Data Analysis after Record Linkage: Sources of Error, Consequences, and Possible Solutions

Martin Slawski & Brady West

George Mason University & University of Michigan

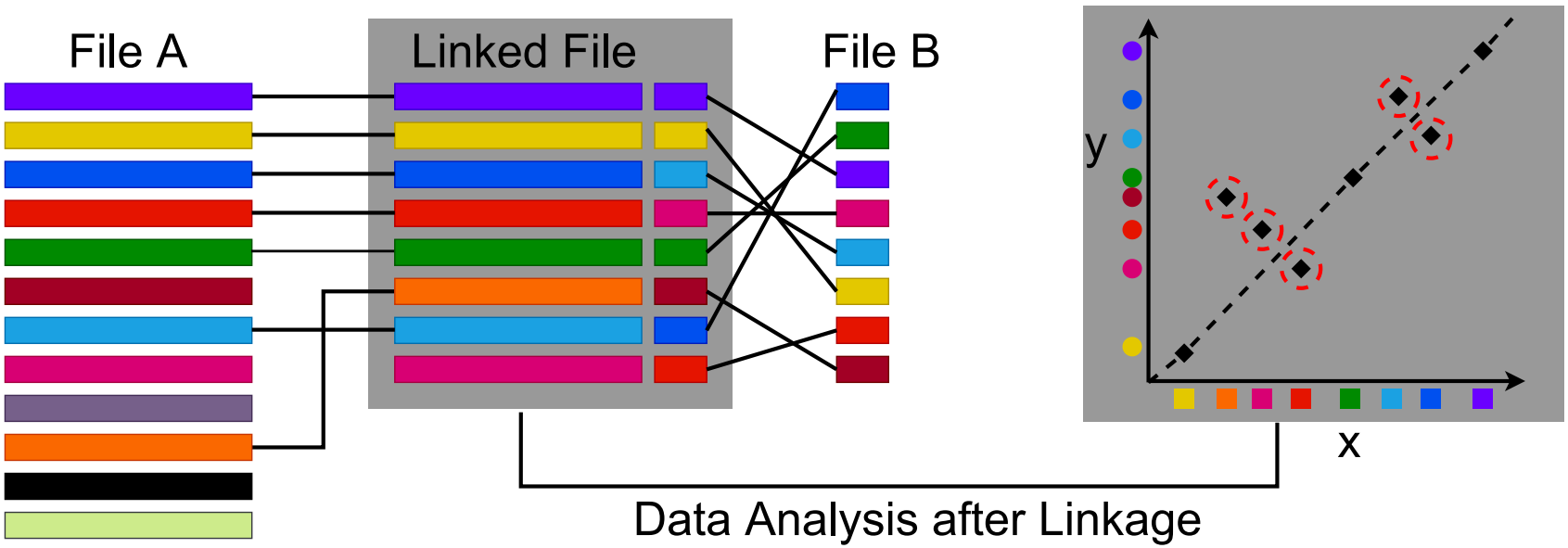
January 24, 2022
ISR Record Linkage Method Seminar Series

Project Overview

sponsored by NSF SES, 09/15/2021 – 08/31/2024.

Project Title:

Computational and Statistical Approaches to Regression Problems in the Presence of Linkage Errors



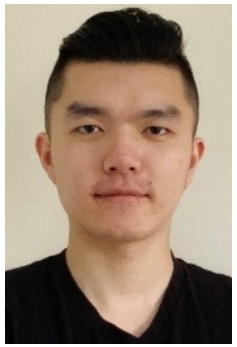
Themes: • Rigorous Statistical Analysis • Modern Computation • Validation & Benchmarking on complex linkage problems.

Project Team

Martin Slawski

Brady West

Emanuel Ben-David
U.S. Census



Zhenbang Wang
Ph.D. student, GMU

new collabarators & students

Linkage Error

For any pair of records (a, b) from file A and file B, record linkage (RL) returns a binary decision:

- **match**,
- **non-match**.

