

# Variable Selection Techniques

Gabriel Ackall and Connor Shrader

June 2, 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 Akaike information criterion (AIC), Bayesian information criterion (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 [4]. Given that in many scenarios, especially those seen in medicine with genomic data or in scenarios where there are more predictors than data samples, there can be many thousands of predictors, and best subset selection becomes impossible.

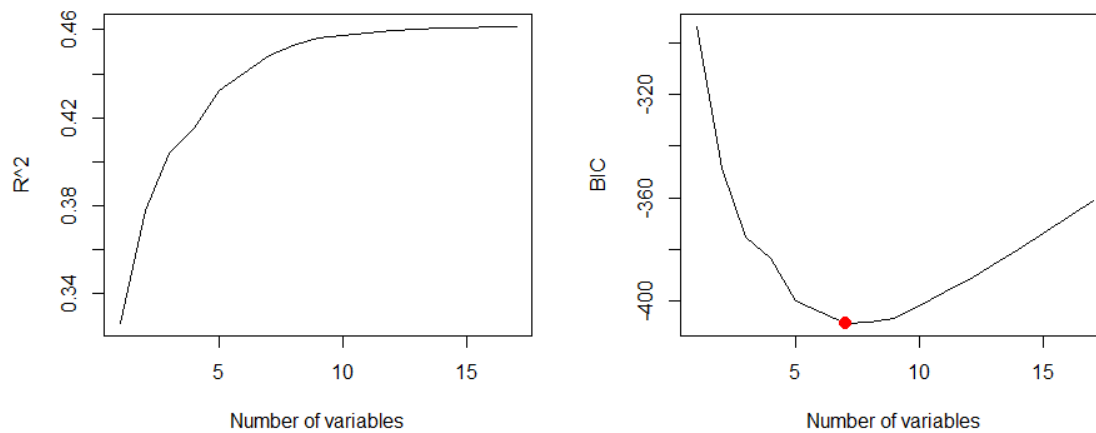
Table 1: Number of fitted models depending on number of predictors (p)

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

Figure 1 below demonstrates how the number of predictors can affect  $R^2$  and BIC when using best subset selection. This plot was created by using the `leaps` library (which provides a function to run best subset selection) [1]. We used the `College` dataset provided by the `ISLR` library [3], and fit models using `Grad.Rate` as the response. We see that as the number of predictors increases,  $R^2$  always increases (as expected). On the other hand, BIC is minimized with a moderate number of variables (between seven and nine). According to the BIC statistic, the best model has seven variables. The code used for this figure is in the `r` folder of the GitHub repository.

## 2 Forward Stepwise Selection

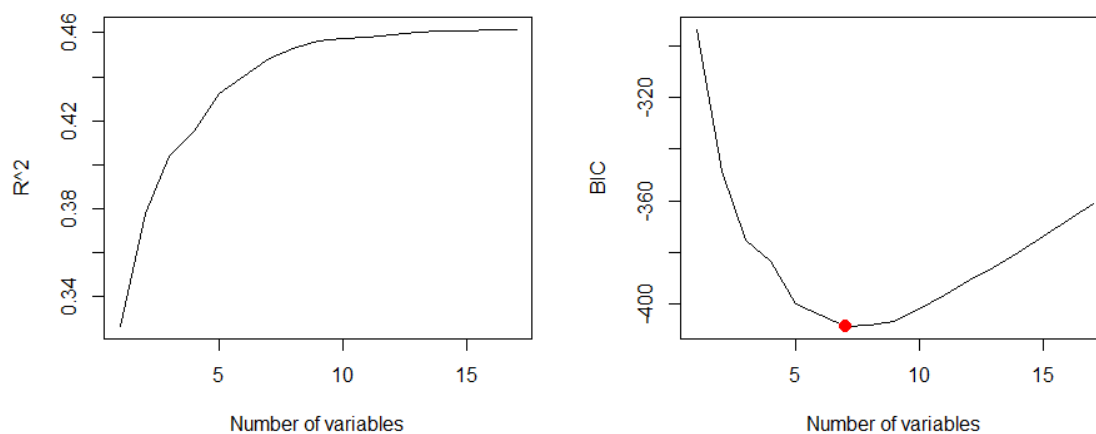
Forward stepwise selection aims to approximate the best combination of predictors in a linear regression model, but with a more computationally efficient method than best subset selection. Forward stepwise

Figure 1:  $R^2$  and BIC when applying best subset selection

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.

This process is much more computationally efficient than best subset selection, but at the cost that it does not necessarily result in the best combination of parameters in the linear regression and is not guaranteed to result in the best model.

