CNC 2023

Methods for High-Dimensional & Big Data

A flexible Record Linkage model

Combining different existing data sources empowers researchers to explore innovative questions, including those raised by tallying casualties and conducting healthcare monitoring studies. However, the lack of availability of a unique identifier often poses challenges. Record linkage procedures identify whether pairs of observations collected on different occasions belong to the same individual (matches) using partially identifying variables. Existing solutions attempt to simplify this task through condensing information but neglect dependencies among linkage decisions and disregard the one-to-one relationship required for building matches. The resulting reduction in computational burden comes at the price of inaccuracies and limited applicability. To avoid those issues, we propose to model the data generating process. We determine the set of matches (linkage) incorporating complex correlation structures based on a latent variable model. We develop an MC-EM algorithm and estimate the linkage using maximum likelihood. Simulations demonstrate the robustness of our model to the linking variables quality and its ability to better connect observations. We illustrate its scalability using the Perinatal Registry of the Netherlands (approximately 500,000 observations per data source) to identify first and second born children belonging to the same mother.