Potential errors:

- False matches **(mismatches, mismatch error)**,
- False non-matches **(missed matches)**.

Focus in this project will be on mismatch error.

Missed matches are not less important, but require a rather different treatment.
We hope to work on this towards the end of the project period.

Common sources of linkage error

- Lack of unique identifiers
- Errors or formatting variations in quasi-identifiers or blocking variables
- Computational bottleneck (it may not be feasible to check all pairs (a, b) for matches)

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

Primary vs. Secondary Analysis

Primary Analysis:

– Access to individual files A and B.

– Record linkage and subsequent data analysis can be performed jointly, with propagation of uncertainty.

Secondary Analysis:

– Access only to the linked file, not the individual files

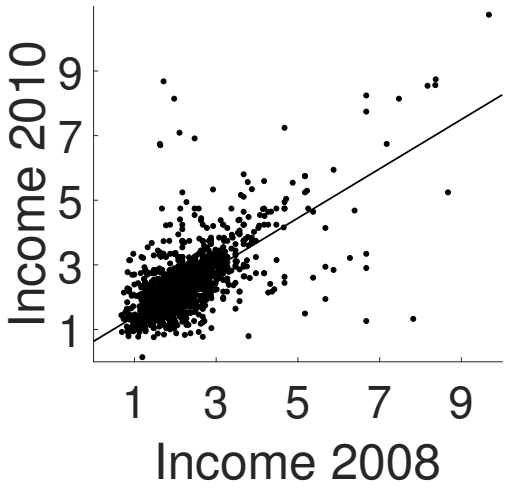
– Information about underlying RL may be available, but limited (e.g., blocking variables used, pair-wise match probabilities, etc.)

The **focus** in this project is on the more challenging **secondary setting**.

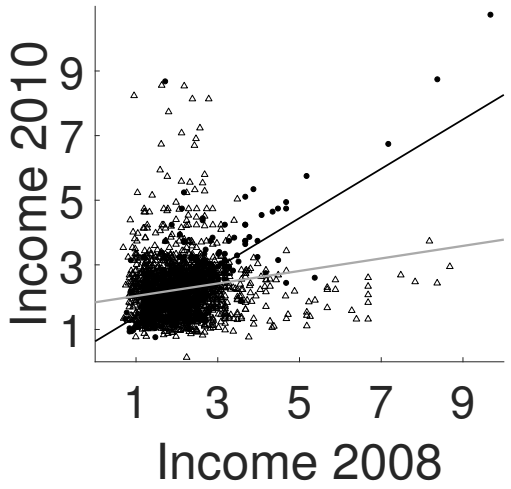
Consequences of ignoring mismatch error

Well-documented for Linear Regression (Neter al., 1965; Scheuren & Winkler, 1997; Lahiri & Larsen, 2001)

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

Compactly summarized in our recent survey (Wang et al. , WIREs Computational Statistics, 2021+) .

Consequences of ignoring mismatch error

Summary of consequences for Linear Regression:

– attenuation bias for regression coefficients $\beta^* = (\beta_1^*, \ldots, \beta_p^*)$; in general, squared bias proportional to

$$\|\beta^*\|_2^2 \times \text{mismatch rate}$$

– inflated standard errors

– Impact more dramatic for "high signal-to-noise" situations with $\|\beta^*\|_2^2/\sigma_*^2$ large

