Recent Developments and Open Problems in Post-Linkage Data Analysis

Martin Slawski

George Mason University

February 21, 2024

JPSM MPSDS Seminar Series

Acknowledgments

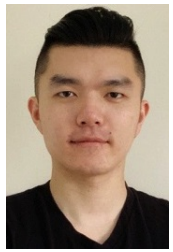
Funding: NSF grants CCF–1849876 & SES–2120318.

Collaborators:

Priyanjali Bukke
Ph.D. student

Emanuel Ben-David
U.S. Census

Guoqing Diao
George Wash



Zhenbang Wang
Ph.D. student

Brady West
U Michigan

Enrico Fabrizi
U Cattolica

Nicola Salvati
U Pisa

Roee Gutman
Brown U

Bodhi Sen
Columbia U

Fadoua Balabdaoui
ETH

Talk Outline

1. Overview and brief literature review
2. Mixture model approach
3. Extension to Small Area Estimation
4. Open Problems / Ongoing Work
5. Post-Linkage Data Analysis without Linkage?

Data Integration & Record Linkage (RL)

Data Integration:
Combination of multiple existing data sources that contain complementary pieces of information.

RL:
Micro-level (i.e., record-by-record) data integration. **Linkage error**, i.e., false matches (mismatches) or false non-matches can occur when variables used for matching do not uniquely identify entities.

HRS Data*

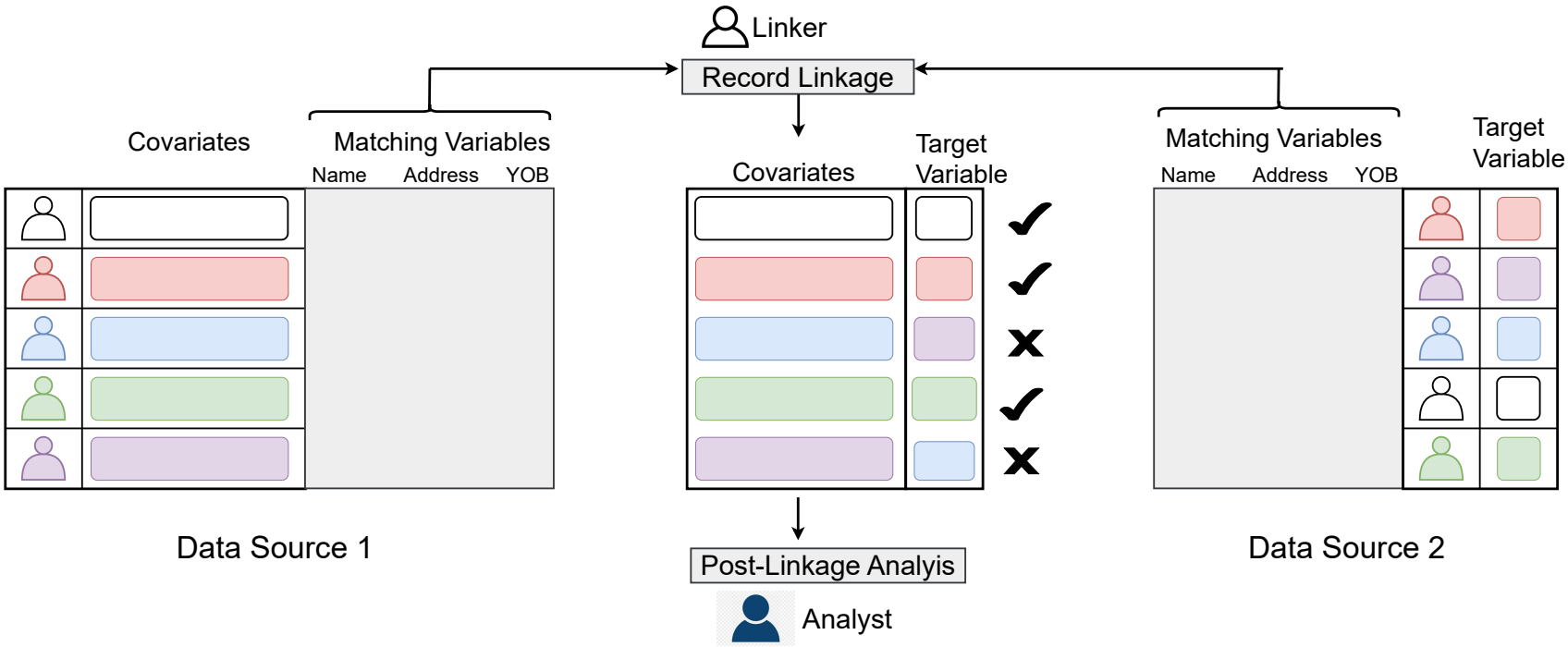
First_Name	Last_Name	Sex	BID	NH_Nights	
William	Smith	M	8LA6-RL1-LE17	1	1
Imari	Vasquez	F	NA	0	2
Morgan	Jones	F	8QP9-RD4-IP64	1	3
Roland	Matthews	M	NA	0	4
Sarah	Begum	F	9YZ3-RZ3-YC19	0	5 X

CMS Data

	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

*: Tables are fake and meant to be illustrative of matching complications.

