# The effect of record linkage algorithms on research results

Flash — Open Science in Demography and Population Studies

Casey F. Breen [1]

[1]University of California, Berkeley | Department of Demography

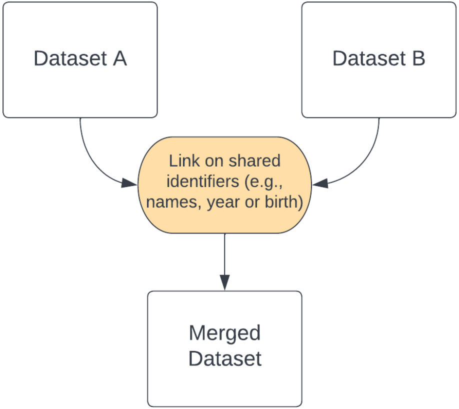
April 13, 2023

# Record linkage

- Identify same person across datasets in absence of a unique identifier (e.g., SSN)
- Wide applications: demography, sociology, computer science, epidemiology, history, medicine, economics, industry, etc.

# Record linkage

- Identify same person across datasets in absence of a unique identifier (e.g., SSN)
- Wide applications: demography, sociology, computer science, epidemiology, history, medicine, economics, industry, etc.



# Background

- Explosion in publicly-available linked census and admin data (Goldstein et al., 2021; Ruggles et al., 2020; Abramitzky et al., 2020)
  - Much lower barriers to entry

# Background

- Explosion in publicly-available linked census and admin data (Goldstein et al., 2021; Ruggles et al., 2020; Abramitzky et al., 2020)
  - Much lower barriers to entry
- Amazing body of methodological research on record linkage (Ruggles, Fitch and Roberts, 2018; Bailey et al., 2020; Abramitzky et al., 2020; Helgertz et al., 2022)

# Background

- Explosion in publicly-available linked census and admin data (Goldstein et al., 2021; Ruggles et al., 2020; Abramitzky et al., 2020)
  - Much lower barriers to entry

- Amazing body of methodological research on record linkage (Ruggles, Fitch and Roberts, 2018; Bailey et al., 2020; Abramitzky et al., 2020; Helgertz et al., 2022)

- **Roadmap** — what are key considerations of working with linked data?

# Key considerations for researchers

### Internal Validity

- False matches threaten internal validity

# Key considerations for researchers

### Internal Validity

- False matches threaten internal validity
- How do false matches affect our descriptive or inferential goals?

# Key considerations for researchers

## Internal Validity

- False matches threaten internal validity
- How do false matches affect our descriptive or inferential goals?

## External Validity

# Key considerations for researchers

## Internal Validity

- False matches threaten internal validity
- How do false matches affect our descriptive or inferential goals?

## External Validity

- Missed matches threaten external validity
- Composition of matched sample similar to population of interest?

# Key considerations for researchers

**Internal Validity**

- False matches threaten internal validity
- How do false matches affect our descriptive or inferential goals?

**External Validity**

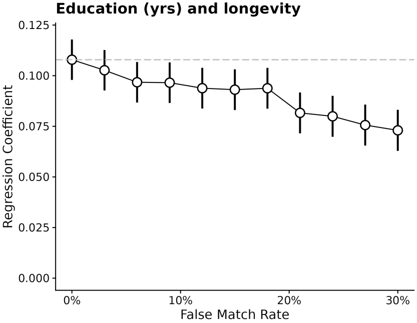
- Missed matches threaten external validity
- Composition of matched sample similar to population of interest?

# How do false matches affect research results?

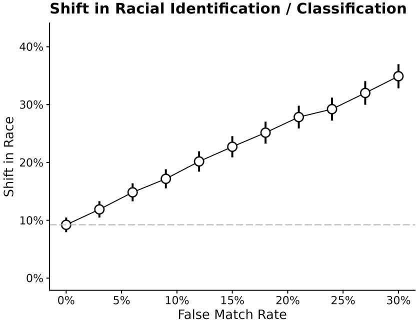
We **falsely** match person A in 1940 Census to Person B in mortality records

# How do false matches affect research results?

We **falsely** match person A in 1940 Census to Person B in mortality records


**Education (yrs) and longevity**

Attenuated regression coef.


