The Limitations of Stepwise Variable Selection and LASSO Regressions

Mary Keonoupheth
Project Advisor: Dr. Bee Leng Lee
August 2022

Abstract

Data analysts are often faced with several regressors in the context of linear regression and resort to using automated methods to select a set of the most important variables. The use of automated variable selection is easy to implement and requires little effort, which makes it a ubiquitous practice in nearly every discipline that uses linear regression. However, several critics have pointed out the flaws in these automated procedures. A fundamental problem with automated variable selection is its failure to select all the regressors that have a significant relationship with the response variable and its fault in including noise variables that may only be significant by chance. Simulation results have shown that this problem is prominent in the presence of multicollinearity, small sample sizes, and when there are several candidate variables under consideration. As a result, the automated methods may select models that are not identical to the true model. Simulation results also reveal that models selected by the automated algorithms tend to produce larger prediction errors than the estimated true model.