Department of Mathematics and Statistics Colloquium

Regression Modeling and File Matching Using Possibly Erroneous Matching Variables

Job Candidate

Abstract: Many analyses require linking records from two databases comprising overlapping sets of individuals. In the absence of unique identifiers, the linkage procedure often involves matching on a set of categorical variables, such as demographics, common to both files. Typically, however, the resulting matches are inexact: some cross-classifications of the matching variables do not generate unique links across files. Further, the matching variables can be subject to reporting errors, which introduce additional uncertainty in analyses. We present a Bayesian file matching methodology designed to estimate regression models and match records simultaneously when categorical matching variables are subject to reporting error. The method relies on a hierarchical model that includes (1) the regression of interest involving variables from the two files given a vector indicating the links, (2) a model for the linking vector given the true values of the matching variables, (3) a measurement error model for reported values of the matching variables given their true values, and (4) a model for the true values of the matching variables. We describe algorithms for sampling from the posterior distribution of the model. We illustrate the methodology using artificial data and data from education records in the state of North Carolina.

Monday, December 12 at 3:45 in Roop 103

refreshments at 3:30