Recent Developments and Open Problems in Post-Linkage Data Analysis

Record linkage and subsequent data analysis of the linked file with suitable propagation of uncertainty can be performed if the analyst also happens to be the linker or at least has comprehensive information about how the data were linked. However, it is rather common that the two processes are considered in a separate fashion, with the analyst being handed a linked file that is possibly subject to substantial linkage error (false matches and missed matches). Ignoring such error can render statistical analysis invalid. At the same time, accounting for linkage error with limited information about the linkage process poses a variety of challenges. This talk will outline a framework based on a mixture model for addressing mismatch error in the secondary analysis of linked files. Its use will be demonstrated in several case studies. Finally, we will present recent extensions, future directions and open problems.

Martin Slawski is an Associate Professor in the Department of Statistics at George Mason University. His research on data analysis after record linkage is currently supported by NSF. His research interests concern topics in computational statistics and applications in various domains. He serves as an associate editor of the Electronic Journal of Statistics. He received his Ph.D. in Computer Science from Saarland University, Germany, and was a postdoctoral associate in Statistics and Computer Science at Rutgers University prior to joining his current institution.