Valid Statistical Inference after Model Selection

Lawrence D. Brown
Statistics Department
Wharton School, Univ of Pennsylvania
lbrown@wharton.upenn.edu

Abstract

Conventional statistical inference requires that a specific model of how the data were generated be specified before the data are analyzed. Yet it is common in applications for a variety of model selection procedures to be undertaken to determine a preferred model followed by statistical tests and confidence intervals computed for this "final" model. Such practices are typically misguided. The parameters being estimated depend on this final model, and post-model-selection sampling distributions may have unexpected properties that are very different from what is conventionally assumed. Confidence intervals and statistical tests do not perform as they should, particularly if the model selection procedures are themselves varied and not fully understood.

We study the commonly used Gaussian linear model. In spite of the pathologies lurking in post model selection inference, we present a procedure that provides valid inference for post model selection parameters. This procedure does not rely on knowledge about the model selection procedure. We also present some results about the performance characteristics of the procedure for some special linear model settings, and some asymptotics for situations involving selection within high dimensional parameter settings.

This is joint work with R. Berk, A. Buja, M. Freiman, K. Zhang and L. Zhao. L. Shepp also collaborated on portions of this research.