Finding Needles in Haystacks: Multiple-Imputation Record Linkage Using Machine Learning

John Abowd, Joelle Abramowitz, Margaret Levenstein, Kristin McCue, Dhiren Patki, Trivellore Raghunathan, Ann Rodgers, Matthew Shapiro, Nada Wasi, and Dawn Zinsser

April 18, 2022

Any opinions and conclusions expressed herein are those of the authors and do not represent the views of the Federal Reserve Bank of Boston, the principals of the Board of Governors, the Federal Reserve System, or the U.S. Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed (release number CBDRB-FY21-CED006-0019).

This research is supported by the Alfred P. Sloan Foundation through the CenHRS project at the University of Michigan with additional support from the Michigan Node of the NSF-Census Research Network (NCRN) under NSF SES 1131500. The HRS (Health and Retirement Study) is sponsored by the National Institute on Aging (grant number NIA U01AG009740) and is conducted by the University of Michigan.

CenHRS overview

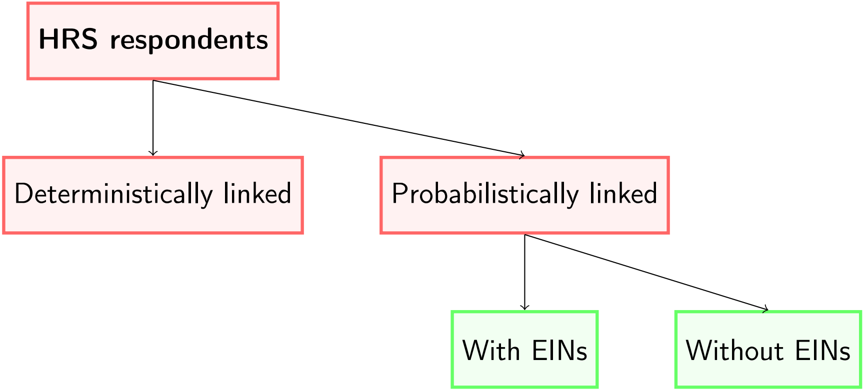
- Health and Retirement Study (HRS) is a longitudinal data set of $\approx$ 20,000 Americans over age 50
- **Our goal:** develop new measures of employer and coworker characteristics of working HRS respondents by linking to the Census Business Register (BR)
- **Challenge:** Lack of common unique employer identifiers in the two data sources
- **Solution:** Use probabilisitic linkage. Account for linkage uncertainty using multiple imputation (MI)

Types of record linkage in the CenHRS

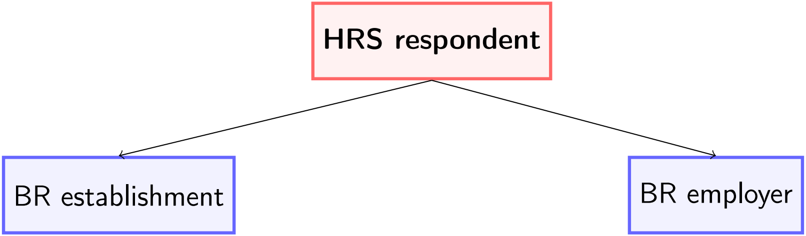
- 70% of respondents consent to SSA linkage - have EINs
- But EIN not always sufficient for 1:1 match

	Share of respondents
Deterministic match, have EIN	0.41
Probabilistic match, have EIN	0.30
Probabilistic match, no EIN	0.29

Types of record linkage in the CenHRS



Linkage targets in the CenHRS



Steps in probabilistic linkage procedure

1. **Blocking:** reduce dimensionality of linkage problem
2. **Training:** learn about true match status for subset of records
3. **Estimation:** estimate model to predict match status
4. **Match assignment:** use estimated model for MI-based assignment of respondents to establishments and employers

Step 1: Blocking

- $N_{HRS} \times N_{BR}$ is of order $10^{10} \implies$ infeasible to consider all pair-wise comparisons
- For each HRS record (i), define all BR candidates (j) that share a common attribute:
  1. EIN
  2. 3 digit zip, area code, city-state, 10 digit phone number
- EIN-based blocking $\implies \approx$ **400** BR candidates per respondent
- Location-based blocking $\implies \approx$ **30,000** BR candidates per respondent

Step 2: Training

- Separately for EIN- and location-based blocking: draw a sample of ≈ 1000 blocked pairs.
- Human reviewers examine pair-level characteristics and score $m_{ij} = 1$ if match, $m_{ij} = 0$ otherwise (separately for employer- and establishment-level match status)
- **Observed variables:** Name, address, phone number, size, industry, occupation, employer provision of health/pension benefits, number of EINs at which respondent is employed

