HW-4-P1

Jeetendra Gan

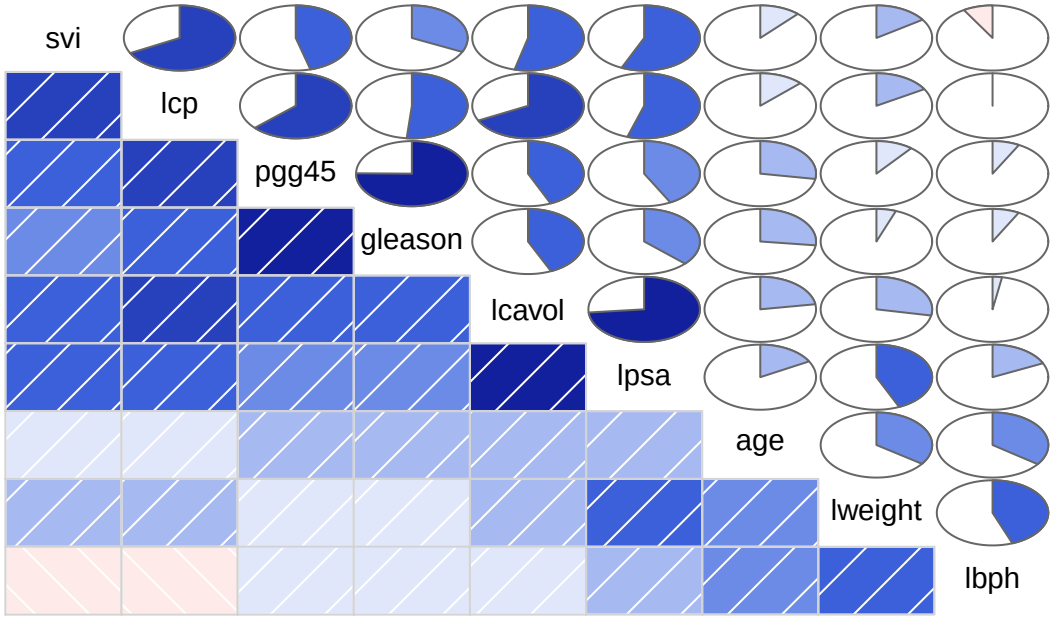
11/27/2019

Data analysis:

Here is the summary of the the prostate data.

```
##      lcavol           lweight           age             lbph        
##  Min.   :-1.3471   Min.   :2.375   Min.   :41.00   Min.   :-1.3863  
##  1st Qu.: 0.5128   1st Qu.:3.376   1st Qu.:60.00   1st Qu.:-1.3863  
##  Median : 1.4469   Median :3.623   Median :65.00   Median : 0.3001  
##  Mean   : 1.3500   Mean   :3.629   Mean   :63.87   Mean   : 0.1004  
##  3rd Qu.: 2.1270   3rd Qu.:3.876   3rd Qu.:68.00   3rd Qu.: 1.5581  
##  Max.   : 3.8210   Max.   :4.780   Max.   :79.00   Max.   : 2.3263  
##       svi              lcp             gleason          pgg45       
##  Min.   :0.0000   Min.   :-1.3863   Min.   :6.000   Min.   :  0.00  
##  1st Qu.:0.0000   1st Qu.:-1.3863   1st Qu.:6.000   1st Qu.:  0.00  
##  Median :0.0000   Median :-0.7985   Median :7.000   Median : 15.00  
##  Mean   :0.2165   Mean   :-0.1794   Mean   :6.753   Mean   : 24.38  
##  3rd Qu.:0.0000   3rd Qu.: 1.1787   3rd Qu.:7.000   3rd Qu.: 40.00  
##  Max.   :1.0000   Max.   : 2.9042   Max.   :9.000   Max.   :100.00  
##       lpsa        
##  Min.   :-0.4308  
##  1st Qu.: 1.7317  
##  Median : 2.5915  
##  Mean   : 2.4784  
##  3rd Qu.: 3.0564  
##  Max.   : 5.5829
```

From the summary, it can be seen that none of the variables have N.A. values.



