analysis on a data set with missing values.

Consequences of false non-matches

A common approach is to avoid false matches (mismatches) at all costs and to restrict attention to “safe” correct matches.

However, it needs to be noted that such an approach comes with the following drawbacks:

- Danger of Selection Bias,
- Loss of Statistical Power.

The approach is often not necessary, since there are effective ways of addressing **mismatch error**.

Consequences of false non-matches

Example: Linkage of the Health & Retirement Study (HRS) and the Census Business Register (BR).

Only 70% of HRS respondents consent to SSA linkage, and hence lack an Employer Identification Number (EIN) – linkage based on the establishment addresses may yield thousands of potential matches in the BR. Abowd *et al.*, 2022.

Despite considerable sophistication in terms of RL strategy, the subsets of linked & unlinked respondents are still different:

	linked (92%)	**unlinked** (8%)
∅Age	57.6	56.9
%White	68	57
%Black	22	24
%Hispanic	14	26
%Native born	87	69
$Earnings	43.2k	33.3k

Source: based on a recent presentation by Dhiren Patki in the ISR record linkage series.

Consequences of false matches

Mismatches tend to introduce data contamination.

Specific consequences are:

- Outliers,
- Attenuated relationships, similar to what is observed in the literature on measurement error,
- Reduced model fit,
- Biased parameter estimates,
- Inflated standard errors.

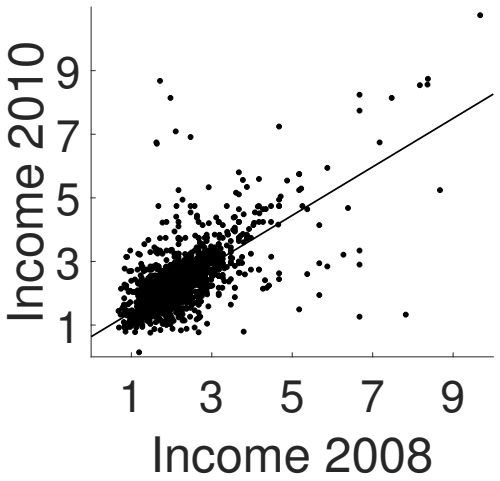
Consequences of Mismatch Error: Linear Regression

Setup:

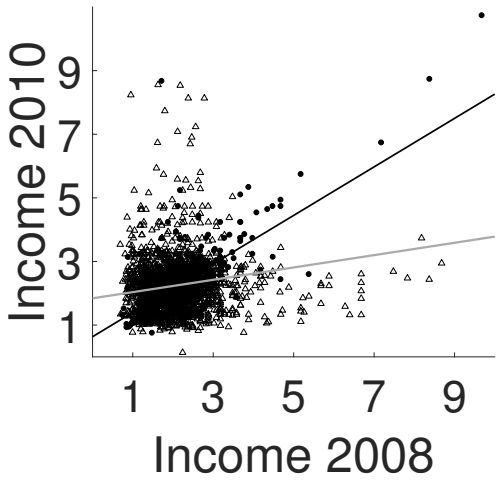
File A: covariates $\mathbf{x}$. File B: response y.

Consequences of mismatch error are well-documented (Neter al., 1965; Scheuren & Winkler, 1997; Lahiri & Larsen, 2005; Wang et al., 2022))

w/o mismatches



w/ mismatches



	w/o	w/
intercept	0.63	1.84
slope	0.76	0.19
residual variance	0.38	0.78
R^2	0.52	0.03

Consequences of Mismatch Error: PCA

Data: Beijing climate data set analyzed in Slawski, Ben-David, Li (2020).

File A: measurements on temperature, air pressure, dew point, precipitation, wind speed, CO.

File B: measurements on PM2.5, PM10, SO_2, NO_2, O_3.

Percentage of variance explained by the top principal components (w/ and w/o mismatches):

# components	1	2	3	4	5
% variance explained w/o	40	66	76	85	90
% variance explained w/	37	62	72	82	87

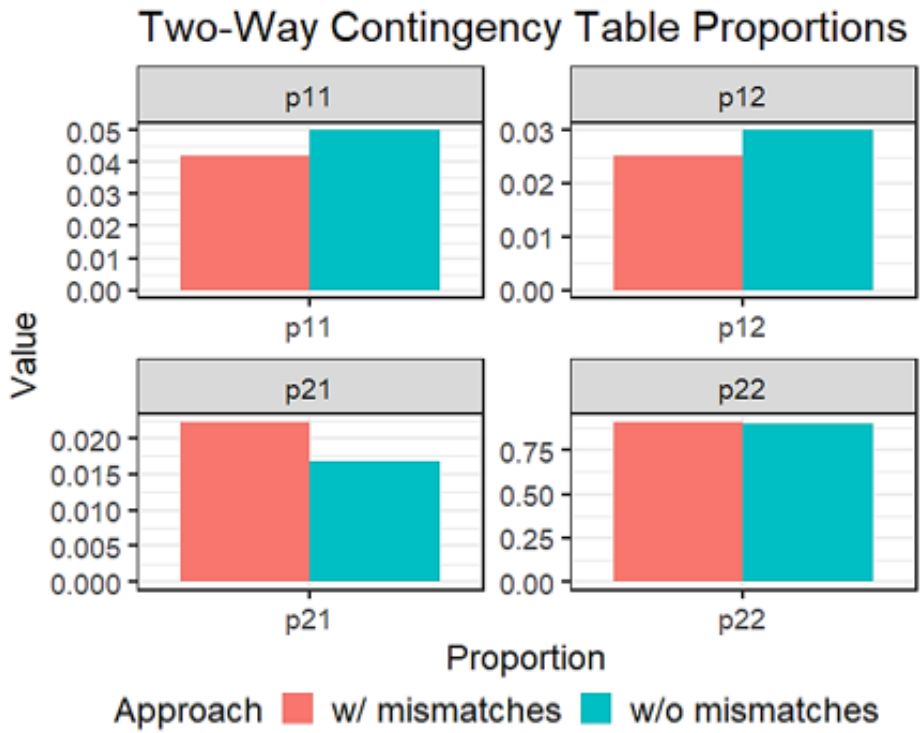
Consequences of Mismatch Error: Contingency Tables

Slawski, West, Bukke, Wang, Diao, Ben-David (2023)

File A: Health & Retirement Study (HRS) data.
File B: Centers for Medicare & Medicaid Services (CMS) data.

Mismatch rate of $\sim \frac{59}{359} \approx 0.164$

		Y Administrative Record (CMS)	
On Nursing Home Residence:		Yes	No
X Self-Report (HRS)	Yes	p_{11}	p_{12}
	No	p_{21}	p_{22}



Evaluation:

Using the known correct matches as a benchmark, we compute

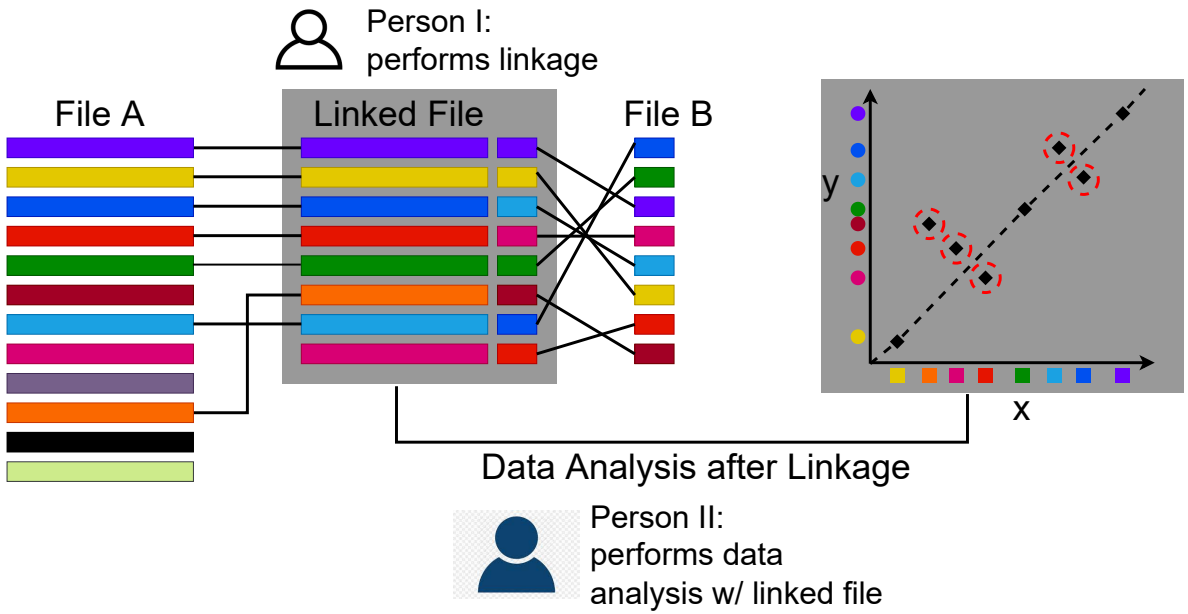
- Mean relative absolute error (MRAE) for the proportions,
- Kullback-Leibler divergence (KLD) as a measure of distance between the w/ vs. w/o proportions,
- One-sample chi-square goodness of fit (GOF) for the proportions

We also compute the chi-square measure of association between the two variables and the Kappa statistic (as a measure of agreement).

	w/o	w/
MRAE	0.0000	0.1685
KLD	0.0000	0.0020
GOF	0.0000	1.4780
Association	146.7991	125.1211
Kappa	0.6569	0.6130

Mitigation Strategies

Primary vs. Secondary Analysis



Primary Analysis:
Access to individual files A and B. RL and subsequent data analysis can be performed jointly, with the propagation of uncertainty.

Secondary Analysis (focus in this short course):
Access only to the linked file, not the individual files. Information about underlying RL may be available, but limited.

Hierarchical Models for Primary Analysis

e.g., Gutman *et al.*, 2013; Tancredi & Liseo, 2015; Steorts *et al.*, 2018.

Let

- $\mathbf{\Gamma}$ denote the set of all pairs of comparison vectors,
- Π^* be the binary matrix encoding the correct matching of records from file A and B,
- $\mathbf{X}$ and $\mathbf{Y}$ represent variables exclusive to file A and file B, respectively, whose relationship is of interest and is characterized via a parameter $\boldsymbol{\theta}$.

Then a basic template is as follows:

$$
\begin{aligned}
&(0) \quad \text{priors for } \Pi^*,\ \boldsymbol{\theta} \\
&(1) \quad \mathbf{\Gamma} | \Pi^* \\
&(2) \quad (\mathbf{X}, \mathbf{Y}) | \Pi^*, \boldsymbol{\theta}
\end{aligned}
$$

We can conduct posterior inference for $\boldsymbol{\theta}$ to account for uncertainty resulting from RL.

Linear Regression w/ Linked Data: Lahiri & Larsen

Lahiri & Larsen (2005) were the first to propose a simple fix to the problem.

