A Social Scientist's Guide to Inference with Linked Data*

Casey F. Breen†

Draft Version: February 23, 2024

Abstract

The explosion in administrative datasets and methods for record linkage has revolutionized social science research. While there has been a proliferation of record linkage algorithms, there is limited guidance for researchers analyzing linked data. In this study, we use a series of simulation studies and empirical examples to illustrate the ways in which *false matches* and *missed matches* can impact research results. For our empirical examples, we investigate three different outcomes—social mobility, shifts in ethnoracial identification, and the educational gradient of longevity—using publicly available linkages from the CenSoc, IPUMS-MLP, and ABE projects. Our paper concludes with a series of practical recommendations for researchers conducting inference with linked data.

*Research reported in this publication was supported by the National Institute of Aging R01AG05894.
†University of Oxford. casey.breen@demography.ox.ac.uk

1 Introduction

Advances in the availability of administrative records and linkage techniques have ushered in a new generation of empirical social science research (Ruggles, Fitch and Roberts, 2018; Bailey et al., 2020; Abramitzky, Boustan and Eriksson, 2014). Past research has investigated technical features of different linkage algorithms, such as overall match rate and number of false matches. Yet less attention has been devoted to providing researchers with a clear statistical framework for analyzing linked data.

In this extended abstract, we use three large-scale linked administrative datasets from the CenSoc Project (Goldstein et al., 2021) to investigate the implications of choice of record linkage algorithm on a simple research question: the association between a covariate and life expectancy. Our results show that the choice of record linkage makes a modest difference, with more conservative algorithms having less attenuated estimates of the association between a covariate and longevity. However, these differences were modest, and results were indistinguishable between our most reliable samples. This suggests that the sensitivity of research results ultimately depends on the specific outcome of interest.

1.1 Data and Methods

For this research, we use the CenSoc datasets – so termed because they link the full-count 1940 Census ("Cen") with Social Security Administration mortality records ("Soc") – a publicly available administrative data resource for researchers studying mortality (Goldstein et al., 2021). These individual-level datasets provide researchers access to millions of mortality records with rich sociodemographic covariates.[1] This allows us to test the effect of record linkage algorithms in a realistic setting, where researchers are already conducting similar analyses (Atherwood, 2022).

CenSoc links the 1940 Census to Social Security mortality records using the ABE exact record linkage algorithm (Abramitzky, Boustan and Eriksson, 2012, 2014, 2016). This linking strategy requires an exact match on first name, last name, and place of birth, while allowing ± 2 years flexibility on year of birth. For additional detail and the specific approach for

[1] These data are available here: https://censoc.berkeley.edu

accounting for women changing their surname during marriage, see (Breen and Osborne, 2022).

1.2 Results

To investigate the effect of record linkage on research results, we first defined three samples from the Numident cohorts of 1900-1920: standard ABE matches, conservative ABE matches, and simple exact matches. We then estimated the association between years of education and longevity in each sample using an OLS regression on age of death. Figure 1 and Figure 2 show that the effect of education and wage and salary income on longevity. Broadly, these figures demonstrate that while samples created by all three different samples are highly comparable, there is a slight regression attenuation bias introduced by false matches in the ABE-Standard sample. However, the results are highly comparable between the ABE-conservative sample and basic exact match algorithm.

Next, for each sample, we defined three subsamples based on middle initial agreement – agree, disagree, or both agree and disagree ("pooled"). In total, this gives nine different samples. On each of the nine samples, we ran separate regression estimating the association between years of education and longevity Figure 3 plots each of the estimated regression coefficients. Several insights emerge from this figure. First, the regression coefficient for the full "pooled" sample was largest for the conservative matches, very slightly attenuated for the standard matches, and substantially attenuated for the standard matches not deemed conservative. Second, when middle initials agree, regression coefficient point estimates are identical across all three samples (conservative, standard, standard not conservative). Third, when middle initials disagree, the estimated regression coefficient is highly attenuated, and is most attenuated for the standard, matches deemed not conservative sample. Finally, for conservative matches, the estimated coefficient is nearly identical for the "pooled" sample and "agree" sample, suggesting that false matches have minimal impact on inference for this sample.

