International Journal of Population Data Science

Journal Website: www.ijpds.org

An instrumental variable approach to estimation of match probabilities or precision in linked data

Doidge, J¹*

¹Intensive Care National Audit and Research Centre (ICNARC)

Background and Rationale

While probabilistic linkage methods are ostensibly based on the probabilities of record pairs being matches (i.e. their marginal precision or positive predictive value), in practice they are used principally for ranking candidate links and fall short of supporting estimation of absolute probabilities. A few variations on Fellegi and Sunter's framework have been proposed to better accommodate the dependencies that limit transformation of match weights into match probabilities, but there are almost no alternative frameworks for match probability estimation.

Main Aim

To explore the feasibility, accuracy and limitations of a novel instrumental variable approach to estimation of match probabilities for use in either probabilistic record linkage or evaluation of linkage error.

Methods and Approach

Using both simulated data and a gold standard (labelled) dataset derived from real-world linked data, I assessed the accuracy of match probability estimation for a range of potential instruments and compared results to estimates produced using conventional probabilistic techniques.

Results

The technique involves trading the potential value of one matching variable in discriminating between candidate links for improved estimation of match probabilities within groups of otherwise similar candidates. Analysis of simulated data confirmed the theoretical validity of the approach in supporting unbiased estimation of match probabilities despite dependencies between other matching variables. Analysis of real-world data demonstrated feasibility in terms of the availability of real-world instruments that provided sufficiently accurate estimation in groups of candidate links above a minimum size.

Invalid instruments produced estimates that could be strongly biased.

Conclusion

These early results are promising but the general availability of valid instruments, their 'affordability' in terms of sacrificed discrimination, and means for identifying valid instruments remain unclear. However, this approach represents a new variety of tool for the data linker's toolkit, which may provide a useful angle on an otherwise difficult-to-estimate parameter and have applications yet to be envisaged.

*Corresponding Author:

Email Address: james.doidge@icnarc.org (J Doidge)