University of California, Irvine
Statistics Seminar

Post-selection Inference with Missing Data and Multiple Imputation

Karen Messer
Professor, Family Medicine and Public Health
University of California, San Diego

Thursday, September 27, 2018
4 p.m., *2011* Bren Hall
(Bldg. #314 on campus map)

It is common to encounter missing data among the potential predictor variables in the setting of model selection. For example, in a recent study we attempted to improve the US guidelines for risk stratification after screening colonoscopy (Liu (2016)), with the aim to help reduce both overuse and underuse of follow-on surveillance colonoscopy. The goal was to incorporate selected additional informative variables into a neoplasia risk-prediction model, going beyond the 3 currently established risk factors, using a large dataset pooled from seven different prospective studies in North America. Unfortunately, not all candidate variables were collected in all studies, so that one or more important potential predictors were missing on over half of the subjects. Thus, while variable selection and risk prediction was a main focus of the study, it was necessary to address the substantial amount of missing data. Multiple imputation can effectively address missing data, and there are also good approaches to incorporate the variable selection process into model-based confidence intervals for risk. However, there is not consensus on appropriate methods of inference which address both issues simultaneously. Our goal here is to study the properties of model-based confidence intervals in the setting of imputation for missing data followed by variable selection. We use both simulation and theory to compare three approaches to such post-imputation-selection inference: a multiple-imputation approach based on Rubin's Rules for variance estimation (Schomaker (2014)); imputation-selection followed by bootstrap percentile confidence intervals; and a new bootstrap model-averaging approach presented here, following Efron (2014). We investigate relative strengths and weaknesses of each method. The `Rubins' Rules' multiple imputation estimator can have severe under coverage, and is not recommended. The imputation-selection estimator with bootstrap percentile confidence intervals works well. The bootstrap-model-averaged estimator, with the `Efron's Rules' estimated variance, may be preferred if the true effect sizes are moderate. We apply these results to the colorectal neoplasia risk-prediction problem which motivated the present work.

For directions/parking information, please visit https://uci.edu/visit/maps.php and http://www.ics.uci.edu/about/visit/index.php