STATISTICAL MODELING AND ESTIMATION FOR LINKED DATA

Partha Lahiri
Joint Program in Survey Methodology
University of Maryland, College Park, U.S.A
plahiri@umd.edu

Computerized record linkage (CRL) methods are frequently used by government statistical agencies to quickly and accurately link two or more large files that contain information on the same individuals or entities using available information, which typically does not include unique, error-free identification codes. Because CRL utilizes already existing databases, it enables new statistical analysis without the substantial time and resources needed to collect new data. The possibility of errors in linkage causes problems for estimating the relationships between variables in the linked dataset. We will present a simple method to correct mismatch biases of standard estimators using an enhancement of the existing mixture models on measurements of the similarity among pairs of records to estimate probabilities used in calculating record linkage weights. We will report findings from a simulation study to compare the alternative estimators. This work is joint with Ms. Judith Law, PhD student, University of Maryland, College Park, USA.