Ann Inst Stat Math (2010) 62:117–125
DOI 10.1007/s10463-009-0234-4

Model selection bias and Freedman's paradox

Paul M. Lukacs · Kenneth P. Burnham · David R. Anderson

Received: 16 October 2008 / Revised: 10 February 2009 / Published online: 26 May 2009

Abstract In situations where limited knowledge of a system exists and the ratio of data points to variables is small, variable selection methods can often be misleading. Freedman (Am Stat 37:152–155, 1983) demonstrated how common it is to select completely unrelated variables as highly "significant" when the number of data points is similar in magnitude to the number of variables. A new type of model averaging estimator based on model selection with Akaike's AIC is used with linear regression to investigate the problems of likely inclusion of spurious effects and model selection bias, the bias introduced while using the data to select a single seemingly "best" model from a (often large) set of models employing many predictor variables. The new model averaging estimator helps reduce these problems and provides confidence interval coverage at the nominal level while traditional stepwise selection has poor inferential properties.

Keywords Akaike's information criterion · Confidence interval coverage · Freedman's paradox · Model averaging · Model selection bias · Model selection uncertainty · Multimodel inference · Stepwise selection