It can also be seen that two paris of features are highly correlated as shown below

- lcavol, lpsa: 0.73
- gleason, pgg45: 0.75

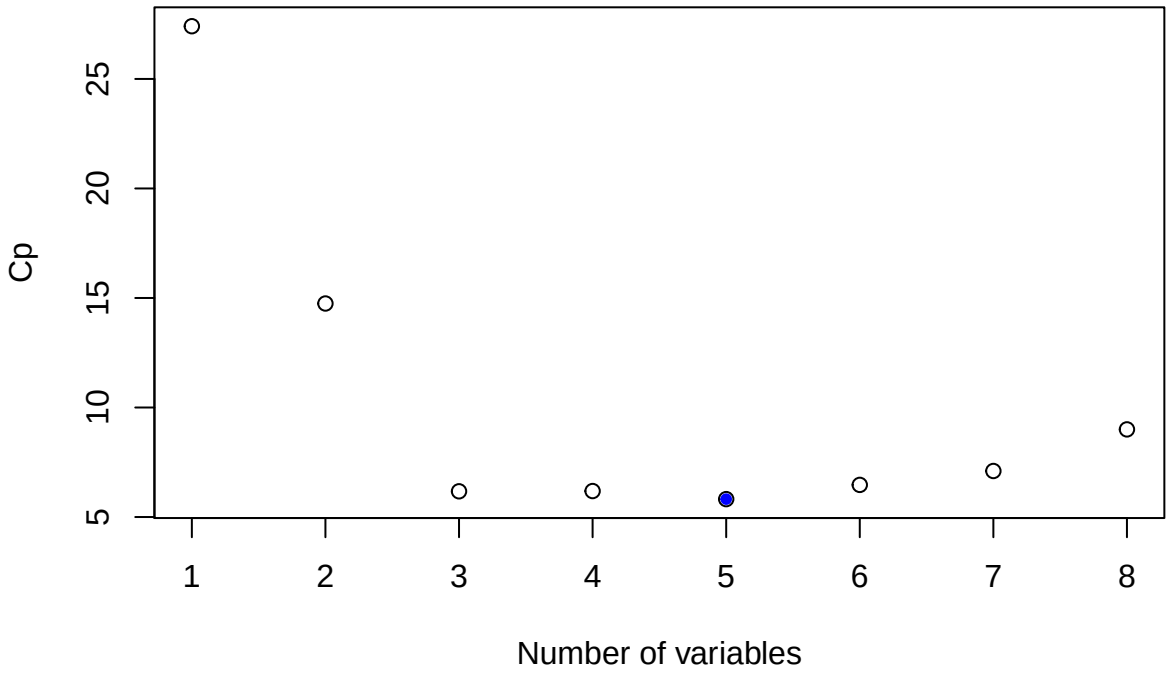
Best subset selection

Here is the summary of best subset selection.

```
## Subset selection object
## Call: regsubsets.formula(lpsa ~ ., data = prostate, method = "exhaustive",
##     nvmax = 9, )
## 8 Variables  (and intercept)
##         Forced in Forced out
## lcavol      FALSE      FALSE
## lweight     FALSE      FALSE
## age         FALSE      FALSE
## lbph        FALSE      FALSE
## svi         FALSE      FALSE
## lcp         FALSE      FALSE
## gleason     FALSE      FALSE
## pgg45       FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          lcavol lweight age lbph svi lcp gleason pgg45
```

```
## 1  ( 1 ) "*"    " "     " " " "  " " " " " " "     " "
## 2  ( 1 ) "*"    "*"     " " " "  " " " " " " "     " "
## 3  ( 1 ) "*"    "*"     " " " "  "*" " " " " "     " "
## 4  ( 1 ) "*"    "*"     " " "*"  "*" " " " " "     " "
## 5  ( 1 ) "*"    "*"     "*" "*"  "*" " " " " "     " "
## 6  ( 1 ) "*"    "*"     "*" "*"  "*" " " " " "     "*"
## 7  ( 1 ) "*"    "*"     "*" "*"  "*" "*" " " "     "*"
## 8  ( 1 ) "*"    "*"     "*" "*"  "*" "*" "*" "     "*"
```

Here is the graph of Cp.