**Shift in Racial Identification / Classification**

Upwardly biased estimates of transitions

# Case study: complete count census linkages

- Data infrastructure projects providing links between complete count census data (Ruggles et al., 2020; Abramitzky et al., 2020)

- Linkages between complete count 1930 Census and 1940 Census

- Create different samples using different linkage algorithms, varying levels of "conservative"

# Regression example – comparable coefficients

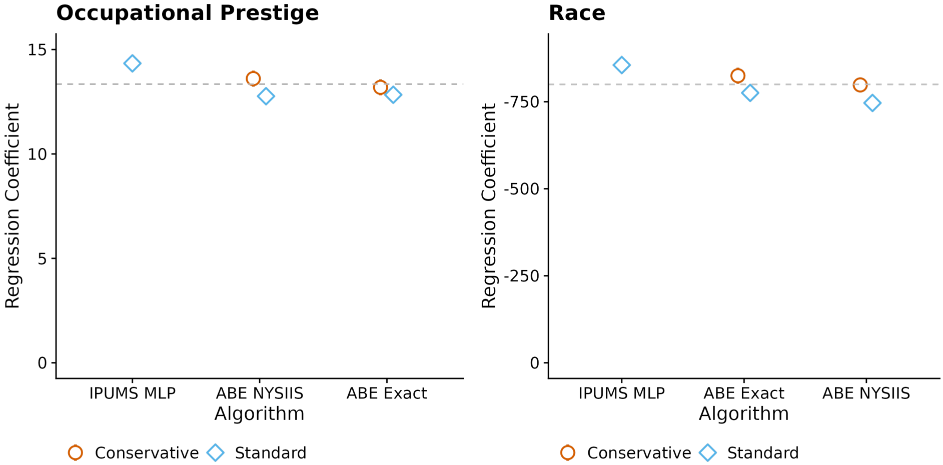


Figure: Association between covariates (1930) and wage and salary income (1940)

# Transition example – larger differences in estimates

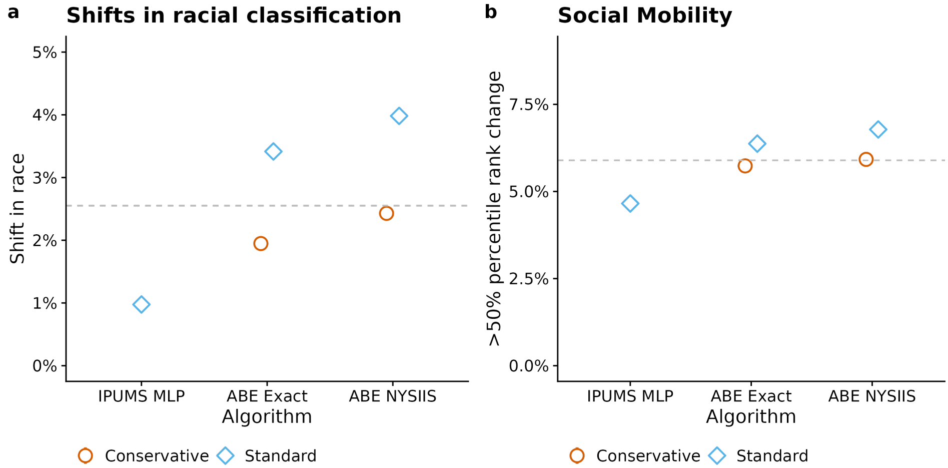


Figure: Transitions, racial classification and occupational prestige percentile rank

# Missed matches may lead to samples that are not representative

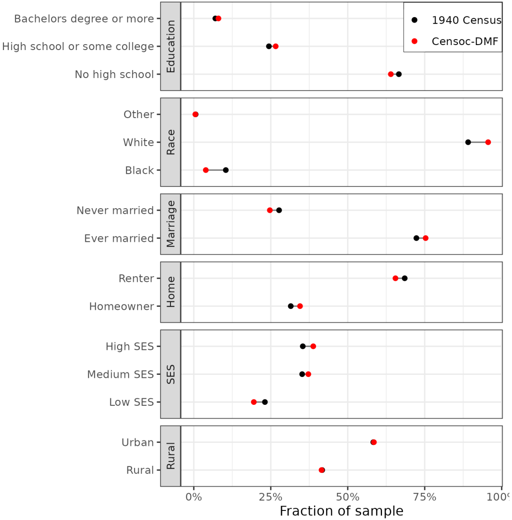
- We should **not** care about the match rate per se