Specifically, by differentiating between the observed y_i's in the linked file and the latent true y_i^*, i.e., true responses associated with $\mathbf{x}_i$'s, it assumes that

$$y_i = \begin{cases} y_i^* & \text{with probability } q_{ii} \\ y_j^* & \text{with probability } q_{ij} \end{cases}$$

In particular, $E(\mathbf{y} \mid \mathbf{y}^*) = Q\mathbf{y}^*$.

Now if $E(\mathbf{y}^* \mid X) = X\beta^*$, then

$$\begin{aligned} E(\mathbf{y} \mid X) &= E\left(E(\mathbf{y} \mid \mathbf{y}^*, \mathbf{X})\right) \\ &= QE(\mathbf{y}^* \mid X) \\ &= QX\beta^*. \end{aligned}$$

This shows that the naive OLS regression with $\mathbf{y}$ results in a biased estimator of β^*.

The same equation shows that $(X^\top Q^\top QX)^{-1}X^\top Q^\top \mathbf{y}$ is an unbiased estimator.

The setting of Lahiri & Larsen, implicitly, assumes that $\mathbf{y} = \Pi^*\mathbf{y}$, where Π^* is generated from a random permutation with the expected value Q.

Chambers (2006) proposes a generalization of the Lahiri & Larsen estimator. Given

$$\mathbf{E}[\mathbf{X}^\top \mathbf{Q}\mathbf{X}\beta^* - \mathbf{X}^\top \mathbf{y}|\mathbf{X}] = \mathbf{0},$$

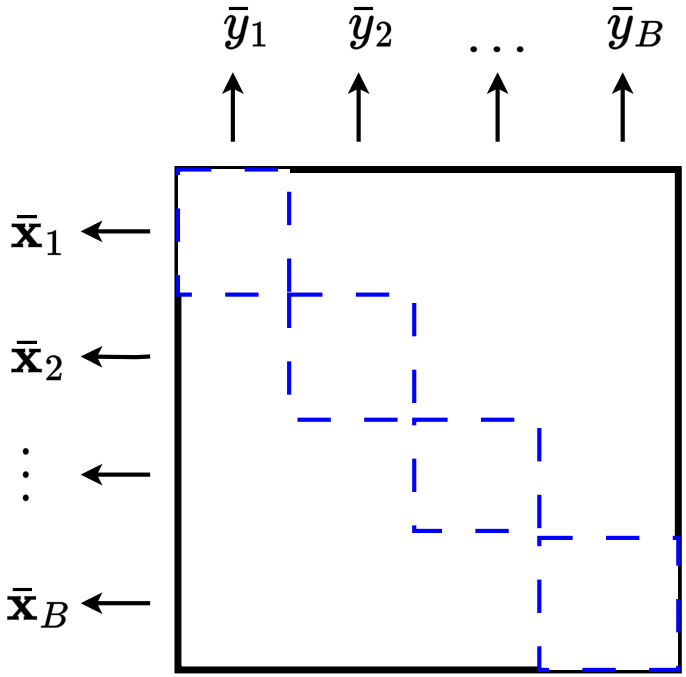
we solve the corresponding unbiased estimation equation.

The assumption that the matrix $\mathbf{Q}$ is known is often unrealistic, particularly in the primary analysis setting.

For this reason, Chambers (2009) proposes the **e**xchangeable **l**inkage **e**rror model (ELE).

In a nutshell, under the ELE, mismatch rates are assumed to be known and constant within blocks of pairs defined by blocking variables.

If the observations can be blocked, a safe strategy is to average the $\mathbf{x}'s$ and y's in each bock, and then perform regression with the resulting averages.



This strategy safeguards against all possible ways of mismatch error, i.e., *regardless of* Π^*, as long as mismatch error does not occur across blocks.

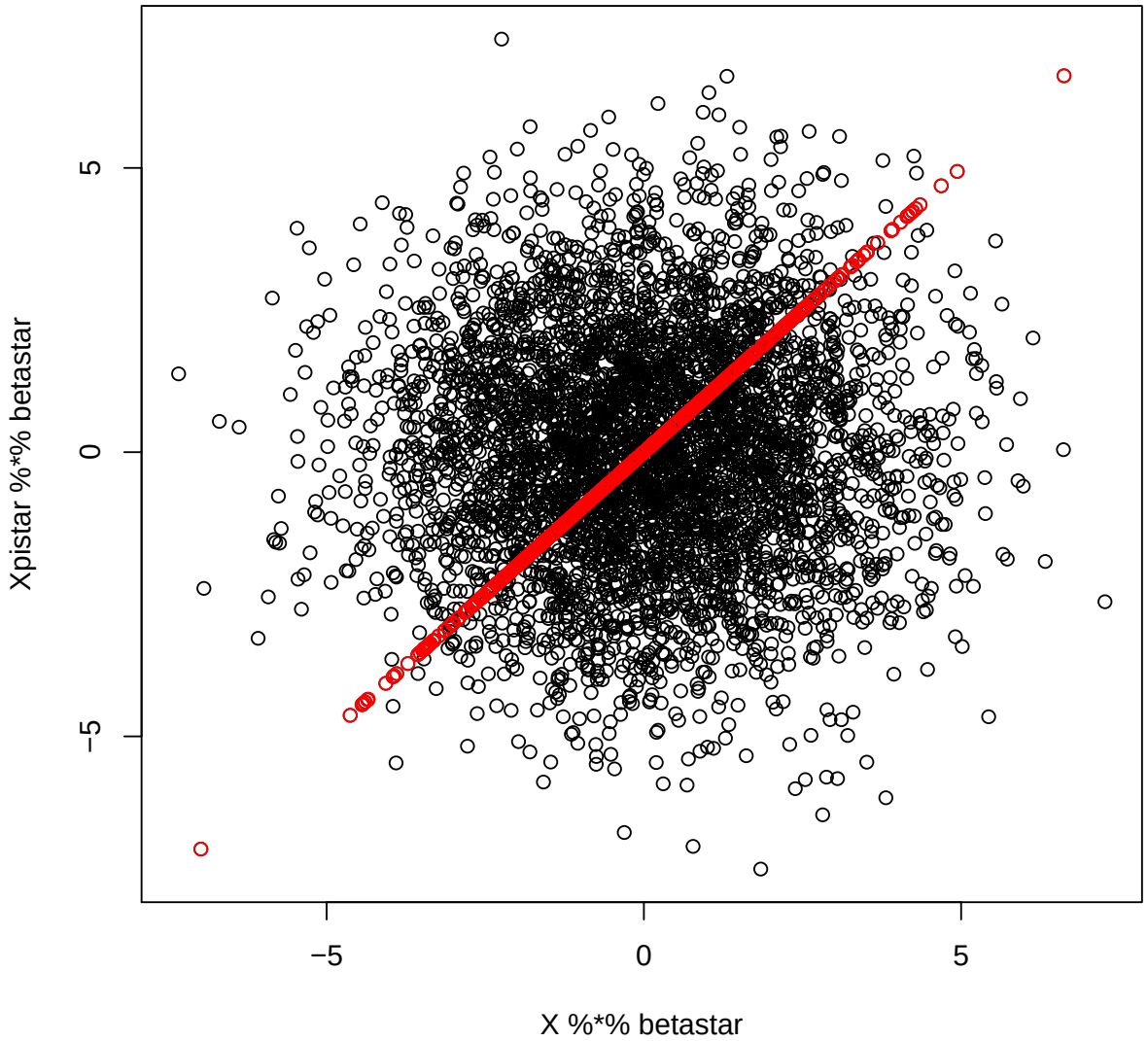
Blockwise averaging can be obtained **as a special case of the Lahiri-Larsen estimator**.

Toy example: blockwise averaging

```
#5k observations, divided into 500 blocks
n <- 5000
blockindex <- sample(1:500, n, replace = TRUE)
d <- 5
X <- matrix(rnorm(n*d), nrow = n, ncol = d)
betastar <- c(1,1,-1,-1,0)
...
...
...
sigma <- 0.2
xi <- sigma * rnorm(n)
y <- Xpistar %*% betastar  + xi
```

Toy example (c'ted)

Response before and after linkage (shuffling):



Toy example (c'ted)

```
source("../code/lahiri_larsen.R")
#approach I (general Q, but slow if Q has block structure)
Q <- generate_Q_block(blockindex)
betaQ <- lahiri_larsen(X, y, Q)


#approach II (tailored to block structure, faster)
betaQcheck <- coef(lahiri_larsen_block(X, y, blockindex))

# naive least squares
coef(lm(y ~ X - 1))
X1          X2          X3          X4          X5
0.07205  0.11612075 -0.10753189 -0.07813942  0.01627887

# Lahiri-Larsen
betaQcheck
XbarX.1     XbarX.2     XbarX.3     XbarX.4     XbarX.5
1.0092      0.9905      -0.9975     -1.0014     -0.0076
```

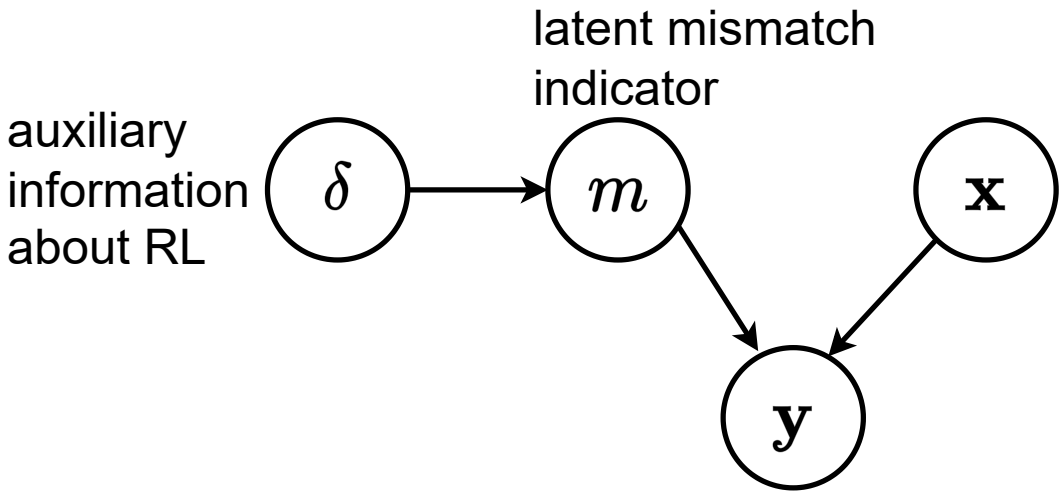
Summary: Lahiri-Larsen-Chambers estimator(s)

The approach is conceptually simple and can be effective depending on the distribution of Π^* and the degree of correct specification of $\mathbf{Q}$, even for high mismatch rates.

- Unbiasedness of the approach requires averaging over Π^*, except in the blockwise averaging approach (which, however, does not yield an efficient estimator in general).
- Inference (standard errors etc.) not clear. Han & Lahiri (2019) propose a jacknife approach that also accounts for potential uncertainty in $\mathbf{Q}$.
- Approach can be (MSE)-inefficient (Slawski, Diao, Ben-David (2020)).