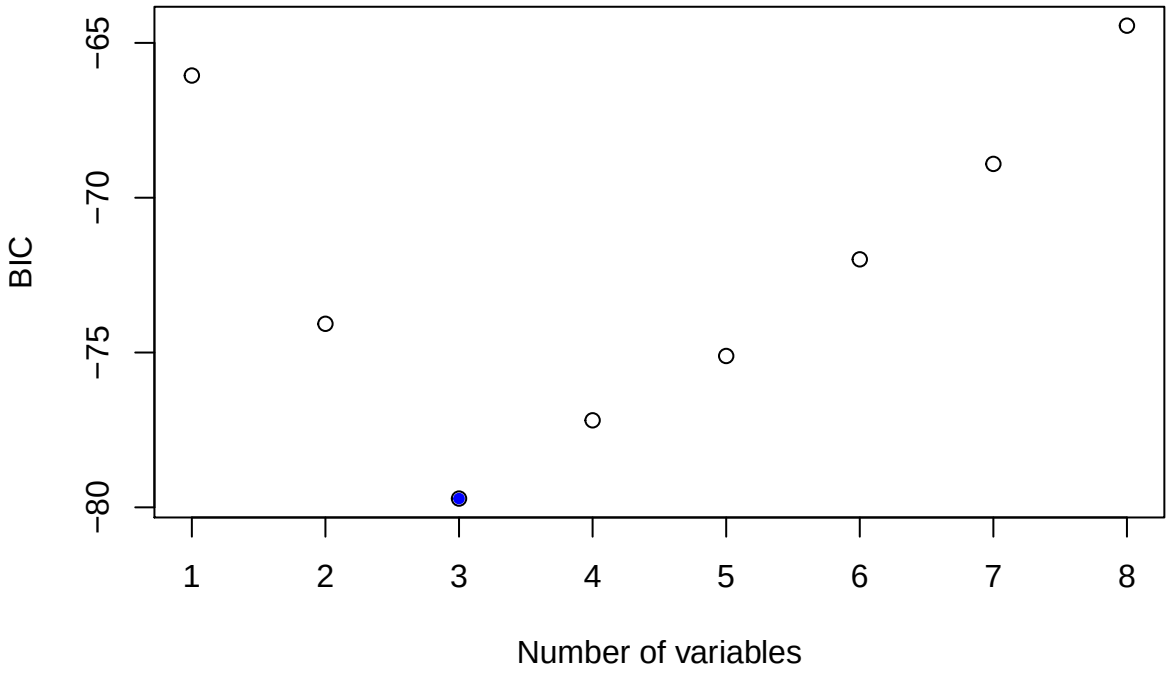


The point with minimun cp is shown as blue. A model with 5 predictors has the lowest Cp. Here are the features and their respective coefficients.

```
## (Intercept)      lcavol     lweight         age        lbph         svi
##  0.49472926  0.54399786  0.58821270 -0.01644485  0.10122333  0.71490398
```

After fitting the linear model to the data with the above features, it can be seen that the feature lcavol is the most significant. Age is the least significant out of all the variables.