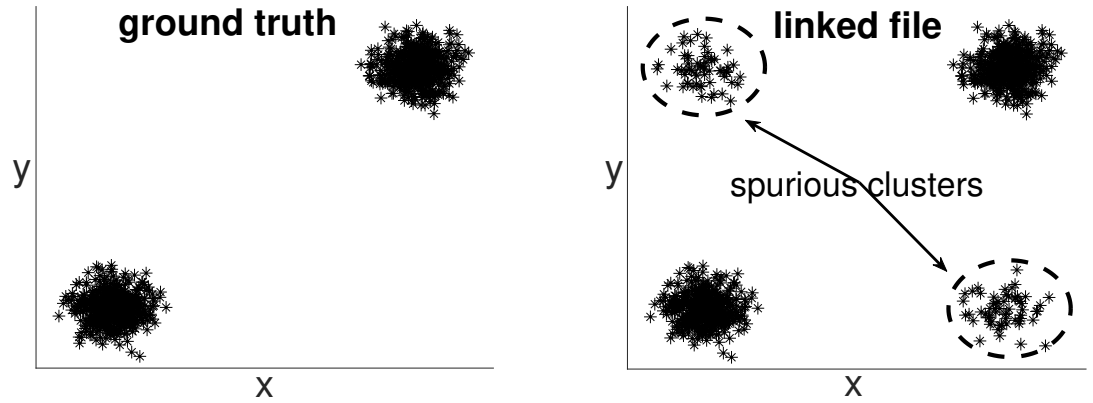
– for "noisy" models and small mismatch rate, mismatch error may be negligible

Consequences of ignoring mismatch error

Beyond the standard linear model, consequences of mismatch error are less well-studied.

Of interest in our project:

- semiparametric models and penalized estimation methods (e.g., lasso),
- unsupervised learning methods (e.g., PCA and clustering).



Mitigation strategies for mismatch error in linear regression

I. The Lahiri–Larsen–Chambers method (Lahiri & Larsen, 2001; Chambers, 2006; Han & Lahiri, 2019)

The model in these works assumes that instead of true response $\mathbf{y}^*$ we observe response $\mathbf{y} = \Pi^* \mathbf{y}^*$, where Π^* is a (generalized) permutation matrix.

Basic idea similar in spirit to instrumental variables:

- Mismatch error yields additional error that depends on covariates $\mathbf{x}$, i.e., regression error no longer uncorrelated w/ $\mathbf{x}$.
- $\rightsquigarrow$ regression on "instrumental variables" $\mathbf{q} = \mathbf{Q}\mathbf{x}$, where $\mathbf{Q} = \mathbf{E}[\Pi^*]$.

Mitigation strategies for mismatch error in linear regression

Pros & Cons of the L-L-C approach:

+ Conceptual simplicity
+ Generalizability beyond the classic linear model via estimating equations (Chambers, 2009; Chambers & DaSilva, 2020)
+ Performs well empirically for "reasonably informative" distributions over the range of Π^* and correctly specified $\mathbf{Q}$, even for high mismatch rates.

– Not conditionally unbiased (for fixed Π^*, bias is unbounded in general),
– Not robust to misspecifications of $\mathbf{Q}$
– Not (fully) clear how to calculate standard errors

Mitigation strategies for mismatch error in linear regression

II. Modern robust regression methods (S. & Ben-David, 2019; Wang et al., 2021+)

+ Explicit bounds on the estimation error
+ No information about RL required
+ Extends to linkage of more than two files

– Requires small mismatch rate and somewhat high signal-to-noise ratio
– Requires hyper-parameter tuning
– Not clear how to calculate standard errors / perform inference

Mitigation strategies for mismatch error in linear regression

III. Missing data approach (Wu, 1998; Gutman et al., 2012; Wang et al., 2021+)

- Unknown (generalized) permutation Π^* as missing data
- Inference via the EM algorithm or data augmentation ($\rightsquigarrow$ MCMC sampling)

\+ Inference about Π^* (in addition to parameters)

– Computational challenge: need to perform sampling over a huge set (all permutations)

– Danger of overfitting (Wang et al., 2021+)

Mitigation strategies for mismatch error in linear regression

IV. Pseudo-likelihood methods (Hof & Zwindermann, 2015; S. et al., 2021)

Basic model:

- Latent indicator variables $\{z_{ij}\}$ indicating match of $\mathbf{x}_j$ and y_i
- Models for

$$\begin{aligned}(y_i, \mathbf{x}_j)|z_{ij} = 1, &\quad \text{(correct match)},\\ (y_i, \mathbf{x}_j)|z_{ij} = 0 &\quad \text{(mismatch)}.\end{aligned}$$