RL and Post-Linkage Data Analysis (PLDA)



Primary Analysis:
Access to individual Data Sources 1 & 2. RL and subsequent data analysis can be performed jointly, with propagation of uncertainty.

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

(Common) Types of PLDA

File A	File B	Task
$\mathbf{x}$	y	Regression
$\mathbf{x}_1$	$\mathbf{x}_2$ y	Regression
$\mathbf{x}$	$\mathbf{y}$	Multivariate Analysis (e.g., PCA)
x $\in \{1, \ldots, L\}$	y $\in \{1, \ldots, M\}$	Two-way contigency table analysis

The top scenario is the most studied. Interestingly, the 2nd scenario appears barely studied. There is some work on the 3rd and 4th scenario.

Consequences of Linkage Error

Consequences of false non-matches
(matching records not identified as such):

File A	File B
a_1	b_1
a_2	b_2
$\vdots$	$\vdots$
$\vdots$	$\vdots$
a_M	b_M

Ideal file w/o
missing any matches

File A	File B
a_1	b_1
a_2	?
$\vdots$	$\vdots$
$\vdots$	$\vdots$
a_M	b_M

File missing match
no. 2

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

$\rightarrow$ Loss of Statistical Power $\rightarrow$ Danger of Selection Bias.

Consequences of Linkage Error

Literature on Secondary Analysis is heavily focused on false matches (mismatches); false non-matches are "argued away" using ignorability assumptions.

Mismatches tend to introduce data contamination.

Specific consequences can be:

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

(Neter et al., 1965; Scheuren & Winkler, 1997; Lahiri & Larsen, 2005; Wang et al., 2022; Chambers et al., 2023)

Very brief literature review: Secondary Analysis

Lahiri & Larsen (2005) study linear regression with predictors $\mathbf{X}$ in file A and responses $\mathbf{Y}$ in file B.

Let Π^* be the binary matrix encoding the correct matching of records from files A and B.

File B

File A

$$\Pi^* = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Given $\mathbf{Q} = \mathbf{E}[\Pi^*]$, they regress $\mathbf{Y}$ on the "instrumental variable" $\mathbf{QX}$.

Han & Lahiri (2019) outline estimation of $\mathbf{Q}$ and uncertainty propagation in the primary analysis setting.

Very brief literature review: Secondary Analysis

Chambers (2009) builds on and generalizes Lahiri & Larsen (2005) using estimating equations. The exchangeable linkage error model (ELE) is introduced to faciliate the estimation of $\mathbf{Q}$ in secondary analysis.

In the ELE model, $\mathbf{Q}$ is assumed to be block-structured according to **blocking variables** used in RL: blocking variables are matching variables required to exhibit exact agreement for matching records (e.g., ethnicity, ZIP code, ...).

Moreover, restricted to each block, the matrix $\mathbf{Q}$ is assumed to be of the form

$$\begin{pmatrix} 1-\alpha_b & \lambda_b & \dots & \dots & \lambda_b \\ \lambda_b & 1-\alpha_b & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \lambda_b \\ \lambda_b & \dots & \vdots & \lambda_b & 1-\alpha_b \end{pmatrix}$$

Very brief literature review: Secondary Analysis

Another approach whose roots go back to DeGroot & Goel (1980) and Wu (1998) is to treat Π^* as **missing data** and then apply established machinery:

- EM algorithm
- full Bayes approach with (Gibbs) sampling of Π^* given data and parameters Gutman et al. (2013).

Wang et al. (2023) argue that the approach is often not suitable in secondary analysis settings since insufficient knowledge about Π^* may lead to overfitting.

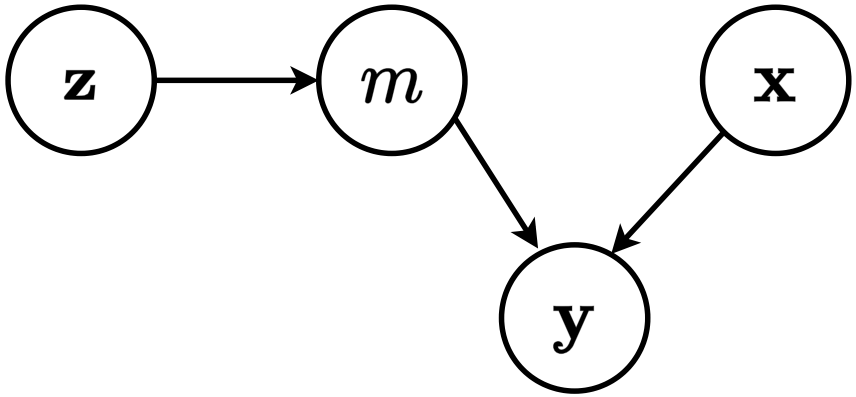
As a remedy, they propose the use of suitable prior distributions for Π^*.

Approach based on a Mixture Model

cf. Slawski et al. (2021, 2023). Related to Hof & Zwinderman (2015) addressing the primary setting.

- Based on a two-component mixture model in alignment with the popular Fellegi-Sunter model for RL,
- Unified framework for various types of **Post-Linkage Data analysis** (PLDA),
- Does not require clerical review or "external" knowledge about RL (but can be incorporated if available),
- Scales linearly in the number of data points,
- Likelihood-based inference,
- Extendable to a Bayesian framework.

