Print

The Stepwise Selection Approach to Model Building

When you have many predictor variables to consider, the number of possible models can become extremely large and fitting all-possible regressions can be prohibitively time consuming. Stepwise selection methods take less computation time, but evaluate fewer models than the all-possible regressions approach. Stepwise selection methods include forward, backward, and stepwise, and you'll use these approaches to select variables based on their p-values.

Forward selection starts with no predictor variables in the model. This method computes an F statistic for each predictor variable not in the model, and examines the largest of these statistics. If it's significant at a specified significance level, the corresponding variable is added to the model. After a variable is added to the model, it stays in, even if it becomes non-significant later. Forward selection keeps adding variables, one at a time, until none of the remaining variables meets the specified level for entry, 0.50 by default.

Backward selection, also called backward elimination, starts with all predictor variables in the model. Results of the F test for individual parameter estimates are examined, and the least significant variable that is above the specified significance level is removed. After a variable is removed from the model, it remains excluded and cannot reenter. Backward selection is repeated until no other variable in the model meets the specified significance level for removal, 0.10 by default.

Stepwise selection combines aspects of both forward and backward selection. It starts with no predictor variables in the model and incrementally builds a model one variable at a time, as in forward selection. However, as in backward selection, stepwise selection can drop non-significant variables. The stepwise selection process terminates if no further variables can be added to or removed from the model, or when the variable to be added to the model is the one just deleted from it. The default p-values to enter and stay in the model are both 0.15. The default p-values to enter or stay in a model for all stepwise selection techniques can be changed based on the research goal or subject-matter expertise.

You might be wondering which automated model selection method is best? In fact, there's no one method that's best, just as there's no one model that's perfect. All models are approximations that are based on a sample from the population of interest. These model selection approaches provide suggestions for a useful approximating model.

When carrying out model selection, you can check the results of multiple approaches. You could try forward, backward, and stepwise selection. If they agree as to the recommended model, you can view that as support for a good model to use. If they suggest different models, you can compare model fit statistics or use subject-matter expertise to make the final choice.

In addition to using different model selection approaches, you can change the significance thresholds as well. Higher significance thresholds can result in more predictors in the recommended model, and lower thresholds will result in fewer.

Close