Model Selection of Generalized Estimating Equations with Multiply Imputed Longitudinal Data

沈仲維

中正大學數學系

Abstract

Longitudinal data often encounter missingness with monotone and/or intermittent missing patterns. Multiple imputation (MI) has been popularly employed for analysis of missing longitudinal data. In particular, the MI-GEE method has been proposed for inference of generalized estimating equations (GEE) when missing data are imputed via MI. However, little is known about how to perform model selection with multiply imputed longitudinal data. In this work, we extend the existing GEE model selection criteria, including the "quasi-likelihood under the independence model criterion" (QIC) and the "missing longitudinal information criterion" (MLIC), to accommodate multiple imputed datasets for selection of the MI-GEE mean model. According to real data analyses from a schizophrenia study and an AIDS study, as well as simulations under non-monotone missingness with moderate proportion of missing observations, we conclude that: (i) more than a few imputed datasets are required for stable and reliable model selection in MI-GEE analysis; (ii) the MI-based GEE model selection methods with a suitable number of imputations generally perform well, while the naive application of existing model selection methods by simply ignoring missing observations may lead to very poor performance; (iii) the model selection criteria based on improper (frequentist) multiple imputation generally performs better than their analogies based on proper (Bayesian) multiple imputation. This is a joint work with Dr. Yi- Hau Chen at Academia Sinica.

Key words: missing data; multiple imputation; variable selection.