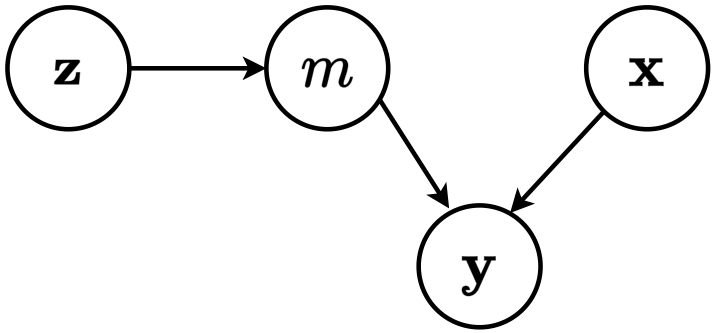
Mixture model approach at a glance



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

- $\mathbf{x}$ in file A, $\mathbf{y}$ in file B; (regression) parameter of interest $\boldsymbol{\theta}$,
- Latent binary mismatch indicator m, (possibly) modeled conditionally on info about RL $\mathbf{z}$,
- "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).

Mixture model approach: assumptions



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

Assumption 1 – Independence for mismatches: $\mathbf{y} \perp\!\!\!\perp \mathbf{x} \,|\, m=1$

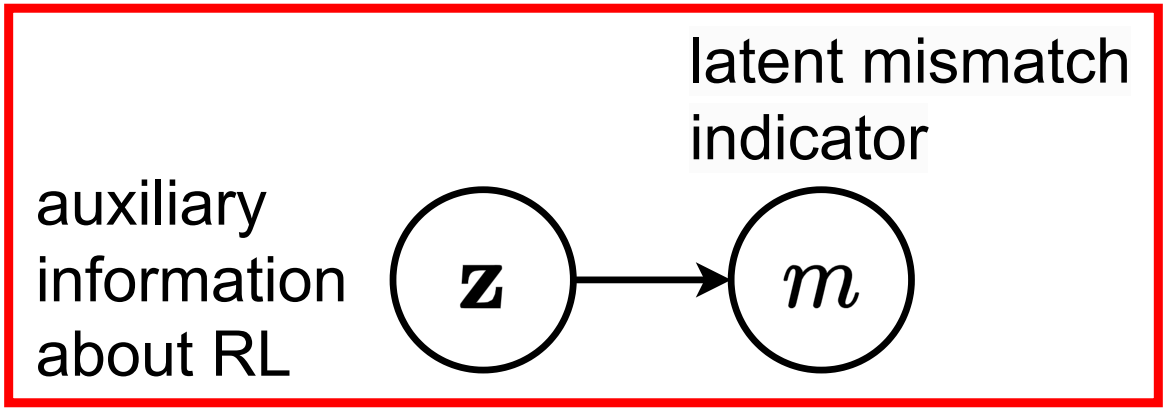
Satisfied if distinct records are independent. Can be violated if mismatches occur within correlated blocks of observations.

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 $\mathbf{z}$ but not on $\mathbf{x}$. This assumption is stronger than those of other methods but renders inference more tractable. In particular, it implies that

$$f(\mathbf{y}|m=1) = f(\mathbf{y}|m=0).$$



The covariates $\mathbf{z}$ 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 $\mathbf{z}$ and m:

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

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.,

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

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

Inference

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

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

Inference (standard errors etc.) via asymptotic theory for composite maximum likelihood estimators (Varin et al., 2011).

Alternative: hierarchical Bayes.

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

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

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

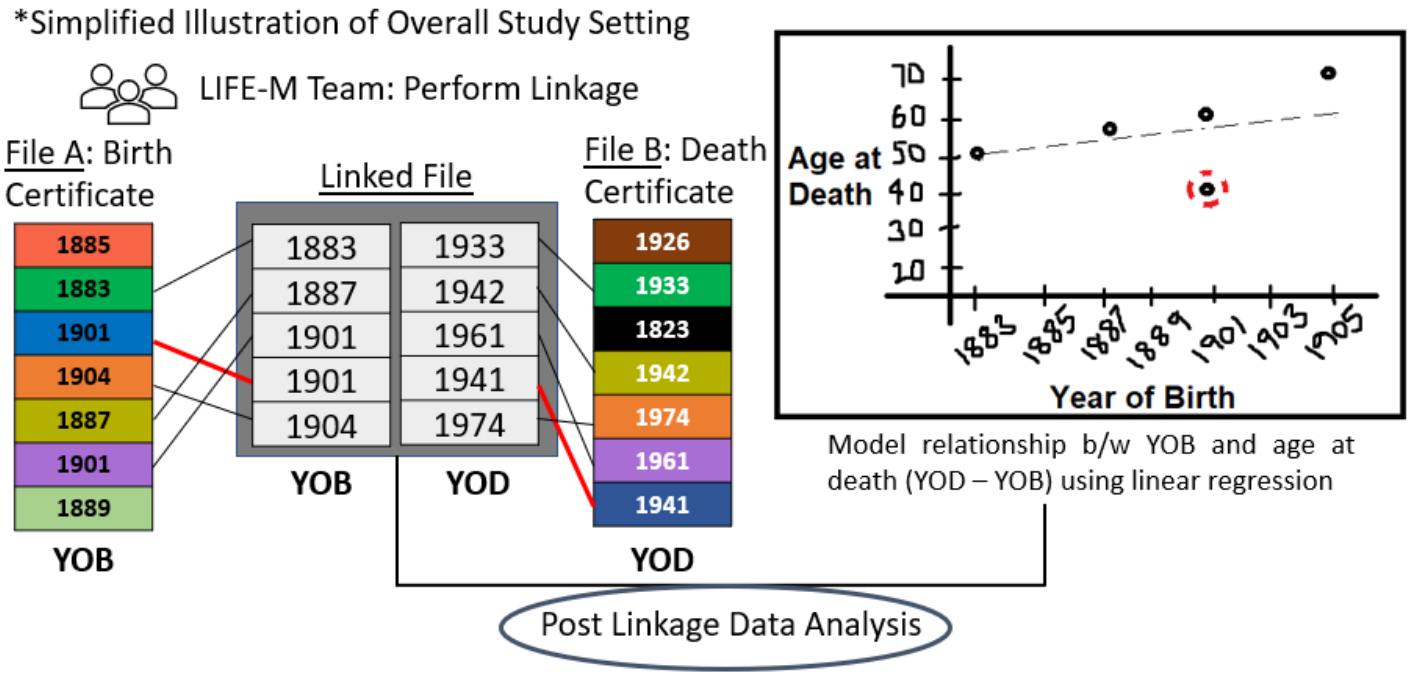
Note: not affected by mismatch error – only involves variables from a single file.

