CNC 2025

Regression methods

False Discovery estimation in Record Linkage

This data era enables combining information to broaden research opportunities without costly new data collection. However, since data are not collected with specific future research questions in mind, and lack unique identifiers for privacy reasons, Record Linkage (RL) algorithms are used to assemble observations. The task poses challenges due to the sub-par reliability of partially identifying variables. Estimating the False Discovery Rate (FDR) associated with RL therefore holds importance for later inference. In particular in healthcare studies, estimating the Type I error of a set of linked records is crucial to determine the reliability of the inference drawn from the linked data. We introduce a new method to estimate the FDR and give guidelines for applying it on any sort of RL algorithm. Our recipe consists in linking records from real and synthesised data, estimating the FDR with the synthetic set. Our procedure enables identifying a threshold on the posterior linkage probabilities for which the RL process may be reliable. We investigate the performance of this methodology with well known RL algorithms and data sets before applying it to the Netherlands Perinatal Registry to show the importance of the FDR in RL when studying children/mother dynamics in healthcare records.