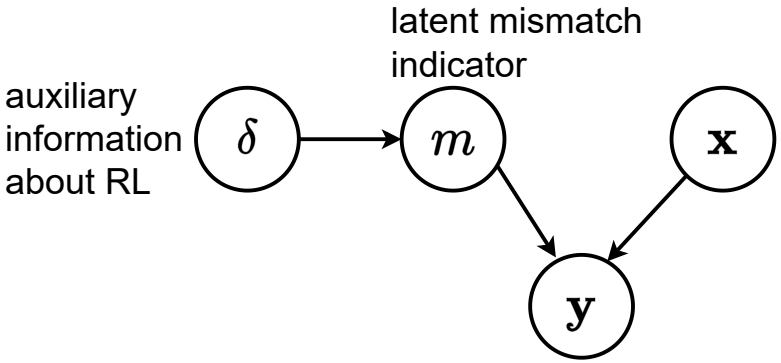
Mixture model approach

Originally proposed in Slawski, Diao, Ben-David '21 and extended substantially in Slawski, West, Bukke, Wang, Diao, Ben-David '23.
Related to Hof & Zwinderman '14 on PLDA in the primary setting.



$$\mathbf{y}|m=1,\mathbf{x} \sim f_{\mathbf{y}} \qquad \mathbf{y}|m=0,\mathbf{x} \sim f(\mathbf{y}|\mathbf{x};\theta)$$

- Latent binary mismatch indicator m, (possibly) modeled conditionally on info about RL δ,
- "Standard model" for pair $(\mathbf{x},\mathbf{y})$ if associated $m=0$ (right),
- Independence model $\mathbf{x} \perp\!\!\!\perp \mathbf{y}$ if associated $m=1$ (left).



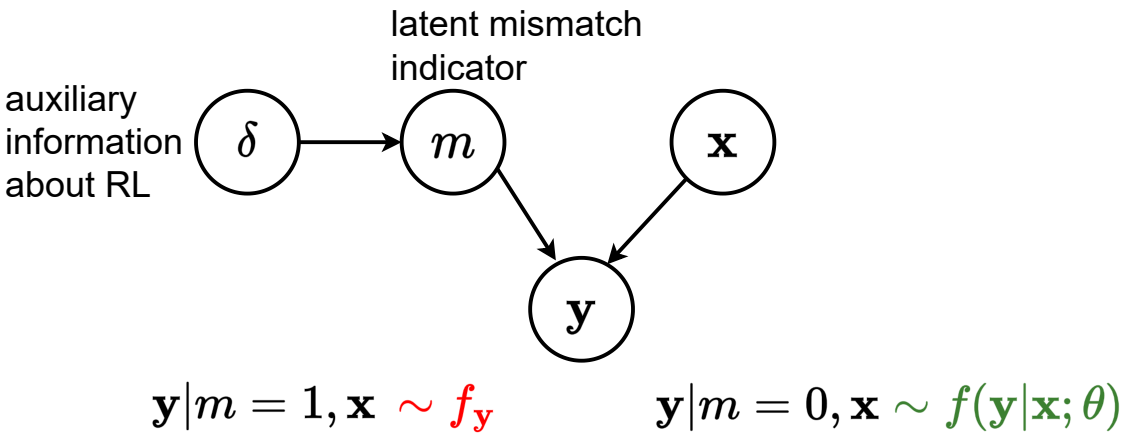
$\mathbf{y}|m=1,\mathbf{x} \sim f_{\mathbf{y}}$ $\quad$ $\mathbf{y}|m=0,\mathbf{x} \sim f(\mathbf{y}|\mathbf{x};\theta)$

Assumption 1 – Independence for mismatches

$\mathbf{y} \perp\!\!\!\perp \mathbf{x} \,|\, m=1$

Satisfied if we are willing to assume that distinct records are independent. Can be violated if mismatches occur within correlated blocks of observations.

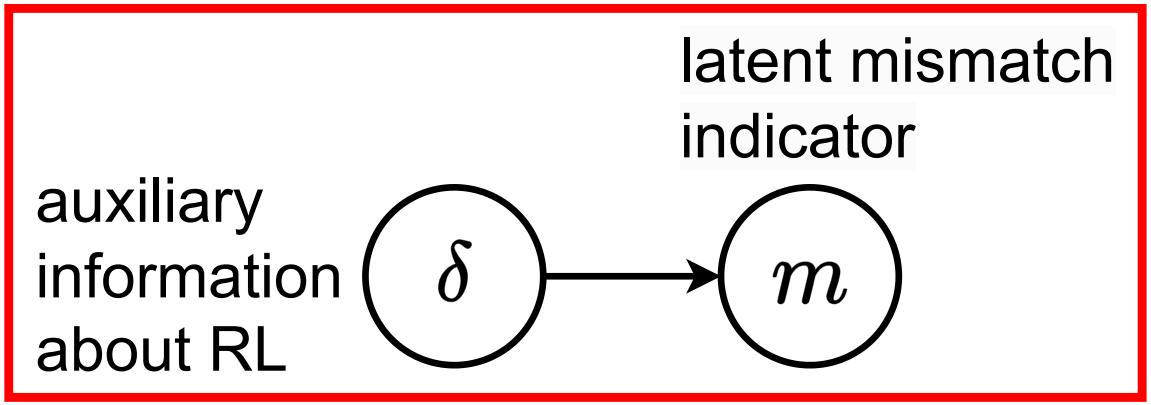
Remark. The above DAG specifically addresses regression setups. Straightforward to generalize when the roles of $\mathbf{x}$ and $\mathbf{y}$ are symmetric.



Assumption 2 – Mismatch error does not depend on $(\mathbf{x}, \mathbf{y})$

The models for m and for $(\mathbf{x}, \mathbf{y})$ are kept strictly separate.

m only depends on δ but not on $\mathbf{x}$. This assumption is stronger than those of other methods but renders inference much more tractable.



The covariates δ for the latent indicator m can be the following:

- . . . An intercept – corresponding to a constant mismatch rate model,
- . . . Block indicators from RL – corresponding to mismatch rates varying across blocks,
- . . . Output from probabilistic RL (e.g., confidence in the correctness of a match),
- . . . Comparison variables used during probabilistic RL.

A standard approach is to use a logistic regression model for the relationship between $\boldsymbol{\delta}$ and m:

$$\mathbf{P}(m_i = 1|\boldsymbol{\delta}_i; \boldsymbol{\gamma}) = \frac{\exp(\gamma_0 + \gamma_1\delta_{1,i} + \ldots + \gamma_q\delta_{q,i})}{1 + \exp(\gamma_0 + \gamma_1\delta_{1,i} + \ldots + \gamma_q\delta_{q,i})}, \quad 1 \leq i \leq n,$$

Note that estimating the parameters of such a model is more challenging than in a vanilla binary regression problem since the mismatch indicators are not observed.

Therefore, it can be helpful to incorporate prior information on the underlying mismatch rate by imposing a corresponding constraint. For computational convenience, such a constraint is imposed on the logit scale, i.e.,

$$\gamma_0 + \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\delta}_i^\top \boldsymbol{\gamma} \leq b,$$

where $b = \text{logit}(\text{mismatch rate})$.

Inference

Maximize the pseudo-likelihood resulting from the postulated model with respect to the unknown parameters:

$$L(\boldsymbol{\theta}, \boldsymbol{\gamma}) = \prod_{i=1}^{n} \left\{ f(\mathbf{y}_i | \mathbf{x}_i; \boldsymbol{\theta}) \, \mathbf{P}(m_i = 0 | \boldsymbol{\delta}_i; \boldsymbol{\gamma}) + f_{\mathbf{y}}(\mathbf{y}_i) \, \mathbf{P}(m_i = 1 | \boldsymbol{\delta}_i; \boldsymbol{\gamma}) \right\}$$

Inference (standard errors) via asymptotic theory for composite maximum likelihood estimators.

Alternative: hierarchical Bayes.

The framework can be applied to various statistical models (GLMs, semi-parametric regression, Cox regression, contingency table analysis, small area models, . . .).

$$L(\boldsymbol{\theta}, \boldsymbol{\gamma}) = \prod_{i=1}^{n} \left\{ f(\mathbf{y}_i|\mathbf{x}_i; \boldsymbol{\theta})\, \mathbf{P}(m_i = 0|\boldsymbol{\delta}_i; \boldsymbol{\gamma}) + \boxed{f_{\mathbf{y}}(\mathbf{y}_i)}\, \mathbf{P}(m_i = 1|\boldsymbol{\delta}_i; \boldsymbol{\gamma}) \right\}$$

Estimation of the Marginal PDF $\boxed{f_{\mathbf{y}}(\mathbf{y}_i)}$:

Note: not affected by mismatch error – only involves variables from a single file.

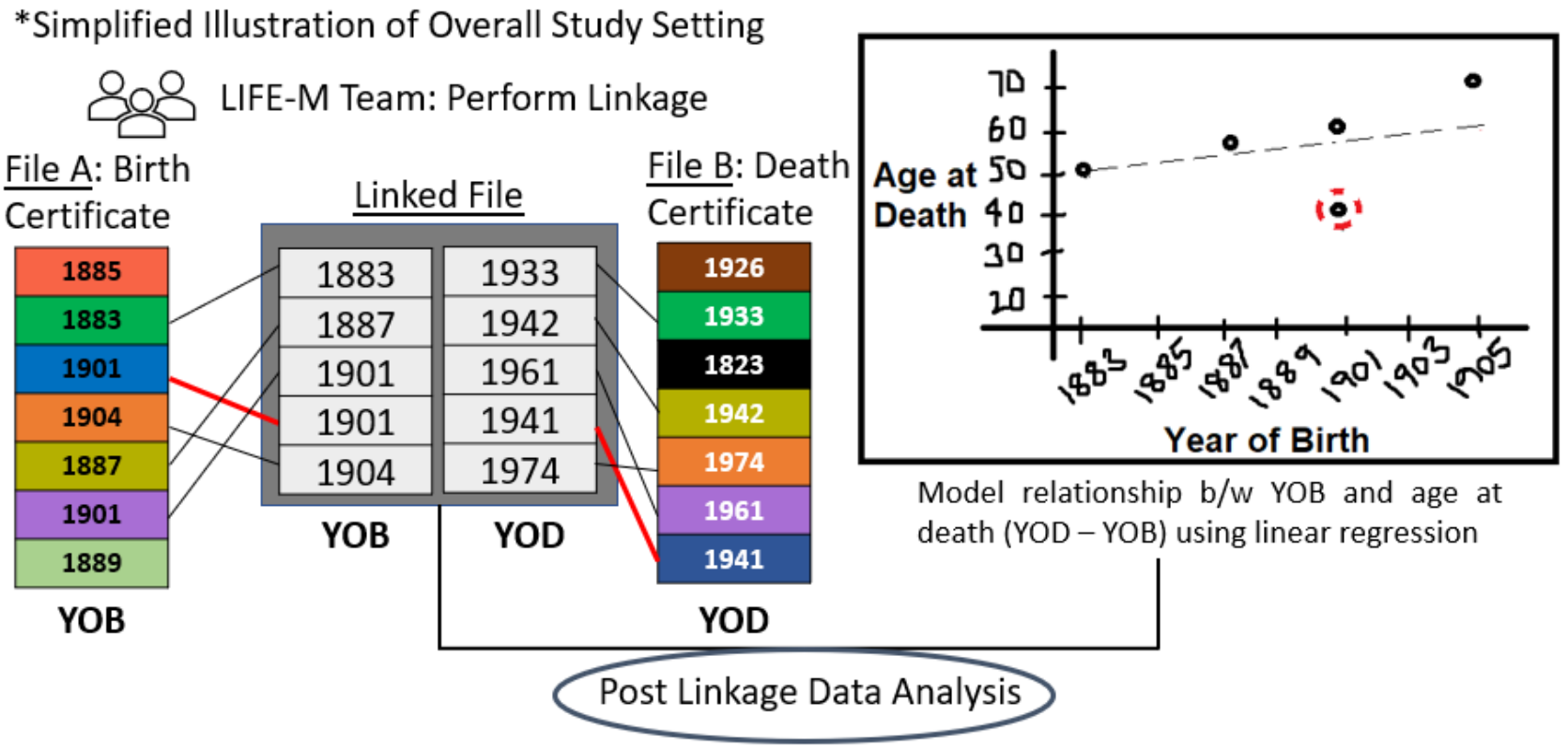
Options:

