Variable Selection Techniques

Gabriel Ackall and Connor Shrader

June 1, 2021

In mathematical modeling, especially when using linear regression, it is often important to reduce the number of predictors, or variables, used to predict an output. This can help to increase the interpretability of a model and reduce its error.

1 Best Subset Selection

Best subset selection is a method for selecting the most influential predictors that minimize error in a least squares linear regression. It does this by fitting a linear regression model to every possible combination of predictors and then choosing the best model of all the possible combinations. The best model is determined through the use of a test error estimate. The most common examples of test error estimation indicators are AIC, BIC, Adjusted R^2, and cross validation.

While best subset selection results in the best possible model given the predictors, it is very computationally expensive. As the number of predictors increases, the number of linear models that best subset selection has to fit increases exponentially. This can be seen in Table 1. Thus, for models with more than 40 predictors, this can become infeasible for most computers to compute [1]. Given that in many scenarios, especially those seen in medicine with genomic data or in scenarios where $p >> n$, there can be many thousands of predictors, and best subset selection becomes impossible.

Table 1: Number of fitted models depending on number of predictors (p)

p	Fitted Models
2	2^2
10	2^{10}
100	2^{100}
k	2^k

2 Forward Stepwise Selection

Forward stepwise selection aims to approximate the best combination of predictors in a linear regression model, but with a more computationally efficient method than best subset selection. Forward stepwise selection starts without using any predictors. It then slowly begins adding the most important predictor to the model. The importance of a predictor is determined by it having the lowest p-value, lowest AIC, lowest BIC, and lowest Adjusted R^2, to name a few. This is repeated until a stopping point is reached which can be defined by p-value, AIC, BIC, and more.

3 Backward Stepwise Selection

details

4 Hybrid Stepwise Selection

details

5 Forward Stagewise Selection

details

References

[1] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.