This analysis demonstrates that false matches systematically introduce measurement error, downwardly biasing the magnitude of estimated regression coefficients (Bailey et al., 2020). However, the attenuation bias in this example is modest—less than 10%. Further,

the direction of the bias is consistent across all samples; the estimated regression coefficients are always biased downwards.

1.3 Next Steps

The full paper will extend this analysis in several ways. First, it will test a much broader range of record linkage algorithms, including probabilistic algorithms (Enamorado, Fifield and Imai, 2019). Second, it will investigate a wider range of outcomes. Finally, our full paper will introduce a set of practical guidelines for researchers working with linked data.

Figure 1: Estimated association between longevity and wage income decile for three different algorithms.

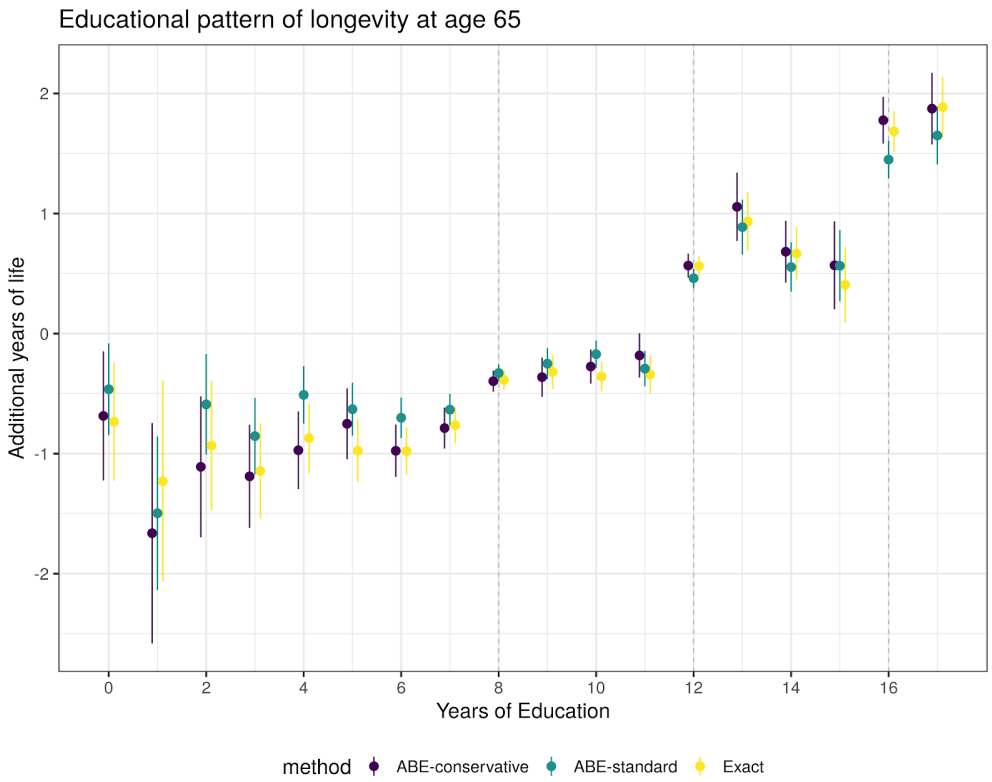

Figure 2: The estimated association between longevity and a given wage income decile.