- Empirical probability mass function (if the cardinality of the range of $\mathbf{y}$ is small),
- Kernel Density Estimation,
- Parametric models,
- Multi-stage (updated with $\boldsymbol{\theta}$).

Case Studies

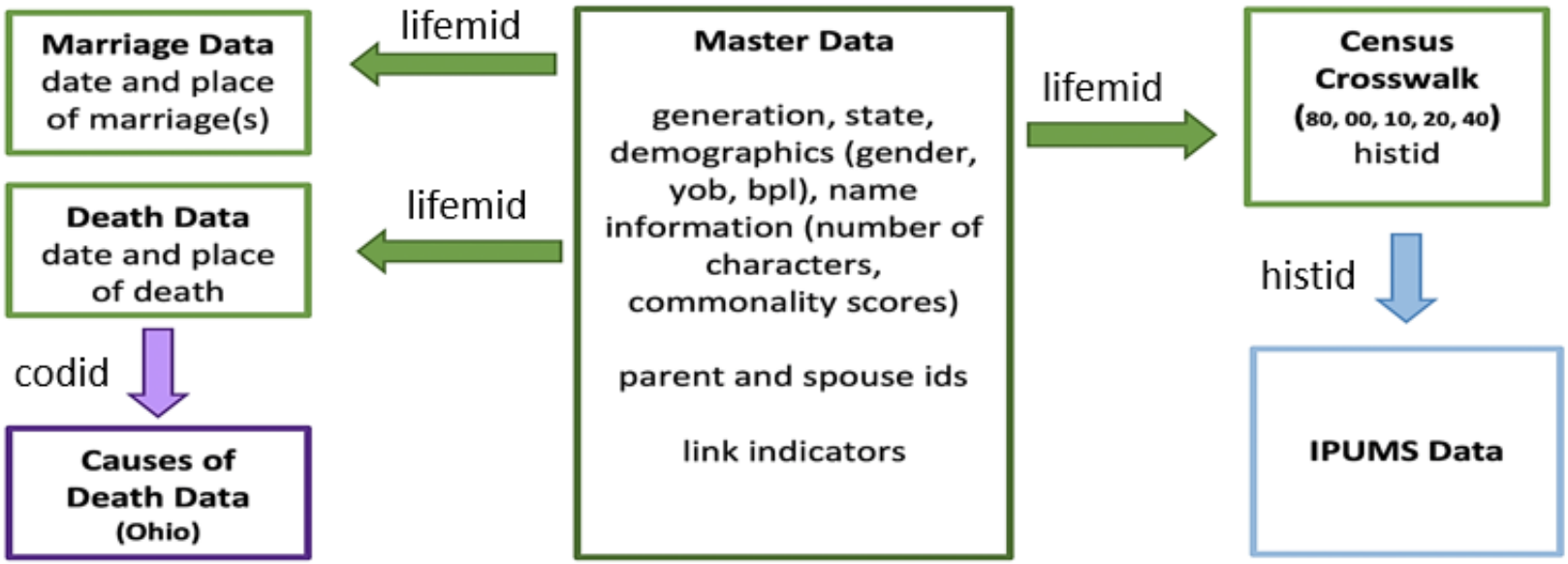
Case Study I: Life-M

Slawski, West, Bukke, Wang, Diao, Ben-David (2023)



LIFE-M Project:

Longitudinal, Intergenerational Family Electronic Micro-Database (`life-m.org`)



Record Linkage Process:

LIFE-M team uses a hybrid of two record linkage procedures:

- "hand-linked" – manual linkage by trained research assistants
- "machine-linked" – probabilistic record linkage without clerical review
  - Anticipate ~5% mismatch rate

	First name	Middle name	Last name	Date of birth	Sex	Race	BPL	F's full name	M's full name	M's maiden name	F's BPL	M's BPL
LIFE-M	X	X	X	X	X		Town or county	X	X	X	Town or county*	Town or county*

Notes: BPL=birthplace, F=Father, M=Mother. IPUMS linkage variables are taken from Ruggles (2002: Table 1). *Not available in some collections/years.

(LIFE-M Documentation)

Linked Data Set of Interest:

156,453 LIFE-M individuals

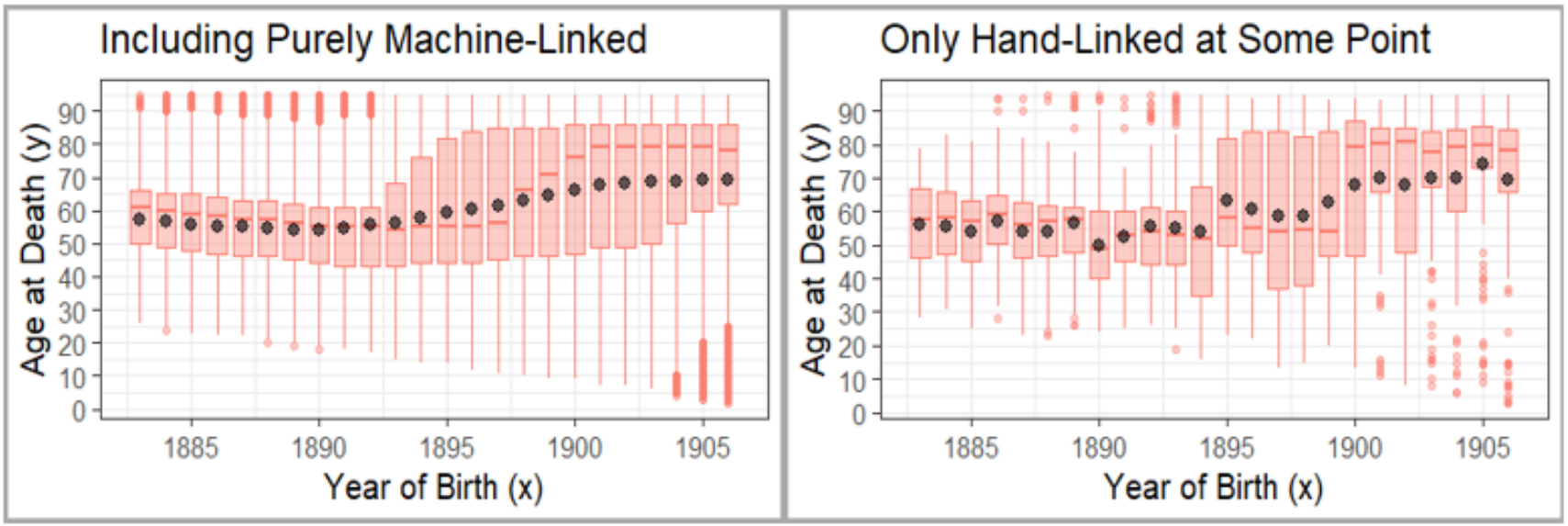
- 2, 159 hand-linked at some level
- 154, 294 purely machine-linked

	hndlnk	yob	yod	age_at_death	commf	comml	...
1	0	1905	1988	83	0.77	0.45	...
2	1	1883	1962	79	0.93	0.08	...
	...	...	...	...	...	...	...
156,453	0	1886	1944	58	0.89	0.80	...

Case Study Set-Up:

Study of the relationship between age at death (y) and year of birth (x)

- Linear regression w/ cubic polynomial

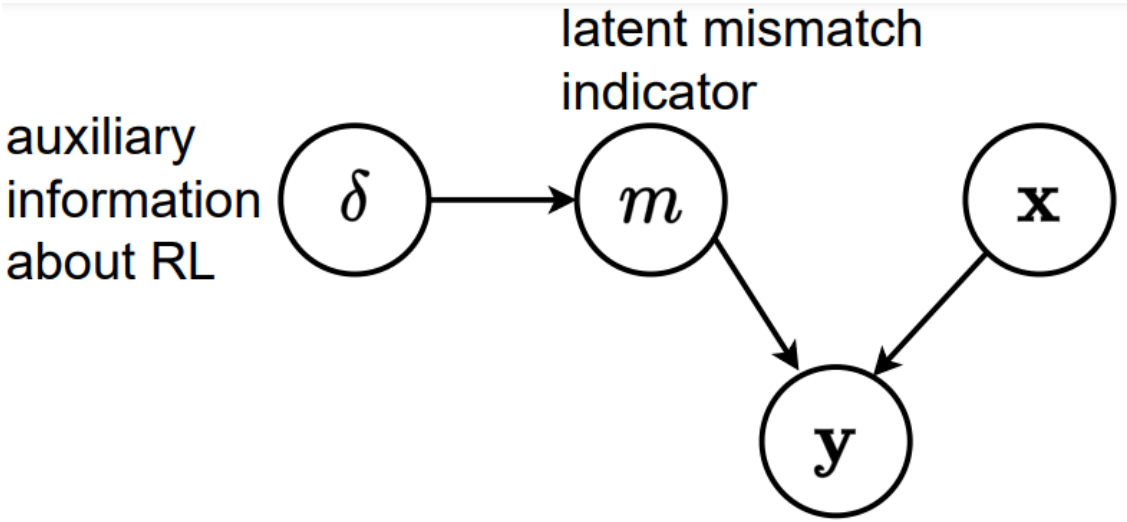


Plain Linear Regression

$y_i \mid x_i \sim N(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3, \sigma^2)$, for $1 \leq i \leq n$.

