Maximum Likelihood Logistic Regression with Auxiliary Information for Probabilistically Linked Data

Gunky Kim*
University of Wollongong, Wollongong NSW 2522 Australia. gkim@uow.edu.au

Ray Chambers
University of Wollongong, Wollongong NSW 2522 Australia. ray@uow.edu.au

Abstract:
Despite the huge potential benefits, any analysis of probabilistically linked data cannot avoid the problem of linkage errors. These errors occur when probability-based methods are used to link or match records from two or more distinct data sets corresponding to the same target population, and they can lead to biased analytical decisions when they are ignored. Previous studies aimed at resolving this problem have assumed that the analyst has access to all the information used in the data linkage process. In practice, however, most analysts are secondary analysts, with only partial access to information about the linkage error structure. As a consequence, our previous research has focussed on using an estimating equations approach to develop bias correction methods for secondary analysis of probabilistically linked data. In this paper we extend this approach to maximum likelihood estimation, using the missing information principle to accommodate the more realistic scenario of dependent linkage errors in both linear and logistic regression settings. We also develop the maximum likelihood solution when population auxiliary information in the form of popuation summary statistics is available. Our simulation results show that an incorrect assumption of independent linkage errors can lead to insufficient linkage error bias correction, while an approach that allows for correlated linkage errors appears to fully correct this bias. We also show that the main advantage from inclusion of population summary information is to correct small sample bias.

Key words: Probabilistic record matching; Linkage errors; Matching errors; Regression modelling; Estimating equations; Auxiliary data; Summary statistics.