Unbiased Regression Estimation for Multi-Linked Data in the Presence of Correlated Linkage Error and Summary Statistics for Secondary Analysis

Linkage errors can occur when probability-based methods are used to link records from two or more distinct data sets corresponding to the same target population. Recent research on methods for modifying standard methods of regression analysis to allow for these errors assumes that when more than two linked data sets are used to generate the data for this analysis, the linkage errors in these different data sets are independent. In this paper we extend these results to accommodate the more realistic scenario of dependent linkage errors. 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 extend our results furthermore to accommodate the small sample bias when the summary statistics are available.

Key words: probabilistic record linkage; correlated linkage errors; linear regression; estimating equations; summary statistics.