- y_i: Age of Death of an individual "i"
- x_i: Year of Birth of an individual "i"

Adjusted linear regression

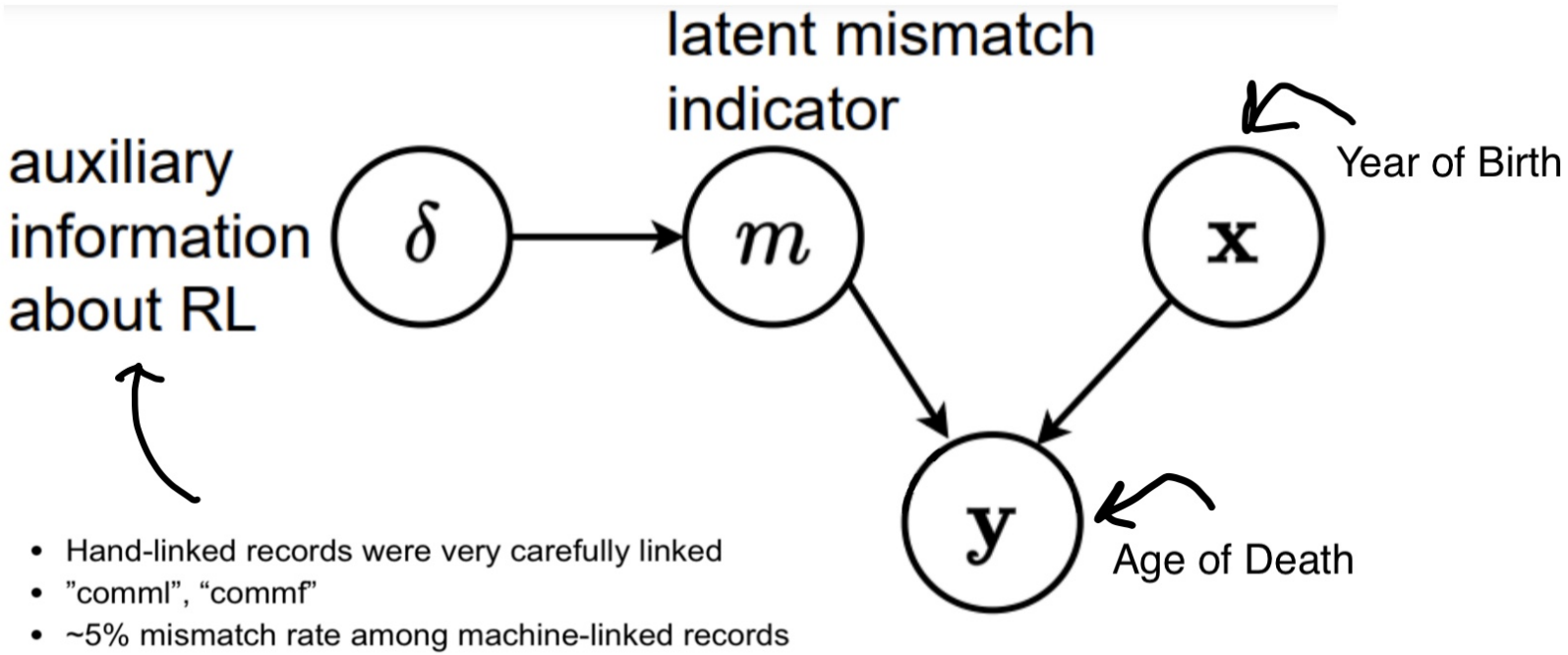


Plain Linear Regression

$y_i \mid x_i \sim N(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3, \sigma^2)$, for $1 \leq i \leq n$.

- y_i: Age of Death of an individual "i"
- x_i: Year of Birth of an individual "i"

Adjusted linear regression

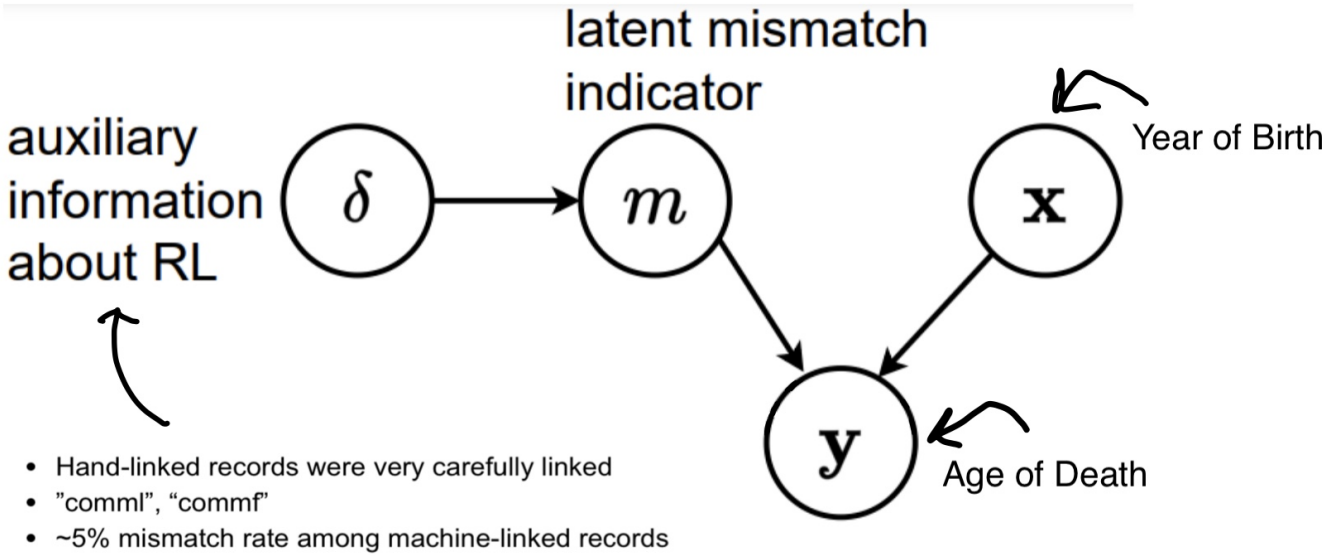


Plain Linear Regression

$\mathbf{y}_i \mid \mathbf{x}_i \sim N(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3, \sigma^2)$, for $1 \leq i \leq n$.

- y_i: Age of Death of an individual "i"
- x_i: Year of Birth of an individual "i"

Adjusted linear regression



$\mathbf{y}_i \mid m_i = 1, \mathbf{x}_i \sim N(\mu, \tau^2)$ $\quad$ $\mathbf{y}_i \mid m_i = 0, \mathbf{x}_i \sim N(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3, \sigma^2)$

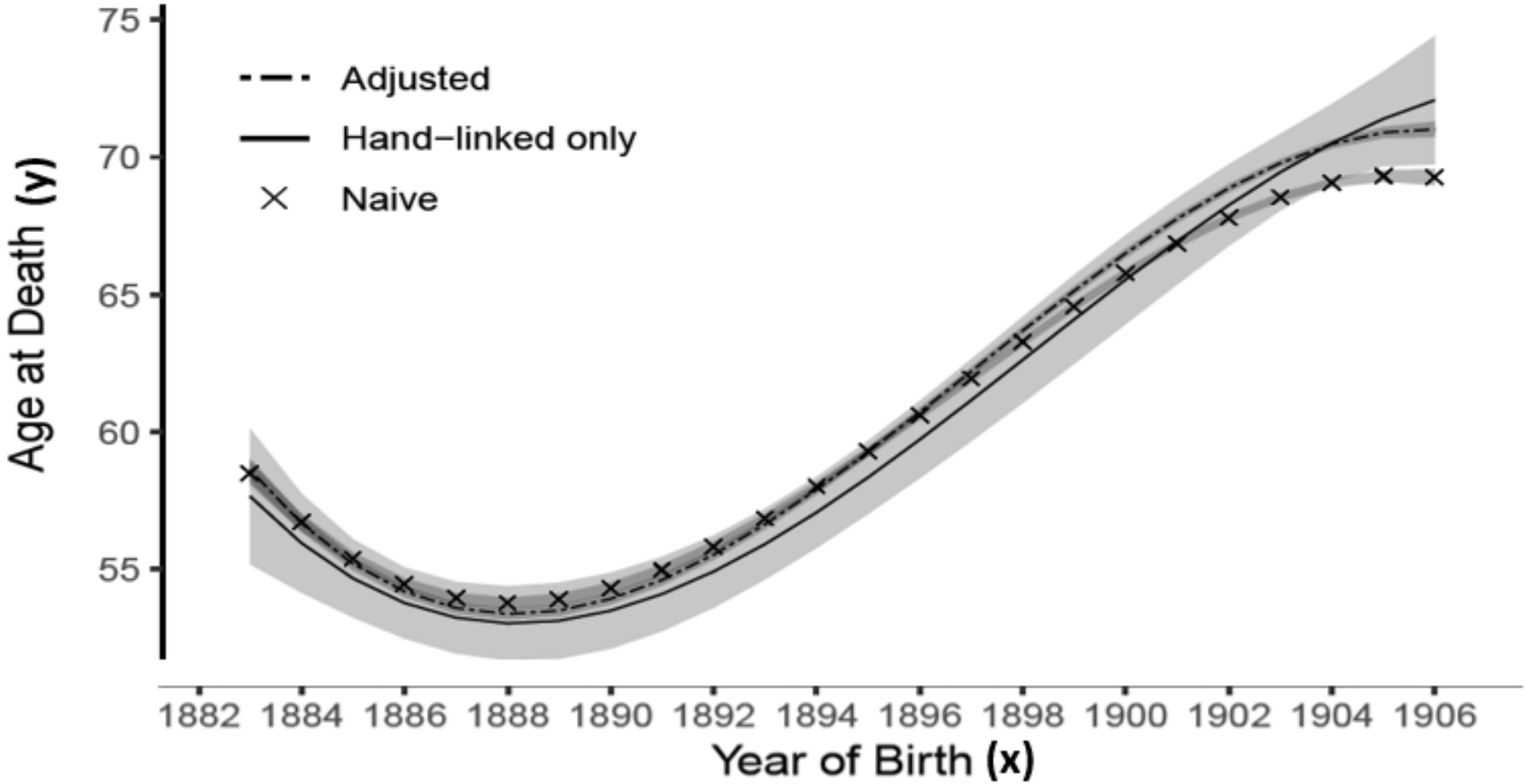
$$\mathbf{m}_i \mid commf_i, comml_i \sim \text{Bernoulli}\left(\frac{\exp(\gamma_0 + \gamma_1 commf_i + \gamma_2 comml_i)}{1 + \exp(\gamma_0 + \gamma_1 commf_i + \gamma_2 comml_i)}\right)$$

Summary of Results:

Parameter	Naïve (n = 156,453)	Adjusted (n = 156,453)	HL-Only (n = 2,159)
$\widehat{\beta_0}$	58.5 (0.2)	58.6 (0.1)	57.7 (1.26)
$\widehat{\beta_1}$	-46.7 (1.8)	-51.0 (1.5)	-44.2 (11.6)
$\widehat{\beta_2}$	130.4 (4.0)	140.3 (3.9)	118.6 (27.9)
$\widehat{\beta_3}$	-72.9 (2.5)	-76.8 (2.6)	-59.9 (18.5)
$\hat{\sigma}$	21.2 (0.04)	20.7 (0.06)	19.0 (0.29)