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 Study I: Life-M project

Life-M project: Longitudinal Intergenerational Family Electronic Micro-Database (`life-m.org`).



The Life-M team used a hybrid of two RL procedures:

- "hand-linked" – clerically reviewed RL,
- "machine-linked" – automated probabilistic RL (anticipated mismatch rate $\sim$5%).

Linked data set: n =156k individuals, about 1.4% hand-linked, rest machine-linked.

handlinked (y/n)	YOB (x)	age of death (y)	commf (z_1)	comml (z_2)
0	1905	83	.77	.45
1	1883	79	.93	.08
...	...	...	...	...
0	1944	58	.89	.80

commf, comml: "commonness" of first name, last name

Model:

- hand-linked records are assumed to be correctly matched ($m_i = 0$)

$$y_i \mid m_i = 1, x_i \sim N(\mu, \tau^2),$$

$$y_i \mid m_i = 0, x_i \sim N(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3, \sigma^2),$$

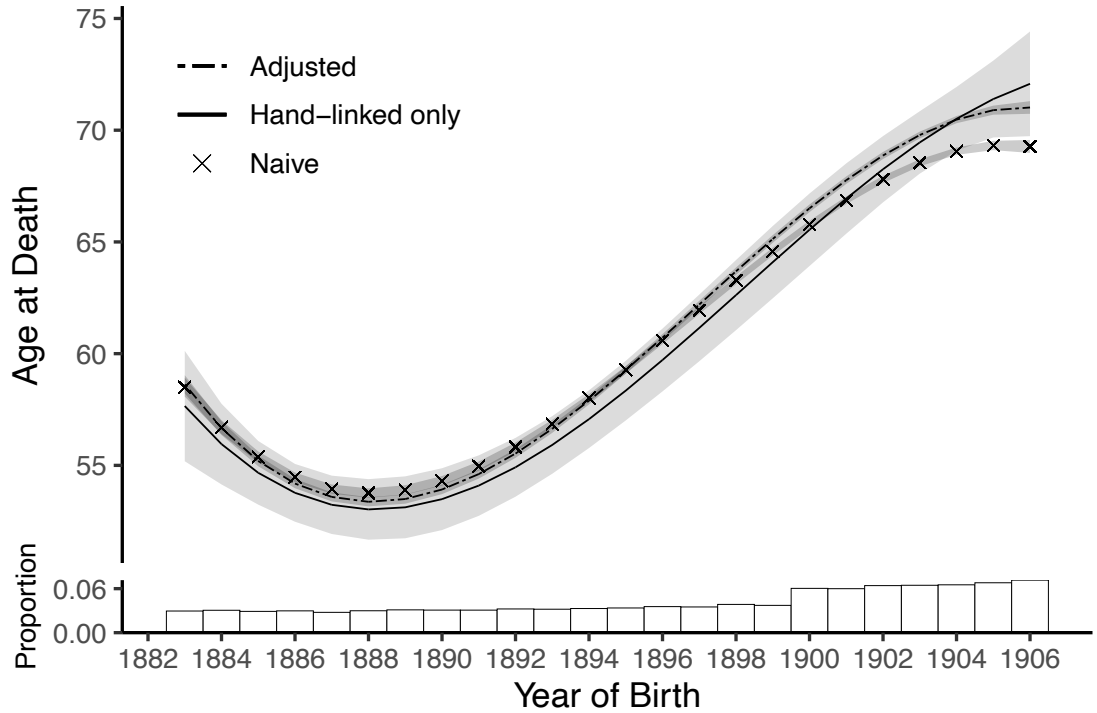
$$m_i \mid \mathsf{commf}_i, \mathsf{comml}_i \sim \text{Bernoulli}\left(\frac{\exp(\gamma_0 + \gamma_1 \mathsf{commf}_i + \gamma_2 \mathsf{comml}_i)}{1 + \exp(\gamma_0 + \gamma_1 \mathsf{commf}_i + \gamma_2 \mathsf{comml}_i)}\right).$$