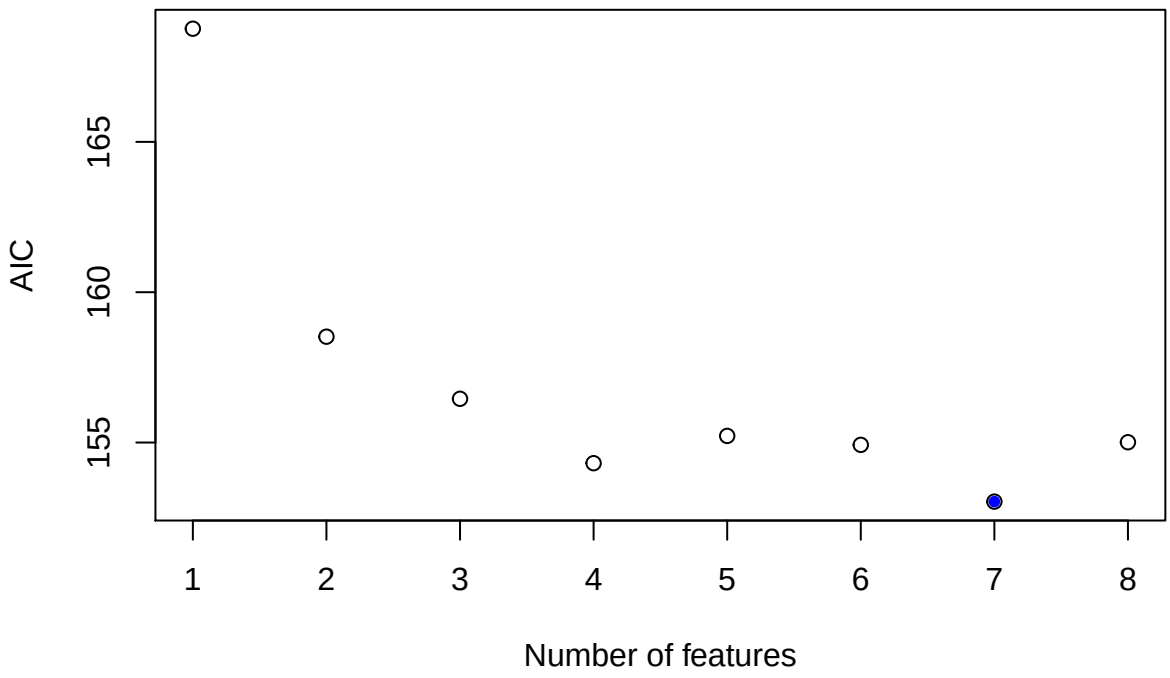
BIC



It can be seen that the model with 3 variables has the lowest BIC.

AIC

I have used all the best 8 subsets selected by regsubsets and calculated their AIC. Here are the results.

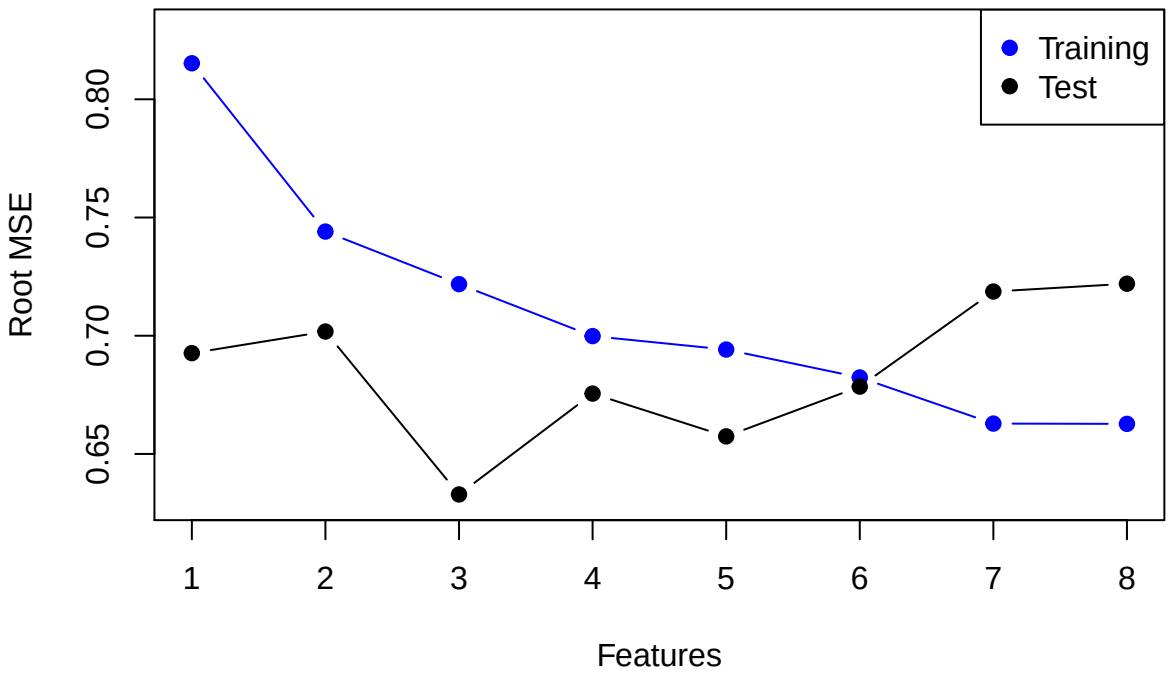


The graph shows that the min AIC is for the model with 7 predictors. Here are the coefficients for the model with 7 predictors.

```
##  (Intercept)       lcavol      lweight          age         lbph          svi
##  0.494154754  0.569546032  0.614419817 -0.020913467  0.097352535  0.752397342
##          lcp        pgg45
## -0.104959408  0.005324465
```