Parameter	Adjusted
$\widehat{\gamma_0}$	-6.0 (0.5)
$\widehat{\gamma_1}$	-1.4 (0.5)
$\widehat{\gamma_2}$	7.2 (0.3)

Predicted Ages at Death for Birth Cohorts:



Demo in R: Load Libraries and Data Preparation

1. Install and load “pldamixture” package

```
install.packages("pldamixture_0.0.0.9000.tar.gz", repos
= NULL, type = "source")
library(pldamixture)
```

2. Read in demo data (2 (HL): 1 (ML) SRS Ratio – 3,238 LIFE-M Individuals)

```
data("demodata")
```

	lifemid	yob	yod	surv_age	hndlnk	commf	comml	uyob
1	0883v	1883	1962	79	Hand-Linked At Some Level	0.93	0.08	0.00000000
2	0qwwh	1886	1967	81	Hand-Linked At Some Level	0.76	0.48	0.13043478
3	6g859	1892	1948	56	Hand-Linked At Some Level	0.81	0.54	0.39130435
4	bz25j	1883	1926	43	Hand-Linked At Some Level	0.79	0.95	0.00000000
5	h2gm2	1891	1976	85	Hand-Linked At Some Level	0.86	0.81	0.34782609

Demo in R: Naïve Analysis

3. Plain Linear Regression: Naïve (n = 3,238)

```
naive_fit <- lm(surv_age ~ poly(uyob,3,raw = TRUE),
                demodata)
summary(naive_fit)
```

Demo in R: Naïve Analysis Results

3. Plain Linear Regression: Naïve (n = 3,238)

```
Call:
lm(formula = surv_age ~ poly(uyob, 3, raw = TRUE), data = demodata)

Residuals:
    Min      1Q  Median      3Q     Max
-67.475 -11.669   3.598  12.956  41.674

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                    57.715      1.137  50.754  < 2e-16 ***
poly(uyob, 3, raw = TRUE)1    -41.687     10.175  -4.097 4.29e-05 ***
poly(uyob, 3, raw = TRUE)2    111.232     23.907   4.653 3.41e-06 ***
poly(uyob, 3, raw = TRUE)3    -56.785     15.578  -3.645 0.000271 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.67 on 3234 degrees of freedom
Multiple R-squared:  0.09362,   Adjusted R-squared:  0.09278
F-statistic: 111.3 on 3 and 3234 DF,  p-value: < 2.2e-16
```

Demo in R: HL-Only Analysis

4. Plain Linear Regression: Hand-Linked Only (n = 2,159)

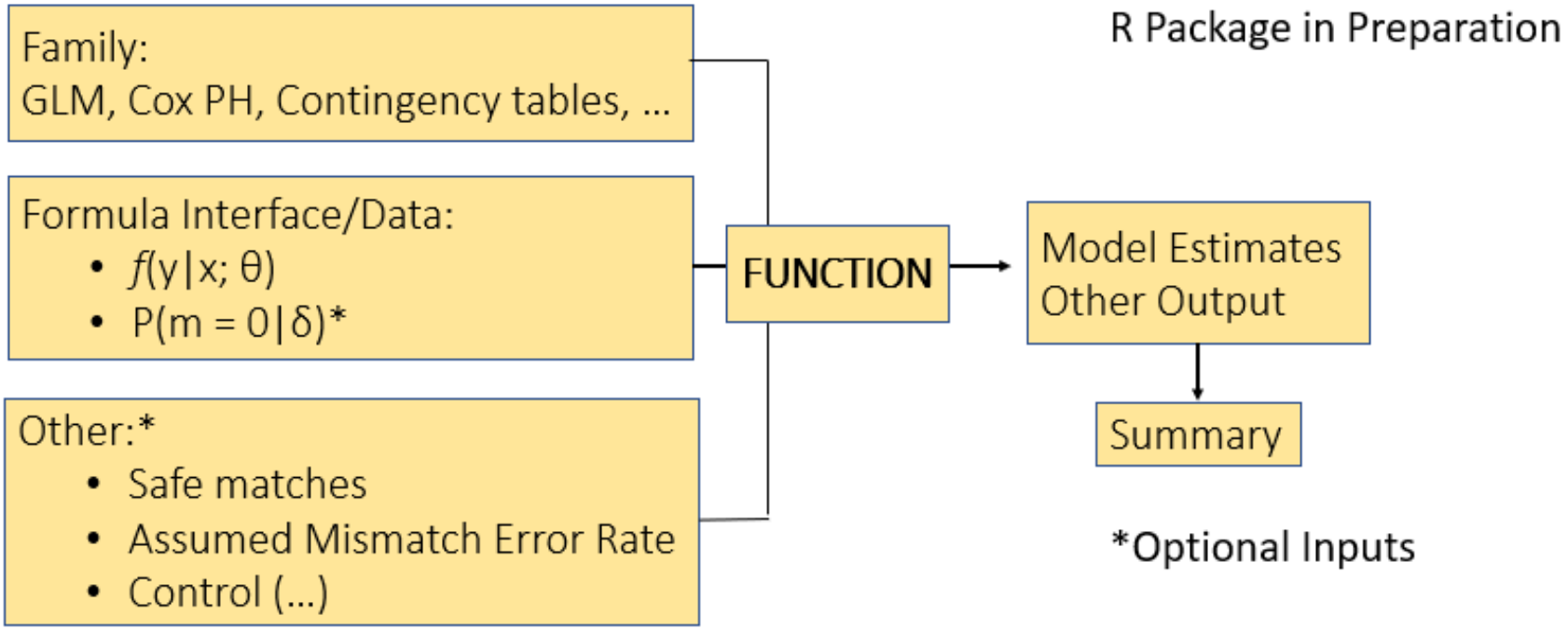
```
hldemodata <- demodata[demodata$hndlnk == "Hand-Linked
                        At Some Level",]
hl_fit <- lm(surv_age ~ poly(uyob,3,raw = TRUE),
             hldemodata)
summary(hl_fit)
```

Demo in R: HL-Only Analysis Results

4. Plain Linear Regression: Hand-Linked Only (n = 2,159)

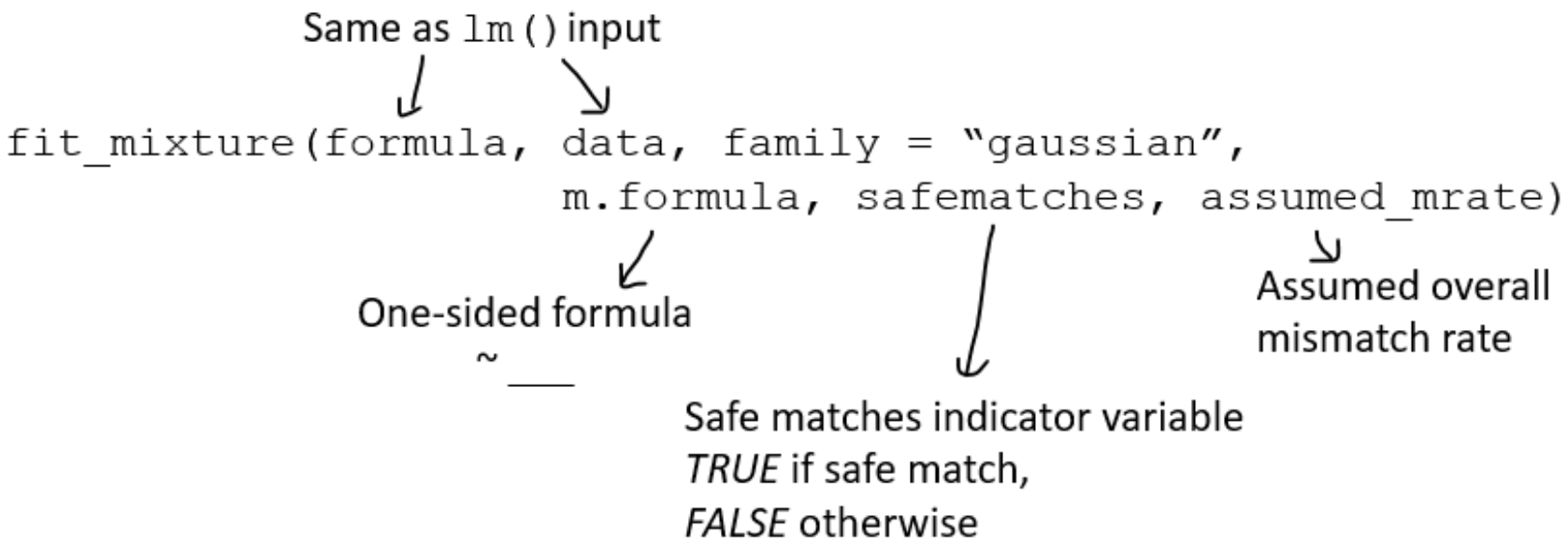
```
Call:
lm(formula = surv_age ~ poly(uyob, 3, raw = TRUE), data = hldemodata)

Residuals:
    Min      1Q  Median      3Q     Max
-69.076 -11.049   3.512  12.048  41.971

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                    57.66       1.26  45.748  < 2e-16 ***
poly(uyob, 3, raw = TRUE)1    -44.23      11.59  -3.815  0.00014 ***
poly(uyob, 3, raw = TRUE)2    118.57      27.92   4.247 2.26e-05 ***
poly(uyob, 3, raw = TRUE)3    -59.92      18.46  -3.246  0.00119 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.98 on 2155 degrees of freedom
Multiple R-squared:  0.1191,    Adjusted R-squared:  0.1178
F-statistic: 97.08 on 3 and 2155 DF,  p-value: < 2.2e-16
```

Adjustment Approach Implementation in R




```
fit_mixture(formula, data, family = "gaussian",
                    m.formula, safematches, assumed_mrate)
```

Adjustment Approach R Function



Demo in R: Adjustment Method

5. Adjustment Method (n = 3,238)

```
am_fit <- fit_mixture(surv_age ~ poly(uyob, 3, raw = TRUE),
          demodata, m.formula = ~ commf + comml,
          safematches = ifelse(demodata$hndlnk ==
                    "Hand-Linked At Some Level", TRUE, FALSE),
          assumed_mrate = 0.05)
summary_fit_mixture(am_fit)
```