Step 3: Estimation

- Fit $p(\boldsymbol{x}_{ij}; \boldsymbol{\beta}) = P(m_{ij} = 1|\boldsymbol{x}_{ij})$ using $m = 1, \dots, M$ Bayesian bootstrap replications of training data
- Obtain $\{\hat{\boldsymbol{\beta}}^{(1)}, \dots, \hat{\boldsymbol{\beta}}^{(M)}\}$
- Elastic Net for model selection; tuning parameters chosen to maximize out-of-sample predictive performance
- Assumption for validity of subsequent MI inference:

$$P(m_{ij} = 1|\boldsymbol{x}_{ij}) = P(m_{ij} = 1|\boldsymbol{x}_{ij}, \boldsymbol{z}_{ij})$$

  - $\boldsymbol{z}_{ij}$ are unobserved determinants of match status

Selected continuous predictors

Predictor	Description
Cubic spline JW score name	Similarity in HRS and BR name
Cubic spline JW score address	Similarity in HRS and BR address
Cubic spline EIN share of earnings	Importance of employer to worker
Cubic spline log employer size	National importance of employer
Full interaction of cubic splines	**Complementarities**
Age, log hourly wage, tenure, schooling	Worker characteristics

Selected binary predictors

Predictor
Employer size-class agreement
Establishment size-class agreement
Industry code agreement
Industry fixed effect
Occupation fixed effect
Survey interview mode and language
Gender, race, ethnicity, nativity, marital status fixed effects

Step 4: Multiply imputed match assignment

- For EIN-blocked cases:
  1. Compute $p(\boldsymbol{x}_{ij}; \hat{\boldsymbol{\beta}}^{(1)})$ for each pair
  2. Select match with probability proportional to $p(\boldsymbol{x}_{ij}; \hat{\boldsymbol{\beta}}^{(1)})$
  3. Repeat M times to create M completed data sets:

$$
\begin{gathered}
p(\boldsymbol{x}_{ij}; \hat{\boldsymbol{\beta}}^{(1)}) \rightarrow \text{implicate } 1 \\
p(\boldsymbol{x}_{ij}; \hat{\boldsymbol{\beta}}^{(2)}) \rightarrow \text{implicate } 2 \\
\vdots \\
p(\boldsymbol{x}_{ij}; \hat{\boldsymbol{\beta}}^{(M)}) \rightarrow \text{implicate } M
\end{gathered}
$$

Linkage uncertainty with MI

- For a given respondent:
  - Concentration of implicates $\implies$ low linkage uncertainty
  - Dispersion of implicates $\implies$ high linkage uncertainty

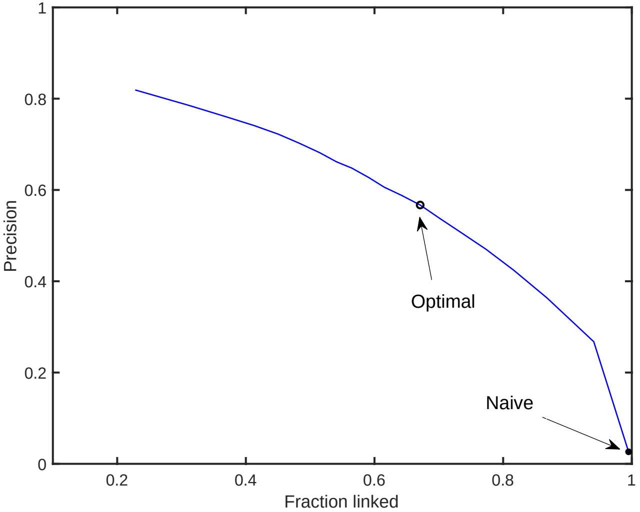
Step 4: Multiply imputed match assignment

- For location-blocked cases: $\approx 30k$ candidates per respondent!
- High chance of selecting false match
- Use deterministically matched sample to find optimal threshold to "cull" candidates before selecting implicates:

$$\hat{p}^{*(m)} = \underset{p \in [0,1]}{\operatorname{argmin}} \left(\left(1 - \underbrace{\mathcal{P}(\hat{\beta}^{(m)}, p)}_{\text{Precision rate}} \right)^2 + \left(1 - \underbrace{\mathcal{L}(\hat{\beta}^{(m)}, p)}_{\text{Link rate}} \right)^2 \right)^{1/2}$$