Summary of Results:

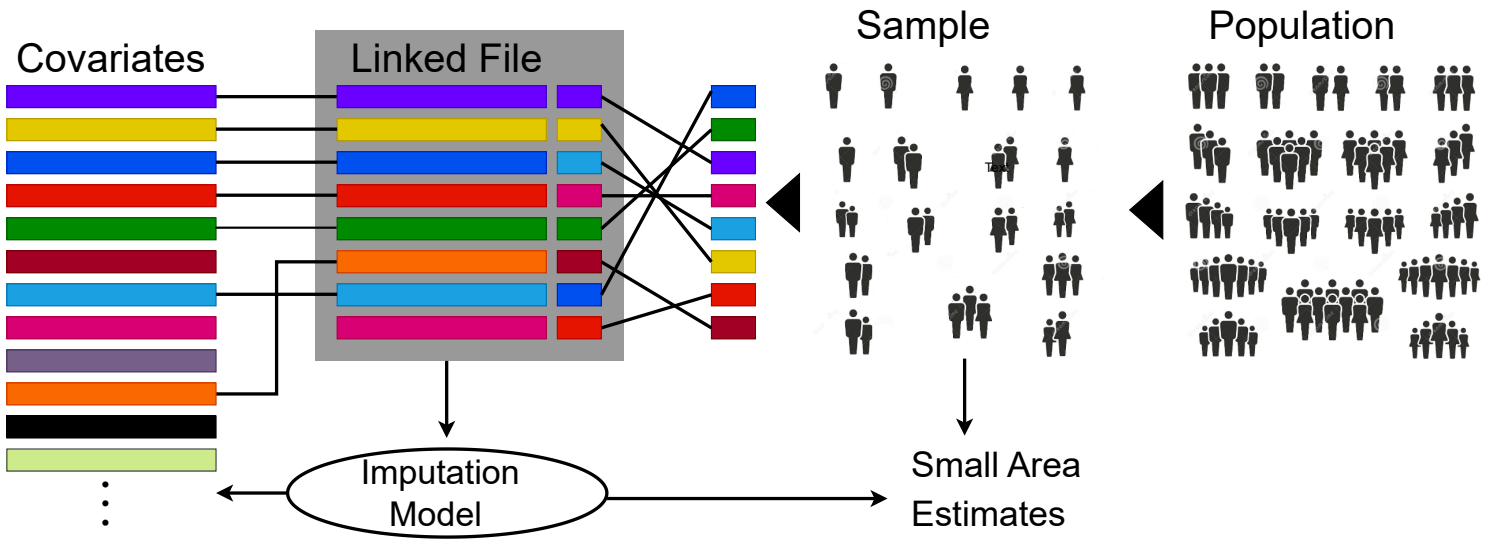
	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\sigma}$	$\widehat{\gamma}_0$	$\widehat{\gamma}_1$	$\widehat{\gamma}_2$
Naive	58.5(.2)	-46.7(1.8)	130.4(4.0)	-72.9(2.5)	21.2(.1)			
Adj‡	58.6(.1)	-51.0(1.5)	140.3(3.9)	-76.8(2.6)	20.7(.1)	-6.0(.5)	-1.5(.6)	7.2(.3)
Adj†	58.7(.2)	-52.5(1.6)	143.2(3.9)	-77.7(2.7)	20.4(.1)	-4.9(.4)	-1.4(.4)	6.1(.3)
HL*	57.7(1.3)	-44.2(11.6)	118.6(27.9)	-59.9(18.5)	19.0(.3)			

Adj‡: proposed, assuming mismatch rate $\leq$5%

Adj†: proposed, assuming mismatch rate $\leq$7.5%, HL: hand-linked only.



Case Study II: Small Area Estimation



- Unit-level Small Area Model (linear mixed effect model) for unit i in area j:

$$y_{ij} = \mathbf{x}_{ij}^{\top}\boldsymbol{\beta} + \gamma_j + \epsilon_{ij}, \quad \gamma_j \sim N(0, \Sigma), \;\; \epsilon_{ij} \sim N(0, \sigma^2).$$

- Mismatch error with some of the sampled units $\{y_{ij}^{(s)}\}$ linked to a non-matching set of covariates

  Han (2018); Salvati et al. (2021).

Goal: obtain EBLUP-style predictions

$$\widehat{y}_{ij} = \mathbf{x}_{ij}^{\top}\widehat{\boldsymbol{\beta}} + \widehat{\gamma}_j$$

The proposed mixture model-based approach remains applicable. However, the pseudo-likelihood is composed of area-level (rather than unit-level) factors.

This renders inference via the EM algorithm with latent variables $\{\mathbf{m}_j = (m_{ij}),\, \gamma_j\}$ much more challenging since there are

$$2^{\#\text{observations in area } j}$$

configurations for each $\mathbf{m}_j$.