Demo in R: Adjustment Method Results

5. Adjustment Method (n = 3,238)

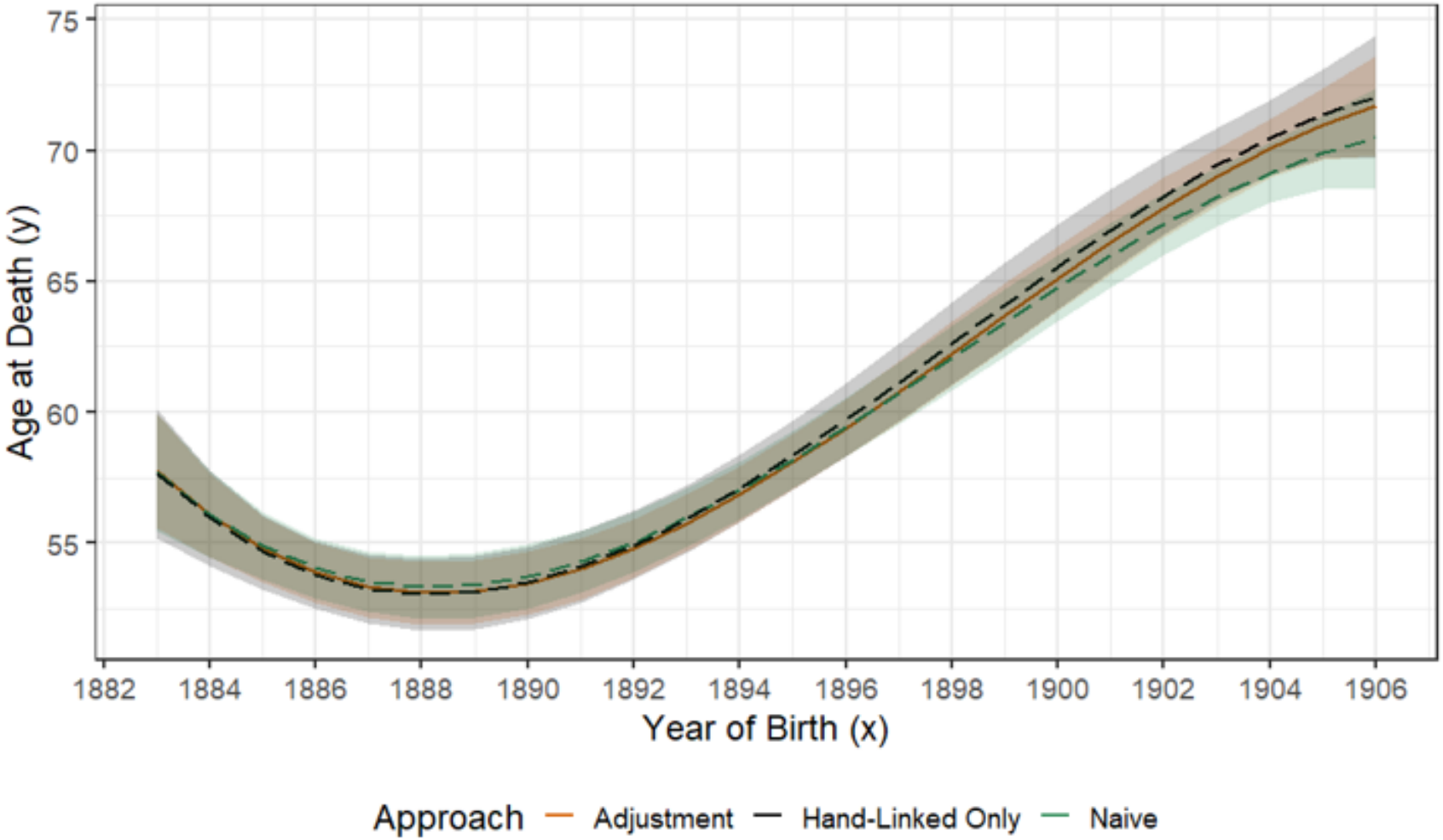
```
Call:
fit_mixture(formula = surv_age ~ poly(uyob, 3, raw = TRUE), data = demodata,
    m.formula = ~commf + comml, safematches = ifelse(demodata$hndlnk ==
        "Hand-Linked At Some Level", TRUE, FALSE), assumed_mrate = 0.05)

Family:
Gaussian
---
Beta Coefficients:
                              Estimate    StdErr
(Intercept)                   57.75130  1.572157
poly(uyob, 3, raw = TRUE)1 -43.73733 18.191100
poly(uyob, 3, raw = TRUE)2 114.87037 45.612635
poly(uyob, 3, raw = TRUE)3 -57.14401 30.476981

Sigma:
 Estimate   StdErr
 19.32071 1.178779

Gamma Coefficients:
             Estimate   StdErr
(Intercept) -7.768067 2.534873
commf        6.926968 2.256127
comml        9.152162 3.252471
---
Average Correct Match Rate Among Not Safe Matches: 0.8538001
```

Summary of Results:

Parameter	Naïve (n = 3,238)	Adjusted (n = 3,238)	HL-Only (n = 2,159)
$\widehat{\beta_0}$	57.7 (1.1)	57.7 (1.6)	57.7 (1.26)
$\widehat{\beta_1}$	-41.7 (10.2)	-43.7 (18.2)	-44.2 (11.6)
$\widehat{\beta_2}$	111.2 (23.9)	114.9 (45.6)	118.6 (27.9)
$\widehat{\beta_3}$	-56.8 (15.6)	-57.1 (30.5)	-59.9 (18.5)
$\hat{\sigma}$	19.7	19.3	19.0

Parameter	Adjusted
$\widehat{\gamma_0}$	-7.8 (2.5)
$\widehat{\gamma_1}$	-6.9 (2.3)
$\widehat{\gamma_2}$	9.1 (3.2)

Predicted Ages at Death for Birth Cohorts:



Case Study II: Healthcare Credential Data

We revisit the data set on nurse licenses studied earlier.

Recall that the substantive question concerns trends over time in the duration of waiting periods between temporary and regular nurse license issuance in WA (2009 – 2021).

In this case study, we construct two linked files:

1) A *restrictively linked file*, enforcing exact agreement of names for linked records:

   $\sim$**61k** records est. mismatch rate: 0.4% to 1%.*

2) A *generously linked file* based on inexact matching, similar to what was shown in our demo on probabilistic RL:

   $\sim$**78k** records est. mismatch rate: 3.7% to 8.1%.*

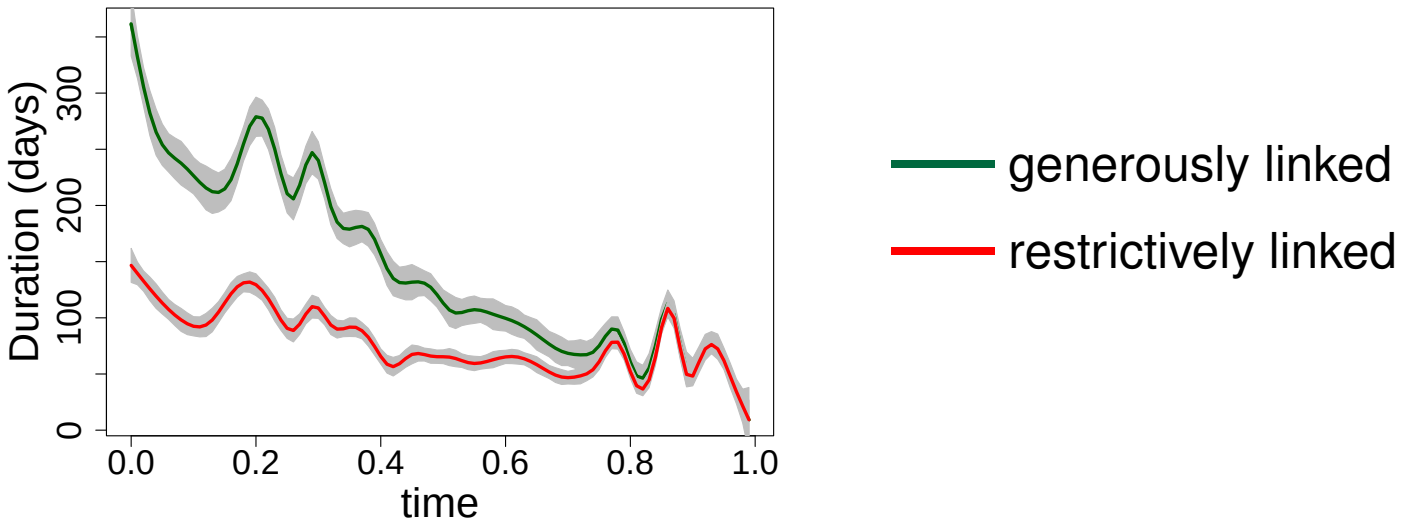
*Records with negative durations were dropped. Estimates are based on the fraction of records with excessive durations ($\geq$ 1.5 years)

Approach: capture time trends in the average processing times via splines.

- Let $t \in [0, 1]$ denote the date of issuance of the temporary licence (scaled, with 0 corresponding to 01/01/2009 etc).
- With y denoting the processing time, we suppose that

$$\mathbf{E}[y|t] \approx s_{\boldsymbol{\beta}}(t), \quad s_{\boldsymbol{\beta}}(t) = \sum_{j=1}^{p} \beta_j B_j(t),$$

where the 2nd expression represents a cubic spline expansion.



Can the results from the two linked data sets be reconciled?

We want to apply the mixture model approach used before, but there is one obstacle, namely a **principled choice** of the **smoothing parameter** for the spline fit.

As a workaround, we cast the approach in a **Bayesian framework**, using a smoothness prior for the spline coefficients:

(i) $p(\alpha) \propto 1, \quad p(\sigma^2) \propto (\sigma^2)^{-1}, \quad p(\tau^2) \propto (\tau^2)^{-1}$

(ii) $\{m_i\} | \, \alpha \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\alpha), \quad p(\boldsymbol{\beta} | \, \tau^2) \propto \exp\left(-\frac{1}{2\tau^2} \boldsymbol{\beta}^\top \mathbf{S} \boldsymbol{\beta}\right)$

