Predicting Future Responses with Possibly Mis-specified Models

Tianxi Cai
Department of Biostatistics, Harvard University, USA

Under a general regression setting, we propose an optimal unconditional prediction procedure for future responses. The resulting prediction intervals or regions have a desirable average coverage level over a set of covariate vectors of interest. When the working model is not correctly specified, the traditional conditional prediction method is generally invalid. On the other hand, one can empirically calibrate the above unconditional procedure and also obtain its cross-validated counterpart. Various large and small sample properties of these unconditional methods are examined analytically and numerically. We find that the K-fold cross validated procedure performs exceptionally well even for cases with rather small sample sizes. The new proposals are illustrated with two real examples, one with a continuous response and the other with a binary outcome.

[Tianxi Cai, Department of Biostatistics, 655 Huntington Ave, Boston, MA 02115, USA; tcai@hsph.harvard.edu]