- Model for $\mathbf{P}(z_{ij} = 1|\mathbf{x}_j, y_i, \ldots) = 1$
- $\rightsquigarrow$ mixture likelihood for each pair

Mitigation strategies for mismatch error in linear regression

Pros & Cons of the pseudo-likelihood approach:

+ Rather flexible model
+ Information about RL can be incorporated, but not required
+ Promising empirical performance
+ Valid inference (standard errors etc.) via asymptotic theory

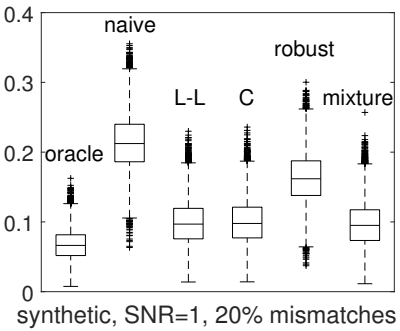
– Computational challenge I:
non-concavity of the pseudo-likelihood $\rightsquigarrow$ dependence on starting values, chance of getting stuck in bad local optima

– Computational challenge II:
intractable pseudo-likelihood for more complex models (such as linear mixed models).

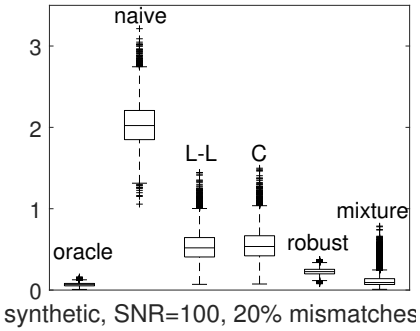
Mitigation strategies for mismatch error in linear regression

Performance of mitigation methods can vary depending on various data-specific characteristics.

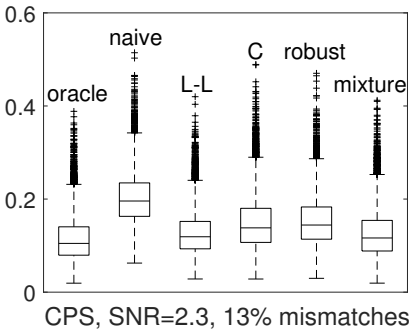
synthetic, low SNR



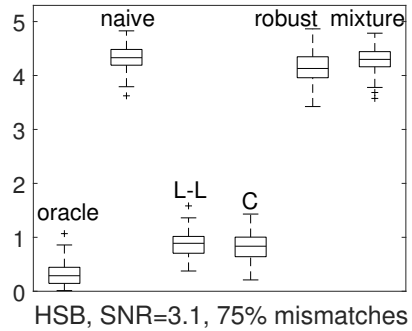
synthetic, high SNR



CPS



HSB



Estimation errors over 10^4 Monte-Carlo/Bootstrap runs of different adjustment methods for synthetic data under the exchangeable linkage model (Chambers, 2009) and semi-synthetic data based on the current population survey (CPS) and educational testing data (HSB, High School & Beyond Study).

Research agenda (other than theory & methods)

- Development of a suite of benchmark problems from real-world linkage problems, to guide model development & validation
- Make those available in suitable form in online repositories
- Disseminate research findings to various stakeholders involved with the analysis of linked data
- Training of undergraduate & graduate students in data science fields

References

[1] **S.** & Ben-David, "Linear Regression with Sparsely Permuted Data", *EJS*, 2019.

[2] Wang, Ben-David, **S.**, "Estimation in exponential family regression based on linked data contaminated by mismatch error", *arXiv*, 2020.

[3] Wang, Ben-David, Diao, **S.**, "Regression with linked data sets subject to linkage error", *WIREs Computational Statistics, to appear*.

[4] **S.**, Diao, Ben-David, "A Pseudo-Likelihood Approach to Linear Regression with Partially Shuffled Data", *JCGS*, 2021.

[5] Wang, Ben-David, **S.**, "Regularization for Shuffled Data Problems via Exponential Family Priors on the Permutation Group", *arXiv*, 2021.

- 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 & 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.

Thanks for your time & attention !