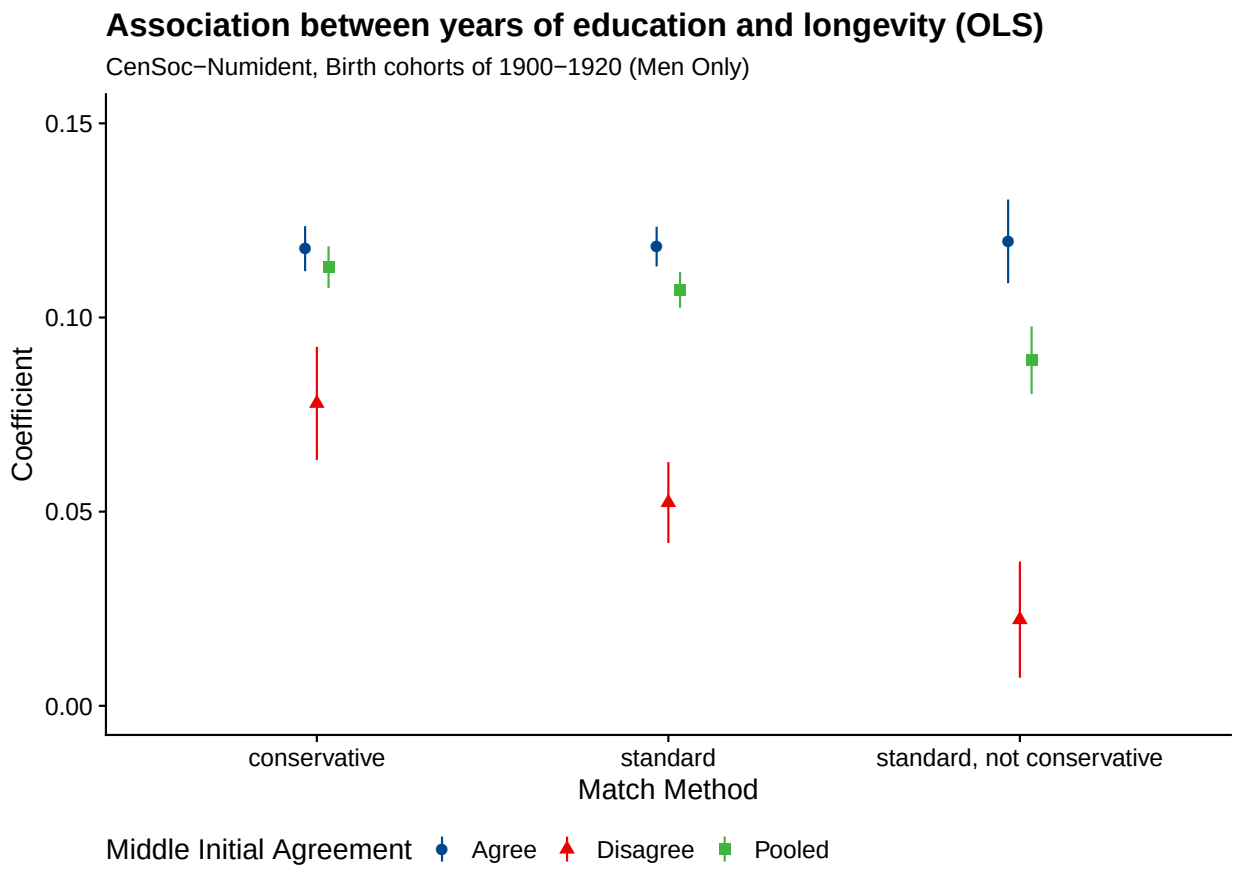

Figure 3: The estimated education gradient using regression on age of death from nine different CenSoc-Numident samples. Blue estimates, where middle initials matched, are nearly identical across samples. Green estimates from the full sample ("pooled") include records where middle initials agree and disagree. Red estimates, where middle initials didn't match are attenuated (biased towards 0).

References

Abramitzky, Ran, Leah Platt Boustan and Katherine Eriksson. 2012. "Europe's Tired, Poor, Huddled Masses: Self-Selection and Economic Outcomes in the Age of Mass Migration." *American Economic Review* 102(5):1832–1856.

Abramitzky, Ran, Leah Platt Boustan and Katherine Eriksson. 2014. "A Nation of Immigrants: Assimilation and Economic Outcomes in the Age of Mass Migration." *Journal of Political Economy* 122(3):467–506.

Abramitzky, Ran, Leah Platt Boustan and Katherine Eriksson. 2016. "To the New World and Back Again: Return Migrants in the Age of Mass Migration." p. 39.

Atherwood, Serge. 2022. "Does a Prolonged Hardship Reduce Life Span? Examining the Longevity of Young Men Who Lived through the 1930s Great Plains Drought." *Population and Environment* 43(4):530–552.

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

Breen, Casey and Maria Osborne. 2022. An Assessment of CenSoc Match Quality. Preprint SocArXiv.

Enamorado, Ted, Benjamin Fifield and Kosuke Imai. 2019. "Using a Probabilistic Model to Assist Merging of Large-Scale Administrative Records." *American Political Science Review* 113(2):353–371.

Goldstein, Joshua R., Monica Alexander, Casey F. Breen, Andrea Miranda-González, Felipe Menares, Maria Osborne and Ugur Yildirim. 2021. CenSoc Mortality File: Version 2.0. Technical report University of California, Berkeley.

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