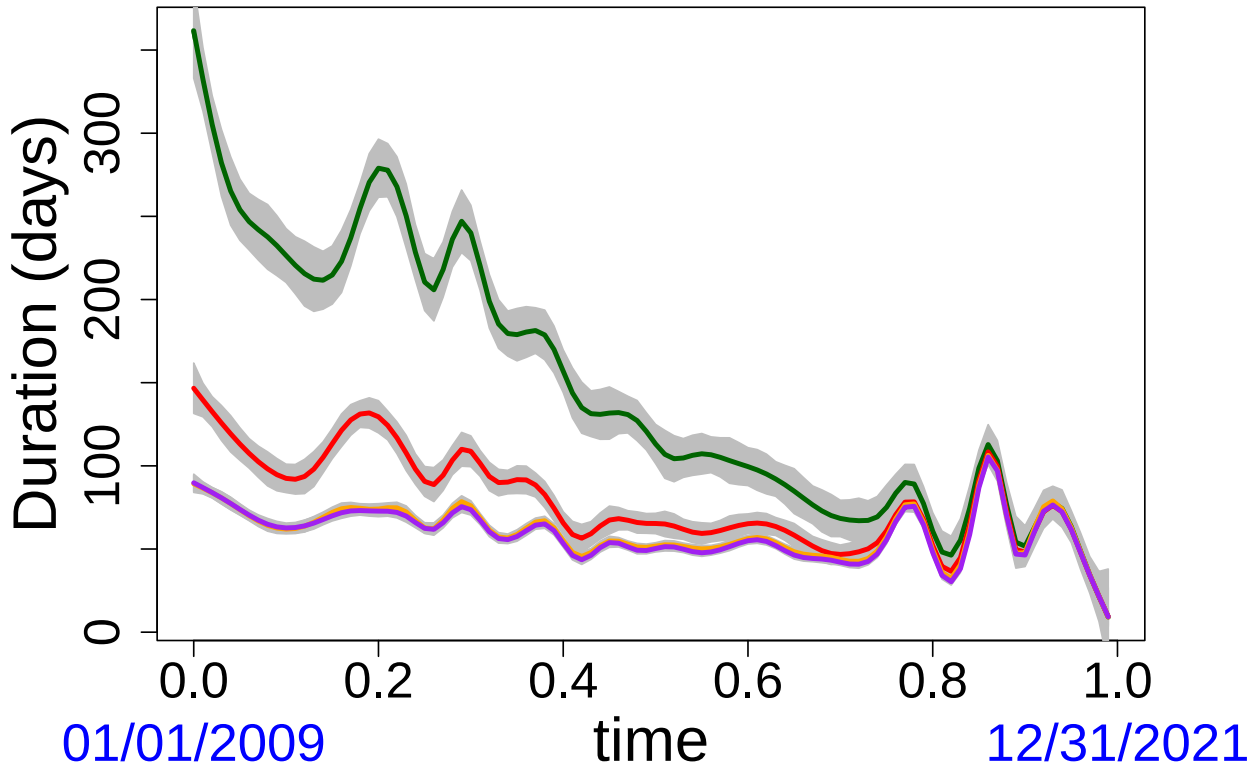
(iii) $y_i | t_i, m_i = 0, \boldsymbol{\beta}, \sigma^2 \sim N\left(s_{\boldsymbol{\beta}}(t_i), \sigma^2\right),$

$y_i | m_i = 1, \mu_0, \sigma_0^2 \sim N(\mu_0, \sigma_0^2), \quad 1 \leq i \leq n.$

We here mimic the situation of not having *any* information about RL, and work with a constant mismatch rate α.

It is straightforward to sample from the posterior using Gibbs sampling.

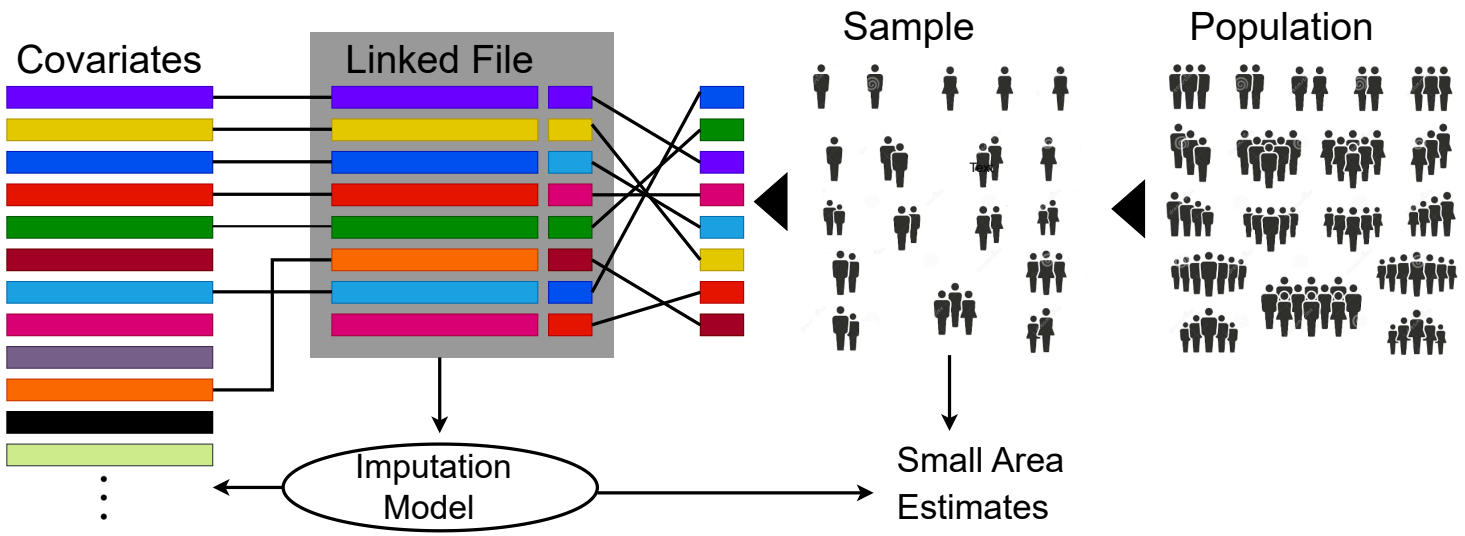
Results.



Ongoing work and concluding remarks

Work in Progress: SAE with linked data

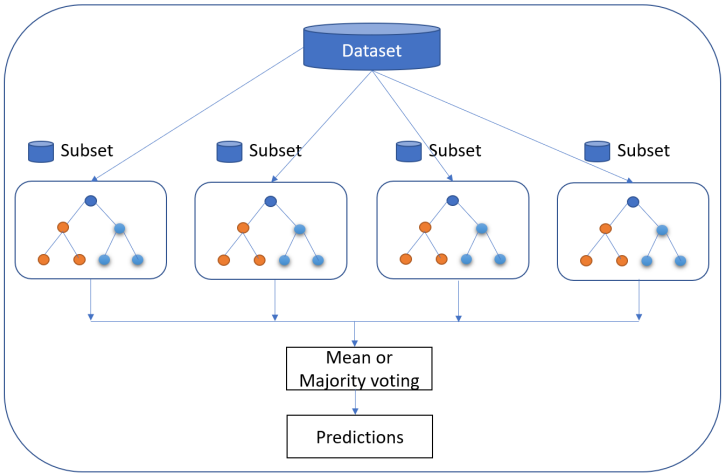
Unit-level Small Area Estimation (SAE) based on linear mixed effect models in the presence of linkage error between covariates and sampled observations.



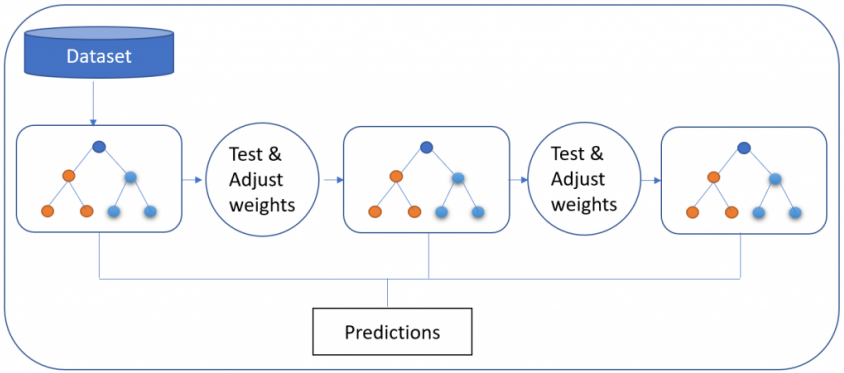
The dependence of observations within a common area breaks the simple structure of the pseudo-likelihood, and a more refined approach is employed. Results to be reported in Fabrizi, Salvati, and S. (2023+).

Work in Progress: Predictive Modeling

The widespread application of Machine Learning predictive models, such as bagging, random forests, boosting, etc, to linked data files, also needs more attention.



Bagging, Random Forest



Boosting

A proposed adjustment method

The outline of a proposed adjustment to bagging trees:

- Adjust the prediction of each decision tree by re-weighting the observations.
- Re-weight the observations using the posterior expectation of the matching indicator in the mixture distribution.
- Apply the bagging process to adjusted decision trees, either by mean or majority voting.
- These predictions can be further improved by using an optimal estimate of mixing coefficients.
- The procedure can be similarly applied to other ensemble methods.

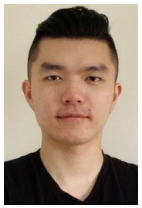
Acknowledgments

Funding:

We acknowledge support by NSF grants CCF-1849876 and SES-2120-318.

Collaborators:

Brady West



Zhenbang Wang

Guoqing Diao

Data:

Life-M team

Jessica Faul

Abraham Flaxman

Supporting materials

Paper:

Slawski, West, Bukke, Wang, Diao, Ben-David (2023).

A General Framework for Regression with Mismatched Data Based on Mixture Modeling

`arXiv:2306.00909`.

Link to GitHub repository:

https://github.com/ehb2126/Data-Analysis-after-Record-Linkage

References I

[1] Slawski & Ben-David, “Linear Regression with Sparsely Permuted Data”, *EJS*, 2019.

[2] Slawski, Ben-David, Li, “Two-Stage Approach to Multivariate Linear Regression with Sparsely Mismatched Data”, *JMLR*, 2020.

[3] Slawski, Diao, Ben-David, ”A Pseudo-Likelihood Approach to Linear Regression with Partially Shuffled Data”, *JCGS*, 2021.

[4] Wang, Ben-David, Diao, Slawski, ”Regression with linked data sets subject to linkage error”, *WIREs Computational Statistics, 2022*.

[5] Wang, Ben-David, Slawski, ”Estimation in exponential family regression based on linked data contaminated by mismatch error”, *SII*, 2023.

[6] Wang, Ben-David, Slawski, ”Regularization for Shuffled Data Problems via Exponential Family Priors on the Permutation Group”, *AISTATS*, 2023.

[7] Fabrizi, Salvati, Slawski, ”Accounting for Mismatch Error in Small Area Estimation with Linked Data”, *in preparation*, 2023.

References II

- Neter, Maynes, Ramanathan, “The Effect of Mismatching on the Measurement of Response Errors”, *JASA*, 1965.
- Scheuren & Winkler, “Regression Analyis of data files that are computer matched”, *Surv Meth*, 1997.
- Lahiri & Larsen, “Regression Analysis with Linked Data”, *JASA*, 2005.
- Han & Lahiri, “Statistical Analysis with Linked Data”, *Int Stat Rev*, 2018.
- Chambers, “Regression Analysis of Probability-Linked Data”, 2009.
- Chambers & DaSilva, “Improved Secondary analysis of linked data”, *JRSS-A*, 2020.
- Gutman et al., “A Bayesian Procedure for File Linking to Analyze End-of-Life Medical Costs”, *JASA*, 2013.
- Hof & Zwinderman, “A mixture model for the analysis of data derived from record linkage”, *Stat Med.*, 2015.
- Tancredi & Liseo, “Regression analysis with linked data: problems and possible solutions”, *Statistica*, 2015.
- Steorts, Tancredi, Liseo, “Generalized Bayesian Record Linkage and Regression with Exact Error Propagation”, *Privacy in Statistical Databases*, 2018.
- Abowd *et al.*, “Finding Needles in Haystacks: Multiple-Imputation Record Linkage Using Machine Learning”, 2022.