CANSSI-CRT Workshop
Modern Methods in Survey Sampling
July 8-10, 2024

Logistic regression on linked data from a secondary analyst perspective

Goldwyn Millar

Statistics Canada

Performing analysis on probabilistically linked data without accounting for the presence of false matches can lead to biased results. We consider the case in which an analyst wishes to conduct logistic regression on linked data in a way that accounts for the presence of false matches. We assume that this analyst is working from a ‘secondary analyst perspective’, meaning that they have access to a set of linked, analytic variables but not to the ‘linkage variables’ that were used to link the files in the first place. We discuss different approaches to this problem available in the literature, including approaches based on Chamber’s exchangeable linkage error assumption [1] and Zhang and Tuoto’s non-informative linkage error assumptions ([2], [3]). We also present the results of a simulation study comparing the performance of several methods of conducting such analysis.

References

[1] R. Chambers and J. O. Chipperfield, Using the Bootstrap to Account for Linkage Errors when Analysing Probabilistically Linked Categorical Data, Journal of Official Statistics, Vol. 31, No. 3, 2015, pp. 397-414.

[2] T. Tuoto and L.-C. Zhang, Linkage Data Linear Regression, Journal of the Royal Statistical Society Series A, Vol. 184, No. 2, 2021, pp. 522-547.

[3] L.-C. Zhang, Secondary Analysis of Linked Categorical Data, Presentation at the International Methodology Symposium, 2022.