Regardless, the E-step can be evaluated in closed form and approximated efficiently via Gibbs sampling within MC-EM (Fabrizi et al., 2023+).

Illustration:

Semi-synthetic problem taken from Salvati et al., 2021 (real data, but linkage/sampling is synthetic) whose study is in turn based on Briscolini et al., 2018.

- Data from Survey on Household Income & Wealth (SHIW), Bank of Italy, $N \approx 26$k.
- $D = 18$ areas (Italian administrative regions)
- $n_j = \max\{0.01 \cdot N_j, 5\}$, $j = 1, \ldots, D$, $n = 267$.
- Variable of interest (y): **annual income**; auxiliary covariate (x): **annual consumption**.
- (Synthetic) probabilistic RL of income to consumption using perturbed (fake) quasi-identifiers (names, gender, DOB).
- Mismatch rate: 15 to 30% (1k replicates via random sampling).

Results:

Relative root mean squared errors for the area means (over 1k replicates):

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Naive	0.04076	0.05631	0.06677	0.07417	0.08026	0.29503
Oracle	0.03327	0.04726	**0.05214**	**0.06053**	0.06048	0.2738
Proposed	0.03250	0.04909	0.05543	0.06035	0.06625	0.15368
Salvati *et al.*	0.0365	0.0486	0.0555	0.0619	0.0659	0.2607

Ongoing work

I) Allowing m to depend on $\mathbf{x}$

We are interested in eliminating the separation into two (independent) sets of covariates $\mathbf{x}$ and $\mathbf{z}$ for the outcome and mismatch indicator models, respectively.

This separation can often be limiting in applications. Its purpose is to achieve that $f(\mathbf{y}|m=0) = f(\mathbf{y}|m=1)$.

For simplicity suppose that $\mathbf{z} = \mathbf{x}$. Observe that

$$\begin{aligned} f(\mathbf{y}|m=1) &= \int f(\mathbf{y}|m=1,\mathbf{x})\, f(\mathbf{x}|m=1)\, d\mathbf{x} \\ &= \int f(\mathbf{y}|\mathbf{x})\, \frac{\mathbf{P}(m=1|\mathbf{x}) f(\mathbf{x})}{\int \mathbf{P}(m=1|\mathbf{x}) f(\mathbf{x})\, d\mathbf{x}}\, d\mathbf{x} \\ &= \sum_{i=1}^{n} f(\mathbf{y}|\mathbf{x}_i;\boldsymbol{\theta})\, \frac{h(\mathbf{x}_i;\boldsymbol{\gamma})}{\sum_{j=1}^{n} h(\mathbf{x}_j;\boldsymbol{\gamma})}, \end{aligned}$$

where the last equality is justified in a "fixed design" scenario (here, $\mathbf{P}(m=1|\mathbf{x}) = h(\mathbf{x};\gamma)$).

Ongoing work

Courtesy: Roee Gutman

Informative vs. **Non-informative** linkage error:

- SN: strongly non-informative linkage error – depends neither on $\mathbf{x}$ and $\mathbf{y}$.
- NL: non-informative linkage error – depends on $\mathbf{x}$ (only).
- WNL: weakly non-informative linkage error – depends on $\mathbf{x}$ and $\mathbf{y}$.
- IL: informative linkage – linkage error depends on on other possibly unobserved variables (correlated with $\mathbf{x}$ and $\mathbf{y}$).

In the following slide, we present coverage rates of the confidence interval for the slope in a simple linear regression model under each scenario for different adjustment methods.

Ongoing work

SE: standard error. CVG: overage rate of confidence intervals.

LEM	SN			NL			WNL			IL		
Meth	Bias	SE	Cvg	Bias	SE	Cvg	Bias	SE	Cvg	Bias	SE	Cvg
Naive	$-.21$	.05	.00	$-.17$	.05	.06	$-.12$	.04	.22	$-.10$	.04	.47
ChR	.004	.08	.97	.035	.08	.96	.11	.08	.73	.20	.08	.11
ChL	.003	.08	.97	.033	.07	.95	.10	.07	.74	.19	.07	.13
ChB	$-.001$	.07	.99	.051	.07	.91	.13	.07	.50	.18	.06	.13
GT	.011	.05	.97	.020	.05	.92	$-.08$	.05	.77	$-.08$	.06	.80
SLW	.000	.04	.95	$-.01$	.04	.91	$-.05$	.04	.73	$-.05$	.04	.71

Naive: no adjustment for linkage error
ChR: Chambers' method.
ChL: Lahiri-Larsen under Chamber's ELE model,
ChB: Chambers' BLUE estimator,
GT: Gutman et al. multiple imputation-based estimator.
SLW: mixture model approach (Slawski, West, et al.)

Ongoing work

II) Framework for Missing links and Mismatches

For example, suppose that some $\mathbf{x}$'s cannot be linked to any of the $\mathbf{y}$'s. Let δ denote the corresponding indicator variable ($\delta = 1$ if linked).