Test and train RMSE for all the subsets

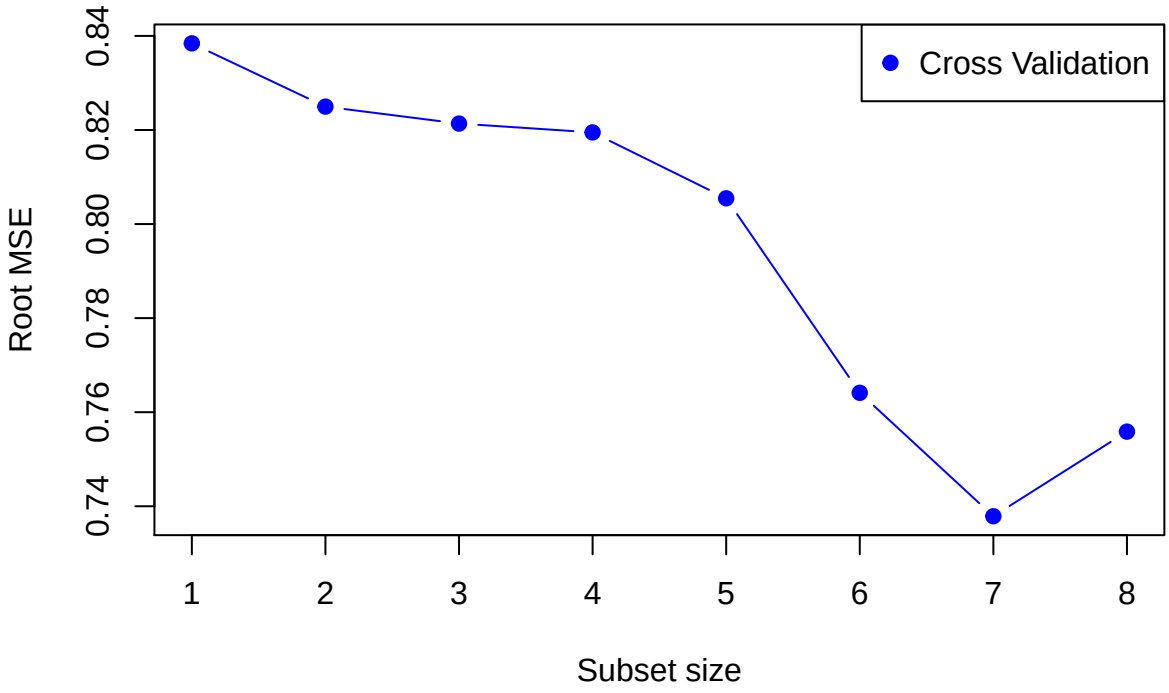
Here is the RMSE for a model selected using subset selection.



It can be seen that the model selected through BIC performs best on the test data.

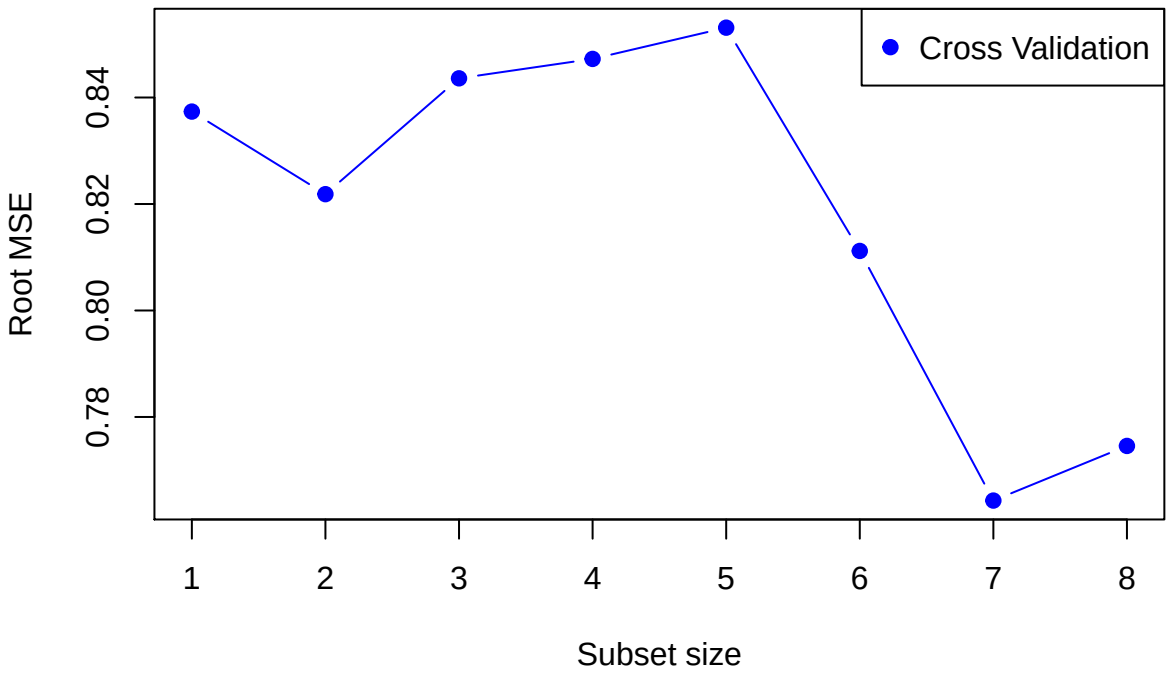
Cross validation

CV for 5 folds.