Figure 2 shows the  $R^2$  and BIC statistics when fitting models using forward stepwise selection. Again, we predicted `Grad.Rate` using the `College` data set using the `leaps` library. The results are almost identical to what we saw for best subset selection. Even though the plots are similar, the specific model chosen by forward stepwise selection is actually different than the one used for best subset selection.

Figure 2:  $R^2$  and BIC when applying forward stepwise selection

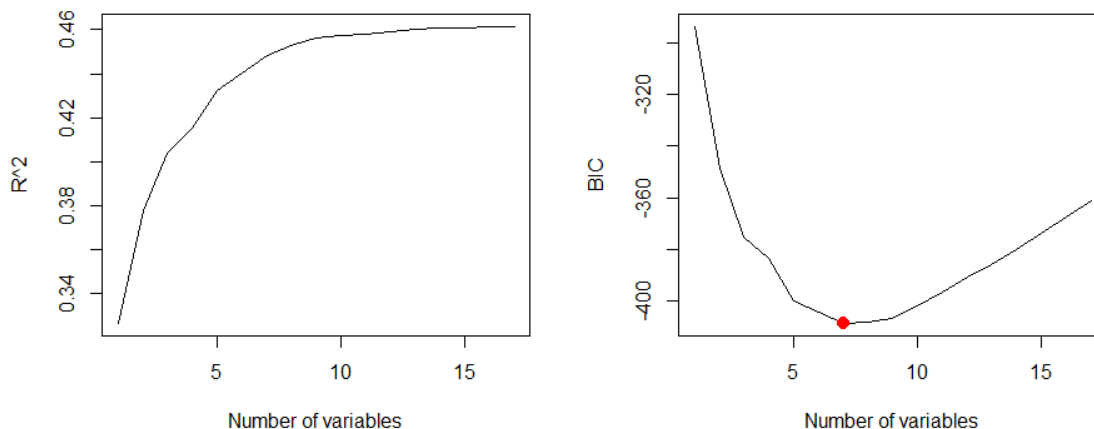
### 3 Backward Stepwise Selection

Backwards stepwise selection works very similarly to forward stepwise selection, except that it starts with every single predictor included in the least squares linear regression. Instead of adding predictors like in forward stepwise selection, backward stepwise selection removes the least important predictor. Similar to the forward method, the importance of a predictor can be determined by its p-value, AIC, BIC, or Adjusted  $R^2$ . This is repeated until a pre-determined stopping point is reached.

Backward stepwise selection can often result in better models than forward stepwise selection because it is guaranteed to test all the predictors together. This is different from forward stepwise selection that can sometimes suppress predictors, especially those that are collinear. For these reasons, when its use is possible, backward stepwise selection is preferred to forward stepwise selection. However, in cases where the number of predictors are greater than the number of samples, backward stepwise selection is impossible. In these case, forward stepwise selection must be used.

Figure 3 shows  $R^2$  and BIC after applying backward stepwise selection to the `College` data set. Again, the results are very similar to Figures 1 and 2, but the particular models chosen by the algorithm were slightly different.

Figure 3:  $R^2$  and BIC when applying backward stepwise selection



### 4 Hybrid Stepwise Selection

One weakness of the forward stepwise and backward stepwise methods is that they are greedy algorithms; in general, they will not find the best model for a given number of predictors. One way to improve model accuracy is to use hybrid stepwise selection, which allows for both forward steps and backward steps [2].

The algorithm could start with either zero predictors or all predictors. In each iteration, the method would either add a new predictor to the model or remove a predictor that does not increase performance. Like the forward and backward stepwise selection methods, this algorithm terminates when the model cannot be improved further; measuring the accuracy of the model can be determined using the AIC or BIC.

Although this strategy is slightly more computationally expensive than forward stepwise or backward stepwise selection, a hybrid approach may improve model results.

## 5 Forward Stagewise Selection

One last method for feature selection is called forward stagewise regression. Like forward stepwise selection, forward stagewise selection starts by fitting a model using none of the predictors. In each iteration, the method chooses the predictor most closely correlated to the residuals of the current model, and fits a simple linear regression using the predictor against the residuals. The coefficient for this predictor in the simple model is then added to the corresponding coefficient in the other model. This process is repeated until none of the predictors are correlated with the residuals.

Note that in each iteration of this algorithm, only one of the coefficients is changed. As a result, this method has a long runtime. In the long run, forward stagewise selection is still competitive compared to the strategies previously discussed.

## References

- [1] Thomas Lumley based on Fortran code by Alan Miller. *leaps: Regression Subset Selection*, 2020. R package version 3.1.
- [2] Jerome Friedman, Trevor Hastie, Robert Tibshirani, et al. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.
- [3] Gareth James, Daniela Witten, Trevor Hastie, and Rob Tibshirani. *ISLR: Data for an Introduction to Statistical Learning with Applications in R*, 2017. R package version 1.2.
- [4] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.