Among the successfully linked data, we might still have mismatches. Assuming (for now) that $\delta \perp\!\!\!\perp m|\mathbf{x}$, we obtain the likelihood contributions

$$(i)\ f(\mathbf{y}, \delta = 1, m|\mathbf{x}) = \boxed{\mathbf{P}(\delta = 1|\mathbf{x}, \mathbf{y}; \boldsymbol{\phi})} \cdot \boxed{f(\mathbf{y}, m|\mathbf{x}; \boldsymbol{\theta}, \boldsymbol{\gamma})},$$

$$(ii)\ f(\delta = 0, \mathbf{x}) = \int \mathbf{P}(\delta = 0|\mathbf{x}, \mathbf{y}; \boldsymbol{\phi}) \cdot f(\mathbf{y}|\mathbf{x})\, d\mathbf{y},$$

The term inside $\boxed{\ldots}$ can be decomposed according to the mixture model as presented earlier.

The new component is the term inside $\boxed{\ldots}$ (with parameter ϕ.

The approach is inspired by models for non-ignorable missing response (e.g., Kim & Shao, 2021).

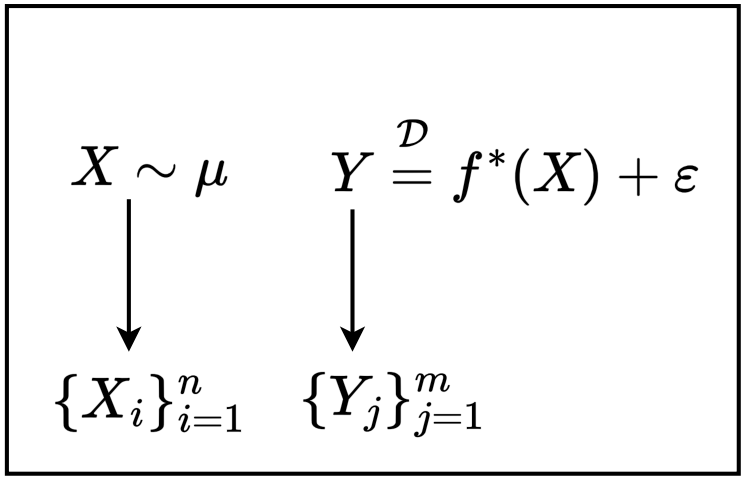
PLDA without RL?

RL may not always be feasible:

- Identifiers may not be shared,
- Respondents do not provide consent for linkage,
- Linkage is not possible for logistical reasons,
- ⋮

Unlinked regression is a recent paradigm (Carpentier & Schlüter, 2016; Rigollet & Weed, 2019; Balabdaoui et al., 2021; Slawski & Sen, 2022; Azadkia & Balabdaoui, 2022) for performing regression without ever linking responses and predictors.

Unlinked Regression



- X's are generated according to some distribution μ.
- Y is equal in distribution $\overset{\mathcal{D}}{=}$ to a transformation f^* of X plus additive noise.

Unlinked linear regression: $f^*(X) = X^\top \beta^*$.

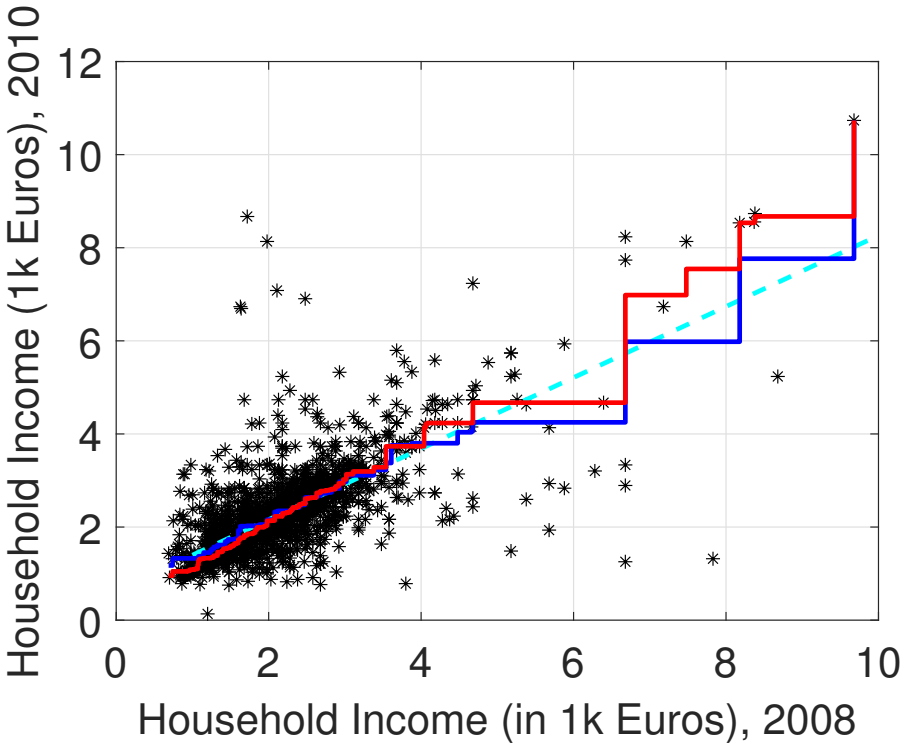
Generally, f^* (or β^* in the linear case) are not identifiable.

A sufficient condition for f^* to be identifiable is monotonicity.
Unlinked regression is closely related to **deconvolution**.