# Missed matches may lead to samples that are not representative

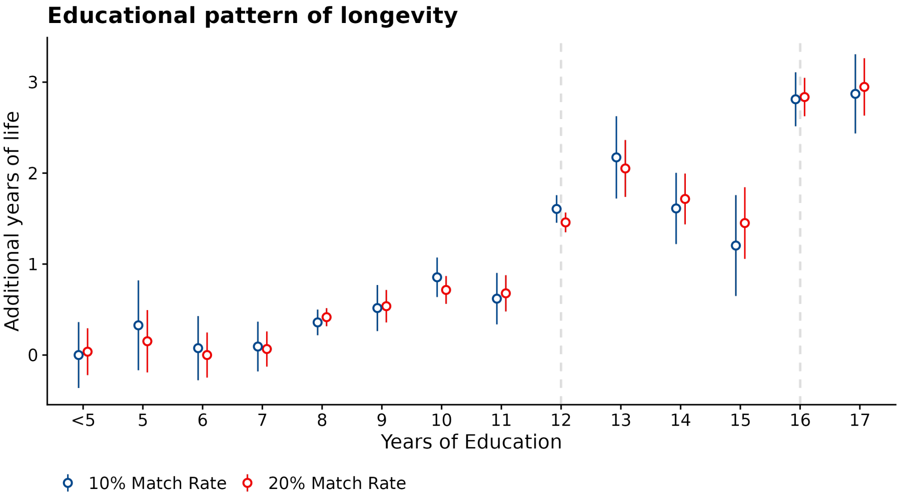
- We should **not** care about the match rate per se
- We should care about the representativeness of the sample

# Missed matches may lead to samples that are not representative

- We should **not** care about the match rate per se
- We should care about the representativeness of the sample
- Compare composition of matched sample to population of interest



# Same general result when match rate is cut in half ...



# Conclusion

The effect of record linkage algorithm on research results ultimately depends on the research questions. Considerations:

# Conclusion

The effect of record linkage algorithm on research results ultimately depends on the research questions. Considerations:

1. **Internal Validity**: Are our research results impacted by **false matches**? If so:

# Conclusion

The effect of record linkage algorithm on research results ultimately depends on the research questions. Considerations:

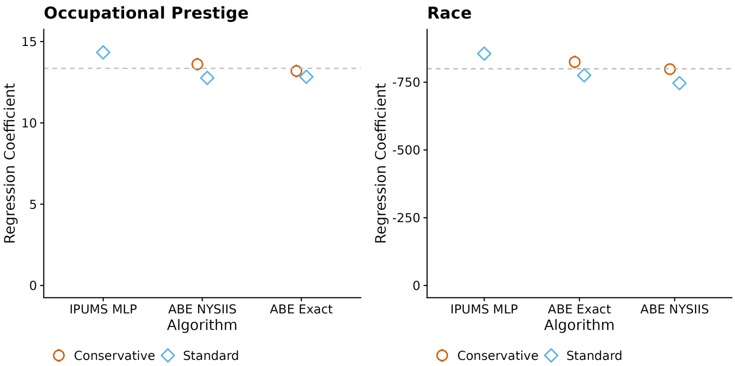
1. **Internal Validity**: Are our research results impacted by **false matches**? If so:
   - Direction/Magnitude of the bias?
2. **External Validity:** Is our matched sample representative of the population we want to learn about? If not:

# Conclusion

The effect of record linkage algorithm on research results ultimately depends on the research questions. Considerations:

1. **Internal Validity**: Are our research results impacted by **false matches**? If so:
   - Direction/Magnitude of the bias?
2. **External Validity:** Is our matched sample representative of the population we want to learn about? If not:
   - How different is the composition of our sample from our population of interest?
   - Can we address this with reweighting?

# Thank You



caseyfbreen

✉ caseybreen@berkeley.edu

# References

Abramitzky, Ran, Leah Boustan, Katherine Eriksson, Santiago Pérez and Myera Rashid. 2020. "Census Linking Project: Version 1.0.".

Bailey, Martha, Connor Cole, Morgan Henderson and Catherine Massey. 2020. "How Well Do Automated Linking Methods Perform? Lessons from U.S. Historical Data." *Journal of economic literature* 58(4):997–1044.

Goldstein, J. R., M. Alexander, C. Breen, A. Miranda González, F. Menares, M. Osborne, M. Snyder and U. Yildirim. 2021. "Censoc Project." *CenSoc Mortality File: Version 2.0. Berkeley: University of California* .

Helgertz, Jonas, Joseph Price, Jacob Wellington, Kelly J Thompson, Steven Ruggles and Catherine A. Fitch. 2022. "A New Strategy for Linking U.S. Historical Censuses: A Case Study for the IPUMS Multigenerational Longitudinal Panel." *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 55(1):12–29.

Ruggles, Steven, Catherine A. Fitch and Evan Roberts. 2018. "Historical Census Record Linkage." *Annual Review of Sociology* 44(1):19–37.

Ruggles, Steven, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas and Mathew Sobek. 2020. "IPUMS USA: Version 10.0 [Dataset]." *Minneapolis, MN: IPUMS. https://doi.org/10.18128/D010.V10.0.* .