- Precision = fraction of respondents correctly linked

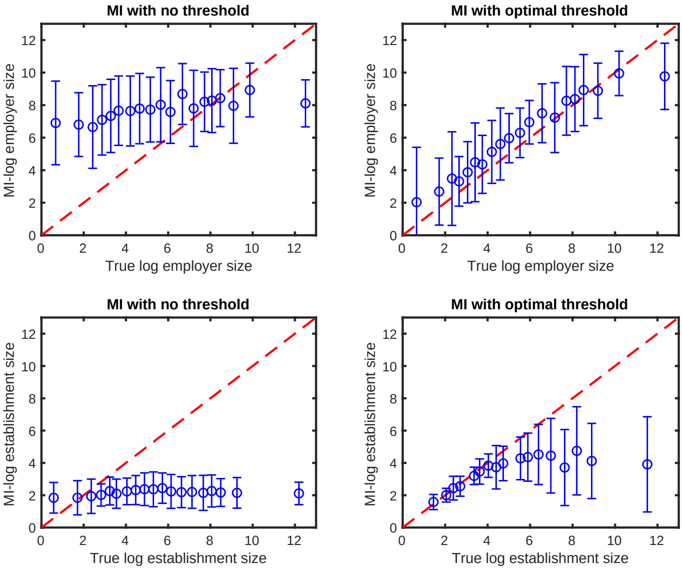
Precision-link rate tradeoff



Optimal thresholds

Employer-level linkage				
	Probability threshold	Link rate	Precision rate	BR candidates per HRS respondent
Naive	0	1	0.026	30,050
Optimal	0.39	0.648	0.587	52.3
Establishment-level linkage				
	Probability threshold	Link rate	Precision rate	BR candidates per HRS respondent
Naive	0	1	0.034	30,050
Optimal	0.095	0.661	0.569	146.8

Optimal thresholds improve imputation quality



Respondent characteristics by linkage status

	Linked	Non-Linked
Share of sample	0.92	0.08
Age	57.6	56.9
White	0.68	0.57
Black	0.22	0.24
Hispanic	0.14	0.26
Native born	0.87	0.69
Annual earnings ($)	43,160	33,330
Public sector worker	0.21	0.03
English interview	0.94	0.81
In-person interview	0.75	0.76

Rubin (1987) combining rules for MI

- For parameter of interest θ the MI estimate is

$$\hat{\theta} = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}^{(m)}$$

- MI variance is

$$\hat{\sigma^2} = \underbrace{\frac{1}{M} \sum_{m=1}^{M} \hat{\sigma^2_m}}_{\text{within variability}} + \left(1 + \frac{1}{M}\right) \underbrace{\frac{1}{(M-1)} \sum_{m=1}^{M} \left(\hat{\theta}_m - \hat{\theta}\right)^2}_{\text{between variability}}$$

Application: Wage-establishment size relationship

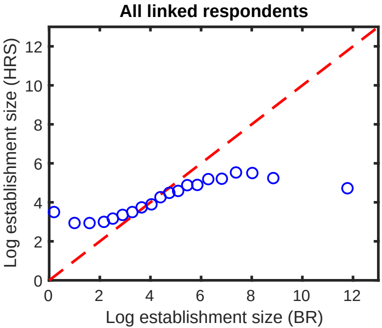
- Robust empirical finding: Larger establishments pay otherwise similar workers higher wages (e.g. Brown and Medoff, 1989; Bloom et al. 2018)
- We show non-classical measurement error in HRS self-reports amplifies this positive gradient

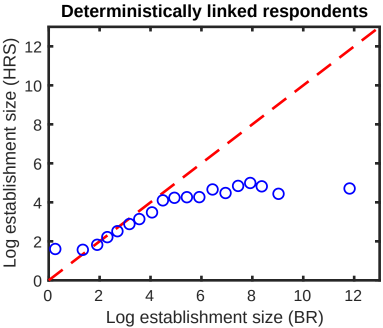
Wage-size gradient estimates

$$\log(\text{wage}_i) = \gamma_0 + \gamma_1 \log(\text{size}_i) + \boldsymbol{\gamma}_2 \boldsymbol{w}_i + \varepsilon_i$$

All linked respondents	
Respondent self-report of size	Imputed size from BR
0.042	0.019
(0.005)	(0.003)
Deterministically linked respondents	
Respondent self-report of size	Imputed size from BR
0.044	0.023
(0.009)	(0.005)

Non-classical measurement error: Reporting error negatively correlated with true value





Summary

- We use probabilistic record linkage to enhance the HRS with administrative data from the Census Bureau
- MI provides a way to incorporate linkage uncertainty in subsequent analysis
- We highlight that household survey reports about employers exhibit non-classical measurement error

Potential research applications of CenHRS

- Effect of trade shocks, mergers, job displacement shocks, other employer-level changes on
  - Retirement decisions
  - Social Security claiming behavior
  - Health and well-being
  - Future career prospects
  - Resource transfers between generations
- How might workers' risk preferences affect sorting to specific employers?
- How do fixed costs like commuting time influence retirement decisions?