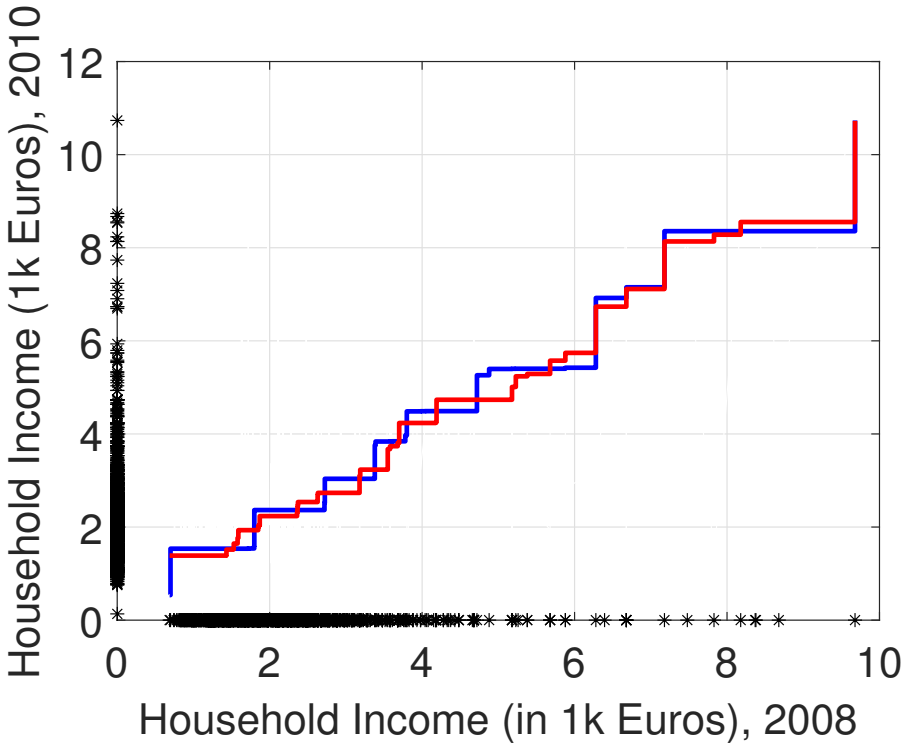
Unlinked Regression: Illustration

Taken from the Italian Survey of Household Income & Wealth (SHIW).
For unlinked regression, we use the method in Slawski & Sen, 2022.

Linked Data (isotonic regression)



Unlinked Data



blue: Assuming Gaussian noise. red: Assuming Laplacian noise.
cyan (dashed): Least squares regression line.

Conclusion and Supporting Materials

There is a good number of open problems in PLDA. Currently, it appears that there is no single approach having all desiderata.

The secondary analysis setting and the presence of linkage error might become even more common in the future given increased considerations for privacy.

There are tendencies not to create a single linked file and propagate uncertainty directly from pair-wise information about match status.

Unlinked regression has emerged as a new paradigm. It is still largely under development.

Papers:

Mixture Model – `arXiv:2306.00909`
Unlinked Regression – `arXiv:2201.03528`

Code:

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

References

• Slawski, Diao, Ben-David, "A Pseudo-Likelihood Approach to Linear Regression with Partially Shuffled Data", *JCGS*, 2021.
• Wang, Ben-David, Diao, Slawski, "Regression with linked data sets subject to linkage error", *WIREs Computational Statistics*, 2022.
• Wang, Ben-David, Slawski, "Regularization for Shuffled Data Problems via Exponential Family Priors on the Permutation Group", *AISTATS*, 2023.
• Fabrizi, Salvati, Slawski, "Accounting for Mismatch Error in Small Area Estimation with Linked Data", *in preparation*, 2023+.
• 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.
• Chambers, Fabrizi, Ranalli, Salvati, Wang, "Robust regression using probabilistically linked data", *WIREs Computational Statistics*, 2023.
• Han & Lahiri, "Statistical Analysis with Linked Data", *Int Stat Rev*, 2019.
• Han "Statistical Inference Using Data From Multiple Files Combined Through Record Linkage", *Ph.D. dissertation, University of Maryland*, 2018.
• DeGroot & Goel "Estimation of the Correlation Coefficient from a Broken Random Sample", *Ann. Stat.* , 1980.
• Wu "A Note on Broken Sample Problem", *Tech. Rep. , Dept. of Statistics, University of Michigan*, 1998.

References (Continued)

• 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.
• Varin, Reid, Firth "An overview of composite likelihood estimation", *Stat Sinica*, 2011.
• Briscolini, Consiglio, Liseo, Tancredi, Tuoto, "New methods for small area estimation with linkage uncertainty", *Int J of Approximate Reasoning*, 2018.
• Kim & Shao. "Statistical Methods of Handling Incomplete Data" *CRC press*, 2021.
• Carpentier & Schlüter. "Learning relationships between data obtained independently." *AISTATS*, 2016.
• Rigollet & Weed. "Uncoupled isotonic regression via minimum Wasserstein deconvolution" *Information & Inference*, 2019.
• Balabdaoui et al. "Unlinked monotone regression" *JMLR*, 2021.
• Azadkia & Balabdaoui "Linear regression with unmatched data: a deconvolution perspective" *arXiv:2207.06320*, 2022.