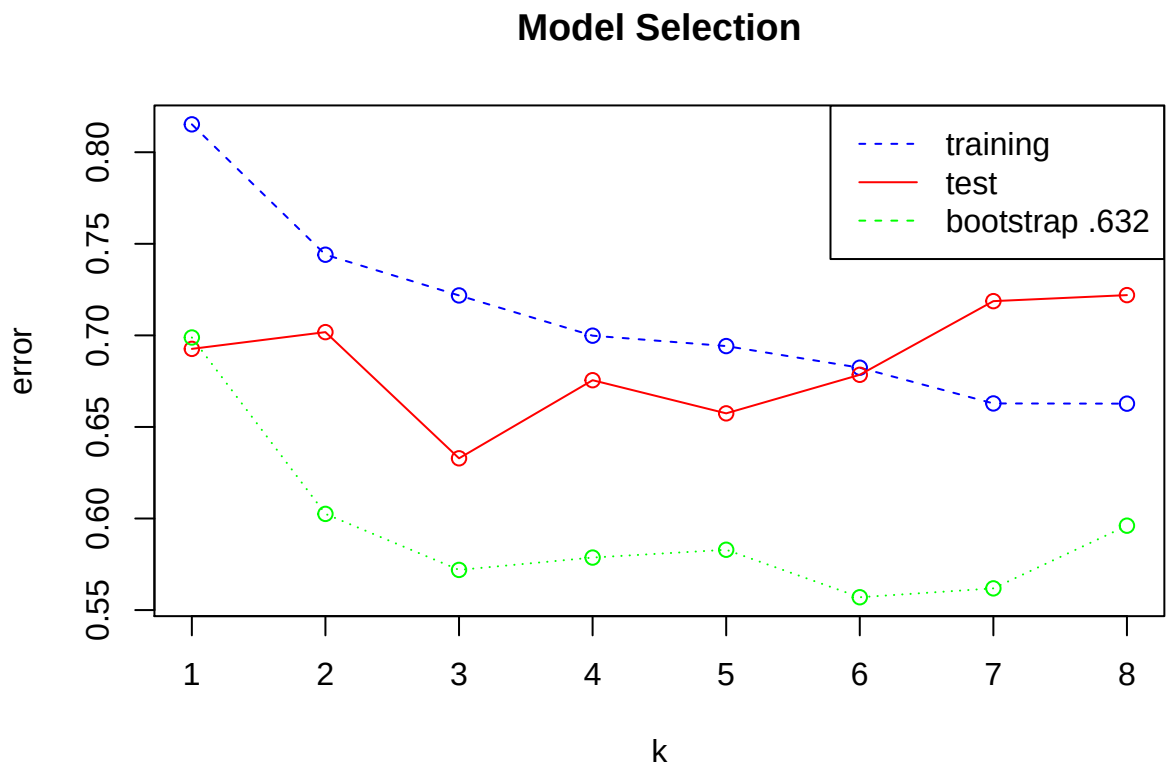


The plot above shows that a model with 7 predictors is better than the rest.

CV for 10 folds.



We can observe the same behaviour for K = 10, as we observed for k = 5



Inference

After applying AIC, BIC, cross-validation(5, 10), bootstrap the results are somewhat differenct. AIC tells us that the best model to choose is with 7 predictors, BIC tells us that the model with 3 predictors is the best. Both cross validation agree with the result of AIC, i.e. the model with 7 predictors is the best. So does bootstrap. The only metric supporting BIC's conclusion is the test error, as it is the least for a model with size 3. But test error should not be relied upon as it can be a case of an unfortunate split. We also have a better measure of how each of our subset-feature models perform, which is through the 10 fold and 5 fold cross validation.

As there is a greater support in favor of a model with 7 subsets, I would select that over a model with size 3.